



MMEDIA 2013

The Fifth International Conferences on Advances in Multimedia

ISBN: 978-1-61208-265-3

April 21 - 26, 2013

Venice, Italy

MMEDIA 2013 Editors

Philip Davies, Bournemouth and Poole College, UK

David Newell, Bournemouth University, UK

MMEDIA 2013

Foreword

The Fifth International Conferences on Advances in Multimedia (MMEDIA 2013), held between April 21st-26th, 2013 in Venice, Italy, was an international forum for researchers, students, and professionals where to present recent research results on advances in multimedia, and mobile and ubiquitous multimedia. MMEDIA 2012 brought together experts from both academia and industry for the exchange of ideas and discussion on future challenges in multimedia fundamentals, mobile and ubiquitous multimedia, multimedia ontology, multimedia user-centered perception, multimedia services and applications, and mobile multimedia.

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

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

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

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

We take here the opportunity to warmly thank all the members of the MMEDIA 2013 Technical Program Committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to MMEDIA 2013. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the MMEDIA 2013 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that MMEDIA 2013 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of multimedia.

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

MMEDIA Advisory Committee:

Dumitru Dan Burdescu, University of Craiova, Romania

Philip Davies, Bournemouth and Poole College, UK

Jean-Claude Moissinac, TELECOM ParisTech, France

David Newell, Bournemouth University, UK

Francisco J. Garcia, Agilent Technologies - Edinburgh, UK

Noël Crespi, Institut Telecom, France

Jonathan Loo, Middlesex University - Hendon, UK

Patrice Rondao Alface, Alcatel-Lucent Bell Labs - Antwerp, Belgium

Trista Chen, Fotologu Inc, USA

Alexander C. Loui, Kodak Research Labs / Eastman Kodak Company-Rochester, USA

MMEDIA 2013

Committee

MMEDIA Advisory Committee

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

MMEDIA 2013 Technical Program Committee

Max Agueh, LACSC - ECE Paris, France
Hakiri Akram, Université Paul Sabatier - Toulouse, France
Musab Al-Hadrusi, Wayne State University, USA
Nancy Alonistioti, N.K. University of Athens, Greece
Giuseppe Amato ISTI-CNR, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" - Pisa, Italy
Maria Teresa Andrade, University of Porto / INESC Porto, Portugal
Marios C. Angelides, Brunel University - Uxbridge, UK
Stylianos Asteriadis, National Technical University of Athens, Greece
Ramazan S. Aygun, University of Alabama in Huntsville, USA
Andrew D. Bagdanov, Universita Autonoma de Barcelona, Spain
Yannick Benezeth, Université de Bourgogne - Dijon, France
Jenny Benois-Pineau, LaBRI/University of Bordeaux 1, France
Sid-Ahmed Berrani, Orange Labs - France Telecom, France
Steven Boker, University of Virginia - Charlottesville, USA
Fernando Boronat Seguí, Universidad Politecnica de Valencia, Spain
Laszlo Böszörményi, University Klagenfurt, Austria
Marius Brezovan, University of Craiova, Romania
Dumitru Burdescu, University of Craiova, Romania
Helmar Burkhart, Universität Basel, Switzerland
Eduardo Cerqueira, Federal University of Para, Brazil
Damon Chandler, Oklahoma State University, USA
Vincent Charvillat, ENSEEIHT/IRIT - Toulouse, France
Bruno Checcucci, Perugia University, Italy
Shu-Ching Chen, Florida International University - Miami, USA
Trista Chen, Fotologu Inc., USA
Wei-Ta Chu, National Chung Cheng University, Taiwan

Antonio d'Acierno, Italian National Council of Research - Avellino, Italy
Philip Davies, Bournemouth and Poole College, UK
Vincenzo De Florio, University of Antwerp & IBBT, Belgium
Manfred del Fabro, Institute for Information Technology, Klagenfurt University, Austria
Vlastislav Dohnal, Masaryk University, Brno, Czech Republic
Jean-Pierre Evain, EBU Technical - Grand Saconnex, Switzerland
Nick Evans, EURECOM - Sophia Antipolis, France
Fabrizio Falchi, ISTI-CNR, Pisa, Italy
Kaori Fujinami, Tokyo University of Agriculture and Technology, Japan
Eugen Ganea, University of Craiova, Romania
Francisco J. Garcia, Agilent Technologies - Edinburgh, UK
Valerie Gouet-Brunet, Conservatoire National des Arts et Métiers - Paris, France
Sasho Gramatikov, Universidad Politécnica de Madrid, Spain
William I. Grosky, University of Michigan-Dearborn, USA
Christos Grecos, University of the West of Scotland, UK
Stefanos Gritzalis, University of the Aegean - Karlovassi, Greece
Angela Guercio, Kent State University, USA
Victor M. Gulias, University of Corunna, Spain
Hermann Hellwagner, Klagenfurt University, Austria
Luigi Iannone, Deutsche Telekom Laboratories, Germany
Razib Iqbal, University of Ottawa, Canada
Dimitris Kanellopoulos, University of Patras, Greece
Eleni Kaplani, TEI of Patra, Greece
Manolya Kavakli-Thorne, Macquarie University - Sydney NSW, Australia
Yasushi 'Yass' Kodama, Hosei University, Japan
Yiannis Kompatsiaris, CERTH-ITI, Greece
Markus Koskela, Aalto University, Finland
Panos Kudumakis, Queen Mary University of London, UK
Mikołaj Leszczuk, AGH University of Science and Technology - Krakow, Poland
Hongyu Li, Tongji University - Shanghai, China
Anthony Y. H. Liao, Asia University, Taiwan
Antonio Liotta, Eindhoven University of Technology, The Netherlands
Alexander C. Loui, Kodak Research Labs, USA
Erik Mannens, Ghent University, Belgium
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Michael Massoth, University of Applied Sciences - Darmstadt, Germany
Mike Matton, VRT research & innovation – Brussel, Belgium
Annett Mitschick, Technical University - Dresden, Germany
Ayman Moghnieh, Universitat Pompeu Fabra - Barcelona, Spain
Jean-Claude Moissinac, TELECOM ParisTech, France
Mario Montagud Climent, Universidad Politecnica de Valencia, Spain
Mireia Montañola, Université catholique de Louvain, Belgium
Michele Nappi, Università di Salerno – Fisciano, Italy
David Newell, Bournemouth University, UK
Petros Nicolitidis, Aristotle University of Thessaloniki, Greece
Vincent Oria, New Jersey Institute of Technology, USA
Jordi Ortiz Murillo, University of Murcia, Spain
Marco Paleari, Italian Institute of Technology / Center for Space Human Robotics - Torino, Italy

Eleni Patouni, University of Athens, Greece
Tom Pfeifer, Waterford Institute of Technology, Ireland
Wei Qu, Graduate University of Chinese Academy of Sciences, China
Piotr Romaniak, AGH University of Science and Technology - Krakow, Poland
Patrice Rondao Alface, Alcatel-Lucent Bell Labs - Antwerp, Belgium
Reza Sahandi, Bournemouth University - Dorset, UK
Susana Sargento, University of Aveiro/Institute of Telecommunications, Portugal
Klaus Schöffmann, Klagenfurt University, Austria
Oliver Schreer, Fraunhofer Heinrich-Hertz-Institute, Germany
Alexei Sourin, NTU, Singapore
Peter L. Stanchev, Kettering University - Flint, USA
Liana Stanescu, University of Craiova, Romania
Cosmin Stoica, University of Craiova, Romania
Yu Sun, University of Central Arkansas, USA
Anel Tanovic, BH Telecom d.d. Sarajevo, Bosnia and Herzegovina
Georg Thallinger, Joanneum Research - Graz, Austria
Daniel Thalmann, EPFL, Switzerland
Christian Timmerer, Alpen-Adria-Universität Klagenfurt, Austria
Chien-Cheng Tseng, National Kaohsiung First University of Science and Technology, Taiwan
Kuniaki Uehara, Kobe University, Japan
Andreas Uhl, Salzburg University, Austria
Binod Vaidya, Instituto de Telecomunicações / University of Beira Interior, Portugal
Andreas Veglis, Aristotle University of Thessaloniki, Greece
Janne Vehkaperä, VTT Technical Research Centre of Finland - Oulu, Finland
Dimitrios D. Vergados, University of Piraeus, Greece
Anne Verroust-Blondet, INRIA Paris-Rocquencourt, France
Giuliana Vitiello, University of Salerno – Fisciano, Italy
Lei Ye, University of Wollongong, Australia
Shigang Yue, University of Lincoln, UK
Sherali Zeadally, University of the District of Columbia, USA
Tong Zhang, Hewlett-Packard Labs, USA
Yang Zhenyu, Florida International University, USA

Copyright Information

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

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

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

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

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

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

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

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

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

Table of Contents

Performance Evaluation of Object Representations in Mean Shift Tracking <i>Peter Hosten, Andreas Steiger, Christian Feldmann, and Christopher Bulla</i>	1
Quantistic approach for classification of images <i>Federico F. Barresi, Giuseppe Battista, Jacopo Pellegrino, and Walter Allasia</i>	7
Efficient and Accurate Label Propagation on Large Graphs and Label Sets <i>Michele Covell and Shumeet Baluja</i>	12
Video Retrieval by Learning Uncertainties in Concept Detection from Imbalanced Annotation Data <i>Kenji Kumabuchi, Kimiaki Shirahama, and Kuniaki Uehara</i>	19
A Statistical Approach for the Automatic Recognition of Traffic Sign Deterioration <i>Sara Lorio, Mario Ferraro, Walter Allasia, and Francesco Gallo</i>	25
A Real-time Video Summarizing Service for Community-contributed Contents of Real-life Events <i>Samantha Vu, Owen Noel Newton Fernando, Mikko Rissanen, Natalie Pang, and Schubert Foo</i>	29
TV6: A Revisit to System Design, User Socialization and Content Recommendation in Social TV <i>Chong Yuan, Zhi Wang, and Lifeng Sun</i>	34
Local Histogram Modification Based Contrast Enhancement with GPU Acceleration <i>Jiang Duan, Min Li, Haiyue Wen, and Yingjie Peng</i>	40
Collaborative Multimedia Platform for Computational Philology - CoPhi Architecture <i>Angelo Mario Del Grosso and Federico Boschetti</i>	46
Efficient, Compact, and Dominant Color Correlogram Descriptors for Content-based Image Retrieval <i>Ahmed Talib, Massudi Mahmuddin, Husniza Husni, and Loay E. George</i>	52
Automatic Aerial Image Alignment for GeoMemories <i>Giuseppe Amato, Fabrizio Falchi, Fausto Rabitti, Andrea Marchetti, and Maurizio Tesconi</i>	62
Impact of Packet Loss on H.264 Scalable Video Coding <i>Siyu Tang and Patrice Rondao Alface</i>	67
News Video Semantic Topic Mining Based on Multi-wing Harmoniums Model <i>Xin Wen Xu, Yu Bo Shen, and Guo Hui Li</i>	74
Video Object Detection by Classification Using String Kernels	82

Wan-Hsuan Yu, Chi-Han Chuang, and Shyi-Chyi Cheng

Robust TV Stream Labelling with Conditional Random Fields 88
Abir Ncibi, Emmanuelle Martienne, Vincent Claveau, Guillaume Gravier, and Patrick Gros

Determinants of Behavioral Intention to Mobile Banking Case From Yemen 96
Abdullah Rashed, Henrique Santos, and Arwa AlEryani

Mobistream: Live Multimedia Streaming in Mobile Devices 100
Chrysa Papadaki and Vana Kalogeraki

H.264 Parallel Optimization on Graphics Processors 109
Elias Baaklini, Hassan Sbeity, and Smail Niar

Development of Context-Aware Real-Sense Services for Multi-Media and Multi-Device Environment 115
Hyunjeong Lee, Jaedoo Huh, Il-Woo Lee, and Sang Ho Lee

Region of Interest Encoding in Video Conference Systems 119
Christopher Bulla, Christian Feldmann, and Martin Schink

Depth Map Compression with Diffusion Modes in 3D-HEVC 125
Yun Li, Marten Sjostrom, Ulf Jennehag, and Roger Olsson

Efficient Stream-Reassembling for Video Conferencing Applications using Tiles in HEVC 130
Christian Feldmann, Christopher Bulla, and Bastian Cellarius

Disocclusion Handling Using Depth-Based Inpainting 136
Suryanarayana Murthy Muddala, Roger Olsson, and Marten Sjostrom

Performance Evaluation of Object Representations in Mean Shift Tracking

Peter Hosten, Andreas Steiger, Christian Feldmann, and Christopher Bulla

Institut für Nachrichtentechnik

RWTH Aachen University

Aachen, Germany

Email: {hosten, steiger, feldmann, bulla}@ient.rwth-aachen.de

Abstract—Mean shift tracking is a real-time capable object tracking approach that is not restricted to a specific object category. Several target object representations based on a feature distribution within an object region have been proposed for mean shift tracking. Quantitative performance metrics for the evaluation of object representations in mean shift tracking are mainly based on a comparison against ground truth data, which is often not available or requires considerable effort for its creation. In this paper, our main contribution is a novel approach for the quantitative evaluation of object representations in mean shift tracking, that does not rely on any ground truth data. Our approach is based on multiple hypotheses for the object location which initialise the mean shift tracking algorithm. The tracking result is then treated as random process and a quantitative metric is derived from its properties. Finally, the evaluation approach is applied to various object representations and test sequences. The findings demonstrate that the usage of multi-part object representations is beneficial if the representation captures the spatial colour distribution of the object.

Keywords- mean shift tracking; multi-part object representation; tracking evaluation

I. INTRODUCTION

The expansion of mobile networks and the spread of mobile devices allow for a universal multimedia access (UMA) in heterogeneous environments [1]. This requires an adaptation of the multimedia content in order to meet the current user situation such as the available data rate or the display device capability. However, considering only the technical requirements in the adaptation process does not necessarily ensure an optimal user experience. Current developments therefore aim to focus on the user and try to adapt the multimedia content with respect to the user preferences as well. The vision of user-centric convergence of multimedia is generally known as universal multimedia experience (UME) [2].

In this context, video adaptation and presentation techniques have become popular that are guided by region of interest (ROI) information. ROI-based video transcoding, for example, allows to reduce the quality of the different regions according to their importance, whereas ROI-aware rich media presentations allow for the interaction with the ROIs [3].

Consequently, these adaptation systems demand automatically created video annotations. Though a clear definition of an ROI cannot be given in general, it is commonly assumed that video objects might be of interest to the user. An automatic detection of arbitrary objects, however, is infeasible in practise. In principle, objects can only be detected when the underlying model assumptions are met. Thus, object detectors that have been trained for a specific appearance of an object, typically have a limited generalisation ability, e.g. they are not able to handle arbitrary deformations or occlusions. Hence, a reliable detection is generally not possible for the complete video, but for certain frames. In order to fill this gap, tracking approaches are necessary that allow to track the detected object and ROI, respectively.

Object tracking comprises an estimation of the target object state based on previous state estimations and the processing of visual information of the current frame. Though several real-time capable tracking methods have been proposed in literature [4], mean shift tracking is of particular interest as it allows for a generic modelling of the object's appearance by a probability density function (PDF) of features [5] and is thus not restricted to a specific object category. It seeks a mode of a similarity function between the target model and a candidate model by iterative computations of mean shift updates. A widespread feature is colour information whose distribution is encoded by a histogram. In order to gain a more distinct object representation, enhanced object representations for mean shift tracking have been proposed in form of multi-part object regions [6] [7].

In order to investigate the suitability of these object representations, in this work a novel quantitative evaluation method is proposed. Common approaches for the evaluation of tracking algorithms and object representations are based on ground truth data such as object centroids or bounding boxes [8] [9]. Object representations for mean shift tracking have been particularly evaluated based on the dice coefficient and the distance of the tracker centroid to the ground truth centroid [10]. The object centroid does, however, not correspond to the mode of the similarity function which is sought by the mean shift tracking. Furthermore, the mode of the similarity function varies dependent on the underlying object representation.

Therefore, we propose an evaluation approach which is independent on ground truth data and focused on the convergence behaviour of the mean shift tracking for different object representations. Based on multiple tracking initialisations drawn from an input random process, the mode to which mean shift tracking converges is treated as random process. Its stochastic properties are used to derive a metric allowing for an analysis of tracking accuracy and robustness of different object representations.

The rest of this paper is organised as follows: In Section II, the mean shift tracking of the target object location is explained and some object representations are presented. In Section III, we present a novel approach for the performance evaluation of object representations in mean shift tracking. Results are provided in Section IV. Finally, Section V concludes and discusses future work.

II. MEAN SHIFT TRACKING

A. Object Representation

The target object is represented by a target model which comprises the PDF of features within an object region. A target candidate at a candidate location is computed according to the same object representation and is evaluated against the reference target model during the course of tracking. Mean shift tracking based on colour features encodes the PDF of colours by a normalised kernel-weighted M-bin histogram [5] at which the weighting kernel $K(\mathbf{x})$ is centred at the target object. The target model $\mathbf{q} = \{q_u\}_{u=1,\dots,M}$ and a candidate model $\mathbf{p}(\mathbf{y}) = \{p_u(\mathbf{y})\}_{u=1,\dots,M}$ at location \mathbf{y} are then computed by:

$$q_u = C \cdot \sum_{n=1}^N K(\mathbf{y}_o - \mathbf{x}_n) \delta(b(\mathbf{x}_n) - u) \quad (1)$$

$$p_u(\mathbf{y}) = C_h \cdot \sum_{n=1}^N K\left(\frac{\mathbf{y} - \mathbf{x}_n}{h}\right) \delta(b(\mathbf{x}_n) - u) \quad (2)$$

Here, u denotes an index of a histogram bin, $b(\cdot)$ yields the bin index of the colour at pixel position \mathbf{x}_n , $\delta(\cdot)$ is the Kronecker delta function, N the number of pixels within the object region, and C and C_h are normalisation constants. Since the scale of the object may vary, the width h of the kernel function must be adapted to the size of the object region.

B. Mean Shift Update

Mean shift has been proposed as technique for seeking the mode of a density estimation [11] based on sample observations $\{\mathbf{x}_n\}$ which may be weighted by weights $\{w_n\}$. In the context of video object tracking, the samples $\{\mathbf{x}_n\}$ represent the pixel positions within the object region of the target and the target candidate, respectively.

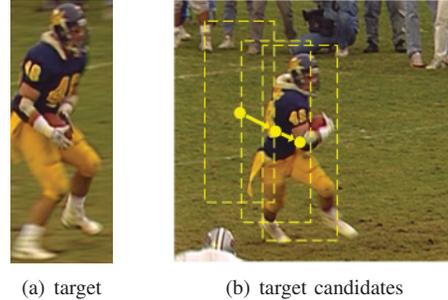


Figure 1. The new location of the target is estimated by iterative mean shift updates until a maximum number of iterations or convergence are reached.

Based on a kernel function $G(\mathbf{x})$ centred at a location \mathbf{y}_{j-1} , a mode estimation \mathbf{y}_j of the weighted kernel density estimation $\hat{f}_{K,h}(\mathbf{x})$ in (3) is provided by the weighted mean shift update in (4).

$$\hat{f}_{K,h}(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N w_n K\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right) \quad (3)$$

$$\mathbf{y}_j = \frac{\sum_{n=1}^N w_n \mathbf{x}_n G\left(\frac{\mathbf{y}_{j-1} - \mathbf{x}_n}{h}\right)}{\sum_{n=1}^N w_n G\left(\frac{\mathbf{y}_{j-1} - \mathbf{x}_n}{h}\right)} \quad (4)$$

The kernel functions $K(\mathbf{x})$ and $G(\mathbf{x})$ are related through their defining profiles $k(x)$ and $g(x)$ at which $g(x) = -k'(x)$ holds and $K(\mathbf{x})$ denotes the shadow of $G(\mathbf{x})$ [11]. The sequence $\{\mathbf{y}_j\}_{j=1,2,\dots}$ converges to the true mode of $\hat{f}_{K,h}(\mathbf{x})$ [12] which indicates the most likely location of the target object. As depicted in figure 1, mean shift location tracking therefore comprises iterative mean shift updates until a maximum number of iterations or convergence are reached.

The actual weight w_n at pixel position \mathbf{x}_n is derived from a Taylor series expansion of the Bhattacharyya coefficient $\rho(\mathbf{q}, \mathbf{p}(\mathbf{y}))$ similarity measure between the target model and a candidate model around a candidate model $\mathbf{p}(\mathbf{y}_0)$ [5]:

$$\rho(\mathbf{q}, \mathbf{p}(\mathbf{y})) = \sum_{u=1}^M \sqrt{p_u(\mathbf{y}) q_u} \quad (5)$$

$$w_n = \sum_{u=1}^M \sqrt{\frac{q_u}{p_u(\mathbf{y}_0)}} \delta(b(\mathbf{x}_n) - u) \quad (6)$$

Different approaches for the setting of weights $\{w_n\}$ are however possible such as target model back-projection [13] or various schemes for background incorporation [14].

The mean shift tracking algorithm can be extended to estimate the scale σ and orientation ϕ by mapping of Cartesian location coordinates \mathbf{x} to a 4-dimensional state space $\Gamma = (\mathbf{x}^\top, \sigma, \phi)^\top$ and computing the mean shift updates in the 4-dimensional state space [15].

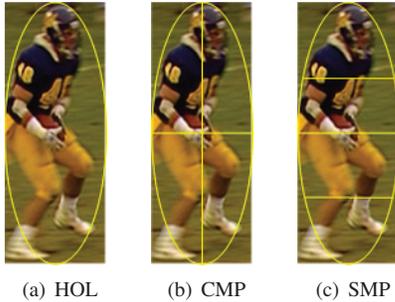


Figure 2. Illustration of the holistic (HOL), cross multi-part (CMP) and stack multi-part (SMP) object representations.

C. Multi-Part Object Representation

Multi-part object representations divide the object region into subregions to provide a more distinct description of the features in the spatial domain. In contrast to holistic (HOL) object representations illustrated in Fig. 2(a) they provide information about the distribution of features for each sub-region of the object region. Various fixed approaches for the spatial division of the object region exist [10]. For our evaluation we consider the cross (CMP) and stack (SMP) approach illustrated in Fig. 2(b) and Fig. 2(c).

III. QUANTITATIVE TRACKING EVALUATION METRIC

The mean shift procedure constitutes a gradient-based mode estimation technique and is therefore only able to locate a local mode of a density function estimation. It is particularly sensitive to different mean shift initialisations which may result in convergence to different modes. Criteria of interest for the evaluation of object models for mean shift tracking include the accuracy of convergence and the robustness of convergence for different initialisations.

In this context, robustness of convergence denotes invariance under poor initialisations and accuracy of convergence denotes the compliance of the estimated mode with the global mode of the multi-modal similarity function. In the following a quantitative metric is derived based on the modelling of the mean shift tracking as random process, which allows for an analysis of the above mentioned criteria.

A. Random Process Modelling

The target object state χ_k to be estimated by the mean shift tracking algorithm is application-specific and may comprise the object location, orientation or scale. Basically, it can be modelled by the following linear system and measurement equations:

$$\chi_k = \chi_{k-1} + \mathbf{n}_k \tag{7}$$

$$\mathbf{y}_k = \chi_k + \mathbf{e}_k \tag{8}$$

The hidden object state χ_k follows from an unknown state transition which is modelled by an additive system

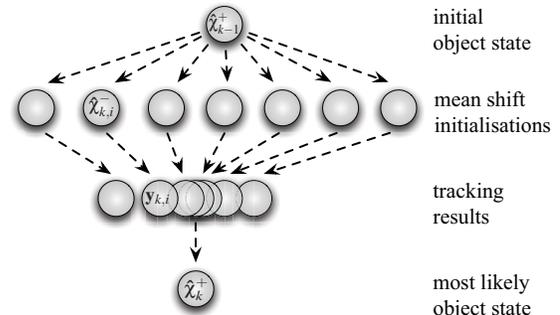


Figure 3. Illustration of Monte Carlo simulation.

noise process \mathbf{n}_k . Thus we are able to model the state uncertainty caused by the object’s motion or deformation, for example. A measurement \mathbf{y}_k of the state χ_k is obtained from the result of the mean shift tracking algorithm and yields a state estimation which ideally resembles the state χ_k . However, as the measurement depends on the mean shift initialisation, a measurement noise process \mathbf{e}_k is introduced representing the error of the mean shift tracking caused by poor initialisations.

The basic idea is to derive a quantitative performance metric from the unknown distribution of the measurement noise process \mathbf{e}_k estimated by a Monte Carlo simulation, which is driven by a predefined system noise process \mathbf{n}_k . Thereby a set of a priori particles $\{\hat{\chi}_{k,i}^-\}$ is predicted from an initial object state $\hat{\chi}_{k-1}^+$ according to the system equation (7):

$$\hat{\chi}_{k,i}^- = \hat{\chi}_{k-1}^+ + \mathbf{n}_{k,i} \tag{9}$$

Note, that the set of particles $\{\mathbf{n}_{k,i}\}$ represents the predefined system noise process \mathbf{n}_k . Each a priori particle $\hat{\chi}_{k,i}^-$ initialises the mean shift tracking, yielding a measurement particle $\mathbf{y}_{k,i}$. Hence the mean shift tracking can be interpreted as non-linear mapping $f(\cdot)$:

$$\mathbf{y}_{k,i} = f(\hat{\chi}_{k,i}^-) \tag{10}$$

That way, we obtain a set of measurement particles $\{\mathbf{y}_{k,i}\}$ approximating the distribution of the measurement process \mathbf{y}_k . As we are interested in the measurement noise process \mathbf{e}_k , we determine the most likely object state $\hat{\chi}_k^+$ by the element-wise median of the set of measurement particles $\{\mathbf{y}_{k,i}\}$:

$$\hat{\chi}_k^+ = \text{median}(\{\mathbf{y}_{k,i}\}) \tag{11}$$

Hence, the measurement noise process \mathbf{e}_k can be approximated by the set of particles $\{\mathbf{e}_{k,i}\}$:

$$\mathbf{e}_{k,i} = \mathbf{y}_{k,i} - \hat{\chi}_k^+ \tag{12}$$

The course of the above described Monte Carlo simulation is illustrated in Fig. 3.



Figure 4. Multiple tracking initialisations.

The random process modelling for mean shift tracking described in this Section effectively resembles a multi-hypotheses tracking approach at which the mean shift tracking algorithm is evaluated for multiple hypotheses drawn from a distribution centred at an initial estimation. This approach is closely related to particle filtering [16] but combined with mean shift tracking.

B. Modelling of System Noise

The system noise process \mathbf{n}_k steers the above described Monte Carlo simulation and consequently has an impact on the estimated measurement noise process \mathbf{e}_k . Particularly, the set of system noise particles $\{\mathbf{n}_{k,i}\}$ controls the quality of the mean shift initialisations $\{\hat{\mathbf{x}}_{k,i}^s\}$ (compare (9)). Thus a large spread of the system noise process increases the occurrence of poor initialisations, leading to erroneous tracking results. For the sake of reproducibility, the system noise particles $\mathbf{n}_{k,i}$ are drawn from a deterministic, zero-mean process, which is derived by regular sampling of a cube with edge length (range) s . The resulting mean shift initialisations are exemplarily illustrated in Fig. 4. That way different noise processes $\{\mathbf{n}_k^s\}$ can be created by varying the parameter s , each resulting in a different measurement noise process \mathbf{e}_k^s .

C. Performance Metric

The distribution of the measurement noise process \mathbf{e}_k^s is approximated by N_e particles $\mathbf{e}_{k,i}^s$. Hence these particles can be used to derive a metric for the evaluation of the tracking performance. We therefore use the mean absolute distance MAD_k^s :

$$\text{MAD}_k^s = \mathcal{E}\{|\mathbf{e}_k^s|\} = \frac{1}{N_e} \sum_{i=1}^{N_e} |\mathbf{e}_{k,i}^s| \quad (13)$$

The defined metric can be used to evaluate the convergence behaviour of different object models for mean shift tracking, i.e. the convergence accuracy and convergence robustness. Possible experiments include the evaluation of the tracking performance for a fixed system noise process \mathbf{n}_k^s across all frames $k \in \{1, \dots, K\}$ of a test sequence or the evaluation for a set of system noise processes $\{\mathbf{n}_k^s\}_{s=1}^S$ and averaging the MAD_k^s over all frames k :

$$\text{MAD}_s = \frac{1}{K} \sum_{k=1}^K \text{MAD}_k^s \quad (14)$$

The latter approach allows an investigation of the robustness towards poor initialisations. Thus a larger value of the parameter s leads to an increased spread of the system noise process, which in turn increases the occurrence of poor initialisations.

IV. EVALUATION

We have implemented a mean shift algorithm for location tracking. The maximum number of mean shift iterations is set to 20 and the convergence bound is set to 0.1 pixels. As recommended in [5], the shadow kernel $K(\mathbf{x})$ is implemented by an Epanechnikov kernel whose bandwidth is equal to the dimension of the target object. Background colour information is not exploited and no update of the target model is performed during the course of tracking.

The presented evaluation results are obtained from three test sequences described by table I and Fig. 7 at which the accuracy and robustness of location tracking is evaluated with regard to the object representations presented in Fig. 2. All computations are based on $N_e = 25$ initialisation samples, which are exemplarily illustrated in Fig. 4. For each sample, a complete sequence is processed. Thereby the initialisation in each frame, that is derived from the tracking result of the previous frame, is shifted according to the current sample. The resulting trajectories are then used to compute the value of MAD_s for each test sequence.

A. Test Sequences

The test sequences feature different characteristics which affect the tracking performance. The *Stefan* sequence comprises tracking a tennis player against background clutter. A small and fast oscillating handbag of a lady is tracked in the *Aëna* sequence where the difficulty lies in the velocity of the target object. A much more distinct and easier target object is given by the pink dressed lady in the *Couple* sequence where, however, partial occlusion occurs.

B. Results

The values of MAD_s for all test sequences are plotted in Fig. 5 to assess the tracking performance across an entire test sequence for different ranges s of the initialisation region. In case of the *Stefan* sequence, the SMP object representation is superior to other object representations for small initialisation regions ($s < 5$) which is confirmed by

TABLE I. TEST SEQUENCES.

Sequence	Size	Target size	# frames
<i>Stefan</i>	355 × 288	70 × 177	300
<i>Aëna</i>	720 × 540	27 × 31	125
<i>Couple</i>	480 × 270	50 × 235	154

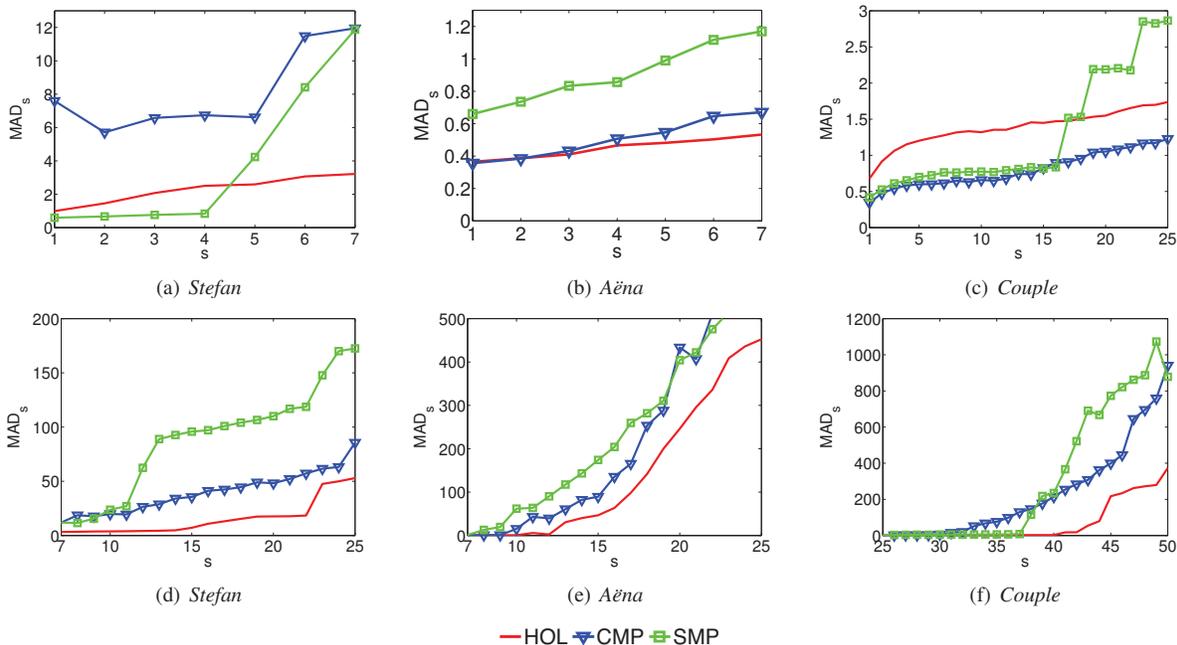


Figure 5. For different object representations the time-averaged MAD_k^s values have been evaluated for different ranges s of the initialisation region.

small MAD_s values indicating a low scatter of the mean shift tracking results (high accuracy).

This result is caused by a superior comprehension of spatial colour information by the SMP object representation in contrast to CMP or the holistic object representation. A drawback of both multi-part object representations is, however, given by a higher sensitivity (less robustness) to larger initialisation regions ($s > 5$).

An advantage of a multi-part object representation is, however, not apparent for the *Aëna* sequence. Fig. 5(b) and Fig. 5(e) demonstrate a superior accuracy and robustness of the holistic object representation for different ranges s of the initialisation region. This outcome is explained by the nearly uniform spatial colour distribution of the target object which allows no exploitation of spatial information by a multi-part object representation. Furthermore, the high velocity of the target object causes a high sensitivity to mean shift initialisations for all object representations which can be observed by the rapid increase of the MAD_s values in Fig. 5(e).

A more representative example for spatial colour information which can be exploited by multi-part object representations is given by the *Couple* sequence. For small initialisation regions ($s < 15$) both multi-part object representations yield a higher tracking accuracy proved by a small scatter of the mean shift tracking results as illustrated in Fig. 5(c) and 5(f). Due to the less distinctive background clutter, the multi-part object representations are more robust to poor mean shift initialisations in the *Couple* sequence than in case of the *Stefan* sequence. The robustness is more distinctive for

the SMP object representation since it subdivides a target object only in vertical direction which is better suited for the target object of the *Couple* sequence.

For the sake of completeness, the temporal MAD_k^s is plotted for a fixed range of the initialisation region ($s = 10$) and a temporal segment of the *Couple* sequence in Fig. 6. This allows to identify key scenes for which certain mean shift object representations perform less accurate or less robust or which are more difficult for mean shift tracking in general. For example, the global peak in Fig. 6 corresponds to the period shortly after a partial occlusion where a more distinct scatter of the mean shift tracking results exists.

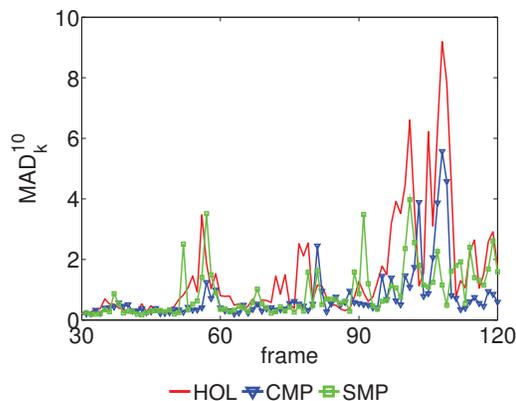


Figure 6. Progress of MAD_k^{10} values over time for the *Couple* sequence and different object representations.

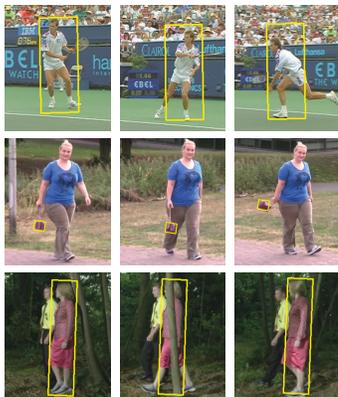


Figure 7. From top to bottom: Exemplary frames of the test sequences *Stefan*, *Aëna* and *Couple*.

V. CONCLUSION AND FUTURE WORK

We have introduced a novel approach for the quantitative evaluation of object representations in mean shift tracking, which does not rely on any ground truth data. Particularly, it has been demonstrated that the usage of multi-part object representations can improve the mean shift tracking accuracy. A drawback of multi-part object representations is, however, their higher sensitivity to poor mean shift initialisations which may occur during the tracking of highly agile target objects. A possible remedy is the combination of mean shift tracking with supportive algorithms, such as Kalman filter or Particle filter, which allow an initial prediction of the target object state.

The used object representation should, however, capture well the spatial colour distribution of the target object. Future work will therefore be focused on the development of an adaptive multi-part object representations that automatically adapts to the varying appearance of the object. As this online learning comprises the risk of a drift towards an invalid object representation a combination with a segmentation approach might also be promising.

ACKNOWLEDGEMENT

This work was co-funded by the German federal state North Rhine Westphalia (NRW) and the European Union (European Regional Development Fund: Investing In Your Future)

REFERENCES

- [1] R. Mohan, J. Smith, and C. Li, "Adapting multimedia internet content for universal access," *IEEE Transactions on Multimedia*, vol. 1, no. 1, 1999, pp. 104–114.
- [2] F. Pereira and I. Burnett, "Universal multimedia experiences for tomorrow," *Signal Processing Magazine, IEEE*, vol. 20, no. 2, 2003, pp. 63–73.
- [3] S. De Bruyne, P. Hosten, C. Concolato, M. Asbach, J. De Cock, M. Unger, J. Le Feuvre, and R. Van de Walle, "Annotation based personalized adaptation and presentation of videos for mobile applications," *Multimedia Tools and Applications*, vol. 55, no. 2, 2011, pp. 307–331.
- [4] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys (CSUR)*, vol. 38, no. 4, 2006, pp. 1–45.
- [5] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, 2003, pp. 564–577.
- [6] E. Maggio and A. Cavallaro, "Multi-part target representation for color tracking," in *Proceedings of IEEE International Conference on Image Processing (ICIP'05)*, vol. 1, 2005, pp. 729–732.
- [7] V. Parameswaran, V. Ramesh, and I. Zoghliami, "Tunable kernels for tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, 2006, pp. 2179–2186.
- [8] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, 2009, pp. 319–336.
- [9] C. J. Needham and R. D. Boyle, "Performance evaluation metrics and statistics for positional tracker evaluation," in *Proceedings of International Conference on Computer Vision Systems (ICVS'03)*, 2003, pp. 278–289.
- [10] D. Caulfield and K. Dawson-Howe, "Evaluation of multi-part models for mean-shift tracking," in *Proceedings of International Machine Vision and Image Processing Conference (IMVIP'08)*, 2008, pp. 77–82.
- [11] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, 1995, pp. 790–799.
- [12] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, 2002, pp. 603–619.
- [13] G. Bradski, "Real time face and object tracking as a component of a perceptual user interface," in *Proceedings of 4th IEEE Workshop on Applications of Computer Vision (WACV'98)*, 1998, pp. 214–219.
- [14] L. Wang, C. Pan, and S. Xiang, "Mean-shift tracking algorithm with weight fusion strategy," in *Proceedings of IEEE International Conference on Image Processing (ICIP'11)*, 2011, pp. 473–476.
- [15] A. Yilmaz, "Kernel-based object tracking using asymmetric kernels with adaptive scale and orientation selection," *Machine Vision and Applications*, vol. 22, no. 2, 2011, pp. 255–268.
- [16] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, 2002, pp. 174–188.

Quantistic Approach for Classification of Images

Federico F. Barresi, Giuseppe Battista, Jacopo Pellegrino

Dipartimento di Fisica

Università di Torino

Torino, Italy

e-mail: federico.barresi@studenti.unito.it,

giuseppe.battista@studenti.unito.it,

jacopo.pellegrino463@studenti.unito.it

Walter Allasia

Research Department

EURIX

Torino, Italy

e-mail: allasia@eurix.it

Abstract—This paper describes a novel approach for managing low level descriptors of images in order to allowing automatic classification and similarity searches. Several works have been made in this field, mostly making use of vector spaces and classical mathematical approaches. The focus of this paper is to investigate the more sophisticated formalisms of Quantum Mechanics that allows to manage images as *quantum states*. In order to check our theoretical method, we have considered a simple set of low level descriptors of images, the Hue Saturation Values (HSV). On the one hand, they are obviously not exhaustive and limited; but, on the other hand, they are enough for demonstrating that low level image descriptors can be represented such as Functions on Hilbert spaces. Since the distance between two colors in HSV space coincides with the human eye perception, its evaluation enabled us to collect results on similarity even if we were making use of few descriptors around HSV such as the MPEG-7 Visual Descriptors ColorLayout and DominantColor. The reader will not find a deep proof of a novel similarity technique applied to large image samples. Instead, she can assess the value of adopting Quantum Mechanics formalisms, translating thresholds of classical vector space distances onto probability density functions of low level image descriptors, such as HSV.

Keywords—Information retrieval; Image indexing; Image feature management; Novel spaces for indexing low level image features; Similarity comparison; Perception of similarity of image features

I. INTRODUCTION

The similarity between images is one of the most important topics in computer vision. Although it is very easy for people to decide if two images are similar, the same can not be said for computers that must learn to see like humans. Each image is perceived by the computer as a set of pixels ordered according to their position but not to their color, each pixel does not take into consideration the color of its neighbors unless any feature is applied. Therefore, we tried to take advantage of the position and color information that the computer is able to provide for each pixel in order to translate them into the kind of information that a human eye would receive observing the image. For this purpose, several MPEG-7 descriptors [12] have been implemented which allow to extract certain features from an image. In our work, we focused the attention to a couple of them, ColorLayout and DominantColor, in order to propose an alternative implementation of these two descriptors based on a new quantistic approach and on the HSV color space. The ColorLayout is rewritten as ColorDistribution to underline

that the main difference is that for each area we consider a distribution of color rather than an average color. The similarity between images is then given by the comparison between the distributions, for each corresponding area of the images to be compared, which returns the percentage of similarity. Concerning the HSVDominantColor, the division of the image into areas is introduced and for each area three dominant colors are estimated. To obtain a percentage of similarity, HSVDominantColor calculates the distance in the HSV cone between colors of the areas of the images to be compared.

The paper is organized as in the following: Section III describes the approach we applied to implement the application, which is presented in Section IV. Section V shows the experimental results achieved with a sample made up of about 3000 images belonging to the collection provided for the research purposes by IRMA project [5].

II. STATE OF THE ART

MPEG-7 [11], [12], [15] formally named *Multimedia Content Description Interface*, is developed by MPEG (*Moving Pictures Experts Group*) [7]. It is one of the most common standard for describing multimedia contents that provides a rich set of multimedia content description tools for applications ranging from content management, organization, navigation and automated processing.

MPEG-7 Visual [8] standardizes the description tools to describe video and image content. The Visual Descriptors are based on visual features that allow to estimate similarity in images or video. Therefore, they can be used to search and filter images and videos based on several visual features like color, texture, object shape, object motion and camera motion. Among Color Descriptors, we have taken into account ColorLayout and DominantColor due to previous research work at EURIX S.r.l. [20].

ColorLayout is a low-level descriptor that extracts information about color and its position within the image. This descriptor divides the image into 64 areas to which associates a representative color and then compares it with the color of the corresponding area of another image by calculating the Euclidean distance in the RGB color space [13]. The representative color can be evaluated with any method. In our work, we followed the [8] standard recommendations using the average of the pixel colors in a block as the corresponding representative color.

DominantColor is also a low-level descriptor that extracts the most present color within the image without caring at its spatial distribution. As **ColorLayout**, comparison between images is due to the calculation of the distance between the two colors in the RGB space multiplied by constant factors taking into account the spatial coherence [18].

In order to apply for a search on images and videos, we have tried to change the nature of these descriptors by introducing the formalism of Quantum Mechanics [16]. In this kind of approach, each object becomes a normalized vector $|x\rangle$ in a real Hilbert space of finite dimension \mathcal{H} . The vector contains the answers to all possible queries [14] [17], each one represented by another vector $|y\rangle$. Usually, in Information Retrieval, similarity matching is accomplished by computing $\langle x|y\rangle$. In the following section, we describe how this approach has been implemented.

III. APPROACH AND METHODS

This work mainly aims at suggesting a novel kind of descriptors whose implementation takes into account the formalism of Quantum Mechanics. Another target is to improve the perceptiveness in order to model the way human eye perceives the similarity between images. It is necessary to switch from global information, such as average values, to more detailed information, as distributions of color occurrences. It is also important to describe colors in a space where the perceptive distance is preserved. We selected the HSV space.

Since we want to apply the formalism of Quantum Mechanics in a HSV space, we have implemented the already mentioned new descriptors **ColorDistribution** and **HSVDominantColor**.

A. The ColorDistribution

As described later in Section IV-B, we have split each image into 64 rectangular cells of equal area. Since we did not consider necessary the use of interpolation or any other solutions, partitioning inevitably means little loss of information because not every pixel of the image can be associated with one of the cells. For each cell, the R, G and B components of all the pixels are extracted, with an appropriate algorithm the conversion is carried out and the values of the components H, S and V are stored into arrays. From these lists, for each cell, relative frequency histograms of the three components are built. It is necessary to highlight that each component spans discreetly in its domain by unitary steps. This choice is reasonable since the variation of one unit of any of the three components is unnoticeable to the human eye and makes the histogram more fittable with a polynomial function. The histogram of H has 360 bins while the ones of S and V have 100 bins because of their percentual variation.

At this point, each histogram is interpolated with a polynomial function whose degree is set to 10 because this order allows us an acceptable level of flexibility. The 11 parameters of the polynomial are calculated from the fit with the method of least squares [10]. Each cell, therefore, is described by 3 polynomial functions of degree 10. Reiterating the process for all the cells of the image 3×64 functions are obtained, that describe the color distribution within the image. This is the

main difference of **ColorDistribution** compared to **ColorLayout** which returns 64 average colors in the RGB space.

Once the color distributions of 2 images to be compared is obtained, it must be expressed as an information about their similarity. The procedure we adopted is to make a comparison between the three distributions of the corresponding cells of the two images, for every cell. At this point, the formalism of Quantum Mechanics is applied: the distributions can be considered as normalized vectors in the Hilbert space and can be compared using the standard scalar product between functions. The result of this comparison is the probability that the two distributions coincide. The comparison through the scalar product is performed cell by cell and component by component in order to obtain, for the whole image, a percentage of compatibility for each component (see: Section IV-C).

B. The HSVDominantColor

As **ColorDistribution**, **HSVDominantColor** makes use of colors of the HSV space as well as the partitioning of the image into cells. In order to decide the actual dominant colors, we divided the HSV space into relevant fields obtained through the division of Hue into 6 areas, Saturation into 4 areas and Value into 5 areas. Each area is identified by its average value. This quantization allows us consider only $6 \times 5 \times 4$ colors which is fundamental for at least two reasons. First of all taking into account all the possible colors does not make sense and it is completely useless since we would find out that identical colors occurs rarely in an image, to gather them into a finite number of ranges seems a valid solution. On the other hand, human eye is not so sensitive to distinguish a unitary variation of any of the three component. For each cell, the 3 most frequent representative colors are calculated and can be considered as the dominant colors. We decided to split the **HSVDominantColor** space into 3 subsets in order to take into account any edges contained in a single cell. If a single dominant color is found this is repeated twice. If two dominant colors are found, the most frequent is repeated once. The procedure is iterated for every cell obtaining 64 triples of dominant colors. The comparison between two images is therefore the calculation of the distance between the corresponding dominant colors of cells that occupy the same position. From the three distances, the average distance for each cell is evaluated and the total percentage of compatibility between the two images is finally obtained as the total normalized average distance.

C. The HSV distance algorithm

In order to calculate the distance [9], [19] between two colors in the HSV space, we have introduced the following Algorithm 1. Firstly, it selects color whose Value is maximum and sets it as $Color_1$. The other color is set to $Color_2$. Then it projects $Color_2$ onto $Color_1$ plane, calculates the distance on that plane and through Carnot's theorem finds the distance in the HSV cone.

IV. IMPLEMENTATION AND TESTS

A. Software architecture

The software architecture diagram in [1] shows the class diagram structure of the application based on elementary classes such as *Point*, *Pixel*, *Color* and *Cell*. Starting from them, more

```

Require:  $Color_1 := (h_1, s_1, v_1), Color_2 := (h_2, s_2, v_2)$ 
if  $v_1 > v_2$  then
  {Colors swap}
   $c' := c_1$ 
   $c_1 := c_2$ 
   $c_2 := c'$ 
end if
 $\Delta h := |h_2 - h_1|$ 
 $\Delta v := \frac{|v_2 - v_1|}{100}$ 
  {Projection on  $v = v_2$ }
   $d_c := \frac{\Delta v}{\cos(\pi/4)}$ 
  {Distance on the plane  $v = v_2$ }
   $d_p := \frac{\sqrt{(\frac{s_1}{100})^2 + (\Delta v)^2 + 2 \cdot \frac{s_1}{100} \cdot \Delta v} + \sqrt{(\frac{s_2}{100})^2 - 2 \cdot (\frac{s_1}{100} + \Delta v) \cdot (\frac{s_2}{100} \cdot (\Delta h))}}{2}$ 
  {Distance between  $Color_1$  and  $Color_2$ }
return  $d := \frac{\sqrt{d_c^2 + d_p^2 - 2 \cdot d_c \cdot d_p \cdot \cos(\pi/4)}}{2}$ 
    
```

Algorithm 1. HSV distance

TABLE I. Waste percentage of pixels

Cells	Loss percentage %
64	1.20
144	1.97
256	2.61
400	3.42
576	3.94
784	4.31
1024	6.00

complex objects and new methods are implemented in order to perform the following steps:

- Step 1* Image management: the application reads two JPEG files as arguments
- Step 2* Partitioning (see: IV-B) and HSV conversion: a *Cell-Builder* object instantiates *Pixel* objects, containing information regarding color and position, and associates them with the corresponding cell according to its location within the image.
- Step 3* Image analysis: ColorDistribution and HSVDominantColor extract the features by iterating over the cells that the picture is divided into.
- Step 4* Comparison: the extracted features of the two images are compared using the scalar product for ColorDistribution and the HSV distance for HSVDominantColor.

B. Image partitioning

Since our descriptors must preserve spatial information it is necessary to split the images into cells. The image partitioning occurs through the division into 64 ($n \times n$) rectangular areas. Neither the vertical nor the horizontal size are in any case integer multiples of n . Hence a certain number of pixels can not be associated with any cell. Partitions with various values of n in steps of 4 are tested on a 36 images sample in order to evaluate and minimize the loss of pixels. Results are reported in Table I.

The choice of considering a number of cells greater than 64, currently used by cell-based descriptors, is due to the need

TABLE II. Survey results regarding Hue

ΔH %	Dissimilarity %
10	96.7
15	98.8
20	99.5
Mean:	98.3

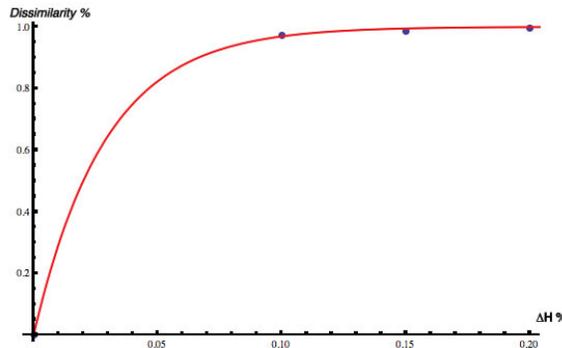


Fig. 1. Graph of Hue perception: It shows dissimilarity between images as function of variation of H

to operate with higher precision. Test demonstrates that a 64 cells partitioning minimizes the loss of pixels.

C. Survey

The ColorDistribution descriptor returns a compatibility percentage for each component of the HSV space. In order to estimate the weight of each component of HSV space we set up a survey asking people to recognize images having HSV slightly modified. The proper relevance of each component in determining the similarity between two images is established by the results of the survey that we have submitted to a sample with more than 600 people asked to indicate, among three or four altered images, the most similar to the original one. Survey results are reported in Figure 1 and in Table II and Table III.

Table II is showing the acquired dissimilarity perception whose mean value amounts to $H = 98\%$. The remaining 2% can be split into S and V according to the results shown in Table III.

According to the collected results, it is possible to assert that a small variation in Hue leads to perceive the image as very different from the original. We have chosen H as the most relevant component and its relevance has been set to 98%. The remaining 2% has been shared between Saturation (8.3%) and Value (91.7%).

D. Multi-threading

Comparison between polynomial functions, performed by the ColorDistribution, has been evaluated through the standard scalar product in Hilbert space. This process requires much more computing power than needed by ColorLayout descriptor. We needed to implement a Thread-Manager which distributes the computation on all the available *cores* of the computer running the application.

TABLE III. Survey results regarding Saturation and Value

Δ %	S Dissimilarity %	V Dissimilarity %
30	3.06	96.94
40	8.29	91.71
50	15.22	84.18
60	3.96	96.04
70	11.01	88.99
Mean:	8.31	91.69

E. Database indexing

Given that the aim of the designed application is to compare a query image to a sample set, it was necessary to implement an index in order to improve the query performances. We coded index choosing the first image and setting it as reference point of the features space in which each coordinate represents the similarity percentage with respect to a certain descriptor, in this case HSVDominantColor and ColorDistribution. For storing the indexed sample we made use of Apache Derby [2] database. The position in the feature space of a query image has been evaluated calculating its similarity with respect to the reference point image. The query returns all the images of the sample included in a range represented by a Gaussian function centered in the query image with standard deviation equal to $1 - t$, where t is the similarity threshold chosen by the user through the Graphical User Interface (GUI) shown in Section IV-G.

F. Total correlation

Taking into account that ColorDistribution has a greater precision with respect to the HSVDominantColor, we define a novel scalar product in the descriptors space according to the following unitary trace matrix:

$$D = \begin{pmatrix} .9 & 0 \\ 0 & .1 \end{pmatrix} \quad (1)$$

which considers the different relevance of each descriptor, defining the total correlation between two images by the formula:

$$\begin{aligned} Similarity = & \sqrt{0.9 \cdot ColorDistribution_{Corr}^2} + \\ & + \sqrt{0.1 \cdot HSVDominantColor_{Corr}^2}. \end{aligned} \quad (2)$$

We decided to make use of this particular scalar product in which HSVDominantColor is considered as correction of ColorDistribution. This scalar product is defined arbitrarily and it does not constitute a constraint to the discussion.

G. User interface

We implemented a simple GUI (Figure 2) to make the program *user-friendly* and let the user choose the sample set of images and the query. Once indexed the sample set of images, many queries can be performed quickly.

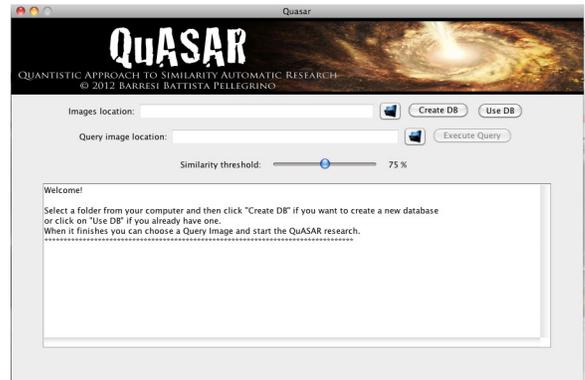


Fig. 2. Graphic User Interface implemented for the application QuASAR [21]

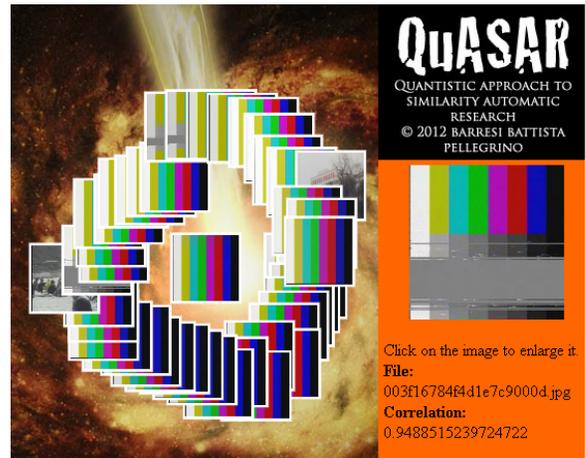


Fig. 3. Graphic interface showing the test results into an interactive HTML5 page

V. RESULTS

Once the software described in Section IV has completed the indexing process, an interactive HTML5 [4] page that we have implemented shows the result's thumbnails around the query image, Figure 4, in a concentric circumference proportional to the total correlation as in Figure 3.

The results are reported in Table IV. Our implementation obtains a recognition rate (*precision*) about 95% and sensitivity rate (*recall*) about 77% with a similarity threshold set to 80% within the 3195 images sampled.

VI. CONCLUSION AND FUTURE WORKS

This paper has proposed a novel technique for performing a similarity search on an indexed sample set of images making use of new low-level color descriptors. We have adopted a quantistic approach for solving the problem of features extraction and the executed tests have demonstrated a potential improvement on efficacy of queries.

Todarello's goal was to test a linear superimposition of n -dimensional tensors [16]. Nevertheless, our work focuses on the projection of visual information onto a Hilbert space whose elements are n -grade polynomial functions.

TABLE IV. Precision and Recall for the 3195 processed images

Query	ActualIm	Threshold	precision	recall
Fig. 4	53	70%	0.42	1.00
		80%	0.95	0.77
		90%	1.00	0.51
		95%	1.00	0.26

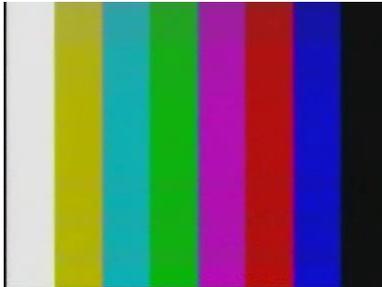


Fig. 4. Query Image used for testing the sample

The intent of this position paper is to put into practice and test the validity of the quantistic approach to similarity search we designed. Further development will consider a larger database of images, tests on the perceptive parameters obtained through survey, other classes of fitting functions and functional spaces, comparison of our results with other similarity techniques in order to define a threshold of performance. It is out of our aim, at this point, to focus onto experimental issues.

The proposed methodology can benefit the CBIR, because compared to the current techniques, making use of vector and the metric spaces where thresholds, usually evaluated experimentally, have to be applied, it enables a probabilistic approach allowing the superimposition of different results. We can foresee an improvement of the pseudo-relevance feedback querying multimedia databases.

Moreover, it is possible to implement other descriptors such as Shape (such as Textures, Edges) or Motion Descriptors in order to add more low level elements to evaluate for better image recognition. Each new descriptor can be represented by an axis in the features space.

Other MPEG-7 descriptors may be reimplemented with the formalism of Quantum Mechanics and the HSV color space described in this paper in order to enable image searches closer to the human being perception of similarity. The authors are analyzing the improvement of the retrieval results adopting more sophisticated visual descriptors as presented in [23].

Furthermore, in order to evaluate the effectiveness of the proposed methodology, it could be useful to make use of a generic publicly-available database [22], where ground-truth is available.

ACKNOWLEDGEMENT

As pointed out in Section I, we acknowledge the IRMA project for providing us with the sample set.

This work was partially supported by the international project FORGET-IT (grant no: 600826), which was funded

by the EC FP7 ICT collaborative research programme, call 2011.4.3.

REFERENCES

- [1] Software architecture diagram of application implemented, newton.ph.unito.it/~barresi/UML.png [retrieved: Feb 2013].
- [2] Apache derby, <http://db.apache.org/derby/> [retrieved: Feb 2013].
- [3] Eclipse, <http://www.eclipse.org/> [retrieved: Feb 2013].
- [4] Html5, <http://dev.w3.org/html5/spec/single-page.html> [retrieved: Feb 2013].
- [5] Irma project, www.progettoirma.it [retrieved: Feb 2013].
- [6] Java SE 1.6, <http://www.oracle.com/it/technologies/java/> [retrieved: Feb 2013].
- [7] Mpeg, <http://mpeg.chiariglione.org/> [retrieved: Feb 2013].
- [8] ISO/IEC 15938-3 fcd information technology - multimedia content description interface - part 3: Visual, Mar. 2003.
- [9] M. K. Agoston, Computer Graphics and Geometric Modeling, Springer, 2005.
- [10] G. Cannelli, Metodologie Sperimentali in Fisica, Edises, 2000.
- [11] B. S. Manjunath, Introduction to MPEG-7, Multimedia Content Description Interface, John Wiley and Sons, Ltd., Jun 2002.
- [12] J.-R. Ohm, L. Cieplinski, H. J. Kim, S. Krishnamachari, B. Manjunath, D. S. Messing, and A. Yamada, The MPEG-7 color descriptors In Introduction to MPEG-7: Multimedia Content Description Interface. Wiley, 2001.
- [13] L. G. Shapiro, and G. C. Stockman, Computer Vision, Pentice Hall, 2001.
- [14] A. Del Bimbo, Visual information retrieval, Morgan Kaufmann, 1999.
- [15] T. Sikora, The mpeg-7 visual standard for content description-an overview, *IEEE Trans. Circuits Syst. Video Techn.*, vol. 11, no. 6, 2001, pp. 696-702
- [16] E. M. Todarello, W. Allasia, and M. Stroppiana, MPEG-7 features in Hilbert spaces: Querying similar images with linear superpositions, Proceedings of the 5th international conference, In Quantum interaction Symposium, Jun. 2011, pp. 223-228, LNCS 7052, Berlin: Springer..
- [17] K. van Rijsbergen, The Geometry of Information Retrieval, Cambridge University Press, 2004.
- [18] S. Wang, L.-T. Chia, and D. Rajan, Image retrieval using dominant color descriptor, In Proceedings of the International Conference on Imaging Science, Systems and Technology, Las Vegas, Nevada, USA, Jun. 2003, pp. 107-110.
- [19] E. W. Weisstein, Cone, MathWorld-A Wolfram Web Resource, May 2012.
- [20] EURIX Group S.r.l., www.eurixgroup.com, [retrieved: Feb 2013].
- [21] QuASAR project, newton.ph.unito.it/~barresi/QuASAR/, [retrieved: Feb 2013].
- [22] P. Budikova, M. Batko, and P. Zezula, Evaluation Platform for Content-based Image Retrieval Systems, In International Conference on Theory and Practice of Digital Libraries 2011, LNCS 6966. Berlin: Springer.
- [23] O. Penati, E. Valle, and R. Torres, Comparative Study of Global Color and Texture Descriptors for Web Image Retrieval, *Journal of Visual Communication and Image Representation*, vol. 23, no. 2, Feb. 2012, pp. 359-380.

Efficient and Accurate Label Propagation on Large Graphs and Label Sets

Michele Covell and Shumeet Baluja

Google Research

Google Inc., Mountain View CA, USA

covell@google.com shumeet@google.com

Abstract—Many web-based application areas must infer label distributions starting from a small set of sparse, noisy labels. Examples include searching for, recommending, and advertising against image, audio, and video content. These labeling problems must handle millions of interconnected entities (users, domains, content segments) and thousands of competing labels (interests, tags, recommendations, topics). Previous work has shown that graph-based propagation can be very effective at finding the best label distribution across nodes, starting from partial information and a weighted-connection graph. In their work on video recommendations, Baluja et al. [1] showed high-quality results using *Adsorption*, a normalized propagation process. An important step in the original formulation of *Adsorption* was re-normalization of the label vectors associated with each node, between every propagation step. That interleaved normalization forced computation of all label distributions, in synchrony, in order to allow the normalization to be correctly determined. Interleaved normalization also prevented use of standard linear-algebra methods, like stabilized bi-conjugate gradient descent (*BiCGStab*) and Gaussian elimination. This paper presents a method that replaces the interleaved normalization with a single pre-normalization, done once before the main propagation process starts, allowing use of selective label computation (*label slicing*) as well as large-matrix-solution methods. As a result, much larger graphs and label sets can be handled than in the original formulation and more accurate solutions can be found in fewer propagation steps. We also report results from using pre-normalized *Adsorption* in topic labeling for web domains, using label slicing and *BiCGStab*.

Keywords—graph propagation, large-scale labeling, stabilized bi-conjugate gradient descent, Gaussian elimination, topic discovery, web domains.

I. INTRODUCTION

Many different approaches have recently been proposed to label propagation across weighted graphs of nodes [1,2,3,4,5,6]. These applications share the characteristics of having a limited amount of label data, often of uneven quality, associated with a large graph of weighted connections between many nodes, some unlabeled and some partially labeled.

We build on the work done by Zhu and Ghahramani [2] and Baluja et al. [1]. The Baluja paper described *Adsorption*, a graph-based approach to estimating label distributions, which was applied to providing YouTube video recommendations. The resulting top-pick recommendation was more accurate than the next-best alternative algorithm

for all users who had watched 3 or more previous videos, with accuracy improvements of up to 100% for the most frequent watchers. In *Adsorption*, each node (e.g., each video for which we are building a recommendation list) has a limited capacity for labels (e.g., the proposed recommendations for that video). Baluja et al. [1] enforce this constraint by interleaving a normalization step at each node, in between every propagation step. Without this normalization, the solution is not guaranteed to converge.

The interleaved normalization step is needed for convergence but prevents label slicing: under the original formulation, we cannot find the estimated distribution of a subset of labels without solving for the full set of labels first. Furthermore, the interleaved normalization prevents the use of most standard linear-algebra techniques, such as Gaussian elimination of nodes that are not of direct interest (though they still are needed for their effect on the remainder of the graph). Additionally, methods for rapid convergence to the final solution, such as stabilized bi-conjugate gradient descent (*BiCGStab*), cannot be used in the original formulation.

This paper presents a formula for pre-normalizing the *Adsorption* graph and label weights, such that there is no need for interleaved normalization (Section III). With this, we can use *BiCGStab* and Gaussian elimination. Our graph size contains more than 10 million nodes and 4 billion inter-connections (i.e., more than 10 million rows and more than 4 billion non-zero entries in the corresponding matrix), which is more than we can reasonably handle in straightforward implementations of these techniques. Instead, we use implementations of *BiCGStab* and Gaussian elimination in the MapReduce framework. We describe these implementations briefly, in Sections IV and V. Finally, in Section VI, we present our results on topic labeling of web domains, using a graph based on shared keywords between pages across the domains. We start the paper with a recap of the original *Adsorption* application and mathematical description, in Section II.

II. ADSORPTION (WITH INTERLEAVED NORMALIZATION)

The original formulation of *Adsorption* [1] can be described as an iteration using two systems of equations:

$$\underline{\tilde{X}}_{n+1} = \sigma \underline{X}_n + \beta \underline{W} \underline{X}_n + \left[\gamma \underline{L} \quad \delta \underline{1} \right] \quad (1)$$

$$\{\underline{X}_{n+1}\}_{j^*} = \{\underline{\tilde{X}}_{n+1}\}_{j^*} / \|\{\underline{\tilde{X}}_{n+1}\}_{j^*}\| \quad (2)$$

where double underlining indicates a matrix of values, a single underline is a vector, not-underlined values are

scalars, and the tilde indicates a not-normalized set of values. The matrix \underline{W} holds the connection weights with row i giving the incoming connections into the i 'th node. This matrix often is symmetric, to start with, but this property is not required and will be given up later to allow for pre-normalization. The matrix \underline{L} holds the weights of the injection label information. These are often noisy or incomplete label sets based on some prior information, with the graph propagation as a way to improve and expand these label sets. In \underline{L} , each label is associated with a column and the weights for the injection labels for the i 'th node of the graph in the i 'th row of the matrix. In addition to the true labels, in \underline{L} , Baluja et al. [1] add an *abandonment label*, represented in (1) by the appended column δ . The scalar δ can be thought of in many different ways: as the loss in certainty about any of the labels that are propagated for one hop in the graph; as the number of random walks through the graph that end with "abandonment", giving no final label set; as the regularization margin in the system of equations. The other scalars (σ , β , and γ) allow graph-wide balancing of the previous (same-node) labels, of the propagated neighbors' labels, and of the injection labels. Finally, the matrix \underline{X}_n is the label distribution estimate, with the i 'th row containing the estimated labels for the i 'th node, including as the last column the abandonment label. In this context, the node's abandonment weight provides a measure, at that node, of the label uncertainty.

Equation (1) creates a new *un-normalized* estimate of the steady-state label distribution across all the nodes using a weighted combination of the previous normalized estimate for the distribution (\underline{X}_n), of a graph-weighted propagated version of that same distribution ($\underline{W}\underline{X}_n$), of injection labels (\underline{L}), and of the abandonment label (δ). Equation (2) provides a normalized estimate of the label distribution, by dividing each row of the estimate from (1) by the L_1 norm of the full label set, including the abandonment label.

Iterating over (1) and (2) together is guaranteed to converge to a stable steady-state solution, as long as δ is greater than 0. Baluja et al. [1] used this algorithm to successfully provide video recommendations that, using a top-pick-accuracy measure, outperformed alternative approaches. Our goal is to provide a formulation for the same Adsorption algorithm that does not require per-propagation-step normalization, allowing us to use label slicing and standard linear-algebra tools.

III. PRE-NORMALIZED ADSORPTION

We achieve our goal of pre-normalized Adsorption by first assuming that all associations in our graph and in our label injection are non-negative. Specifically: $sign(\{\underline{X}_n\}_{ij}) \geq 0$, $sign(\{\underline{W}\}_{ij}) \geq 0$, and $sign(\{\underline{L}\}_{ij}) \geq 0$.

This non-negative assumption works well with the partial-information applications that are the most common ones in large-graph labeling formulations: for example, in video recommendation, we can say that two videos are often watched together, within a single viewing session, but it is

much more difficult to say that two videos are negatively associated (that watching one means you are significantly less likely to watch the other), since we seldom have enough training data to make such an assertion with any confidence.

For those applications where we do have confidence in negative label-to-node associations (negative values in \underline{L}), we can handle these by introducing a negated label column and using positive associations with the negated label where we would have otherwise used negative associations with the positive label. Handling negative node-to-node connections (negative values in \underline{W}) is also possible but we omit it here, since it is an uncommon use case (and is much more complicated).

Assuming we have non-negative values in our component matrices, we can consider the denominator of (2) in more detail:

$$\|\{\tilde{\underline{X}}_{n+1}\}_{i^*}\|_1 = \left\| \left\{ (\sigma \underline{I} + \beta \underline{W}) \underline{X}_n + \left[\gamma \underline{L} \quad \delta \right] \right\}_{i^*} \right\|_1 \quad (3)$$

$$= \sum_j \left(\sum_k \{ \sigma \underline{I} + \beta \underline{W} \}_{ik} \{ \underline{X}_n \}_{kj} \right) + \sum_j \left\{ \left[\gamma \underline{L} \quad \delta \right] \right\}_{ij} \quad (4)$$

$$= \sum_k \{ \sigma \underline{I} + \beta \underline{W} \}_{ik} \sum_j \{ \underline{X}_n \}_{kj} + \gamma \sum_j \{ \underline{L} \}_{ij} + \delta \quad (5)$$

$$= \sum_k \{ \sigma \underline{I} + \beta \underline{W} \}_{ik} \|\{ \underline{X}_n \}_{i^*}\|_1 + \delta \quad (6)$$

$$= \sigma + \beta \|\{ \underline{W} \}_{i^*}\|_1 + \gamma \|\{ \underline{L} \}_{i^*}\|_1 + \delta \quad (7)$$

Equation (3) simply provides the expansion of the L_1 row norm using the propagation (1). Equation (4) makes use of the non-negativity conditions that we are requiring, in order to remove the absolute values implied by the L_1 norm and expands the norm summation, as well as the summation implicit in the $\underline{W}\underline{X}_n$ matrix multiply. Equation (5) swaps the order of summation, allowing us to make use of the unit L_1 row norm for \underline{X}_n in (6). Simplifying the summations and noting the use of the row-norm definitions for \underline{L} and \underline{W} finally results in (7).

The useful property of (7) is that $\|\{\tilde{\underline{X}}_{n+1}\}_{i^*}\|_1$ depends only the initial combination weights and the row norms of \underline{L} and \underline{W} . We can use this property to pre-normalize by first defining

$$\lambda_i = \sigma + \beta \|\{ \underline{W} \}_{i^*}\|_1 + \gamma \|\{ \underline{L} \}_{i^*}\|_1 + \delta \quad (8)$$

$$\hat{\sigma}_i = \sigma / \lambda_i \quad \hat{\underline{\sigma}} = \text{diag}(\hat{\sigma}_i) \quad (9)$$

$$\hat{\beta}_i = \beta \|\{ \underline{W} \}_{i^*}\|_1 / \lambda_i \quad \hat{\underline{\beta}} = \text{diag}(\hat{\beta}_i) \quad (10)$$

$$\hat{\gamma}_i = \gamma \|\{ \underline{L} \}_{i^*}\|_1 / \lambda_i \quad \hat{\underline{\gamma}} = \text{diag}(\hat{\gamma}_i) \quad (11)$$

$$\hat{\delta}_i = \delta / \lambda_i \quad \hat{\underline{\delta}} = \text{vec}(\hat{\delta}_i) \quad (12)$$

and then using these new quantities in a pre-normalized Adsorption algorithm.

$$\underline{X}_{n+1} = \hat{\underline{\sigma}} \underline{X}_n + \hat{\underline{\beta}} \underline{W} \underline{X}_n + \left[\hat{\underline{\gamma}} \underline{L} \quad \hat{\underline{\delta}} \right] \quad (13)$$

Note that direct use of (13) is exactly the power-iteration approach to finding the solution (used in [1]) and will give the same solutions at every iteration as the combination of (1) and (2): the pre-normalization has the exact same effect, even though it is only done once, as the interleaved normalizations. Equation (13), therefore, also is guaranteed to converge to a stable solution, just as the original Adsorption algorithm is guaranteed. The advantage is that we do not need to normalize at each step and, as a result, we can compute an incomplete set of labels, while still deriving the benefits of the full label set to limit belief within the set of labels that are interested in. This slicing directly reduces the computational costs by the same percentage as the percentage of dropped labels. Furthermore, with the use of (13) as the system of equations for which we want a solution, we can use standard linear-algebra tools, like BiCGStab (for faster convergence) and Gaussian elimination (for shrinking our graph matrix). We discuss these algorithms and their large-graph implementations next.

IV. MAP-REDUCE FORMULATION OF STABILIZED BI-CONJUGATE GRADIENT DESCENT (BiCGSTAB)

In [1], Baluja et al. implicitly use power iteration to solve their system of constraints. For symmetric systems of constraints, gradient-descent methods can find solutions in fewer iterations, for any given level of accuracy (as measured by the average residual error). However, due to the pre-normalization of Adsorption, we no longer have a symmetric matrix, and must move to bi-conjugate gradient approaches. Since the most direct generalization (biconjugate gradient descent) is not numerically stable, we focus on stabilized biconjugate gradient descent [7], which has been shown to converge more uniformly than power iteration, without the numerical issues (not-stabilized) bi-conjugate gradient descent. We ran several simulations using power iteration and BiCGStab, based on random graph matrices with the same level of regularization as we expect to see through the abandonment variable in our true graphs. In these tests, when the graph matrix and the beginning label estimates were non-sparse, on average, BiCGStab converged to the correct solution 12 times faster than the power-iteration method (e.g., BiCGStab would converge in two iterations, requiring only 5 graph-matrix multiplies, while power iteration would require 60 iterations, needing 60 graph-matrix multiplies to converge to the same level of accuracy).

When the graph matrix and the beginning label estimates were sparse, there were similar differences in the rate of convergence, away from the “wavefront boundary”. We use the term *wavefront* to emphasize that (for both power iteration and BiCGStab), updates are done in such a way that non-zero values propagate through the graph according to the neighborhood connections. When the labels are sparsely injected, non-zero values move in a “wave”, outward from non-zero areas into areas that were zero (due to sparseness). Both power iteration and BiCGStab rely on the graph matrix to determine the label-estimate update, so both have their non-zero wavefronts progress in the same way.

Due to the size of the graph over which we will be operating, we implemented BiCGStab using three MapReduce [8] stages per iteration. Using the notation from the Wikipedia article on BiCGStab [9], we have a distinct set of vectors for each of the labels on which we want to estimate the final distribution. We arrive at the BiCGStab components \underline{A} and \underline{b} (at least conceptually) by separating $\hat{\underline{y}}\underline{L}$ into columns corresponding to \underline{b} , by separating \underline{X}_n into columns corresponding to \underline{x}_n and by using

$$\underline{A} = \underline{I} - \hat{\underline{\sigma}} - \hat{\underline{\beta}}\underline{W} \quad (14)$$

We select an initial *shadow direction* $\hat{\underline{r}}_0$ for each column aligned with its first-pass residual vector, \underline{r}_0 . Note that computing the first-pass residual vector takes one MapReduce to compute $\underline{r}_0 = \underline{b} - \underline{A}\underline{x}_0$. (For our applications, \underline{b} itself is often a good initial estimate for \underline{x} .) It is this separate estimation of each column (where each column corresponds to a single label) that makes label slicing so simple and powerful in combination with BiCGStab.

Unlike [9], we mark all our auxiliary variables with the iteration on which they were computed, since this makes our Reduce processing more uniform and reliable: therefore, we use α_n , \underline{s}_n and \underline{t}_n here (instead of their un-versioned form from [9]). To allow the remaining framework to operate smoothly, starting from the initialization (the 0'th pass), we also use the settings for our auxiliary variables that are suggested in [9], namely: $\rho_0 = \alpha = \omega_0 = 1$ and $\underline{v}_0 = \underline{p}_0 = \underline{0}$.

For all iterations after this initialization, there are 3 MapReduce stages: (A) updating the search direction and its projection through \underline{A} ; (B) updating the shadow direction and its projection through \underline{A} ; and (C) combining the computed components to give a new state estimate and residual.

For all three MapReduce stages, the reduce processing is the same: from the set of inputs computed in the Map stage, as well as the inputs passed directly through to the Reducer from previous stages or iterations, keep and combine the results for each variable (auxiliary variables, residual, and state estimate) that is marked with the highest iteration number observed for that variable, and throw away earlier versions.

A. Updating the search direction and its projection

1) Map (shared) context:

- a. From initial selection: $\hat{\underline{r}}_0$
- b. From previous iteration:

$$\rho_{n-1}, \alpha_{n-1}, \omega_{i-1}, \underline{r}_{n-1}, \underline{v}_{n-1}, \underline{p}_{n-1}$$

- c. From pre-map computation:

$$\rho_n = \langle \hat{\underline{r}}_0, \underline{r}_{n-1} \rangle$$

$$\underline{p}_n = \underline{r}_{n-1} + \left(\frac{\rho_n}{\rho_{n-1}} \right) \left(\frac{\alpha_{n-1}}{\omega_{n-1}} \right) (\underline{p}_{n-1} - \omega_{n-1} \underline{\eta}_{n-1})$$

2) Map computation:

$$\text{For each row in } \underline{A}, \text{ compute } \{ \underline{\eta}_n \}_i = \{ \underline{A} \}_i^* \underline{p}_n$$

B. Updating the shadow direction and its projection

1) Map (shared) context:

- a. From initial selection: \hat{r}_0
- b. From previous iteration: r_{n-1}
- c. From previous stage of current iteration:
 ρ_n, η_n
- d. From pre-map computation:
 $\alpha_n = \rho_n / \langle \hat{r}_0, \eta_n \rangle$
 $s_n = r_{n-1} - \alpha_n \eta_n$

2) *Map computation:*

For each row in \underline{A} , compute $\{t_n\}_i = \{\underline{A}\}_{i^*} s_n$

C. Combining components for residual and state estimates

1) *Map (shared) context:*

- a. From previous iteration: x_{n-1}
- b. From previous stages of current iteration:
 α_n, s_n, t_n, p_n

2) *Map computation: For each label, compute*

$$\omega_n = \langle s_n, t_n \rangle / \langle t_n, t_n \rangle$$

$$x_n = x_{n-1} + \alpha_n p_n + \omega_n s_n$$

$$r_n = s_n - \omega_n t_n$$

V. MAPREDUCE FORMULATION OF GAUSSIAN ELIMINATION

Label slicing allows us to compute our distributions on the subset of labels that are of most interest, while still benefiting from the constraints effectively imposed by the full label set. In a similar way, Gaussian elimination allows us to compute our distribution on a subset of nodes (domains), while still benefiting from the indirect interconnections that are formed through the nodes that we do not want to explicitly include in our calculation. The computational savings provided by Gaussian elimination is linear with the percentage reduction in the number of graph connections. In addition, Gaussian elimination can speed up convergence, by effectively increasing the wavefront-propagation speed through those parts of the graph that were originally connected via the eliminated nodes.

Gaussian elimination is much simpler to implement in the MapReduce framework than BiCGStab, requiring only a single stage and capable of handling elimination of multiple nodes per run. The Reduce processing in the MapReduce is a straight pass-through of the outputs from the map stage.

To make the description more concise, define

$$\underline{A}_{keep} = \{\underline{A}\}_{i^*} \quad \underline{L}_{y_{keep}} = \{\hat{y}L\}_{i^*} \quad i \in \left\{ \begin{array}{l} \text{nodes} \\ \text{to be kept} \end{array} \right\}$$

$$\underline{A}_{remove} = \{\underline{A}\}_{j^*} \quad \underline{L}_{y_{remove}} = \{\hat{y}L\}_{j^*} \quad j \in \left\{ \begin{array}{l} \text{nodes to be} \\ \text{eliminated} \end{array} \right\}$$

Using this notation, the map processing is

1) *Map (shared) context:*

From stored representation:

$$\underline{A}_{remove}, \underline{L}_{y_{remove}}$$

2) *Map computation: For each row, i, in \underline{A}_{keep} and $\underline{L}_{y_{keep}}$*

a) Initialize

$$\underline{\tilde{A}}_{keep} = \underline{A}_{keep}, \quad \underline{\tilde{A}}_{remove} = \underline{A}_{remove}$$

$$\underline{\tilde{L}}_{y_{keep}} = \underline{L}_{y_{keep}}, \quad \underline{\tilde{L}}_{y_{remove}} = \underline{L}_{y_{remove}}$$

b) Compute the pivot strength, π_{ij} , for each $j \in \{\text{nodes to be eliminated}\}$:

$$\pi_{ij} = \left\{ \underline{\tilde{A}}_{keep} \right\}_{ij} / \left\{ \underline{\tilde{A}}_{remove} \right\}_{jj}$$

and select the elimination node, \tilde{j} , with the smallest amplitude $|\pi_{ij}|$

c) Eliminate all non-zero entries in the \tilde{j} 'th column in $\left\{ \underline{\tilde{A}}_{keep} \right\}_{i^*}$ and $\underline{\tilde{A}}_{remove}$, with matched operations on $\left\{ \underline{\tilde{L}}_{y_{keep}} \right\}_{i^*}$ and $\underline{\tilde{L}}_{y_{remove}}$:

$$\left\{ \underline{\tilde{A}}_{keep} \right\}_{ik} \leftarrow \left\{ \underline{\tilde{A}}_{keep} \right\}_{ik} - \pi_{ij} \left\{ \underline{\tilde{A}}_{remove} \right\}_{jk}$$

$$\left\{ \underline{\tilde{L}}_{y_{keep}} \right\}_{ik} \leftarrow \left\{ \underline{\tilde{L}}_{y_{keep}} \right\}_{ik} - \pi_{ij} \left\{ \underline{\tilde{L}}_{y_{remove}} \right\}_{jk}$$

$$\left\{ \underline{\tilde{A}}_{remove} \right\}_{nk} \leftarrow \left\{ \underline{\tilde{A}}_{remove} \right\}_{nk} - \tilde{\pi}_{nj} \left\{ \underline{\tilde{A}}_{remove} \right\}_{jk} \quad \forall n \neq \tilde{j}$$

$$\left\{ \underline{\tilde{L}}_{y_{remove}} \right\}_{nk} \leftarrow \left\{ \underline{\tilde{L}}_{y_{remove}} \right\}_{nk} - \tilde{\pi}_{nj} \left\{ \underline{\tilde{L}}_{y_{remove}} \right\}_{jk} \quad \forall n \neq \tilde{j}$$

with $\tilde{\pi}_{nj} = \left\{ \underline{\tilde{A}}_{remove} \right\}_{nj} / \left\{ \underline{\tilde{A}}_{remove} \right\}_{jj}$

d) Remove row \tilde{j} from $\underline{\tilde{A}}_{remove}, \underline{\tilde{L}}_{y_{remove}}$

e) Repeat (b), (c), and (d), until there are no more rows (nodes) to be removed.

f) Output $\left\{ \underline{\tilde{A}}_{keep} \right\}_{i^*}$ and $\left\{ \underline{\tilde{L}}_{y_{keep}} \right\}_{i^*}$

VI. LARGE-SCALE DOMAIN-LEVEL TOPIC LABELING

Baluja et al. [1] already showed the usefulness of the Adsorption approach in video recommendations. The pre-normalized Adsorption algorithm provides identical results at a fraction of the computational cost using the new formulation with label slicing, Gaussian elimination, and BiCGStab. The final computational cost is reduced by the product of the savings of all three approaches (label slicing, BiCGStab and Gaussian elimination).

For this paper, we explored using pre-normalized Adsorption for topic labeling on web domains, for search and advertising. Many page urls, and even whole domains, are poorly classified by standard topic-analysis approaches, due to having little in the way of machine-understandable content to classify. A standard example of this problem are domains that primarily host images or video – while the page url can be examined for clues to the topic, as well as the linked-to urls, the results are impoverished and noisy. If we can improve the topic labeling, we could more accurately index these pages for search and for content-matched advertisement.

Specifically, we created a graph with domains as nodes and a measure of shared searches for cross-domain pairs of urls as weighted connections between nodes. Our measure looked at, for each search term, the click rates for each url served in the results and set the strength of the url-url-term triple to the lower of the click rates between the paired urls. The connection weight between pairs of urls is the sum over all triples that terminate at those two urls. To aggregate from url-pair connections, up to domain-pair connections, we sum across those url-pair connections where the first of the pair of urls is from the first domain and the second is from the second domain. Similarly, our injection labeling is based on combining topic analysis of the urls within the domain, dropping those topics that were based on keywords that showed too much within-domain variance in their strength. We aggregate the link and topic-label strength up to the domain level to improve coverage and reliability of our graph connections. Even with this aggregation of urls to domain-level nodes and filtering of keyword labels to within-domain-stable sets, our initial data provides a graph of about 13 million domains (nodes), with about 4 billion node-to-node connections based on analysis of more than 253 million search terms. Our topic analysis provides more than 4,500 general topics, using traditional text-based classification.

From this set of 4,500 topics, we focused on 71 commercial topics (see Fig. 1 for examples). The computational savings (over the original Adsorption approach) for the label slicing alone was a factor of 63 times. We do not include this savings in the remainder of this discussion, since it is available to both power iteration and BiCGstab, as long as we are using the pre-normalized Adsorption formulation. That said, it is the most significant source of computational savings, compared to the original work [1].

We ran this set of 71 labels through two iterations of BiCGstab (5 graph-matrix multiplies) and through 70 iterations of the power method, both starting from the same initial estimate. Fig. 2 shows the size of the per-node residual for BiCGstab on these labels (using an L_1 norm). As with our small-scale simulations, at the end of our second iteration, the not-insignificant residuals occurred at the 3% of the nodes that were at the “wavefront boundary” of one or more of the topic labels. This level of convergence, with just 5 matrix multiplies, is not seen in the power-iteration

- Clothing
 - Women’s, Men’s, Children’s
 - Athletic, Casual, Formal, Outerwear, Sleepwear
 - Shoes, Boots
- Accessories
 - Jewelry, Watches, Purses
- Toys
 - Building Toys, Dolls, Stuffed Animals, Ride-on Toys
- Gifts
 - Flowers, Cards, Party Items, Holiday Items
- Discounts
 - Coupons, Loyalty Cards

Figure 1: Examples from selected 71 commercial topics.

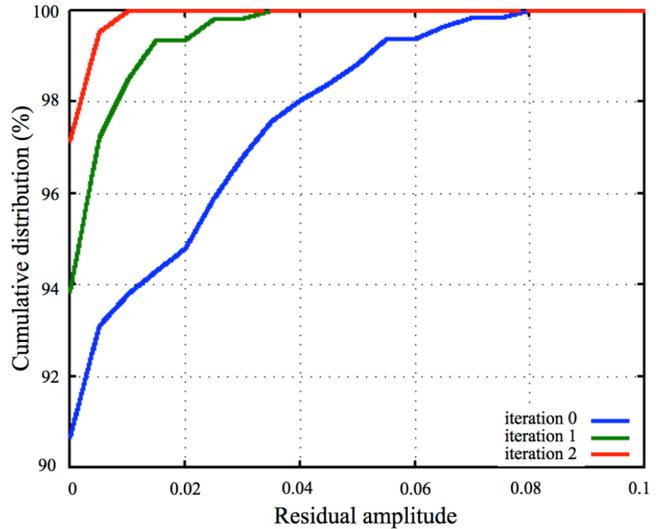


Figure 2: Cumulative residual distribution (by iteration).

solution until the 62th iteration (an additional savings of nearly 12.5 times).

Since the goal of our label propagation is to increase the richness and extent of the topic labeling on poorly labeled (or unlabeled) domains without over-extending into domains that are not related to our commercial subset, it is helpful to look at the statistics summarized in Fig. 3 through 5.

Fig. 3 gives a measure of the richness of our labels on commercial domains and how that richness increases as a function of iteration. The plot shows the percentages of domains by how many commercial-topic labels are seen on that domain. If a domain is commercial, the more commercial labels that are associated with the domain, the richer the topic description. As shown by the plots, our injection labels (those given by topic analysis) within each domain provides sparse topic labels, with the largest percentage of commercial domains having only one label.

Distribution of number of labels on each domain (by iteration)

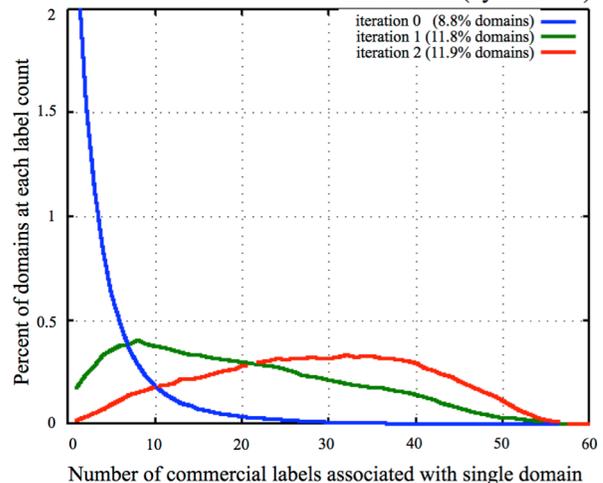


Figure 3: Node-level coherence of commercial labels.

Since our 71 commercial topics are actually a hierarchical set, this sparseness is unlikely to be correct for most domains. By the end of the second iteration, the mode of that distribution has moved to around 30 topic labels per commercial domain.

Also, the legend in Fig. 3 gives us the information needed to check that we are not just expanding the support of our commercial-topic labels indiscriminately across the full domain graph. The first iteration extends the support of the commercial labels by a third, from just under 9% of all domains to just under 12%, suggesting the addition of a subset of the unlabeled domains within the graph. After the first iteration, the support of the commercial-label set is effectively unchanged. This can be traced back to the effect of pre-normalizing on the full set of topic labels. Even though the non-commercial topics are not being explicitly computed in our iterations, they still have an effect, keeping the commercial labels from spreading onto distant (in the graph-connection sense) domains, as they otherwise would as the commercial wavefront progressed. This highlights both one of the main advantages of the original Adsorption as well as the most compelling advantage of the pre-normalized Adsorption. With the original Adsorption, each node has a limited capacity for supporting labels, thereby limiting propagation – but enforcing that limited capacity forced computation of all label distributions, not just the labels of interest. With pre-normalized Adsorption, there is still the per-node limited capacity for supporting labels, but we achieve that capacity limit by pre-normalizing, freeing us to compute only at that subset of labels that we are interested in, without having those labels spread unchecked.

Up to now, our analysis of our results has focused on the richness and extent of our commercial labels but not on the likely quality of the mix of labels that we are introducing onto commercial nodes. Since our topics are structured into a hierarchical framework, intuitively what we would like is to have each commercial site labeled mostly by closely related subsets of the available topics. We can use *dendrite distances* between the labels to capture this sense of closeness among the sets of labels associated with each domain node. As with standard dendrite measures, for each pair of labels on a domain, we count the number of hierarchical topic links that we have to go across in order to travel from one topic label to the other. We lengthen that distance by one for each generation that *both* labels have to travel back through, in order to penalize siblings more than grandparent-grandchild relations. As an example, if we need to calculate the distance between women’s jewelry and men’s clothing and we have the two tree branches “Jewelry -> Women’s Accessories -> Apparel” and “Men’s Clothing -> Apparel”, our dendrite distance measure would be 4: two (for “Women’s Jewelry” to “Apparel”) plus one (for “Men’s Clothing” to “Apparel”) plus one (for the one generation removal from direct descendent connection).

As a way to evaluate our label distributions on domains with 2 to 6 labels, we computed all pairwise dendrite distances within each domain and averaged them (again, on a per-domain basis). Due to the use of the topic hierarchy in our dendrite-distance measure, smaller distances amongst the

labels on a single domain correspond to more believable topic mixes. Fig. 4 shows our results, as function of iteration. When the initial topic labeling provides more than one label, it includes many dissimilar labels, with the mode of the dendrite average distance being up between 6 and 7. Our propagation reduces that average distance, filling in parent and children nodes, to give a mode that is just above one. While parents could always be filled in by knowing the hierarchical structure of our topic labels, the propagation graph is doing this without that knowledge – it is finding these associations purely through propagation of neighbor labels. (Furthermore, we could not use the tree-structure meta-information to fill in the correct children labels – if we blindly used the tree structure, we would get numerous nearby but irrelevant labels.) For this set of nodes, we are enriching the topic description without introducing unrelated labels. This measure of quality is a stringent one, since at no point do we use the dendrite structure to limit our propagation.

Fig. 5 shows a similar measure, for domains with more than 6 labels, again averaging the dendrite distances within each node. We did this separation between Fig. 4, for domains with 2-6 commercial labels, and Fig. 5, for domains with more than 6 commercial labels, since the dendrite distances across larger sets of labels, taken from the same hierarchy will have a larger minimum-average distance than will smaller sets of labels. For small sets, you can often find 2-6 labels, with all parent-child or sibling relationships with one another but, for large sets of labels, this is not possible and first and second cousin relationships become a major part of even the most compact set of labels. Same as with Fig. 4, Fig. 5 shows that the average dendrite distance decreases with each iteration, even on nodes with more than 6 labels. Since closely related sets of topic labels are more likely to be a full and accurate description of the domain topic, our topic labeling seems to be improved by our graph

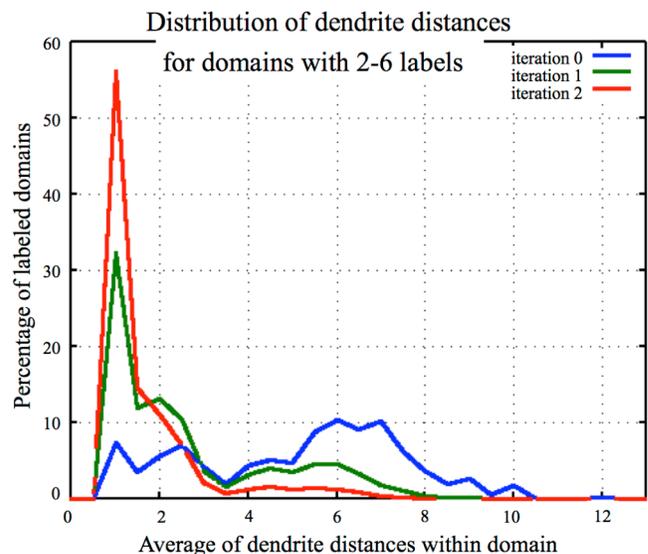


Figure 4: Dendrite topic-label distance on domains with 2-6 labels (by iteration).

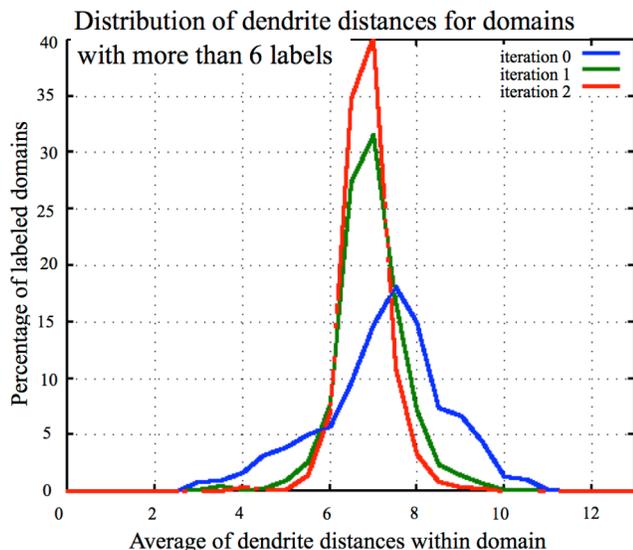


Figure 5: Dendrite topic-label distance on domains with more than 6 labels (by iteration).

propagation work.

All of the measurements conducted on the propagation of web labels on this large set of domains indicate an improvement in search indexing and content-matched advertising. In the future, we will expand these experiments in two directions. First, we will run live trials, with full user-facing experiments, to determine the quality improvement in the user experience. Second, we will increase our graph size and specificity by including individual urls, for those sites that have enough textual information to support that level of analysis.

VII. CONCLUSIONS

This paper improves the computational efficiency of Adsorption, a graph-based labeling approach that has already been shown to be highly effective. We do so by replacing propagation-interleaved normalization with pre-normalization, without changing the results provided by Adsorption. Specifically, if the power-method approach to finding a solution is used, as it was with Adsorption, the answers at every iteration will be exactly the same using either the original or the pre-normalized Adsorption. The advantage of the pre-normalized Adsorption is computational efficiency in determining the label distribution. With the pre-normalized version, we can use label slicing, to compute only those labels that are of direct

interest, without losing the beneficial belief-limiting characteristics of the full label set. Label slicing reduces the computational cost linearly with the percentage of dropped labels. Similarly, we can use Gaussian elimination, to compute the labels only on those nodes that are of direct interest, without losing the effects of the connections that occur indirectly through currently not-of-interest nodes. Finally, we can speed up convergence to the steady-state solution by a factor of 12 (in numbers of graph matrix multiples), by using stabilized biconjugate gradient descent, instead of power iteration. We also applied pre-normalized Adsorption to a new, large-scale application area, topic labeling on web domains, with promising results.

REFERENCES

- [1] S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly, "Video suggestion and discovery for YouTube: taking random walks through the view graph," Proc. International Conference on World Wide Web, ACM, April 2008, pp. 895-904.
- [2] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," CMU tech report, CMU-CALD-02-107, 2002.
- [3] P.P. Talukdar, J. Reisinger, M. Pasca, D. Ravichandran, R. Bhagat, and F. Pereira, "Weakly-supervised acquisition of labeled class instances using graph random walks," Proc. Conf. on Empirical Methods in Natural Language Processing, Assoc. Computational Linguistics, October 2008, pp. 582-590.
- [4] Y. Jing and S. Baluja, "Visual Rank: applying Page Rank to large-scale image search," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 30, November 2008, pp. 1877-1890.
- [5] J. Liu, W. Lai, X S. Hua, Y. Huang, and S. Li, "Video search re-ranking via multi-graph propagation," Proc. International Conference on Multimedia, ACM, September 2007, pp. 208-217.
- [6] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge, "Twitter polarity classification with label propagation over lexical links and the follower graph," Proc. Workshop on Unsupervised Learning in NLP, Assoc. Computational Linguistics, July 2011, pp. 53-63.
- [7] H. A. Van der Vorst, "Bi-CGSTAB: A Fast and Smoothly Converging Variant of BiCG for the Solution of Nonsymmetric Linear Systems," SIAM Journal on Scientific and Statistical Computing, vol. 13, March 1992, pp. 631-644.
- [8] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Symposium on Operating System Design and Implementation, December 2004, pp. 137-150.
- [9] Wikipedia, "Biconjugate Gradient Stabilized Method," http://en.wikipedia.org/wiki/Biconjugate_gradient_stabilized_method [retrieved February, 2013].

Video Retrieval by Learning Uncertainties in Concept Detection from Imbalanced Annotation Data

Kimiaki Shirahama
College of Information and Systems
Muroran Institute of Technology
shirahama@mmm.muroran-it.ac.jp

Kenji Kumabuchi and Kuniaki Uehara
Graduate School of System Informatics
Kobe University
kumabuchi@ai.cs.kobe-u.ac.jp, uehara@kobe-u.ac.jp

Abstract—Concept-based video retrieval retrieves shots relevant to a query based on detection results of concepts, such as *Person*, *Building* and *Car*. However, concept detection is ‘uncertain’ because even state-of-the-art methods cannot accurately detect various concepts. Thus, we introduce a video retrieval method, which models the uncertainty in the detection of each concept using ‘plausibilities’. A plausibility represents an upper bound of probability that the concept is present (or absent) in a shot. Using such plausibilities, false positive and false negative detections of the concept can be effectively managed. We derive plausibilities by estimating the density ratio between shots annotated with the concept’s presence and absence. However, annotating randomly sampled shots does not lead appropriate plausibilities due to the ‘imbalanced problem’. This means that the number of shots where the concept is present is generally much smaller than the number of shots where it is absent. To overcome this, a selective sampling method is developed to preferentially sample unannotated shots, which are similar to shots already annotated with the concept’s presence. Experimental results on TRECVID 2009 video data validates the effectiveness of derived plausibilities.

Keywords—Video retrieval; Uncertainty in concept detection; Dempster-Shafer theory; Imbalanced problem; Density ratio;

I. INTRODUCTION

Concept-based video retrieval is an approach which retrieves shots relevant to a query based on detection results of concepts, such as *Person*, *Car* and *Building*. Fig. 1 illustrates an overview of concept-based video retrieval. First of all, a *concept detector* is built to detect a concept’s presence in shots. Using such detectors, a shot is represented as a multi-dimensional vector consisting of *concept detection scores*, as shown in Fig. 1 (b). Each detection score represents the probability of a concept’s presence. Based on this shot representation, given example shots for a query, a retrieval model is constructed to discriminate between relevant and irrelevant shots to the query. In other words, detection scores for multiple concepts are fused into a single *relevance score*, which indicates the relevance of a shot to the query. Since the detector of a concept is built using a large amount of training shots, the concept can be robustly detected irrespective of its size, position and direction on the screen. Using concept detection scores as ‘intermediate’ features, concept-based video retrieval can achieve state-of-the-art retrieval

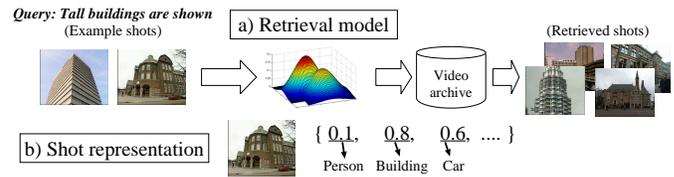


Figure 1. An overview of concept-based video retrieval.

performance [1], [2], [3].

However, even using most effective detectors, it is difficult to accurately detect any kind of concept. For example, TRECVID is an annual competition where concept detectors developed all over the world are benchmarked using large-scale video data [1]. At TRECVID 2012, the top-ranked detectors achieved high performances for concepts such as *Male_Person* and *Walking_Running* (with average precisions greater than 0.7). On the other hand, the detection of concepts like *Bicycling* and *Sitting_down* was difficult (with average precisions less than 0.1). Thus, relying on such *uncertain* concept detection significantly degrades retrieval performance.

We have been exploring a method which manages uncertainties in concept detection based on *Dempster–Shafer Theory* (DST) [4]. DST is a generalization of Bayesian theory, where a probability is not assigned to a variable, but instead to a subset of variables [5]. Specifically, we consider two singletons $\{P\}$ and $\{A\}$, which represent the presence and absence of a concept in a shot, respectively. In addition, $\{P, A\}$ represents the uncertainty of whether the concept is present or not. For the above three subsets, a *mass function* m defines masses $m(\{P\})$, $m(\{A\})$ and $m(\{P, A\})$. Here, $m(\{P\})$ and $m(\{A\})$ denote the probability that the concept is certainly present in a shot, and the probability that it is certainly absent, respectively, while $m(\{P, A\})$ denotes the probability that the concept is possibly present in the shot. Using these masses, DST can represent uncertainties much more effective than Bayesian theory, where the only way to represent an uncertainty is to assign the probability 0.5 to both variables P and A .

One big difficulty of DST is how to define a mass function. In our case, deriving the mass $m(\{P, A\})$ is substantially infeasible because it is very subjective to annotate shots with $\{P, A\}$ (*i.e.*, a concept's presence is uncertain). Thus, we avoid the mass function derivation by deforming the construction of a retrieval model based on the set-theoretic operation [6]. The retrieval model is constructed based on a *plausibility functions* pl , which is defined by combinations of masses: $pl(\{P\}) = m(\{P\}) + m(\{P, A\})$ and $pl(\{A\}) = m(\{A\}) + m(\{P, A\})$. The plausibility of a concept's presence $pl(\{P\})$ and the one of its absence $pl(\{A\})$ represent upper bound probabilities that it is present and absent in a shot, respectively. Thus, $pl(\{P\})$ is useful for recovering false negative detection of a concept, while $pl(\{A\})$ is useful for alleviating false positive detection.

We mainly address how to derive a plausibility function for each concept. Here, plausibilities of the concept's presence and absence of a shot are obtained based on the detection score of this shot. In our previous work [4], a plausibility function is derived by simple line approximation. However, plausibilities cannot be accurately characterized by lines. Thus, we develop a method which derives a plausibility function by estimating the *density ratio* [7] between shots annotated with a concept's presence and absence on the axis of detection scores. Intuitively, a large plausibility of the concept's presence (absence) should be associated with a detection score, around which the number of shots annotated with its presence (absence) is much larger than that of shots annotated with its absence (presence). Also, plausibilities of the concept's presence and absence should be similar at a detection score, around which numbers of shots annotated with its presence and absence are similar.

However, density estimation involves the *imbalanced problem* [8], meaning that the number of shots where a concept is present is generally much smaller than the number of shots where it is absent. Thus, when annotating randomly selected shots, almost all of them are annotated with the concept's absence, and their detection scores are nearly 0. As a result, an estimated density ratio is much biased towards the detection score 0. To balance numbers of shots annotated with the concept's presence and absence over detection scores, a *selective sampling* method is developed to preferentially select unannotated shots, which are similar to shots already annotated with the concept's presence.

This paper is organized as follows: The next section compares our method to existing ones, in terms of mass and plausibility derivation, and management in data uncertainty. Section 3 presents our video retrieval method, consisting of retrieval model construction based on DST, plausibility function derivation based on density estimation, and selective sampling. Experimental results in section 4 shows the effectiveness of plausibility functions derived by our method. Section 5 concludes this paper.

II. RELATED WORK

Although several methods for deriving mass and plausibility functions have been proposed, most of them assume special kinds of data like multivariate (transactional) data [9] and data with nested structures [10], or assume an underlying data distribution like Gaussian distribution [11]. Compared to this, we target multi-dimensional categorical data where each dimension represents a concept's presence (P) or absence (A), and does not have any prior knowledge about the data distribution. Hence, we derive plausibility functions in a 'data-driven' approach, where detection scores of shots for the concept are used as source data, and a part of these shots are manually annotated to indicate its presence or absence. In addition, none of existing methods consider the imbalanced problem.

Although an uncertainty in data is addressed in fields of data mining and machine learning, it is defined as a variance of observed values [12]. Compared to this, we define an uncertainty as the inaccuracy of determining the class label of a shot (*i.e.*, a concept's presence or absence). Thus, most of data mining and machine learning methods for uncertain data like [12], cannot be used to deal with uncertainties in this paper.

In concept-based video retrieval, many researchers have explored how to use concept detection scores to achieve accurate retrieval. For example, weighted linear combination is used in [2], [3], where the relevance score of a shot is computed as the sum of weighted detection scores for multiple concepts. Popular weighting methods use the lexical similarity between query terms and a concept, their co-occurrence, and detection scores of the concept in example shots. In [2], a discriminative classifier (*e.g.*, SVM) is built based on the shot representation with concept detection scores. Furthermore, in [13], shots are retrieved based on their similarity to example shots in terms of concept detection scores. To the best of our knowledge, except for our previous work [4], no existing works explicitly address uncertainties in concept detection.

Some researchers addressed uncertainties in combining concept detection results on different features (or modalities) [14], [15]. Such an uncertainty arises when conducting concept detection only using a single feature. In [14], concept detection results on different features are combined based on Portfolio theory, so that for each feature, the expected detection accuracy is maximized and the uncertainty is minimized. Note that this uncertainty is defined as the variance of the detection accuracy on the feature. Compared to this, an uncertainty in this paper means the inaccuracy of detecting a concept's presence or absence. Also, although DST is used in [15], mass function are hand-crafted, so their appropriateness for representing uncertainties is not guaranteed. In this paper, a plausibility function is derived by estimating the density ratio between shots annotated with a

concept's presence and absence. This statistically represents the uncertainty of the concept's presence or absence.

III. VIDEO RETRIEVAL BY MODELING UNCERTAINTIES IN CONCEPT DETECTION

This section describes our video retrieval method based on DST. First of all, detectors of various concepts are assumed to be already built using a large amount of shots annotated with various concepts' presence and absence. Under this condition, in order to derive a plausibility function for each concept, an additional set of annotated shots are created. In particular, considering the imbalanced problem, our selective sampling method is used to preferentially sample unannotated shots, which are similar to shots already annotated with the concept's presence. Then, a plausibility function is derived by estimating the density ratio between shots annotated with the concept's presence and absence. Finally, given example shots for a query, a retrieval model is constructed by incorporating plausibility functions of different concepts into maximum likelihood estimation.

Below, we first present our video retrieval model where a mass function is transformed into a plausibility function based on DST's set-theoretic operation. Then, our plausibility function derivation and selective sampling methods are described sequentially.

A. Video Retrieval Model based on DST

Our video retrieval model is constructed in the framework of Expectation-Maximization (EM) algorithm [6]. Let $x_i = (x_i^1, \dots, x_i^M)$ be the 'complete' vector representation of the i -th example shot ($1 \leq i \leq N$). Here, the j -th dimension x_i^j ($1 \leq j \leq M$) represents the presence or absence of the j -th concept with no uncertainty (i.e., $x_i^j \in \{P, A\}$). Assume that x_i^j follows a probability distribution with the parameter θ^j , that is, $p(x_i^j = P; \theta^j)$ and $p(x_i^j = A; \theta^j)$. However, since the detection of the j -th concept is uncertain, $p(x_i^j = P; \theta^j)$ and $p(x_i^j = A; \theta^j)$ incur uncertainties, which are modeled by a mass function m^j . To implement this, based on [6], the following likelihood function $L(\theta; m)$ is used where each example shot and each dimension are assumed to be independent:

$$L(\theta; m) = \prod_{i=1}^N \prod_{j=1}^M \left(\sum_{S \subseteq \{P, A\}} m^j(S) \sum_{x_i^j \in S} p(x_i^j; \theta^j) \right) \quad (1)$$

where $\theta = \{\theta^1, \dots, \theta^M\}$ is a set of parameters for probability distributions for M dimensions (concepts), and $m = \{m^1, \dots, m^M\}$ is a set of mass functions for M concepts. In addition, S is any subset of $\{P, A\}$, that is, $\{P\}$, $\{A\}$ or $\{P, A\}$. Equation (1) means that $p(x_i^j = P; \theta^j)$ for the complete j -th concept's presence is weighted by masses, which are associated with subsets including P . Similarly, $p(x_i^j = A; \theta^j)$ is weighted by masses, associated with subsets including A . Based on this inclusive relation, the

term surrounded by big parenthesis in equation (1) can be expanded and deformed as follows:

$$\begin{aligned} & m^j(\{P\})p(x_i^j = P; \theta^j) + m(\{A\})p(x_i^j = A; \theta^j) \\ & + m(\{P, A\}) \left(p(x_i^j = P; \theta^j) + p(x_i^j = A; \theta^j) \right) \\ = & p(x_i^j = P; \theta^j) \left(m^j(\{P\}) + m(\{P, A\}) \right) \\ & + p(x_i^j = A; \theta^j) \left(m^j(\{A\}) + m(\{P, A\}) \right) \\ = & p(x_i^j = P; \theta^j)pl^j(\{P\}) + p(x_i^j = A; \theta^j)pl^j(\{A\}) \\ = & \sum_{x_i^j \in \{P, A\}} p(x_i^j; \theta^j)pl^j(x_i^j) \end{aligned} \quad (2)$$

Therefore, the estimation of θ^j does not require the mass function m^j , but requires the plausibility function pl^j . We rewrite $L(\theta; m)$ as $L(\theta; pl)$ where $pl = \{pl^1, \dots, pl^M\}$ is a set of plausibility functions for M concepts. Estimating θ , which maximizes $L(\theta; pl)$ is equivalent to maximizing the agreement between the probabilistic model $p(x_i^j; \theta^j)$ and uncertain concept detection $pl^j(x_i^j)$.

In our implementation, $p(x_i^j; \theta^j)$ is modeled as a simple discrete probability distribution with two parameters, each of which represents the probability that the j -th concept is present or absent. That is, $\theta^j = \{\alpha^{jP}, \alpha^{jA}\}$. Considering equation (1) and (2), $L(\theta; pl)$ is written as follows:

$$L(\theta; pl) = \prod_{i=1}^N \prod_{j=1}^M \left(\alpha^{jP} pl^j(x_i^j = P) + \alpha^{jA} pl^j(x_i^j = A) \right) \quad (3)$$

Please refer to [4], [6] for the detailed computation process of the estimation of θ . Finally, after θ is obtained using example shots for a query, the relevance score of a test shot x' is computed as follows:

$$rel(x') = \prod_{j=1}^M \left(\alpha^{jP} pl^j(x'^j = P) + \alpha^{jA} pl^j(x'^j = A) \right), \quad (4)$$

where $rel(x')$ represents the agreement between plausibilities of each concept's presence and absence in x' and the probabilistic distribution parameterized by $\theta^j = \{\alpha^{jP}, \alpha^{jA}\}$. The set of 1,000 test shots with the largest $rel(x')$ is returned as a retrieval result.

B. Plausibility Function Derivation by Density Estimation

For a shot x_i , we compute plausibilities of the j -th concept's presence and absence, $pl^j(x_i^j = P)$ and $pl^j(x_i^j = A)$, based on the detection score of x_i , s_i^j . These plausibilities are defined by the density ratio between two probability distributions, $p_{pr}(s_i^j)$ and $p_{ab}(s_i^j)$. The former represents the probability of the j -th concept's presence at the detection score s_i^j , while the latter represents the probability of its absence at s_i^j .

To compute s_i^j , a concept detector is built as follows: First, each shot is represented using the 1,000-dimensional Bag-of-Visual-Words representation, where each dimension

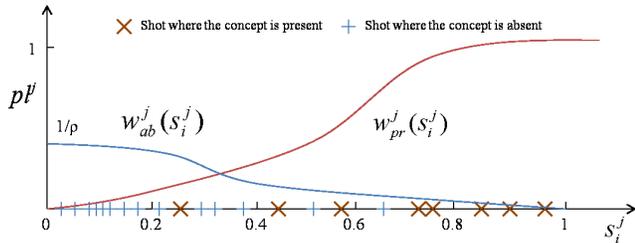


Figure 2. Plausibility computation using density ratio functions

represents the frequency of a characteristic local shape in the keyframe of the shot. Using training shots annotated with the j -th concept's presence and absence, a Support Vector Machine (SVM) is built as a concept detector. The detection score s^j is computed as the SVM's probabilistic output, which approximates the distance between x_i and the detection boundary using a sigmoid function [16].

Fig. 2 illustrates how to compute $pl^j(x_i^j = P)$ and $pl^j(x_i^j = A)$ based on s_i^j . The horizontal axis represents detection scores where \times s represent detection scores of shots annotated with the j -th concept's presence, and $+$ s represent detection scores of shots annotated with its absence. The vertical axis represents plausibilities defined by the following density ratio functions:

$$pl^j(x_i^j = P) = w_{pr}^j(s_i^j) = p_{pr}(s_i^j)/p_{ab}(s_i^j) \quad (5)$$

$$pl^j(x_i^j = A) = w_{ab}^j(s_i^j) = p_{ab}(s_i^j)/p_{pr}(s_i^j) \quad (6)$$

As shown in Fig. 2, $pl^j(x_i^j = P)$ becomes large as a detection score where the number of \times s is larger than the number of $+$ s. On the other hand, $pl^j(x_i^j = A)$ becomes large as a detection score where the number of $+$ s is larger than the number of \times s.

To estimate the density ratio functions $w_{pr}^j(s_i^j)$ and $w_{ab}^j(s_i^j)$, we use the method called *unconstrained Least-Squares Importance Fitting* (uLSIF) [7]. Using uLSIF, $w_{pr}^j(s_i^j)$ is estimated without estimating $p_{pr}(s_i^j)$ or $p_{ab}(s_i^j)$. Instead, it is modeled as the following linear combination of basis functions:

$$w_{pr}^j(s_i^j) = \sum_{l=1}^b \alpha_l^j \phi_l(s_i^j), \quad (7)$$

where a weight α_l^j for the l -th basis function $\phi_l(s_i^j)$ is estimated using shots annotated with the j -th concept's presence and absence. We define ϕ_l as a gaussian function. Please refer to [7] for the estimation of α_l^j . Finally, $w_{ab}^j(s_i^j)$ can be obtained in the same way to $w_{pr}^j(s_i^j)$.

C. Sampling from imbalanced data

For appropriate density ratio estimation, we need to solve the imbalanced problem between shots where a concept is present and shots where it is absent. To this end, we present

Algorithm 1 k-NN based Selective sampling method for Imbalanced Data (kNNSID)

Input: D : A set of unannotated shots,
 P : A set of shots annotated with a concept's presence,
 N : Number of shots to be sampled
Output: R : Array that contains sampled shots
1: $D' \leftarrow \text{findUniqueDetectionScores}(D)$
2: **while** $|R| < N$ **do**
3: **for all** $x \in D'$ **do**
4: $score \leftarrow \text{computePriorityScore}(x, R, P)$
5: **end for**
6: $R \leftarrow \text{getTopScoreShot}(D', score)$
7: **end while**
8: **return** R

Figure 3. k-NN based Selective sampling method for Imbalanced Data

k-NN based Selective sampling method for Imbalanced Data (kNNSID). Shots selected by kNNSID are annotated by a user, and used in the density estimation.

Fig. 3 shows a pseudo code of kNNSID, consisting of the following three steps: The first step at line 2 in Algorithm 1 creates a set of unannotated shots, where only one shot is retained for a unique detection score. In the second step at line 7, for each shot, the *priority score* which represents the priority of sampling is calculated. The third step at line 10 samples the shot with the highest priority score. As shown in lines from 4 to 11, the second and third steps are repeated by re-calculating the priority score of each shot until the number of sampled shots reaches the specified number.

The second step calculates the priority score of an annotated shot x , $p(x)$, using the following equation:

$$p(x) = \frac{1}{k_1} \sum_{i=1}^{k_1} d(x, X_i) - \frac{1}{k_2} \sum_{j=1}^{k_2} d(x, Y_j), \quad (8)$$

where $X = \{X_1, X_2, \dots, X_{k_1}\}$ is a set of already sampled shots that are similar to x . On the other hand, $Y = \{Y_1, Y_2, \dots, Y_{k_2}\}$ is a set of shots that are similar to x and already annotated with the concept's presence. The function d represents the Euclidean distance between two shots in terms of their detection scores. The first term in equation (8) computes the average distance between x and X . This is useful for collecting shots with a diversity of detection scores. The second term computes the average distance between x and Y . This gives high priorities to shots, which are similar to shots already annotated with the concept's presence. Hence, by annotating sampled shots, we can examine inaccuracies of different detection scores, which are similar to those of shots already annotated with the concept's presence. As a result, we can accurately estimate the density ratio function by alleviating the influence of too many shots where the concept is absent.

IV. EXPERIMENTAL RESULTS

This section evaluates our video retrieval method. First of all, we use 346 concepts defined in Large-Scale Concept Ontology for Multimedia (LSCOM) [17]. These concepts are defined based on their ‘utility’ for classifying content in videos, their ‘coverage’ for responding to a variety of queries, their ‘feasibility’ for automatic detection, and the ‘availability’ (or ‘observability’) for a large amount of training shots. We collect training shots via our online video annotation game [18], which is being developed in parallel with this paper. The game aims to efficiently annotate a large amount of shots with various concepts’ presences and absences, with the help of numerous online game users. Specifically, 292,911 shots in TRECVID 2011 development videos are targeted by the game, and annotated shots are used as training shots to build concept detectors.

The following experiment is conducted by applying the above concept detectors to TRECVID 2009 video data, consisting of 36,106 shots in 219 development videos, and 97,150 shots in 619 test videos. For each concept, a plausibility function is derived by the density ratio estimation on 1,000 shots, annotated with the concept’s presence or absence. These shots are collected from development videos using our selective sampling method. Our video retrieval method are tested on the following three queries: (1) “A view of one or more tall buildings and the top story visible”, (2) “One or more people, each at a table or desk with a computer visible”, and (3) “An airplane or helicopter on the ground, seen from outside”. For each query, a retrieval model is constructed using 10 example shots selected from development videos, and used to retrieve relevant shots in test videos. Here, concepts unrelated to the query are ignored to improve the retrieval performance. In other words, concepts related to the query are selected as the ones, for which average detection scores in example shots are larger than the threshold. The retrieval is conducted using detection scores and plausibility functions for selected concepts.

In order to examine the effectiveness of plausibility functions, the above retrieval method denoted by *PL* is compared to a method, which is denoted by *Direct* and constructs a retrieval model directly from concept detection scores. In other words, the model in *Direct* is constructed by replacing $pl^j(x_i^j = P)$ in equation (3) with the detection score s_i^j ($pl^j(x_i^j = A)$ is replaced with $1 - s_i^j$). Fig. 4 shows a performance comparison between *PL* and *Direct* in terms of their precisions. A precision represents the probability of relevant shots in 1,000 retrieved shots. In each bar graph in Fig. 4, white-colored and black-colored bars represent precisions obtained by *PL* and *Direct*, respectively. In addition, the white-colored and black-colored bars at the top respectively present precisions obtained by plausibility functions (*PL*) and detection scores (*Direct*) for ‘ALL’ concepts. Each of the other bars presents the precision obtained by the plausibility

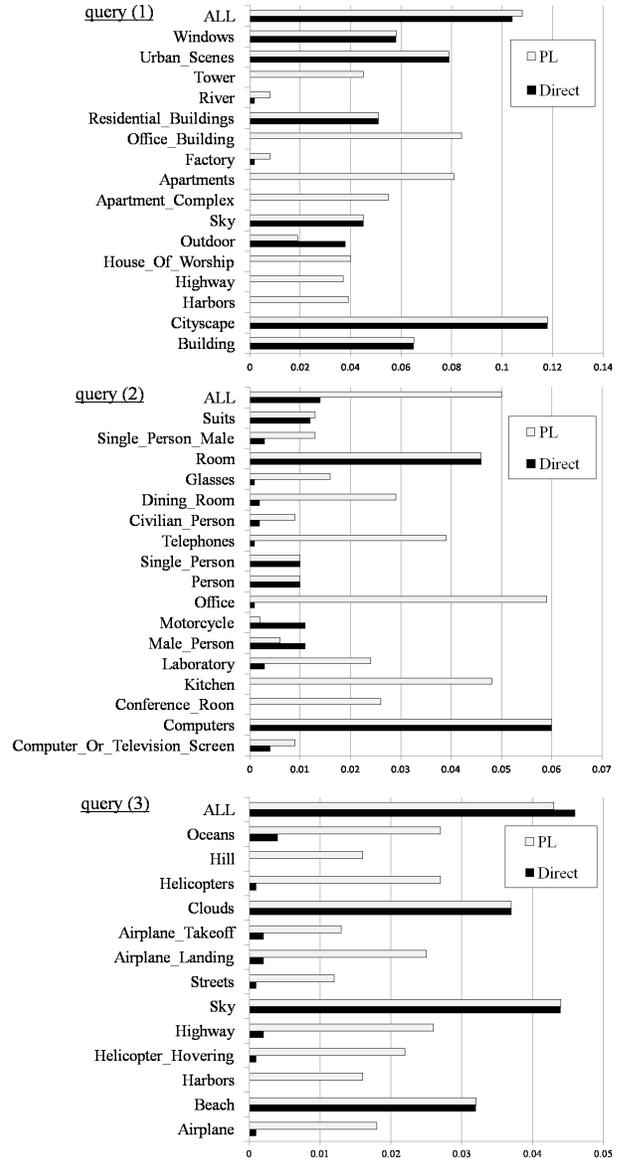


Figure 4. Performance comparison between *PL* and *Direct*

function (or detection scores) for a single concept. Its name is shown in the left side of the bar graph.

As can be seen from Fig. 4, for query (1) and (2), *PL* is superior to *Direct* in the case of using all concepts. Regarding cases of using single concepts, for almost all concepts where precisions of *Direct* are very low, *PL* achieves much higher precisions. It can be said that detecting such concepts involves much uncertainties, which are effectively modeled by plausibility functions.

However, for query (3), *PL* is outperformed by *Direct* in the case of using all concepts, although precisions of the former are much higher than those of the latter in cases of using single concepts. This means that *PL*’s advantage

over *Direct* in cases of using single concepts is weakened in the case of using combinations of these concepts. One main reason is the simplicity of our video retrieval model, where relevant shots to a query are characterized only by a single combination of concepts' presences and absences (see equation (3)). But, actually, relevant shots show different combinations of concepts' presences and absences depending on varied camera techniques. Thus, we plan to incorporate a mixture model into our video retrieval model, or adopt another method, which can extract a non-linear classification boundary between relevant and irrelevant shots based on plausibility functions [19].

V. CONCLUSION AND FUTURE WORKS

In this paper, we introduced a concept-based video retrieval method where uncertainties in concept detection are modeled using plausibility functions. Each of them is derived by estimating the density ratio between shots annotated with a concept's presence and absence. In particular, to solve the imbalanced problem between the number of shots where the concept is present and that of shots where it is absent, the selective sampling method *kNNSID* is developed to preferentially sample unannotated shots, which are similar to shots already annotated as the concept's presence. Experimental results on TRECVID 2009 video data show that derived plausibility functions effectively manage uncertainties in concept detection. In the future, we plan to improve the retrieval performance in the case of combining plausibility functions for multiple concepts. To this end, our video retrieval method will be extended by incorporating a mixture model, or adopting a method which extracts a non-linear classification boundary between relevant and irrelevant shots to a query using plausibility functions [19].

REFERENCES

- [1] A. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVID," in Proc. of MIR 2006, October, 2006, pp. 321–330.
- [2] A. Natsev, A. Haubold, Tešić, L. Xie, and R. Yan, "Semantic Concept-based Query Expansion and Re-ranking for Multimedia Retrieval," in Proc. of ACM MM 2007, September, 2007, pp. 991–1000.
- [3] X. Wei, Y. Jiang, and C. Ngo, "Concept-driven multi-modality fusion for video search," IEEE Transactions on Circuits and Systems for Video Technology, vol. 21, no. 1, January, 2011, pp. 62–73.
- [4] K. Shirahama, K. Kumabuchi, and K. Uehara, "Video retrieval by managing uncertainty in concept detection using dempster-shafer theory," in Proc. of MMEDIA 2012, April, 2012, pp. 71–74.
- [5] G. Shafer, A Mathematical Theory of Evidence. Princeton University Press, 1976.
- [6] T. Denœux, "Maximum likelihood estimation from uncertain data in the belief function framework," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 1, January, 2013, pp. 119–130.
- [7] T. Kanamori, S. Hido, and M. Sugiyama, "A least-squares approach to direct importance estimation," Journal of Machine Learning Research, vol. 10, no. 7, July, 2009, pp. 1391–1445.
- [8] H. He, and E. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, September, 2009, pp. 1263–1284.
- [9] H. Wang, and S. McClean, "Deriving evidence theoretical functions in multivariate data spaces: A systematic approach," IEEE Transactions on Systems, Man and Cybernetics - Part B, vol. 38, no. 2, March, 2008, pp. 455–465.
- [10] A. Aregui, and T. Denœux, "Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities," International Journal of Approximate Reasoning, vol. 49, no. 3, November, 2008, pp. 575–594.
- [11] M. Zribi, "Parametric estimation of dempster-shafer belief functions," in Proc. of ISIF 2003, July, 2003, pp. 485–491.
- [12] C. Aggarwal, and P. Yu, "A survey of uncertain data algorithms and applications," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 5, May, 2009, pp. 609–623.
- [13] X. Li, D. Wang, J. Li, and B. Zhang, "Video search in concept subspace: A text-like paradigm," in Proc. of CIVR 2007, July, 2007, pp. 603–610.
- [14] X. Wang, and M. Kankanhalli, "Portfolio theory of multimedia fusion," in Proc. of ACM Multimedia 2010, October, 2010, pp. 723–726.
- [15] R. Benmokhtar, and B. Huet, "Perplexity-based evidential neural network classifier fusion using MPEG-7 low-level visual features," in Proc. of MIR 2008, October, 2008, pp. 336–341.
- [16] C. Chang, and C. Lin, "Libsvm : A library for support vector machines," ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, April, 2011, pp. 1–27.
- [17] M. Naphade, J. Smith, J. Tešić, S. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," IEEE Multimedia, vol. 13, no. 3, July, 2006, pp. 86–91.
- [18] Y. Watanabe, K. Shirahama, and K. Uehara, "Online video annotation game with active learning and tag ranking," IEICE Technical Report, vol. 112, no. 346, December, 2012, pp. 75–80 (In Japanese).
- [19] T. Denœux, "A k-nearest neighbor classification rule based on dempster-shafer theory," IEEE Transactions on Systems, Man, and Cybernetics, vol. 25, no. 5, May, 1995, pp. 804–813.

A Statistical Approach for the Automatic Recognition of Traffic Sign Deterioration

Walter Allasia, Francesco Gallo

Research Department
EURIX

Torino, Italy

e-mail: allasia@eurix.it, gallo@eurix.it

Mario Ferraro

Dipartimento di Fisica
Università di Torino

Torino, Italy

e-mail: ferraro@to.infn.it

Sara Lorio

LREN, Departement des Neurosciences Cliniques
CHUV, University of Lausanne

Lausanne, Switzerland

e-mail: sara.lorio87@gmail.com

Abstract—This paper describes a software application based on statistical methods for the automatic recognition of traffic sign deterioration. The evaluation of traffic sign degradation is usually performed by devices applied on top of the road sign surface, measuring color parameters such as chromatic coordinates and the luminance factor. Moreover, the devices can only check a small fraction of the traffic sign surface at a time, requiring several acquisitions on the same traffic sign. In order to reduce the costs related to monitoring and have a periodic control of the traffic sign status, we propose a fast automatic method based on video acquisition and processing that can be easily operated in patrolling vehicles provided with a camera. A pattern detection algorithm based on color and texture features is applied to the images extracted from the acquired videos in order to detect the traffic signs ROIs, which are analyzed using a statistical approach based on the Kullback-Leibler divergence and Kolmogorov-Smirnov test. Making use of a control sample of not deteriorated traffic sign images, a comparison between the acquired and the reference images is performed. Both statistical methods have been used to compare 150 pairs of traffic signs, achieving high precision and recall, proving that the proposed approach can be a good candidate solution for automatic traffic sign deterioration analysis.

Keywords—*traffic sign recognition; automatic video processing; image deterioration; Kullback-Leibler divergence; Kolmogorov-Smirnov statistical test*

I. INTRODUCTION

Usually, quality of the traffic sign surface concerning the reflectance and color value is evaluated remeasuring color parameters such as chromatic coordinates and the luminance factor. This requires the application of measurement devices and equipments in contact with the traffic sign surface, which cannot be performed automatically.

Moreover, these devices can only check a small fraction of the traffic sign surface at a time, requiring several acquisitions on the same traffic sign. In order to reduce the costs related to monitoring and have a periodic control of the traffic sign status, we propose a fast automatic method based on video acquisition and processing that can be easily operated in patrolling vehicles provided with a camera.

Many approaches have been proposed and are available in the literature that aim at automatically recognize the traffic sign patterns within digital images making use of well established image processing and pattern recognition techniques; see, for example, [1][2] and references therein. In [3], a method for

the measurement of degradation based on retroreflectivity has been proposed, where the authors have computed the photometrical response function for each pixel in order to establish a threshold for identification of damaged traffic signs. The statistical measurement we are presenting here enables the calculation of a threshold at a more global level, measuring differences between two probability density functions.

In this paper, we present two methods to assess signal degradation by comparing degraded traffic signs with a reference sample of non-deteriorated signs. This comparison has been carried out with two methods: a divergence measure based on the Kullback-Leibler [4] distance and the Kolmogorov-Smirnov [5][6] statistical test that, to the best of our knowledge, are the only available in literature for this kind of analysis and comparison.

We have taken into account the protocols for evaluating the deterioration of traffic signs provided by Italian public institutions such as the Ministry of Public Works and Transport [7], and ANAS [8], responsible for the traffic sign maintenance and control.

This paper is organized as in the following: Section II describes the image processing approach, whilst Section III reports the statistical methods analyzed and applied; Section IV provides the experimental results based on a sample of 150 traffic signs; finally, Section V summarizes the results and possible future work.

II. IMAGE PROCESSING

As mentioned in the introduction, we have made use of two separate image sets: one set of *non-deteriorated* traffic signs adopted as reference sample and one set of traffic signs whose degradation must be evaluated.

The reference sample has been created by manually acquiring digital images traffic signs that appeared in good conditions, and examples of degraded traffic signs have been from roads in Turin. A patrolling vehicle with a camera has provided a series of images of traffic signs (mjpeg [9]) with the associated geospatial references stored in the EXIF [10] image format. Hence a software already developed (and tuned for high recall) [1] has recognized the traffic signs, making use of Haar cascades [11][12] for *regions of interest* (ROI) detection and a MPEG-7 [13] feature-based classifier for traffic sign identification.

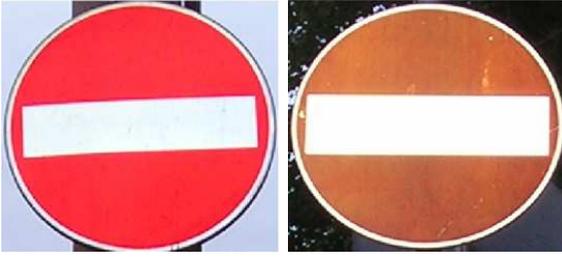


Fig. 1. Sample (left) and deteriorated (right) traffic signs

For each recognized traffic sign, the ROI has been extracted and normalized to a reference size. Finally, it has been split up into three color channels (RGB), analyzed separately in order to evaluate the pixel color level distribution.

The application has been written in C++ language and using the OpenCV [14] libraries for image manipulation.

Summarizing, the image processing flow is made up of the following five steps:

- 1) Image acquisition by a patrolling vehicle provided with a camera.
- 2) Traffic sign recognition for the acquired images
- 3) Image cropping for ROI selection and normalization
- 4) Traffic sign ROI split into RGB planes
- 5) Pixel color level distribution evaluation

In order to have the statistical distributions of *not deteriorated* traffic signs to be used as **control sample**, we have performed the overall processing chain to the new and unused traffic signs images acquired manually.

The information about the pixel color level distribution for both samples has then been used as input for the statistical analysis.

III. ANALYSIS METHODS

In order to evaluate the level of degradation of traffic signs, we have taken into account the color distributions of the pixels. We adopted a statistical approach for the analysis, based on the Kullback-Leibler [15] divergence and Kolmogorov-Smirnov [16] test. A comparison of the two methods has been performed in order to select the most appropriate one as a good solution for automatic recognition of road signs degradation.

The probability $\mathcal{P}(px, ch)$ that a generic pixel px belonging to a $N \times M$ image has a given color level ℓ (with $\ell = 0, \dots, 255$) in a specific color channel ch (where $ch = R, G, B$) is given by:

$$\mathcal{P}(px, ch) = \frac{(n_\ell)_{ch}}{(N \times M)_{ch}} \quad (1)$$

where $(n_\ell)_{ch}$ is the number of pixels having color level ℓ .

In order to compute the probability function of the **control sample**, an average value of \mathcal{P} is evaluated using all the images representing a specific traffic sign. These values are compared with the probability function computed for each actual traffic sign image acquired.

Here, we have compared the probability functions associated to the same color channel of the traffic signs acquired as mentioned above and the control sample.

A. Kullback-Leibler divergence

The Kullback-Leibler divergence \mathcal{D}_{kl} , associated with two probability distributions $f(x)$ and $g(x)$, is defined as follows:

$$\mathcal{D}_{kl}(f \parallel g) = \int f(x) \log \frac{f(x)}{g(x)} dx \quad (2)$$

where x is a random continuous variable.

In our analysis, the random variable x is associated to the digital color levels, moving from continuous to discrete, and $f(x), g(x)$ are the probability distributions associated to the same color channel of the two images to be compared (the sample and deteriorated sign). Moreover, the equation (2) is not symmetric.

In order to use discrete variables and symmetric functions, the Kullback-Leibler divergence definition given in (2) can be rewritten in a symmetric form \mathcal{S}_{kl} as:

$$\begin{aligned} \mathcal{S}_{kl}(f \parallel g) &= \mathcal{S}_{kl}(g \parallel f) = \\ &= \mathcal{D}_{kl}(f \parallel g) + \mathcal{D}_{kl}(g \parallel f) = \\ &= \sum_{x=0}^N \left[(f(x) - g(x)) \log \frac{f(x)}{g(x)} \right]. \end{aligned} \quad (3)$$

In order to remove the singularity for $g(x) = 0$, every $g(x)$ is added a small value ϵ chosen as:

$$\epsilon = \frac{gmax}{N} = \frac{1}{256} \quad (4)$$

where N is the total number of color levels. Using the transformation $g'(x) = g(x) + \epsilon$, equation (3) becomes:

$$\mathcal{S}_{kl}(f \parallel g') = \sum_{x=0}^N \left[(f(x) - g'(x)) \log \frac{f(x)}{g'(x)} \right] \quad (5)$$

\mathcal{S}_{kl} is calculated for each probability distribution pair related to the same color channel. The three evaluated divergences are summed up to a total divergence \mathcal{S}_{kl}^{tot} . If the overall amount exceeds a threshold value T , the traffic sign is declared *deteriorated* ($\mathcal{S}_{kl}^{tot} > T$).

B. Kolmogorov-Smirnov test

The other method used is the Kolmogorov-Smirnov test for two samples, which compares two *cumulative distribution functions*, related to two data sets, in order to estimate if they belong to the same distribution which can be unknown.

We have experienced that applying the test to all the pixels of the image, we overestimate the degree of freedom of the test because many pixels are correlated to their neighbors. Pixel correlation is evaluated to the images converted to greyscale values and applying the following algorithm:

$$\begin{aligned} g(x, y) &= w(x, y) \cdot f(x, y) = \\ &= \sum_{s=-a}^a \sum_{t=-b}^b w(s, t) \cdot f(x+s, y+t) \end{aligned} \quad (6)$$

where $f(x, y)$ is the greyscale value of the pixel (x, y) , $w(x, y)$ is the weight of the $m \times n$ correlation mask, $g(x, y)$ is the correlation level, $a = \frac{m-1}{2}$ and $b = \frac{n-1}{2}$

If $g(x, y)$ is *null* there is no variations between pixel at (x, y) and its neighbors.

In order to reduce number of correlated pixels, we have applied the *Canny* [17] edge detector and considered solely filtered pixels. Canny edge detector provides greyscale conversion, noise reduction, correlation evaluation, minimal gradient suppression, hysteresis threshold.

The two cumulative distribution functions for each image have been calculated using only the pixels belonging to the edges, separately for each color channel:

$$\mathcal{F}_{ch} = \{\mathcal{F}_c, c = 0, \dots, 255\}_{ch} \quad (7)$$

$$\mathcal{G}_{ch} = \{\mathcal{G}_c, c = 0, \dots, 255\}_{ch} \quad (8)$$

$$\mathcal{F}_c = \sum_{\ell=0}^c h_{\ell} \quad (9)$$

$$\mathcal{G}_c = \sum_{\ell=0}^c h'_{\ell} \quad (10)$$

where h_{ℓ} and h'_{ℓ} are the color level frequencies. Given $\mathcal{F}_c h$ and $\mathcal{G}_c h$, the maximum distance can then be calculated:

$$\mathcal{D}_{ks} = \max | \mathcal{F}_{ch} \mathcal{G}_{ch} | \quad (11)$$

The output of (11) can be compared to a threshold value λ , thus determining whether the two distributions are significantly different. Giving \mathcal{J} as the following value:

$$\mathcal{J} = \frac{N_c N_d}{N_c + N_d} \quad (12)$$

where N_c is the total number of pixel in a good image channel and N_d is the total number of pixel in a deteriorated image channel, we can compare the equation (11) to:

$$\mathcal{D}_{ks} \geq \frac{\lambda}{\sqrt{\mathcal{J}}} \quad (13)$$

If the inequality (13) is not satisfied, we can say that the distributions are significantly different with a level of significance associated to the threshold level λ . It means that the evaluated traffic sign is deteriorated and should be replaced.

IV. EXPERIMENTAL RESULTS

Both methods described in the paper, the Kullback-Leibler divergence and the Kolmogorov-Smirnov test, have been applied on 150 traffic signs pairs (the sample and the deteriorated). According to the analysis procedure described in Section III-A, the threshold T has been chosen evaluating the distribution of S_{kl}^{tot} corresponding to the 150 traffic signs pairs. Fig. 2 shows the Gaussian fit to the data, whose mean μ and standard deviation σ are calculated. The threshold value has been chosen as $T = \mu - \sigma = 2.5$. We obtained a χ^2 value equal to 4.618 with 11 degrees of freedom and the Gaussian distribution hypothesis can be accepted with a significance level of $\alpha = 0.05$. We also tried to fit the data with a Poisson distribution, which gives a $\chi^2 = 3.201$ with 12 degrees of freedom ($\alpha = 0.05$) and the obtained threshold $T = 2.8$ does not change significantly.

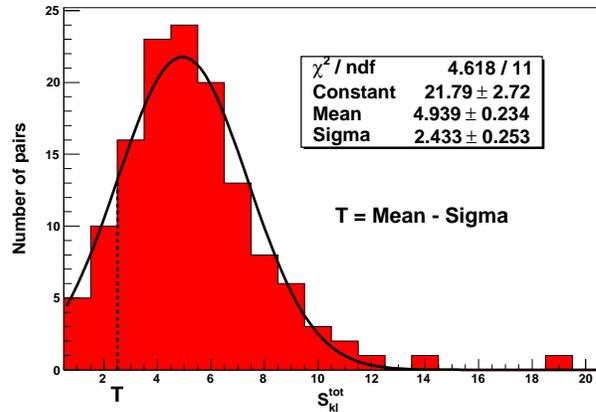


Fig. 2. S_{kl}^{tot} with fit results using a Gaussian distribution and the chosen threshold T .

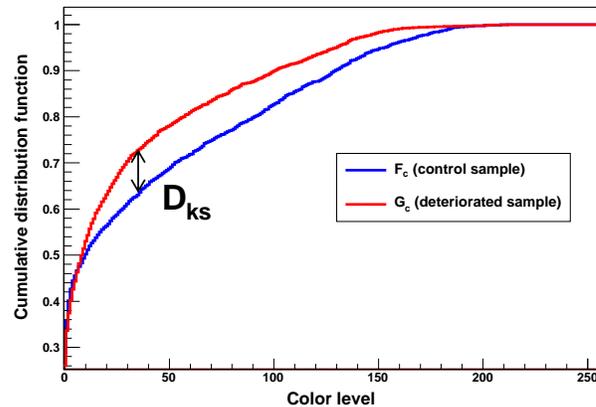


Fig. 3. Cumulative distribution functions \mathcal{F}_{ch} and \mathcal{G}_{ch} for the control sample (blue) and deteriorated sample (red) with the chosen \mathcal{D}_{ks} after applying the edge filter

According to the analysis method discussed in Section III-A, Fig. 3 shows the cumulative distribution functions \mathcal{F}_{ch} and \mathcal{G}_{ch} for the traffic sign samples with the application of the edge filter where the selected value of \mathcal{D}_{ks} is also displayed.

Precision, recall and accuracy are evaluated according to the following equations:

$$precision = \frac{tp}{tp + fp} \quad (14)$$

$$recall = \frac{tp}{tp + fn} \quad (15)$$

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (16)$$

where tp are the true positive, tn are the true negative, fp are the false positive and fn are the false negative values.

The calculated values are reported in Table I. It shows that both methods have good accuracy, precision and recall.

TABLE I. RESULTS FOR A SAMPLE SET OF 150 TRAFFIC SIGN PAIRS

Method	Accuracy	Precision	Recall
Kullback-Leibler	0.80	0.88	0.79
Kolmogorov-Smirnov	0.88	0.93	0.89

The Kolmogorov-Smirnov method appears to give a better performance, with 89% recall and 93% precision.

V. CONCLUSION AND FUTURE WORK

Values in Table I show that both Kullback-Leibler divergence and Kolmogorov-Smirnov test achieve high accuracy, precision and recall parameters, so both methods can, in principle, be used.

However, Kolmogorov-Smirnov test results in the highest values of statistical parameters; indeed the Kullback-Leibler divergence gives only a measure of the distance between two distributions, and cannot capture the statistics of the samples as done by the Kolmogorov-Smirnov test.

The method presented here could be, for instance, applied to automatic monitoring of traffic sign deterioration: a camera could be installed onto public transportation vehicles patrolling the roads and acquiring the digital images [1] that then could be processed. A traffic signs catalogue could thus be generated enabling to determine in advance (before their scheduled expiring time) the traffic signs to be replaced.

A possible improvement of the proposed approach that the authors would like to implement as future work, is the use of HSV (Hue Saturation Value) color space instead of RGB, because the distance between two colors in HSV dimensions is closer to the human eye color perception. Hence, the presented statistical methods should result in a better evaluation of color deterioration. Since HSV has a clear separation between luminance and colors, it is more suitable to appraise the color variations, even in case of large changes of luminance. Nevertheless, it should be considered that the hue (H) component which holds color information in HSV color space depends on distance, weather conditions, and age of traffic sign [18].

REFERENCES

- [1] W. Allasia, C. Culeddu, M. Ferraro, F. Gallo, and M. Vigilante, Automatic video processing for traffic sign recognition, In Proceedings of the 5th IASTED European Conference on Internet and Multimedia Systems and Applications EuroIMS A 2009, Cambridge, UK, ACTA Press, Jul. 2009, pp. 99-103
- [2] A. Soetedjo and K. Yamada, A new approach on red color thresholding for traffic sign recognition system, Journal of Japan Society for Fuzzy Theory and Intelligent Informatics, volume 19, number 5, Oct. 2007, pp. 458-465.
- [3] P. Siegmann, R. Javier Lopez-Sastre, P. Gil-Jimnez, S. Lafuente-Arroyo, and S. Maldonado-Bascn, Fundaments in Luminance and Retroreflectivity Measurements of Vertical Traffic Signs Using a Color Digital Camera, IEEE Transactions on Instrumentation and Measurement, Vol. 57, No. 3, Mar. 2008
- [4] S. Kullback and R.A. Leibler, On Information and Sufficiency, Annals of Mathematical Statistics, Vol. 22, No. 1, 1951, pp. 7986
- [5] M. Hazewinkel, Kolmogorov-Smirnov test, Encyclopedia of Mathematics, Springer, 2001
- [6] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006, pp. 55-58
- [7] Ministero dei Lavori Pubblici, Decreto Ministeriale 31 marzo 1995, n. 1584 (g.u. n. 106 del 9.5.1995), 1995.
- [8] Centro Sperimentale Stradale di ANAS SPA, Segnaletica stradale verticale: normativa, materiali, caratteristiche tecniche e metodologie di controllo, Nov. 2007.
- [9] Joint Photographic Experts Group, ISO/IEC 15444-3:2007, JPEG 2000 image coding system: Motion JPEG 2000, 2007.
- [10] Japan Electronics and Information Technology Industries Association, Exif 2.2 (exchangeable image file format), <http://www.exif.org/Exif2-2.PDF>, [retrieved: March, 2013].
- [11] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection, In Proc. International Conference on Image Processing, volume 1, 2002, pp. 900-903.
- [12] P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features, In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 1, 2001, pp. 511-518.
- [13] Moving Picture Expert Group, MPEG-7 ISO/IEC 15938, Information Technology Multimedia Content Description Interfaces, <http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm> [retrieved: March, 2013].
- [14] G. Bradski and A. Kaehler, Learning OpenCV: Computer Vision with the OpenCV Library, O'Reilly Media, 1st edition, Sept. 2008.
- [15] P. Piro, S. Anthoine, E. Debreuve, and M. Barlaud, Image retrieval via Kullback-Leibler divergence of patches of multiscale coefficients in the knn framework, In CBMI 2008. Proceedings of the International Workshop on Content-Based Multimedia Indexing, IEEE, London, UK, Jun. 2008, pp. 230-235.
- [16] H. Srinivasan, S. N. Srihari, and M. J. Beal, Signature verification using kolmogorov-smirnov statistic, 2005.
- [17] R. Sun, J. Jia, H. Li, and M. Jaeger, Image-based lightweight tree modeling, In VRCAI '09: Proceedings of the 8th International Conference on Virtual Reality Continuum and its Applications in Industry, New York, USA, ACM, 2009, pp. 17-22.
- [18] Z. T. Kardkovcs, Z. Parczy, V. Endre, A. Siegler and P. Lucz, Real-time traffic sign recognition system, In Proceedings of CogInfoCom2011, 2nd International Conference on Cognitive Infocommunications, Budapest, Hungary, Jul. 2011, pp. 1-5.

A Real-time Video Summarizing Service for Community-contributed Contents of Real-life Events

Samantha Vu, Owen Noel Newton Fernando, Mikko Rissanen, Natalie Pang, Schubert Foo

Nanyang Technological University, Singapore

14 Nanyang Drive, Singapore 637332

{sgtvu, ofernando, mjrissanen, nlspong, sfoo}@ntu.edu.sg

Abstract—Using the example of a group of spectators watching a live street dance, this paper presents a work-in-progress concept of real-time video summarizing service by 'jumping' in different point of views. This innovative service takes up the challenges from various aspects of multimedia: mobile multimedia, authoring, real time interactive multimedia applications.

Keywords—multimedia entertainment; community-contributed content; real-time; real-life events

I. INTRODUCTION

The increasing popularity of mobile live event capturing is undeniable. More and more people are using their phones for recording video clips and for editing activities. Video recording quality of mobile phones has reached a level whereby the phones are becoming a serious tool for on-the-move content creators. To share these self-edited videos is to share experience which is a social norm in modern communities. However, according to Juhlin et al. [1], people utilizing mobile broadcasting are still struggling with the technology despite all advances in modern camera-equipped smartphones and mobile live video broadcasting services.

In this paper, we focus on two problems in the current state of art in multimedia editing. The first problem is lack of motivation for editing. It is known that people recording the videos rarely watch them because of the amount of work for exploring, and annotating; especially since mobile videos are rarely edited after recording [2]. Kirk et al. [2] take a holistic user centric view on people's practices around home videos. Their results reveal useful information about people's motivations and practices for editing home videos. One of their results is that people do not find any reason to do editing of the short video clips they had taken. To overcome this, our proposed service (which will be explained in details in part II) would make producers more motivated to create the summary video by offering ability to participate in the live event and switch from one point of view to another in real time.

The second problem is a lack of good quality summary videos or remixing videos of real-life events. For instance, a live concert usually attracts a good number of video recordings by members of the audience but there are few compact and coherent videos that can capture the highlights of the concert. Similarly, if a search is done on Youtube about a certain event, for example, Obama's victory speech, there is usually an extremely long list of videos found. Even

though the list is sorted corresponding to relevance, this is not a proper results returned for such a question. We argue that a preferred way is to get a compact presentation of a predefined length, which gives a summary, composed from the views of many people that have witnessed the event. This is not necessarily the best view, but the view that can be created based on the information people provided when uploading the content. However, this "summary" video hardly exists in most cases. Thus, with the motivation to supply such a video, in this paper, we will present the concept of a ubiquitous service for summarizing live event.

Making available a large pool of snapshot digital videos taken by the audience in the same concert can result in higher value material than individual video clips. The individual digital video clips can be remixed into compilations that potentially enhance the perceived value of the event, are useful for various stakeholders such as the artists, and the fans of the artists.

According to Vihavainen et al. [3], remixing can also give the fans the possibility to become creators and not just receivers, and enhance the community feeling between the performers and the fans. Engström et al. [4] analyzed how video jockeys in dance clubs work and suggested that mobile video could enhance the interaction between the club visitors and VJs (Video Jockeys). Engström et al. [5] continued their studies and presented the SwarmCam prototype for video capture and live transmission of mobile video. Club visitors film their dancing on the dance floor and stream the video live to the VJ, who possibly broadcasts it to a big display screen. From our paper's point of view, the study is interesting, given that it concentrates on the interaction. In our usage scenario, the interaction between audiences and artists and distance audiences (remote participants) is emphasized as well. This is also why we explain the proposed service using an example of a street performance.

Our concept emphasized the element of real time editing which allows the producer to create summaries but also participate in the event live. The participation factor has never been discussed before in the area of video summarization of events.

The various aspects of multimedia discussed above (automatic summarizing, video editing and authoring, multi-camera video production, live events, user point of view, real time participation) have been researched in several well-established studies but never have all been integrated in a single service. This is what the proposed service in this paper aims to do.

The next part of the paper will describe the usage scenario and technical architecture of the proposed service in details. Discussions on advantages and disadvantages of the service will be covered afterwards and followed by conclusion.

II. THE CONCEPT

A. Usage scenario

Fig. 1 shows the proposed real-time video summarizing service designed for three types of users: spectators, performers, and producers. The spectators record the event using their camera-equipped smartphones with the intention of documenting the performance as well as their experiences during it. Pointing a smartphone to record the event is a natural action for them. The performers are equipped with wearable cameras and optional dataglasses. The remote participants become either normal passive spectators or they act as producers who are in charge of producing a complete documentation of the event for the community. For each event, the community can utilize one or more producers. They are located away from the local site and equipped with computers more powerful than smartphones, i.e. laptop or desktop computers. The producers are able to 'jump' into either the spectators or the performers to share live audio connection and viewpoints through the cameras. The producers are presented as augmented reality avatars connected to the person they are sharing viewpoint with. This is done to emphasize their presence to the spectators, performers and to themselves, which subsequently enhance their participation in the event. User interfaces are demonstrated in a practical scenario shown in Fig. 2.

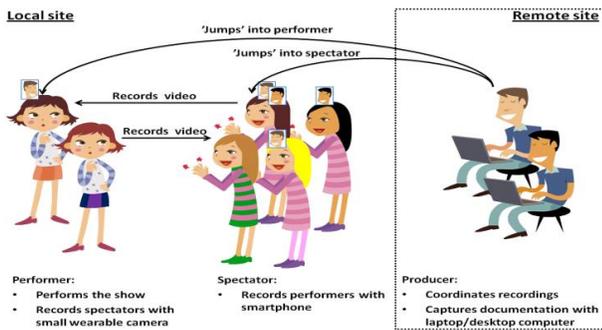


Figure 1. Participation in live events by 'jumping' into others.

Time	Sam the Spectator	Pete the Performer	Stacy the Spectator
1	Ryan jumps to record		Rita jumps to record
2	(Ryan keeps recording)	Ryan jumps to record	(Rita keeps recording)
3	(Ryan keeps recording)	(Ryan keeps recording)	Ryan jumps to record (Rita keeps recording)

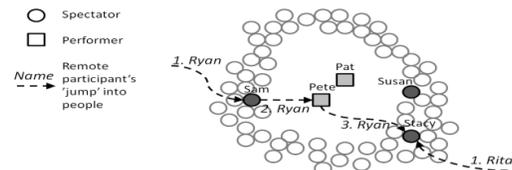


Figure 2. Setup of a live street dance scenario with 7 community members. Ryan captures 3 viewpoints and Rita 1 viewpoint of video recordings.



Figure 3. Ryan recording through Sam's smartphone.



Figure 4. Ryan recording through Pete's dataglasses.



Figure 5. Ryan and Rita recording through Stacy's smartphone.

There are a) 2 street dance performers called Pat and Pete, b) 3 spectators called Sam, Susan and Stacy, and c) 2 remote participants Ryan and Rita who act as the producers. Fig. 3, 4, and 5 illustrate how Ryan 'jumps' first into Sam, then Pete and finally to Stacy who is already visited by Rita. The recordings Ryan captures from these 3 viewpoints (shown in Table 1) show others and himself appearing in the video. Default scene selections are done automatically based on the viewpoints Ryan has collected by 'jumping' into others. These can be dropped out or included in the summary video. Yet, Ryan has the option of mixing content from Rita's view or adding them later if post-processing is desired.

TABLE I. VIDEO STREAMS RECORDED BY PRODUCERS (PR) THROUGH SPECTATOR OR PERFORMERS (SP). STACY AND SAM RECORDED RYAN'S 'JUMPS' WHICH RYAN CAN VIEW AS INSTANT REPLAY.

Pr	Sp	Timepoint 1	Timepoint 2	Timepoint 3
Ryan	Sam			
	Pete			
	Stacy			
Rita	Stacy			

B. Architecture

Fig. 6 represents the overall architecture of the ubiquitous service of live events for remote participants. This architecture seeks to address the problem of sharing experience using mobile devices [6] with remote participants in a real world context with practical and technical challenges and social implications on design. This includes: the accessibility of hardware devices, wireless connectivity, current low-tech approaches, cost, and the habits and preferences of the end-users. Live video and audio from smartphones or other portable devices are streamed over a 3G or wireless connection while face recognition algorithm on smartphones processes the data stream [7]. GPS coordinates and orientation data from each spectator's and performer's location are used to decide which image frame is needed for the face recognition and tracking algorithms in order to reduce computational cost. Such tracking system, which combines vision-based and inertial trackers, provides the speed, precision, and stability to support outdoor mobile augmented reality systems [8]. In a crowded festival or indoors, mere GPS location and viewpoint orientation may not be adequate for recognizing the exact position of a community member. Advanced face detection and tracking is incorporated in the processing layer of the service to triangulate the exact position of a specific person appearing in a video stream, which will subsequently help to produce accurate augmented reality avatars with name labels.

Recognition results are then returned in real-time to the mobile devices and augmented with text and images. The mobile device is mostly responsible for sending live video and audio streams to the remote participants for recognition processing, which in turn provides the computed results. Both the spectator and the performer will be able to see augmented video on smartphones or data glasses, and the remote participants can see augmented video while experiencing spatial sound capabilities. The service delivers real-time face recognition information, text and image based information, and it enables functionalities of augmented reality.

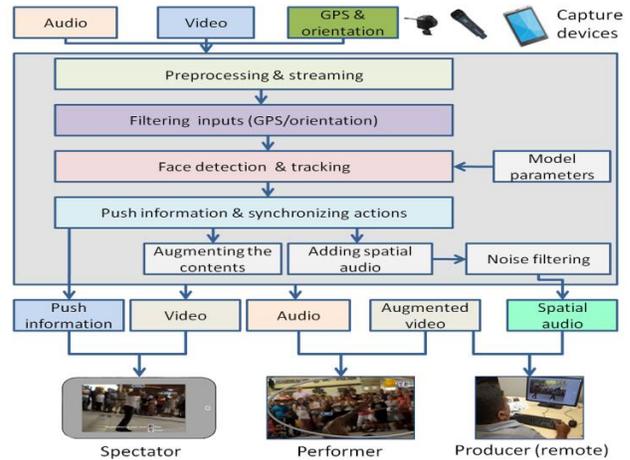


Figure 6. Service architecture.

Additionally, in terms of participation, we integrate spatial audio capabilities together with noise filtering to bring a natural experience of 'being there' with the spectators and performers. The architecture as a whole requires intensive research and development both in the communication and in the processing layers. As of present, the project is still in its initial stage of concept. However, we foresee some potential technical implementation issues e.g network connectivity will dictate the real-timeliness and quality of service. The service is conceptualised with the assumption of stable 3G and 4G networks in Singapore context. In case of connectivity failure, an automatic video recording will be activated to record the happenings from different viewpoints. This acts as a backup which the producer (remote) can still refer to after the real-life event.

III. DISCUSSION

There are previous studies on automatic summarizing systems or applications for community-contributed contents of a real-life event. Kenney et al. [9] proposed an automatic video mixing of concerts using audio fingerprints. Vihavainen et al. [3] presented an automatic remixing system for community contributed content from music concerts whereby users record and upload videos during live events and the shared content is then synchronized based on the creation timestamps. The single audio tracks of the synchronized videos are compiled to produce a master audio track. Through automatically detecting regions of interest in this master track, video remixes of a concert are automatically produced. However, as Gunes et al. [10] pointed out, for any event, elements like human emotions are still very difficult to be recognized automatically. Furthermore, the fully automatic video authoring methods are not in existence yet. Recent studies like Fabro et al. [11] focus on how to crawl community-contributed multimedia content from Youtube, Flickr to make video summaries of social events (using the Royal wedding of William and Kate as case study). The summaries produced are results of an algorithm which combines the multimedia content based on three criteria: quality, diversity and coverage. It is basically

data crawling without human acting as the producer. In our proposed service, we believe that authoring is best left to the human producers rather than algorithms. Furthermore, the previously developed systems enable only asynchronized summarizing whilst our service provides real-time synchronized summarization through the ability to jump across the different viewpoints.

Similar to our proposed service in terms of real-time editing, Engstrom et al [5] presented a live video broadcast video system whereby the camera operators interact and decide filming angles to collaboratively create a coherent narrative of an event. In this system, the different camera operators are the producers who collectively produce one summary video. On the other hand, in our concept, any single producer does not have to negotiate with another producer about angles, storyline, etc since the producer(s) can be the participant(s) anytime by jumping from one view point to another view point, including another producer's viewpoint. Thus, each producer creates their own summarized video which can be more coherent than that produced collaboratively by producers in Engstrom et al. [12].

There are a few areas that would eventually complete our service in the future. Current video broadcasting services with mobile wearable cameras do not provide any virtual embodiment for the remote participants (producers) yet as there is no display with augmented reality capabilities. Current smartphones support augmented reality using applications such as Layar but lack the functionality of producing audio and video stream (e.g. FaceTime on iPhone) while recording. Nor it is usually possible to use both of the smartphone cameras (one pointing forward and the other at the user's face) simultaneously, which would enable the producer to get further indications about what the spectators and performers are feeling. Discussion on hardware components is beyond the scope of this project.

Recognition of joyous moments e.g smiling and laughter could be a valuable support for selecting views where interesting things might be happening. According to Jacucci et al. [13], many of the most interesting moments that people wish to document happen spontaneously. Therefore, it would be also essential to provide a flexible and highly automated instant replay function that is already segmented to appropriate size for the producer to view it and clip it in the documentary while watching the primary view of the person 'jumped' into. However, this replay function is subjected to the capabilities of the devices used to record. Devices like Samsung Galaxy Note already has the instant replay function while other lower end devices like Aakash tablet does not.

IV. CONCLUSION AND FUTURE WORK

We presented a real-time video summary service that enables producers, who can be remote participants, to switch from one viewpoint to another viewpoint to create summary videos of live events. This service provides added value to the current literature of video summarizing applications by proposing the idea of viewpoint 'jumping' and how it can be

incorporated in accordance with other elements in a basic video summarizing and broadcasting system.

User studies will be conducted to test the proof of concept. A mockup dancing performance with a group of minimum five people including performers, spectators and remote audience will be conducted to evaluate the switching point-of-view interaction, level of participation felt by the producer, how the summaries are perceived by people other than the producer who have or have not attended the live events, and whether task allocation does result in the intended quality of experience.

Future work can explore how to perform this type of real time video summaries on longer events such as Olympic Games with many parallel sub-events.

ACKNOWLEDGMENT

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

REFERENCES

- [1] O. Juhlin, A. Engström, and E. Reponen, "Mobile broadcasting – The whats and hows of live video as a social medium," Proc. MobileHCI'10. The 12th international conference on Human computer interaction with mobile devices and services, ACM Press, Sept. 2010, pp. 35-44, doi: 10.1145/1851600.1851610.
- [2] D. Kirk, A. Sellen, R. J. Harper, and K. Wood, "Understanding videowork, " Proc. CHI 2007, ACM Press, 2007, pp. 61-70, doi: 10.1145/1240624.1240634.
- [3] S. Vihavainen, S. Mate, L. Seppälä, F. Cricri, and I. D. Curcio, "We want more: human-computer collaboration in mobile social video remixing of music concerts," Proc. CHI'11. The 2011 annual conference on Human factors in computing systems, ACM Press, May 2011, pp. 287-296, doi: 10.1145/1978942.1978983.
- [4] A. Engström, M. Esbjörnsson, and O. Juhlin, "Nighttime visual media production in club environments," Night and darkness: interaction after dark. Workshop at CHI 2008.
- [5] A. Engström, M. Esbjörnsson, and O. Juhlin, "Mobile collaborative live video mixing," Proc. MobileHCI 2008. The 10th international conference on Human computer interaction with mobile devices and services, ACM Press, Sept. 2008, pp. 157-166, doi: 10.1145/1409240.1409258.
- [6] S. Järvinen, J. Peltola, J. Lahti, and A. Sachinopoulou, "Multimedia service creation platform for mobile experience sharing," Proc. MUM'09. The 8th International Conference on Mobile and Ubiquitous Multimedia, ACM Press, Nov. 2009, pp. 1-9, doi: 10.1145/1658550.1658556.
- [7] B. Chen, J. Shen, and H. Sun, "A fast face recognition system on mobile phone," Proc. IEEE. International Conference on Systems and Informatics (ICSAI 12), IEEE Press, May 2012, pp. 1783-1786, doi: 10.1109/ICSAI.2012.6223389.
- [8] K. Satoh, M. Anabuki, H. Yamamoto, and H. Tamura, H, "A hybrid registration method for outdoor augmented reality," Proc. IEEE and ACM Symp. International Symposium on Augmented Reality (ISAR'01), IEEE Press, Oct. 2001, pp. 67-76, doi: 10.1109/ISAR.2001.970516.

- [9] L. Kennedy, and M. Naaman, "Less talk, more rock: Automated organization of community-contributed collections of concert videos," Proc. WWW 2009. The 18th international conference on World wide web (WWW 2009), ACM Press, Apr. 2009, pp. 311-320, doi: 10.1145/1526709.1526752.
- [10] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," Proc. FG'11. IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG'11), IEEE Press, Mar. 2011, pp. 827-834, doi: 10.1109/FG.2011.5771357.
- [11] M. Fabro, A. Sobe, and L. Boszormenyi, "Summarization of Real-life events based on community-contributed content," Proc. MMEDIA 2012. The 4th international conference on Advances in Multimedia (MMEDIA 12), IARIA, May 2012, pp. 119-126.
- [12] A. Engström, M. Perry, and O. Juhlin, "Amateur vision and recreational orientation: Creating live video together," Proc. CSCW. The ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12), ACM Press, Feb. 2012, pp. 651-660, doi: 10.1145/2145204.2145304.
- [13] G. Jacucci, A. Oulasvirta, and A. Salovaara, "Active construction of experience through mobile media: a field study with implications for recording and sharing," Personal and Ubiquitous Computing, vol. 11(4), Apr. 2007, pp. 215-234, doi: 10.1007/s00779-006-0084-5.

TV6: A Revisit to System Design, User Socialization and Content Recommendation in Social TV

Chong Yuan, Zhi Wang and Lifeng Sun
 Department of Computer Science and Technology
 Tsinghua University
 Beijing, China

{yuanc10, wangzhi04}@mails.tsinghua.edu.cn, sunlf@tsinghua.edu.cn

Abstract—The move to social TV has challenged the traditional TV experience by allowing users to interact with their friends and receive social multimedia contents when they are watching videos. Social TV has made the social experience equally important to users as the video contents. The rapid emergence of online social network service and video sharing service in today’s Internet, as well as the open interfaces (e.g., Open API) make it possible for a third party to build a social TV system, based on existing social network and online video service infrastructures. In this paper, we share our experience of designing, developing, deploying and operating TV6—a social TV system that embeds multiple online social network systems and online video sharing systems. Through detailed analysis of TV6, we present our observations and insights on provisioning superb social TV experience to users, including social content provision and traditional video content provision.

Keywords-social TV; video sharing; recommendation algorithm;

I. INTRODUCTION

Recent years have witnessed a rapid convergence of online social network service and online video service. Due to this convergence, Social TV, a new TV service, has emerged as important video experience in recent years (e.g., CollaboraTV [1] and Social TV from Motorola [2]). In Social TV service, users receive not only the traditional video contents from the video systems, but also the social multimedia contents from the social networks. Social TV allows users to interact and share their feelings with their friends while watching videos.

Numbers of works have been devoted to the system implementation of Social TV. Coppens et al. have designed AmigoTV, which allows users to communicate with each other using voice, text and animated emoticons. Which communication form is the best one has been fiercely debated (e.g., [3]). Boertjes et al. [4] have designed AOLTV, a Set-Top Box (STB) which offers users the ability to send and receive instant messages and emails on their TV. The commercial social Web TV Joost [5], enables users to do text chatting while watching online videos. Traditional user-generated content (UGC) providers like YouTube are also trying to allow users to communicate with each other

through video conference while watching the same video. D. Shamma et al. [6] have studied the integration of Twitter updates during live video streaming.

Previous works are mainly focused on designing a new Social TV system which includes building a new social network system. This can be too expensive to realize especially when there are dominant online social network systems such as Facebook and Twitter. To overcome this problem and build a cheap Social TV system, we propose to implement TV6—a social TV system based on the existing online social networks and online video sharing systems. We only control the logical information flow about how videos and social contents reach different users. The name “TV6” stands for the “Six-degree of separation”, which means users are closely connected in the system with diverse video and social sources. As TV6 embeds multiple online social network systems, we can use their social data, such as social graph and users’ profiles, to provide superb social experience to users. Our goal is to create a socially watching environment.

In this paper, we share our experience of designing, developing, deploying and operating TV6. A recommendation system is implemented in TV6 to recommend videos for both group and individuals. Our contributions can be summarized as follows: (1) System design and analysis of TV6, without owning any video service system and social network system, are presented in detail; (2) A variety of social experiences are provided to users by using the valuable social relationship and context from the social networks; (3) We present how we recommend videos to users when they are consuming the videos in the social network, e.g., they join social groups to watch a video.

The rest of the paper is organized as follows. We review the related works in Section II. Our system design is shown in Section III. Then social experiences provision and video contents provision are explained in Section IV and Section V respectively. Finally, we conclude the paper in Section VI.

II. RELATED WORK

People may feel closer to their friends when they are synchronously watching TV with them [7]. Traditional web video does not support synchronous viewing. In order to

address these explicit challenges, researchers have developed different solutions. Recently, D. Shamma et al. [7] propose three practical prototypes—“Messenger Zync”, “Web Zync” and “Invisible Zync”, which enable people to watch web video in synchronization. But these solutions have some limitations. The “Messenger Zync” is required to use Yahoo! Messenger for Windows where group conversations are not possible. Though the “Web Zync” and “Invisible Zync” provide group conversation, they do not embed social relationship in the system. It is not convenient for users to start a watching activity through the two prototypes. For example, user has to copy and paste the URL of video to every friend who he/she decides to invite.

Watching TV program with friends or families is considered to be a preferred social viewing experience. People like to share videos with friends and suggest interesting videos to others. The traditional TV or video service does not support video suggestion. When users want to recommend a funny clip to others, they have to send the video’s address through SMS or Instant Messaging(IM). Ambient TV [8] enables users to send video suggestion to others through the TV directly. G. Harboe et al. [8] also show that video suggestion can prompt communication between users and some people even use suggestion as a conversation starter. Explicit suggestions are playing important roles in social influence and pulling participants into a shared TV experience.

III. SYSTEM DESIGN AND IMPLEMENTATION

In this section, we present how we build TV6 as a third-party application, based on three large online social network systems and several large online video sharing systems, without real control of these systems.

A. System Architecture and Data Flow

Social network sites, such as Facebook [9] and Twitter [10] have begun to offer APIs to developers to access users’ profile data and social graph in order to create their own applications. Leveraging the APIs, we can quickly construct a social TV application.

TV6 is based on the design philosophy to incorporate social relations and user profiles from social network sites (SNS), and videos (only the title, URL and tag information) from video sharing systems. In particular, the social network sites used in TV6 are illustrated as below:

- Weibo [11]: A twitter-like service of Tencent company (Tencent is one of China’s largest Internet service portal)
- QQ [12]: An instant messaging service of Tencent company
- QZone [13]: A friend social network of Tencent company

TV6 server can retrieve the social graph and user profile from these three SNS after obtaining users’ authorization.

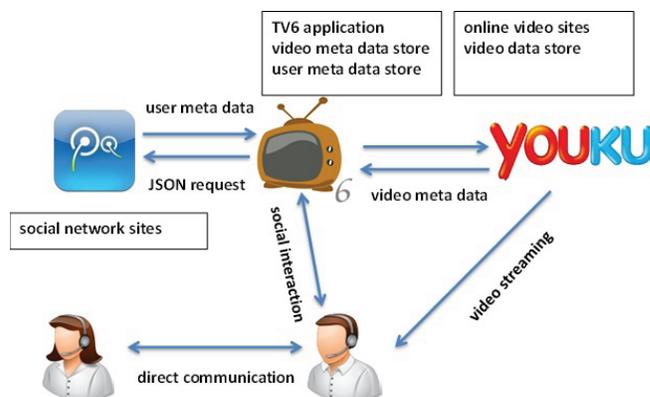


Figure 1. Overview of system architecture

The video sharing systems used in TV6 contain dominant video sharing sites in China, such as Youku [14] and Tencent video [15].

Fig. 1 provides an overview of the system architecture which is basically consisted of three parts. The actual TV6 application is set up on our own server which occupies the central position in the architecture. The application collects user meta data from the server of social network system and video meta data from the server of video sharing system periodically. These meta data are stored in the database of TV6 application. When a user performs an action, such as sharing a video or posting a comment, the TV6 server handles the request. Then the application generates appropriate answers which are JSON or XML responses and pushes these to the SNS server. The chat message are actually sent through the SNS server or application server. But it seems users can communicate with each other directly (described in Section IV-A). Both the TV6 server and the SNS server do not handle the requests of video steaming. Once a user begins to watch a video, the video data is loaded from the server of video sharing system. The system’s architecture is divided into three parts. The TV6 server plays a central role in the system and controls the logical data flow in the system. The SNS stores the user meta data and handles request from TV6 server. The video sharing system stores the video data and directly provides the video streaming for the user.

B. Social Information Collection

Our system is designed to make use of user profiles hosted by the existing online social networks. The open protocol OAuth is used in TV6 to access users’ profiles. When a user uses TV6 for the first time, he/she is asked to authorize our system to use his/her social profile via Open APIs provided by these online social networks. After the authorization, TV6 is able to access these information: (1) user’s social connections, including which users they are following and which users are following them; (2) users’ posted microblogs , including the ones with video links; (3) the number of

shares of a particular video link, which can later be used for video content provision.

C. Video Information Collection

As video links imported from different video sharing sites are collected in the microblogs from users' profiles, we are able to retrieve the video information from these video sharing sites. TV6 only collects the meta data, i.e. titles, URLs and tags of the videos, to reduce bandwidth and save storage. When users play videos in TV6, they will receive streaming data from the original video servers of the video sharing sites. After obtaining video titles and tags, we will use the word segmentation method to find patterns of words in them. These patterns of words can be clustered into several topics which can be used to describe the content of the videos. This is very important for the recommendation system, which will be described in Section V.

D. Content Presentation

After collection of social profiles and video meta data, the contents are presented back to users by TV6 as below. (1) *Video sharing list*: This page contains videos that users' friends have shared. Because users are probably interested in the content generated by their friends [16], they may want to watch the videos in the video sharing list. (2) *Watching list*: Users can see what their friends are watching now and viewing histories of their friends. (3) *Personal page*: It contains user's personal information such as his/her profile photo, viewing history and sharing history. (4) *Player page*: User can watch TV programs or videos on this page and chat with friends through text or voice at the same time.

IV. SOCIAL EXPERIENCE PROVISION

In this section, we will present how social information and interactions are provided to users.

A. Talking to Friends

Social communication is the basic feature in TV6. Rich communication tools with different purposes are provided in TV6.



Figure 2. An example of public chat room in TV6.

1) *IM-like Communication*: The most important facility of the Internet-based Social TV is that it connects friends located distantly, allowing them to watch videos simultaneously, as if they were sitting on the same couch [17], [18]. TV6 offers various ways of communication to supply the distant communication among different users. (1) Single chat via voice and/or text is offered between users in pairs. (2) Group voice chat is also offered in this system. Participants can speak simultaneously just like they were talking face to face. (3) Users in (1) and (2) are both chatting in privacy. TV6 also provides public chat rooms for users watching the same video to discuss and comment on it. Each video has its own chat room which is separate from other's. The new comments will be pushed to all users who are watching the same video by using the AJAX technology, as illustrated in Fig. 2.

2) *Email-like Communication*: The IM-like communication requires users to be online to join the discussions timely. We also provide the Email-like communications to sustain longer period responding. In particular, we use the *Tencent Weibo* and *Tencent QQ* to achieve the email-like communication. A TV6 user can leave comments to his friends by posting a message. The message will be sent to *Weibo* and *QQ* by the TV6 server. With the "@" function, the designated friend will notify the comments immediately when he/she login *Weibo* or *QQ*.

B. Knowing about Friends

In a Social TV system, it is important for users to be aware of their friends. In TV6, a user knows not only about what videos his friends have watched and are watching, but also about the exact playback position that his friends are currently at.

1) *Knowing Videos Watched by Friends*: When a user watch a video, TV6 stores the video record into the user's viewing history list. He/She can add his/her favourite videos to favorites list. The user's viewing history and favorites list are public to his/her friends. Users can figure out what their friends have watched in the past and what they like. This facility allows users to learn more about others' video viewing habits and preferences, and fosters a sense of connectedness.

video watching list

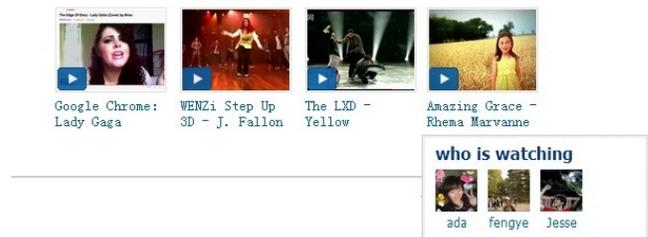


Figure 3. Knowing videos being watched by friends.

2) *Knowing Videos Being Watched by Friends:* In TV6, we also extended user's awareness to what videos are being watched now by his/her friends. TV6 provides a "video watching list", which contains the videos that the user's friends are watching, as illustrated in Fig. 3. Moving the mouse pointer upon the title of video in the watching list, we can see profile photos and names of users who are watching this video.

3) *Knowing the Exact Playback Position of Friends:* In TV6, user knows the exact playback position of his/her friends. There is a module embedded in our video player that can display the playback position of user's friends. When user moves mouse pointer upon the seek bar of the player, he/she can see photos of his/her friends who are watching the same video currently upon the seek bar, as illustrated in Fig. 4. The position of a friend's profile photo indicates his/her playback position. User's playback position is available to his/her friends by default and user can hide his/her playback position by change personal setting.

4) *Privacy Protection in TV6:* The privacy issues arise when users know everything about their friends' viewing status [19]. In TV6, we allow users to choose if their status is available to their friends. Every user has two states: on-line and away. The away state indicates that this user is not watching TV now. When he logs out, his state will be set to be away automatically. User's state is public to his friends by default. If user wants to watch video in privacy, he can set himself to be invisible and his friends won't know his status.

C. Watching with Friends

1) *Following A Friend's Playback Progress:* In Social TV, users are interested in sharing the same playback progress with their friends. The most popular Internet application, web video, does not support synchronous social conversation. People have to watch videos alone and the most common interaction in video sharing websites confines to asynchronous comment. But many people want to watch some TV programs or videos, such as football games, with their friends synchronously. Their demand may be difficult to be satisfied. When a user's friend provides a feedback or a commentary on the video's interesting moment, it is hard for the user to figure out which part of the video attract his friends. From our survey, in order to achieve the goal of synchronous watching, people may send the URL of a video to their friends and appoint a time when they start to watch it. Apparently, this is inconvenient.

In TV6, users can "follow" the playback progress of their friends by just clicking on friend's name or profile photo. Then the user will have a synchronized viewing experience with friends. The user can interrupt the synchronization at any time.

2) *Making Any Video Live in A Social Group:* Furthermore, our system also provides a functionality that users



Figure 4. Friend's profile photo indicates playback position.

can create activities to invite their friends to watch videos synchronously. Assuming that there is an important match tomorrow and you want to watch it with your close friends, you can create a private activity to invite your friends. After the activity is created, all the invited participants will get notifications about the activity. There is a conceptual "room" for every activity and the creator of the activity is the room's "master". The playback progress of participants in one activity are the same and synchronous experiences are shared. Since the number of participants may be very large, nobody except the "master" is authorized to control the playback progress, i.e., to pause, rewind or load a new video (change channel). At the same time, participants can communicate with each other through text and voice. Owing to this technique, remotely located friends and families can share watching experience together.

In order to guarantee the synchronization, a server was used to control the playback progress of participants. If the "master" changes playback status (pause, rewind or change channel), the server will inform other participants and change their status synchronously. All the other participants' progress synchronizes with the master's progress. The server checks every participant's progress regularly to guarantee the synchronization.

The synchronization module supplies a virtual co-present "watch together" and "on-the-couch" viewing experience. This feature offers social experience combining online video and traditional TV program which might become a very important characteristic of future social TV.

V. VIDEO CONTENT PROVISION

A. Group Recommendation

As described in Section IV-C2, users may form social group to enjoy co-present viewing experience. As a result, the concept of group interactions in such systems has been boosted. Such group viewing provides great potential for users to find videos that interest members in the group, namely, group recommendation [20].

In TV6, group characteristics are affected not only by inside group members, but also outsiders (e.g., followings). Furthermore, in our system, the relation within group is sparse. The information that we can use is not enough. So the recommendation that is only based on information within group is not accurate. According to [20], using information from *external experts* ("External Experts" indicate

the people who are friends of the group members but do not belong to the group), can help us make more accurate recommendation. Our system adopts this method which has proved its availability.

The database stores information of more than 30,000 videos which contains URLs, titles and tags. We use *Topic Model clustering algorithm* [21] to find out videos' *eigenvectors* by using their titles and tags. After words segmentation by *LDA* and *data training*, we obtain 10 topics by Topic Model clustering algorithm which have probability statistical significance and each video has a 10-dimension vector on them.

We calculate the characteristics of external experts by making use of their history behavior. If the video viewing history of an external expert contains N videos denoted as $v_i (i = 1, 2 \dots N)$ and we use V_i to represent the eigenvector of v_i . Then the preference of the external expert, denoted as P , can be calculated by

$$P = \frac{\sum_{i=1}^N V_i}{N} \quad (1)$$

Users of online social networks are often very interested in the content generated by their friends and there is strong relation between friends' preference and user's interests. So the group members' interests can be represented by their external experts' preference. As different external expert has different level of impact on group members, we need a weight to describe their different contribution to group members' interests. Here we set two kinds of weights. One is Weight of Friends Number since we know that friendship relation can get people with similar interest together. The other is Weight of Common Behavior as we have measured that common behavior is an important indicator of similar interest. So we count total two kinds of weights between an external expert and inner group members. Then summate them to be Weight of External Expert, denoted as w , to describe the contribution of the external expert to group members' interests.

Now we can calculate profile to group G from the preference of group members' external experts. We assume group G has M external experts. The preference of the external expert j is denoted as P_j and the weight j is denoted as w_j . We carry out the profile of group G , represented by F_G , as below:

$$F_G = \frac{\sum_{j=1}^M P_j \times w_j}{\sum_{j=1}^M w_j} \quad (2)$$

We get a "virtual external expert" profile for each group with Group Preference Model. We can use Topic Model to find out the *eigenvector* of each video in candidate list. Then we calculate the similarity, denoted as $sim(G, i)$, between each group profile and each video in the candidate list. We

carry out the similarity computing formulation as below:

$$sim(G, i) = \frac{V_i \cdot F_G}{\|V_i\| \times \|F_G\|} \quad (3)$$

We can sort the videos in the candidate list by $sim(G, i)$ and the sorted video list is the final group recommendation results.

With the above steps, we can get a comprehensive group recommend proposal which can fully meet needs of a more open social network with features of great dynamic and sparse tightness.

The algorithm of individual recommendation is almost the same as group recommendation.

B. Recommendation Experiments

We conduct experiments based on real-world group to evaluate the performance of our recommendation algorithm. We compare our algorithm with three other algorithms—Average satisfaction, Least Misery and Most Pleasure [22]. We invited 18 groups to form various scale of groups including 8 groups of 3 people, 6 groups of 5 and 4 groups of 8. These participants are college students and graduates with various profession backgrounds. People in one group enter into TV6 to watch videos together and our server automatically collect their meta data from SNS. When video play ends, TV6 recommends videos for them using the four recommendation algorithms. The group members mark scores ranging from 1 to 5 for the results. Then we calculate the average scores for each algorithm. The higher scores indicate better performance. As we have three group size:3, 5, 8, we compare the four algorithms under different group size. Fig. 5 illustrates the results. Apparently, our algorithm performs better than other algorithms.

In face of sparse relation among group members, traditional algorithms have a poor performance. In contrast, our algorithm concerning the influence of external experts is more effective. Because group members are very interested in content generated by their friends, using external experts' history behavior can well reflect group members' interests. Our algorithm can also be used in other situation, such as e-commerce if we use tweets with e-commerce content instead of video content.

VI. CONCLUSION AND FUTURE WORK

We discuss how to design and deploy TV6, a novel social TV system, embeds multiple social network systems and online video sharing systems in this paper. The TV6 application is publicly available now and more than five thousand people have installed it. A number of users give us their experience and some valuable advice. Based on the users' feedback, we learn about how users use the social TV system. TV6 is also currently used as an experiment platform in a course of Tsinghua University and students are conducting experiments on it.

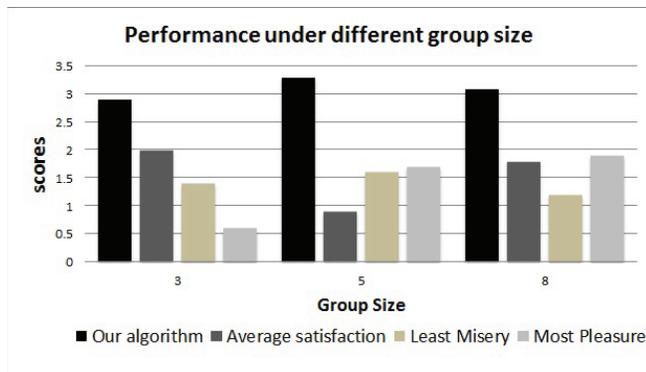


Figure 5. Performance under different group size.

We found that the awareness of friends in TV6 can make users more engaged into the social television experience and foster a sense of connectedness through television. Some users say that they like to watch videos in their friends' viewing history and favorites. After they login TV6, they may go directly to the video watching list to view videos that their friends are watching currently. Users are often interested in the content generated or shared by their friends. Some users even use their friends' viewing history and favorites as schedules of program that they plan to watch. Provided viewing history and favorites of users' friends can help users to learn about their friends' interests. Given the access to know what others are watching and their playback progress, users find social TV viewing an opportunity to communicate with each other. Users may feel closer to others while interact with their friends and the sociability among users is also enhanced.

For future work, we will combine the internal and external factors to see whether the comprehensive method can enhance the group recommendation. We will also do some experiments on larger groups as the group size are relatively small now.

REFERENCES

- [1] M. Nathan, C. Harrison, S. Yarosh, L. Terveen, L. Stead, and B. Amento, "Collaboratv: making television viewing social again," in *Proceedings of the 1st international conference on Designing interactive user experiences for TV and video*, ser. UXTV '08. New York, NY, USA: ACM, 2008, pp. 85–94.
- [2] C. Metcalf, G. Harboe, J. Tullio, N. Massey, G. Romano, E. M. Huang, and F. Bentley, "Examining presence and lightweight messaging in a social television experience," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 4, no. 4, pp. 27:1–27:16, Nov. 2008.
- [3] J. Tullio, G. Harboe, and N. Massey, "Investigating the use of voice and text chat in a social television system," *Changing Television Environments*, pp. 163–167, 2008.
- [4] E. Boertjes, J. Klok, O. Niamut, and M. Staal, "Connectv: Share the experience," in *Adjunct Proceedings of EuroITV*, 2007, pp. 139–140.
- [5] <http://www.joost.com>. October 2012.
- [6] D. Shamma, L. Kennedy, and E. Churchill, "Tweet the debates: understanding community annotation of uncollected sources," in *Proceedings of the first SIGMM workshop on Social media*. ACM, 2009, pp. 3–10.
- [7] D. A. Shamma, M. Bastea-Forte, N. Joubert, and Y. Liu, "Enhancing online personal connections through the synchronized sharing of online video," in *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '08. New York, NY, USA: ACM, 2008, pp. 2931–2936.
- [8] G. Harboe, C. J. Metcalf, F. Bentley, J. Tullio, N. Massey, and G. Romano, "Ambient social tv: drawing people into a shared experience," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '08. New York, NY, USA: ACM, 2008, pp. 1–10.
- [9] <http://developers.facebook.com>. February 2013.
- [10] <https://dev.twitter.com>. February 2013.
- [11] <http://t.qq.com>. February 2013.
- [12] <http://im.qq.com>. February 2013.
- [13] <http://qzone.qq.com>. February 2013.
- [14] <http://www.youku.com>. February 2013.
- [15] <http://v.qq.com>. February 2013.
- [16] P. Cesar and D. Geerts, "Past, present, and future of social tv: A categorization," in *Consumer Communications and Networking Conference (CCNC)*. IEEE, 2011, pp. 347–351.
- [17] D. Geerts and D. De Grooff, "Supporting the social uses of television: sociability heuristics for social tv," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '09. New York, NY, USA: ACM, 2009, pp. 595–604.
- [18] P. Cesar and D. Geerts, "Understanding social tv: a survey," in *Proceedings of the Networked and Electronic Media Summit (NEM Summit 2011)*, Torino, Italy, September 2011, pp. 94–99.
- [19] G. Harboe, N. Massey, C. Metcalf, D. Wheatley, and G. Romano, "The uses of social television," *Comput. Entertain.*, vol. 6, no. 1, pp. 8:1–8:15, May 2008.
- [20] X. Wang, L. Sun, Z. Wang, and D. Meng, "Group recommendation using external follower for social tv," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, July 2012, pp. 37–42.
- [21] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 113–120.
- [22] M. Gartrell, X. Xing, Q. Lv, A. Beach, R. Han, S. Mishra, and K. Seada, "Enhancing group recommendation by incorporating social relationship interactions," in *Proceedings of the 16th ACM international conference on Supporting group work*. ACM, 2010, pp. 97–106.

Local Histogram Modification Based Contrast Enhancement with GPU Acceleration

Jiang Duan, Min Li, Haiyue Wen
School of Economic Information Engineering
Southwestern University of Finance and Economics
Chengdu, P. R. China
duanj_t@swufe.edu.cn, torrentlee@hotmail.com,
wenhaiyue@gmail.com

Yingjie Peng
Cavendish Laboratory
University of Cambridge
Cambridge, United Kingdom
y.peng@mrao.cam.ac.uk

Abstract – This paper presents a novel local contrast enhancement algorithm based on local histogram modification. The computation of local contrast enhancement operators is usually slow though they produce better local contrast and details. We have addressed this issue by subtly designing a highly parallel algorithm, which could be easily implemented on Graphics Processing Units (GPU) to harvest high computational efficiency. Our method is fast and easy to use, and the experiment results show that the technique can produce good results on a variety of images.

Keywords – GPU; contrast enhancement; histogram modification.

I. INTRODUCTION

Contrast enhancement is an important step in digital image processing when visual perception of information is limited by small differences in gray levels in the image. Several reasons, such as the limitation of the imaging devices and the adverse capturing conditions, make an image or a video have low contrast that the intensity levels of the pixels reside densely in a narrow range in the histogram of the image. In this case, the available dynamic range is not fully utilized. As a result, such images or videos may have a washed-out and unnatural look and lose details of the original scenes. Contrast enhancement techniques redefine pixel values in order to make full use of dynamic range and render various contents of images easily distinguishable. This improves the image quality of a display and visual perception of human beings. Contrast enhancement is useful and widely used in many applications, such as digital photography, medical image analysis, remote sensing and scientific visualization.

This paper will address the issue by presenting a novel local contrast enhancement algorithm based on local histogram modification with mechanism of parallel computation, which will then be accelerated using GPU. The organization of the paper is as follows. In the next section, we briefly review previous work of histogram modification based contrast enhancement methods. We describe our algorithm in detail in Section III. Section IV describes the implementation of the designed algorithm. Section V presents experimental results and Section VI concludes the paper.

II. REVIEW OF HISTOGRAM MODIFICATION BASED CONTRAST ENHANCEMENT METHODS

Several contrast enhancement techniques have been introduced to improve the contrast of an image, among which histogram modification techniques receive the most attention due to straightforward and intuitive implementation qualities. These methods modify the image through some pixel mapping such that the histogram of the processed image is more spread than that of the original image.

Histogram modification techniques are usually classified as either global or local. Global histogram modification techniques derive a single mapping from the image and apply it to every pixel across the image. They do not involve spatial processing and are therefore computationally very simple. Histogram Equalization (HE) is one of the most commonly used algorithms [1]. The mechanism of HE is to transform the gray levels of an image to a uniform histogram based on the probability of occurrence of gray levels in an input image. However, HE without any modification may result in an excessively enhanced output image and cause unacceptable visual artifacts. Various methods have been proposed to improve HE. Bi-Histogram Equalization (BHE) is proposed to overcome the brightness preservation problems [2]. BHE divides the input histogram into two sub-histograms based on the mean brightness and the two sub-histograms are then manipulated by HE individually. A similar method in [3] creates the two separate histograms using the median intensity instead of the mean intensity. Other sub-histogram methods include [4-8]. They mainly differ in how to separate the input histogram.

Global contrast enhancement is less than optimal, especially when the image contains large areas with substantially different average gray levels, and the contrast within each part is low. In such a case the original histogram is already fairly flat and any further global equalization does little to improve the local contrast. This problem can be addressed by local histogram modification techniques which use a spatially varying mapping based on local pixel statistics and contexts. Local methods can make premium contrast enhancement effect but at higher computational cost. In local histogram modification (LHM) methods in [9, 10], a square neighborhood is defined around each pixel, over which the histogram is equalized.

Adaptive histogram equalization (AHE) [11] divides the whole image into square regions and a histogram equalizing transformation is found for each region separately. The modified value for every pixel is then calculated by bilinear interpolation between the transformations given for the four neighboring regions. In [12], a low-pass filter-type mask is used to get a non-overlapped sub-block histogram-equalization function to produce the high contrast associated with local HE but with the simplicity of global HE. Other local methods include [13-15].

III. ALGORITHM

Our method divides images into non-overlapping regular rectangular blocks and reproduces the contrast and brightness in each of them simultaneously using a very parallel global contrast enhancement operator. Finally, a weighting scheme is used to eliminate the boundary artifacts caused from the contrast adjustment in different blocks. This method subtly addresses the issue that local operators are hard to be paralleled and provide promise for GPU acceleration.

A. Global Contrast Enhancement

Min-max linear contrast stretch and HE are two commonly used contrast enhancement methods. When using the linear contrast stretch, the intensity values of the original image are linearly mapped to a newly specified set of values, usually the full range of available brightness values. Consider an image with a minimum and maximum value of D_{min} and D_{max} . If this image is displayed on a visualization device with minimum and maximum displayable levels P_{min} and P_{max} (which are usually 0 and 255), the range of $[P_{min} to D_{min}]$ and $[D_{max}, P_{max}]$ will be not displayed and thus wasted. In this case, the dynamic range of the display device are not made full use of. Linear contrast stretch targets to expand the narrow range of $[D_{min}, D_{max}]$ to $[P_{min}, P_{max}]$ as:

$$I(x, y) = \frac{D(x, y) - D_{min}}{D_{max} - D_{min}} * (P_{max} - P_{min}) \quad (1)$$

where $D(x, y)$ and $I(x, y)$ are input and output pixel values; (x, y) is pixel coordination; D_{min} and D_{max} are minimum and maximum gray levels of the original image; P_{min} and P_{max} are minimum and maximum displayable levels of the visualization device. It's more convenient to regard linear contrast stretch as a mapping function LC, which divides compact range $[D_{min}, D_{max}]$ into 256 equal length intervals using cutting points l_n and maps pixels falling into the same interval to the same integer display level d . This process can be visually demonstrated by Fig. 1(a).

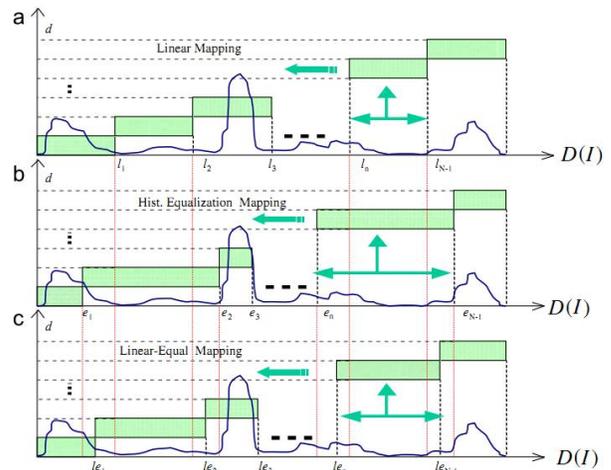


Figure 1. Contrast enhancement as mapping function. (a) min-max linear contrast stretch; (b) HE; (c) our algorithm.

HE transforms the gray levels of an image to a uniform histogram. Consider a digital image $D(x, y)$ which has the total number of S pixels with gray levels in the range $[0, L-1]$. The probability density function (PDF) $P(k)$ of the image is defined as:

$$P(k) = \frac{n_k}{S}, \text{ for } k = 0, 1, 2, \dots, L-1 \quad (2)$$

where L is the maximum gray level of image; n_k is the total number of pixels in the image with gray level k . The cumulative distribution function (CDF) of the image is then obtained by:

$$C(k) = \sum_{i=0}^k P(i), \text{ for } k = 0, 1, 2, \dots, L-1 \quad (3)$$

HE will map an input gray level k into an output gray level $EC(k)$ using the following mapping function:

$$EC(k) = (L-1) * C(k) \quad (4)$$

The mapping process is demonstrated in Fig. 1(b). HE divides $[D_{min}, D_{max}]$ into 256 intervals such that the number of pixels falling into each interval is the same. All pixels falling into the same interval are mapped to the same integer display level d . e_n are the cutting points.

Linear enhancement and HE have their own disadvantages. Linear contrast stretch is done purely on the basis of the actual pixel values without taking into account the image's pixel distribution characteristics. As a consequence, in densely populated intervals, too many pixels are squeezed into one display level, resulting in a loss of detail and contrast, while in sparse population intervals, too few pixels occupy quite a few valuable display levels thus resulting in the under utilization of display levels. HE takes into account pixel distribution and

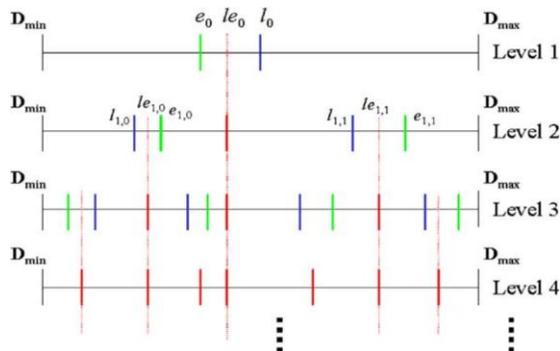


Figure 2. Recursive binary cut approach for LEC.

the display levels can be fully utilized. However, densely populated intensity intervals can result in the exaggeration of contrast while in sparsely populated luminance intervals mapping is too aggressive that the mapped gray levels will be very different from original values.

Fortunately, the drawbacks of the two methods are compensated by one another. In order to achieve the desirable results, our designed enhancement method strikes a balance between the linear contrast stretch and HE as shown in Fig. 1(c) using cutting points le_n , which is computed as:

$$le_n = l_n + \beta * (e_n - l_n) \tag{5}$$

where β is a controlling parameter. If $\beta=0$, the enhancement is linear; $\beta=1$, the enhancement is HE. In a sense, β controls the contrast enhancement level in the mapping process. Setting $0 \leq \beta \leq 1$, we can strike a balance between the two extreme forms. We call this mapping function *LEC*.

To implement *LEC*, we have developed a highly efficient recursive binary cut approach as illustrated in Fig. 2. This recursive binary cut approach first divides the range of $[D_{min}, D_{max}]$ into two segments according to (5) on level 1. Then these two segments are each independently divided into 2 segments similarly on level 2. The process is then applied recursively onto each resultant segment until level 8 with 256 segments created, each of which will be then allocated one corresponding displayable value between 0 and 255.

Fig. 3 shows the original image and the result from our global contrast enhancement method, and their corresponding luminance histograms. It's obvious that the histogram of the processed image is more stretched out than that of the original image, which means that the available gray levels of the display device are in better use and this makes the resultant image more visually pleasing. However, some regions lose local contrast and detail instead, like the wall of the building in the center of the image. This is because global contrast enhancement methods have the problem that it cannot improve the regional contrast since it uses only one mapping curve for the entire image as discussed in Section II. In order to address this problem, we

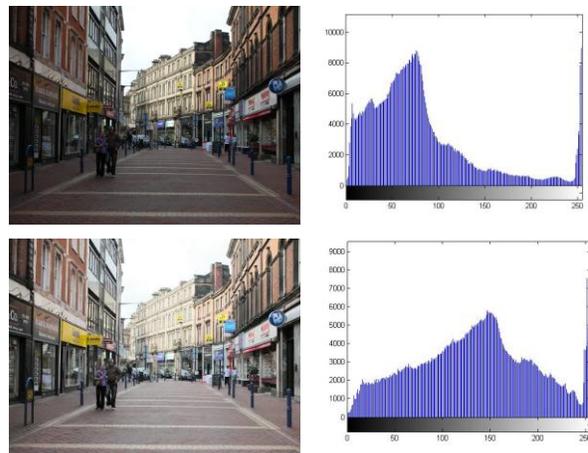


Figure 3. First row: original image and its luminance histogram; second row: processed result from our global contrast enhancement method and its luminance histogram.

extend our global method to a local one in the following section.

B. Global contrast enhancement in local regions

Following previous local contrast enhancement methods like [11], we segment image into non-overlapping rectangular regions, in which we compute local LEC_n ($1 \leq n \leq R$, where R is the number of segmented regions) based on the pixel statistics in each region in the same way as in the global case described in Section III(B). We use a common parameter $\beta=0.6$ for all the regions in our local method. If we regard LEC_n as the mapping function, for an individual pixel luminance value $D(x,y)$, output integer display level $d(x,y)$ is given by

$$d(x,y) = LEC_n[D(x,y)] \quad (x,y) \in n \tag{6}$$

Fig. 4 shows the result directly from local LEC. Obviously, the image shows more details and local contrast in either dark or bright regions in comparison with the global case shown in Fig. 3.

However, the direct application of LEC in each independent local area causes sharp jumps among different regions. The result is the boundary artifacts shown in Fig. 4, making the mapped images unacceptable despite of the improvement in detail visibility and local contrast. This is due to the fact that LEC_n are computed based on different luminance distributions. Pixels with similar values but on different sides of the local regions boundaries can be projected to have very different values and thus lead to boundary artifacts.



Figure 4. Result directly from LEC in local regions.



Figure 5. Final result after eliminating boundary artifacts and considering local contrast enhancement adjustment.

C. Boundary artifacts elimination

To eliminate the boundary artifacts, we introduce a weighting scheme. For each pixel value $D(x, y)$ in the image, the final mapped pixel value is the weighted average of the results of N nearest regions according to a distance weighting function:

$$d(x, y) = \frac{\sum_{n=1}^{n=N} LEC_n[D(x, y)] \cdot w_d(n)}{\sum_{n=1}^{n=N} w_d(n)} \quad (7)$$

$$w_d(n) = e^{-(d_n/\sigma_d)} \quad (8)$$

where N is the number of blocks used; w_d is the distance weighting function; d_n is the Euclidean distance between the current pixel position and the center of each of the used regions. σ_d controls the smoothness between blocks. Larger values of N and σ_d facilitate the elimination of boundary artifacts but will produce image with less local contrast. Fig. 5 shows the final result until this step. We can see all undesirable artifacts have been removed.

IV. GPU IMPLEMENTATION

GPU has of late gained considerable computational power and the introduction of programmability has enabled its use outside the original application domain of computer graphics for more general purposed computing tasks. In the field of image processing, some researchers have already considered to take the advantage of CPU implementation [16]. CUDA is often mentioned among them [17].

A. Basics of CUDA

CUDA is a newly emerged scalable parallel programming model and a software environment for parallel computing on GPU [18]. It allows almost the direct translation of C code onto the GPU, with the syntax consisting of minimal extensions of the C language.

CUDA programmers launch kernels to accomplish computation tasks on GPU. One important way in which kernels differ from normal C functions is that they are executed in parallel, over a large number of CUDA threads. Individual threads execute in parallel the same kernel program on different data. Threads are organized into blocks and blocks make up grids, as shown in Fig. 6. Built-in variables threadIdx, blockIdx and gridIdx, up to three dimensions, help locate each thread and determine what data to work on. The tricky parts of CUDA programming are to decide the grid and block size, and identify target data using the mentioned ID variables. Kernel program is launched as:

```
kernel<<<grid_size, block_size>>>( arguments );
```

We will describe their CUDA implementation in detail in the next two sections.

B. Accelerating local mapping function construction

In addition to simultaneous deriving local mapping function in each region, we also propose a parallel implementation of LEC operator as shown in Fig. 2. This recursive binary cut approach first divides the range of $D(I)$ into two segments according to (5) on level 1. Then these two segments are each independently divided into 2 segments similarly on level 2. The process is then applied recursively onto each resultant segment until level 8 with 256 segments created, each of which will be then allocated one corresponding displayable value between 0 and 255. On level i , 2^{i-1} cuts are created independently and thus could be calculated on GPU simultaneously. We launch one kernel program for each level as:

```
DeriveLEC_i<<<Grids, 2i-1>>>( arg ); ( i = 1, 2 ... 8)
```

DeriveLEC_i is the calculation on level i (5). Grids is a two dimensional variable with each component equal to the number of regions in the image vertically and horizontally. Kernels are so launched to make sure one CUDA block is responsible for constructing mapping function in one local zone and each thread serves to create a new cut between segments.

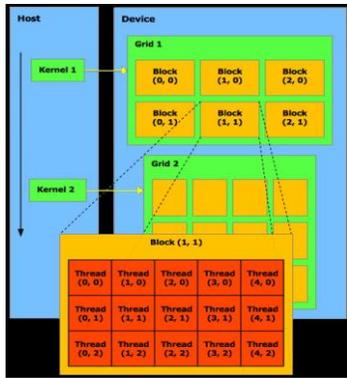


Figure 6. CUDA threads organization. Figure is courtesy of NVIDIA.

C. Accelerating weighting process

As discussed in Section III(B), we could conduct computation according to (7) for each pixel across the image concurrently. To put it into practice, we pre-calculate the distance weighting function and the similarity function, and then launch just one kernel as:

Weighting <<<Grids, Blocks>>>(arg);

Weighting is the kernel program to calculate (7). Grids is the same as that in the previous section. Blocks is a two dimensional variable. Its first and second dimension size is equal to the number of pixels of a local zone horizontally and vertically respectively. In this manner, each CUDA thread is in charge of the weighting process for one pixel to get the final mapping result.

V. EXPERIMENTAL RESULTS AND DISCUSSION

As discussed in Section III, our local contrast enhancement method controls enhancement level using several parameters, namely contrast controlling parameter β in the global method, the number of segmented regions R , the number of regions used to eliminate boundary artifacts N , smoothness control between blocks σ_d . It is intuitive that larger β means more local contrast enhancement. In term of region number R , mapping results of image segmented into more regions obviously have more local contrast since the full dynamic range of the display can be better utilized in local areas. Large N and σ_d result in an image free from boundary artifacts but with less local contrast. For most cases in our experiments, setting β to 0.6, R to 32×32 , N to 7×7 and σ_d to 18.0 leads to good results and therefore we choose these values as our default parameters in order to overcome the difficulty of too many parameters for users to set.

Fig. 7 shows resultant images from different contrast enhancement methods. The resultant image of our algorithm is comparable to those of HE, and contrast limited adaptive histogram equalization CLAHE [13]. All the results from local methods are produced using default parameters. In the bottom left image produced by HE, there is less local contrast like the wall of the building in the center of the image. In the bottom right image from CLAHE, the contrast is so strong that noises are presented, such as the road. Fig. 8 shows more results of our algorithm.



Figure 7. Resultant images from different contrast enhancement methods. From left to right, top to bottom: original image, result from our algorithm, result from HE and result from CLAHE [13].

To demonstrate the computational efficiency of the proposed method, we implemented it on both CPU and GPU. For the $768 * 1024$ pixel test image, it takes 1.477s for an i5-2410M CPU @ 2.30Hz with 4GB RAM running 64-bit Windows 7 Ultimate to compute the final result. The mapping function construction and weighting process occupy 0.392s and 0.899s respectively. The GPU experimental platform is NVIDIA GeForce GT 550M with 2 multiprocessors. Without considering careful optimizations of memory use and cooperation between CPU and GPU, CUDA codes could shorten the time to 0.358s to compute the same test image above. Specifically, mapping function construction time has been reduced to 0.172s while weighting process time to 0.103s, from which we could experience about 2 and 9 times speedup for each part. The reason that the weighting process has gained higher ratio of acceleration is because it has more parallelism we could utilize. In other words, there are more computations with potential to be paralleled, $768 * 1024$ in this case.



Figure 8. More results of our algorithm.

VI. CONCLUSION AND FUTURE WORK

In this paper, we present a novel local contrast enhancement method based on local histogram modification with mechanism of parallel computation. It can be further accelerated by using GPU. We describe not only the detailed algorithm of the method, but also the implementation of it. The experiment results show that the method has been demonstrated fast and effective for enhancing images.

Future work will focus on optimizing CUDA implementation, obtaining a better acceleration from the perspective of algorithm design [19][20] and using a better GPU to render video in real time.

ACKNOWLEDGEMENT

Images used in this paper courtesy of corresponding author(s). This project is sponsored by National Natural Science Foundation of China (Grant No. 60903128), the Scientific Research Foundation for the Returned Overseas Chinese Scholars and the Program for New Century Excellent Talents in University of State Education Ministry, and Excellent Youth Foundation of Sichuan Scientific Committee (2012jq0017).

REFERENCES

- [1] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. New Jersey: Prentice-Hall, Inc., 2001.
- [2] Y. T. Kim, "Contrast enhancement using brightness preserving bi-histogram equalization," *IEEE Trans. Consum. Electron.*, vol. 43, no. 1, Feb. 1997, pp. 1-8.
- [3] Y. Wang, Q. Chen, and B. Zhang, "Image enhancement based on equal area dualistic sub-image histogram equalization method," *IEEE Trans. Consum. Electron.*, vol. 45, no. 1, Feb. 1999, pp. 68-75.
- [4] S. D. Chen and A. R. Ramli, "Minimum mean brightness error bi-histogram equalization in contrast enhancement," *IEEE Trans. Consumer Electron.*, vol. 49, no. 4, Nov. 2003, pp. 1310-1319.
- [5] S. D. Chen, and A. R. Ramli, "Contrast enhancement using recursive mean-separate histogram equalization for scalable brightness preservation," *IEEE Trans. Consumer Electron.*, vol. 49, no. 4, Nov. 2003, pp. 1301-1309.
- [6] K. S. Sim, C. P. Tso, and Y. Y. Tan, "Recursive sub-image histogram equalization applied to gray scale images," *Pattern Recognition Letters*, vol. 28, no. 10, Jul. 2007, pp. 1209-1221.
- [7] M. Abdullah-Al-Wadud, M. H. Kabir, M. A. A. Dewan, and O. Chae, "A dynamic histogram equalization for image contrast enhancement," *IEEE Trans. Consumer Electron.*, vol. 53, no. 2, May 2007, pp. 593-600.
- [8] H. Ibrahim and N. S. P. Kong, "Brightness preserving dynamic histogram equalization for image contrast enhancement," *IEEE Trans. Consumer Electron.*, vol. 53, no. 4, Nov. 2007, pp. 1752-1758.
- [9] D. J. Ketcham, R. W. Lowe and J. W. Weber, "Image enhancement techniques for Cockpit Displays" No. TR-P74-530R Display Systems Laboratory, Hughes Aircraft Co., Culver City, CA, USA, 1974.
- [10] R. Hummel, "Image enhancement by histogram transformation," *Comput. Graphics and Image Process.* vol. 6, no. 2, Apr. 1977, pp. 184-195.
- [11] S. M. Pizer, J. B. Zimmerman, and E. V. Staab, "Adaptive grey level assignment in CT scan display," *Journal of Computer Assisted Tomography*. vol. 8, no. 2, Apr. 1984, pp. 300-305.
- [12] J. Y. Kim, L. S. Kim, and S. H. Hwang, "An Advanced Contrast Enhancement Using Partially Overlapped Sub-block Histogram Equalization," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 4, Apr. 2001, pp. 475-484.
- [13] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," In *Graphics gems IV*, Paul S. Heckbert (Ed.). Academic Press Professional, Inc., San Diego, CA, USA, 1994, pp. 474-485.
- [14] J. A. Stark, "Adaptive image contrast enhancement using generalizations of histogram equalization," *IEEE Trans. Image Process.*, vol. 9, no. 5, May 2000, pp. 889-896.
- [15] T. Iwanami, T. Goto, S. Hirano, and M. Sakurai, "An adaptive contrast enhancement using regional dynamic histogram equalization", *IEEE International Conference on Consumer Electronics (ICCE)*, Jan. 2012, pp. 719-722.
- [16] T. Scheuermann and J. Hensley. "Efficient histogram generation using scattering on GPUs," *Proceedings of the 2007 symposium on Interactive 3D graphics and games*. ACM, pp. 33-37.
- [17] Z. Yang, Y. Zhu, and Y. Pu. "Parallel image processing based on cuda," *Computer Science and Software Engineering, 2008 International Conference on*. vol. 3. IEEE, 2008, pp. 198-201
- [18] <http://developer.nvidia.com/object/cuda.html> [Retrieved: Feb. 2013]
- [19] K. E. van de Sande, T. Gevers, and C. G. Snoek. "Empowering Visual Categorization with the GPU," *Multimedia, IEEE Transactions on*, 2011, 13(1), pp. 60-70.
- [20] W. T. Chu and S. C. Tseng, "GPU-Accelerated Scene Categorization under Multiscale Category-Specific Visual Word Strategy," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*. 2012, pp. 885-888.

Collaborative Multimedia Platform for Computational Philology

CoPhi Architecture

Angelo Mario Del Grosso
Istituto di Linguistica Computazionale “A. Zampolli”
Consiglio Nazionale delle Ricerche
Pisa, Italy
angelo.delgrosso@ilc.cnr.it

Federico Boschetti
federico.boschetti@ilc.cnr.it

Abstract— This paper aims at illustrating a collaborative and modular web platform in the domain of digital and computational philology. The proposed work deals with parallel multilingual and multimedia resources. Two case studies are discussed in order to show the flexibility of the designed platform. The reusability of the components in different projects is achieved by abstract modeling and through the application of effective design patterns. The platform deals with textual resources and associated multimedia content, which can be retrieved by the metadata and shown in parallel (e.g., the page image of a manuscripts and the related transcription). The library of components will distribute under GPL 3.0 license and available at <https://github.com/CoPhi>.

Keywords—computational philology; digital philology; multilingualism; parallel multimedia; enterprise systems.

I. INTRODUCTION

A general introduction provides an overview on initiatives and projects related to the specific domain of digital philology. The proposed methodology, illustrated in the second section, is based on the application of design patterns. Two pilot projects constitute the results shown in the third section: the former is an application for digital epigraphy and the latter is an application to manage manuscripts. Finally, conclusions are discussed, in order to point out the flexibility and reusability of the system.

Current trends in digital and computational philology are focused on the implementation of collaborative environments, in order to provide the international scholarly community with a suitable infrastructure to share and reuse scientific products, such as digital critical editions, commentaries, linguistic and stylistic analyses, annotations on manuscript page images, descriptions of archaeological and epigraphical artifacts, etc. [13-17][19][23]

The main efforts to achieve this aim are (1) the standardization of data formats for literary and philological studies [2][16][19][23], (2) the modelling of domain ontologies [4-6], (3) the interconnection with disciplines focused on text-bearing objects, such as Digital

Epigraphy [13][23] or artifacts mentioned in literary works, such as Digital Archaeology and, eventually, (4) the management of data by web-services through common protocols, such as OAI-PMH [1].

The standardization of data formats for literary and philological studies is the mission of the Text Encoding Initiative (TEI [2]). The TEI provides XML schemes and guidelines for text, extra-text, and para-text encoding with bibliographical, linguistic and philological meta-information.

Domain ontologies are provided by various institutions. Europeana [3], a platform for information and knowledge exchange in the domain of Digital Humanities, has formalized a data model that is becoming a de-facto standard (EDM: Europeana Data Model [4]).

The CIDOC-CRM [5], on the other hand, provides the conceptual reference model in the domain of Digital Archaeology and a joint effort between CIDOC-CRM and the Functional Requirements for Bibliographic Records has created FRBRoo [6]: an ontology intended to link bibliographical and museographic information.

Web-services allow the data exchange among working groups distributed world-wide. In the field of classical literature and philology, Bamboo [7] and Interedition [8] aim at providing web-services to make critical editions in collaborative environments.

The Perseus Project [9] (Tufts University) is the leading initiative that provides scholars with the suitable cyberinfrastructure: Philologist, powered by Son of Suda (SoSol [10]), and Alpheios [11] are the web applications that allow the version-controlled editing of texts and linguistic analyses associated to them. The identification of textual units is formalized by the Canonical Text Service (CTS [12]), which associate a URN (Uniform Resource Name) to every word of any specific edition.

As shown above, standard data formats on one hand and web-services on the other hand, highly promote the interoperability. But in the field computational philology it is necessary to improve also the software reusability. Whereas libraries and API for information retrieval, such as Lucene or linguistic analysis, such as LingPipe exist and are

maintained, libraries devoted to the specific field of computational philology are wanted: computational linguistics analyzes a single text flow associated to single linguistic analyses (e.g., syntactic and semantic analyses), whereas computational philology must deal with multiple versions of the same text (due to variants in the manuscripts or conjectural emendations provided by the scholars) and multiple interpretations at each level of analysis (due to the disagreement of authoritative scholars recorded in several commentaries along the centuries).

The main purpose of our work is the constitution of a library of components (the CoPhi Beans library) focused on philological activities, such as the alignment of complex textual objects (e.g., the alignment of variants according to their semantic similarity, not only according to the edit distance of the inflected forms), the extension of levels of analysis (e.g., metrical and colometrical analysis) or editing and retrieval of multiple, concurrent annotations. Furthermore, the linkage of textual resources to multimedia sources, such as the manuscript page images, must be taken into account in the new paradigm of philology in the digital age, which pays increasing attention to the disintermediation between the philologist and the (digital representation of) the primary sources.

Due to the abstract modeling and the modular design, our library and the platform based on, even if it is still in alpha version, is already used in a number of national and international projects devoted to manage parallel multimedia resources, such as text, image and music combined together.

Three main areas are involved for developing a platform based on the CoPhi Beans library able to deal with such spread needs:

- Acquisition of resources by Optical Character Recognition (OCR) or information extraction and document transformation from semi-structured resources to structured ones (ETL: Extract, Transform and Load);
- Text processing and indexing;
- Collaborative Enterprise Application designing and developing.

Eventually, we are designing and developing components, modules and plug-ins for a collaborative enterprise web-based system in order to build a suitable environment to analyze, on one hand, manuscripts and printed documents and, on the other hand, to produce new critical editions.

II. METHODOLOGY

Flexibility and reuse is achieved by a modular and abstract design of the components.

The core of the platform is: (1) the view resources component, (2) the search and index component, (3) the analysis component, (4) the comment and collaborative component (5) the editorial component.

The architecture of our platform is based on the MVC (Model View Controller) pattern, which separates the

business model from the GUI (Graphical User Interface) and from the objects devoted to the behavioral aspects. Fig. 1 shows the class model design involved in commenting results and the design model involved in the analysis process.

Comments written by the users in natural language and micro-annotations automatically produced by morpho-syntactic parsers must be editable versioned and searchable. Furthermore Textual components can be composite with linguistic analyses and multimedia resources (images, audio, etc.) in a flexible way, at different levels of granularity (single words, sentences, paragraphs, documents, etc.). In order to achieve this goal, compound patterns have been used: (1) Composite (2) Typed Relationship (3) Factory Method.

The whole system has been developed according to the Java Server Faces 2 specification (JSR-314) and according to the general JSR-316 specification (Java Enterprise Edition 6). Persistence is obtained using a native XML-DB, eXist, which stores and retrieves documents encoded in TEI or other XML compliant documents.

The system, currently in alpha release, is producing promising results and several deployments in real projects, such as the Res Gestae Divi Augusti Web App and the Saussure Web App, which will be illustrated below.

III. RESULTS

A. Res Gestae Divi Augusti Web Application

The Res Gestae Divi Augusti Web Application handles the Mommsen's edition of the well known bilingual epigraph (Fig. 2). The aligned texts are shown in parallel and the granularity of the information is flexible: the units can be paragraphs for a coarse alignment, words for the morphological analysis and characters for the annotation of the status on the stone, such as readable or unreadable.

The comment component allows scholars to record exegetical annotations, commenting fragments on a selected chunk of text, which can be labeled. Index and Search component is suitable for advanced text retrieval based on metadata produced both by automatic processes and by scholar studies. For example in this project it is possible to perform queries for chunks in both languages (Greek and Latin) and word status on the available support (e.g., attested or conjectured).

B. A digital edition of F. de Saussure's Manuscripts

A prototype of the digital edition of Saussure's texts, based on a representative selection of his manuscript images and transcriptions, has been the focus of a Research Programs of Relevant National Interest (PRIN2008). This project has been a test for evaluating the platform with text and image resources. The text has been extracted from semi-structured electronic documents and transformed in a TEI-compliant format. Fig. 3 shows the component that manages parallel resources, in order to browse and search both texts and images. The links are referred to annotations that author of the manuscripts has done and to which editor refers in the critical apparatus. As shown in Fig. 1, the platform is based

on component aggregation (e.g., comments), or on the extension of components (e.g., indexes). Many indexes (one per language used in Saussure’s quotations) are required in this project. Consequently, the component is able to handle indexes and concordances for each language.

The comment component is useful to enhance collaborative annotations and arrange canonical linkage between selected text, such as named entity and authoritative resources on the web infrastructure.

IV. CONCLUSION

The new paradigm of computational philology in the current web generation is the collaborative philology, based on crowdsourcing both for software development and for the acquisition and generation of data. Even if many libraries of components have been implemented for computational linguistics, which can be reused in the domain of philology, computational philologists must afford specific issues due to multiple readings of the same text and multiple interpretations of the same reading. Furthermore, collaborative philology is open to other disciplines, such as palaeography, codicology, musicology, which heavily involve multimedia. For this reason we have illustrated how we are developing a library of java components that constitutes the core of a web platform, focused on the domain of collaborative philology to deal with texts and related multimedia sources. As seen above, the modularity of the system promotes the reuse of philological components in many applications. This aim is achieved both by aggregation and by extension.

Future work gives priority to the multimedia aspects of the philological studies. The same framework used to create a new textual edition from many corrupted witnesses can be applied to restore a sound track from many low quality, noisy or incomplete recordings of the same concert. APIs developed by third parties devoted to specific tasks, such as sound track alignment, will be studied and interfaced to our platform.

In conclusion, collaborative philology is experiencing a double transition: on one hand it is moving from single, desktop applications towards web applications, web services and distributed programming, on the other hand it is widening its attention from text to multimedia.

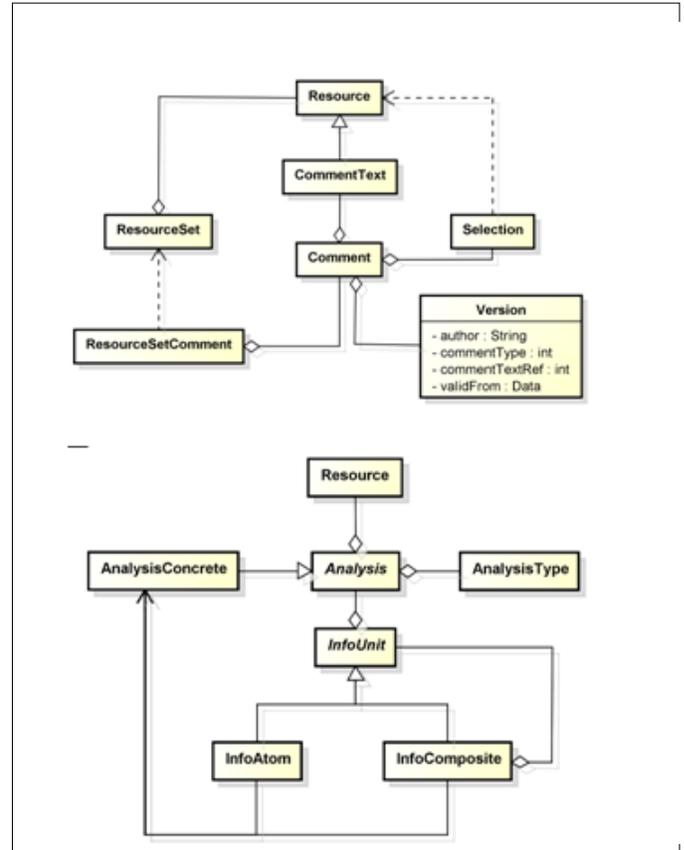


Figure 1. Class Model Diagram for Comment and Analysis components

Res Gestae Web Application v.0.3.21

Home View parallel pericopes Search Manage pericopes Manage witnesses Order by Greek Order by Latin Comment Linguistic Analysis

(1 of 10) 1 2 3 4 5 6 7 8 9 10

Latin	Latin Pericope	Greek Pericope	Greek
TIT	Rerum gestarum divi Augusti, quibus orbem terra[rum] imperio populi Rom. subiecit, § et impensarum, quas in rem publicam populumque Ro[ma]num fecit, incisarum in duabus aeneis pilis, quae su[n]t Romae positae, exemplar sub[sc]riptum.	Μεθρημνευμένοι υπεγράφησαν πρόξεις τε και δωρεαι Σεβαστου θεου, ως απελπεν επί Ρώμης ενκεχαραγμένες χαλκαϊς στήλαις δυοι.	TIT
1.1	Annōs undēvīginti natus exercitum privato consilio et privatā impensā comparāvī, [§] per quem rem publicam [do]minatione factionis oppressam in libertatē vindica[vi].	Ετών δεκαε[ν]νέα άν τὸ στράτευμα ἐμῆ γνώμη και ἐμοῦ ἀν[α]λόγῳ αἰσιν ἠτοί[μωσα], δι' οὗ τὰ κοινὰ πράγματα [ἐκ τῆς] τῶν συνο[μωσα]μένων δουλη[γ]ας [ἤλευ]θέρωσα.	1.1
1.2	Ob quae sen[atus] decrevit honor[um] in ordinem suum m[e] adlegit C. Pansa A. Hirri[us] consulib[us], c[on]sularem locum s[imul] dans sententiae ferendae, et im[perium] mihi dedit. [§]	Ἐπ' οὗτ' ἡ σύνκλητος ἐπανέστασά [με ψηφίσασα] προκατέλεξε τῆ βουλή Γάϊο Πά[ν]σο [Αἰλίω Ἡρτίῳ ἀ]νητό[τε]ς, ἐν τῇ τῶν ἀπα[ρ]κτώ[ν] [ἄρα τ]ῶ σ[υ]μβου[λί]ῳ δόσσα, ράβδου[ς] τ' ἐμοί ἔδωκεν.	1.2

Latin Selected Text: Res publica n[e] quid detrimenti caperet, me] pro praetore simul cum consulibus pro[videre] iussit.

Greek Selected Text: [Περ] τὰ δημόσια πράγματα μή τι βλαβῆ, ἐμοί με- [τὰ τῶν ὑπό]των προνοεῖν ἐπέτρεψεν ἀντί στρατηγο[ύ]. [.....

Latin Text Analysis			Greek Text Analysis		
Word Form	Word Lemma	Status	Word Form	Word Lemma	Status
RES	REOR RES	att*	ΠΕΡΙ	ΠΕΡΙ	partatt*
PUBLICA	PUBLICA PUBLICO PUBLICUM PUBLICUS	att*	ΤΑ	Ο*	att*
NE	NE NEO	partatt*	ΔΗΜΟΣΙΑ	ΔΗΜΟΣΙΟΣ	att*
QUID	QUIS	notatt*	ΠΡΑΓΜΑΤΑ	ΠΡΑΓΜΑ	att*
			ΜΗ	ΜΗ*	att*

Comments: [AN]: la factio si ri...

Annōs undēvīginti natus exercitum privato consilio et privatā impensā comparāvī, [§] per quem rem publicam [do]minatione factionis oppressam in libertatē vindica[vi].

Ετών δεκαε[ν]νέα άν τὸ στράτευμα ἐμῆ γνώμη και ἐμοῦ ἀν[α]λόγῳ αἰσιν ἠτοί[μωσα], δι' οὗ τὰ κοινὰ πράγματα [ἐκ τῆς] τῶν συνο[μωσα]μένων δουλη[γ]ας [ἤλευ]θέρωσα.

[do]minatione factionis oppressamin libertatē vindica[vi].

τῶν συνο[μωσα]μένων δουλη[γ]ας [ἤλευ]θέρωσα.

latin selection greek selection

la factio si riferisce a Marco Antonio tuttavia come nota Lucio Canfora il Greco segnala che la schiavitù era imposta dai congiurati. Il nome di Marco Antonio è volontariamente omissso.

- literal translation
- free rendering
- amplification
- misunderstanding
- interpolation
- glossary
- additional note

new delete literal translation submit clear

Word Index: ΤΡΙΑΚΟΝΤΑ - 2, ΤΡΙΑΚΟΣΤΟΣ - 1, ΤΡΙΗΡΗΣ - 1, ΤΡΙΣ - 6, ΤΡΙΣΚΑΔΕΚΑΤΟΣ - 3, ΤΡΙΣΜΥΡΙΟΙ - 1, ΤΡΙΣΧΕΛΙΑΙΟΙ - 1, ΤΡΙΣΧΙΛΙΑΙΟΙ - 2, ΤΡΙΤΟΣ - 3, ΤΡΟΠΑΙΟΦΟΡΟΥ - 2

SEARCH GREEK: Word A, Word B, Word C. lemma, form, form. ΤΡΟΠΟΣ, Search for..., Search for..., Every Status, Every Status, Every Status. OR. Save Parameters, Clear Parameters.

results: 6.1 Consulibus M. Vinucio et Q. Lucretio et postea P.] et Cn. Lentulis et tertium Paulo Fabio Maximo et Q. Tuberone senatu populoque Romano consentibus

6.1 Ὑπότος Μάρκου Οἰνουκίου και Κοίντου Λ[ουκρη]τίου και μετὰ τ[ού]το Πουλλῶ και Πάυλο Αδντῆλιος και τρίτον Παύλιον Φαβίῳ Μαξιμῶ και Κόντινῳ Τουβέρωνι § τῆς [τε σ]υνκλητου και τῶ δῆμου τοῦ Ρωμικῶν ὁμολογ[ού]ντων, ἵνα ἐπιμε[ληθῆς] τῶν τε νόμων και τῶν [ἐ]πιμε[ληθῆς] [ἐ]ξουσιῶν μὴ τῶς χαριστονεῖσθαι §, ἀρχὴν αὐδε- μ[ω]ν παρὰ τὰ νότρησι] ἐξέθη φερόμενη ἀνεδε- ἔμην: §

Εἰ δὴ ταῦτα ὀρθῶς λέγεται, ἴσονται ἂν ἴδῃ αἱ ἀπορίαι

Figure 2. Res Gestae Divi Augusti deployment

“REFERENCES”

- [1] <http://www.openarchives.org/pmh> [retrieved February, 2013]
- [2] <http://www.tei-c.org/index.xml> [retrieved February, 2013]
- [3] <http://www.europeana.eu/portal/> [retrieved February, 2013]
- [4] <http://pro.europeana.eu/edm-documentation> [retrieved February, 2013]
- [5] <http://www.cidoc-crm.org/> [retrieved February, 2013]
- [6] <http://www.ifla.org/node/928> - http://www.cidoc-crm.org/frbr_intro.html [retrieved February, 2013]
- [7] <http://www.projectbamboo.org/> [retrieved February, 2013]
- [8] <http://www.interedition.eu/> [retrieved February, 2013]
- [9] <http://www.perseus.tufts.edu/hopper/> [retrieved February, 2013]
- [10] <http://idp.atlantides.org/trac/idp/wiki/SoSOL/Overview> [retrieved February, 2013]
- [11] <http://alpheios.net/> [retrieved February, 2013]
- [12] <http://www.homermultitext.org/hmt-doc/cite/index.html> [retrieved February, 2013]
- [13] A. Berra, “Exploitation de la matière épigraphique dans un espace numérique, Edition savante et humanités numériques”, <http://philologia.hypotheses.org/648> [retrieved February, 2013].
- [14] A. Bozzi, M. M. Morales, and M. Rufino, “Imago et umbra: Programma di digitalizzazione per l’Archivio storico della Pontificia Università Gregoriana: criteri, metodi e strumenti,” in *Digitalia*, Anno V, Numero 2, Roma, 2010, pp 79-99.
- [15] A. Bozzi, V. Sandrucci, “Uno strumento al servizio dell’archiviazione, lo studio, l’edizione e l’interrogazione di documenti digitali,” in Carmen Alén Garabato, Mercedes Brea, Xosé Afonso Álvarez (a cura di), *Quelle linguistique romane au XXIe siècle?*, Paris: L’Harmattan, Langue et parole, 2010, pp. 27-40.
- [16] A. Bozzi, “Edizione elettronica e filologia computazionale”, in A. Stussi (a cura di), “Fondamenti di critica testuale”, Il Mulino Manuali, Bologna, 2006, pp. 207-232.
- [17] A. Bozzi, “Towards a philological workstation,” in *Revue informatique et statistique dans les Sciences humaines*, XXIX, 1993, pp. 33-49.
- [18] S. Burbeck, "Applications Programming in Smalltalk-80TM: How to use Model-View-Controller (MVC)," 1992 <http://st-www.cs.illinois.edu/users/smarch/st-docs/mvc.html>. [retrieved February, 2013].
- [19] G. Crane, B.Seales, and M. Terras, “Cyberinfrastructure for Classical Philology,” *Digital Humanities Quarterly*, 3 (1), 2009 URL: <http://www.digitalhumanities.org/dhq/vol/3/1/000023/000023.html> [retrieved February, 2013].
- [20] M. Fowler, “Analysis Patterns: Reusable Object Models”. Menlo Park, Calif. ; Harlow : Addison Wesley. 1996.
- [21] E. Gamma, R. Helm, R.Johnson, and J. Vlissides, “Design Patterns: Elements of Reusable Object-Oriented Software”. Reading, Mass: Addison-Wesley, 1995.
- [22] G.Hohpe, B. Woolf, “Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions”. Boston: Addison-Wesley, 2004.
- [23] M. Lamé, V. Valchera, and F. Boschetti, “Epigrafia digitale. Paradigmi di rappresentazione per il trattamento digitale delle epigrafi,” *Epigraphica*, LXXIV vol. 1-2, 2012, pp. 331-338
- [24] M. McCandless, E. Hatcher, and O. Gospodnetić, “Lucene in action”, Manning, 2010.
- [25] G. Stewart, G. Crane, and A. Babeu: A New Generation of Textual Corpora. *JCDL 2007*, pp. 356–365.

Efficient, Compact, and Dominant Color Correlogram Descriptors for Content-based Image Retrieval

Ahmed Talib, Massudi Mahmuddin, Husniza Husni
 Computer Science Dept., School of Computing,
 University Utara Malaysia,
 06010 Sintok, Kedah, Malaysia
 s91707@student.uum.edu.my, {ady, husniza}@uum.edu.my

Loay E. George
 Computer Science Dept., College of Science,
 Baghdad University,
 10071 Al-Jadriya, Baghdad, Iraq
 loayedwar57@yahoo.com

Abstract— Color is one of the most important and widely used cues in content analysis and retrieval. However, most promising color descriptors consume massive amounts of computation and storage, which is a serious drawback. One of these promising color techniques in image retrieval is the color correlogram, but the technique also suffers from the aforementioned drawbacks. In this paper, we present two compact representations of the color correlogram. The first representation is the compact-generalized correlogram, which compresses colors and generalizes the distances of the original correlogram descriptor. The second representation is the dominant color-based correlogram, which is also a compact and conceptual correlogram descriptor. This representation computes the spatial correlations of the dominant colors of a few images instead of a large number of quantized colors used by the original descriptor. The two representations are integrated. The experimental results prove the high effectiveness and feasibility of the proposed descriptors through two large image databases (i.e., Corel-10K and Cartoon-11K) using ARR, ANMRR, P(10), and MAP metrics.

Keywords—Color Correlogram; Large database; Dominant Color; Compact descriptor; Content-based Image Retrieval

I. INTRODUCTION

Color descriptors play an important role in reducing the gap between low-level features, such as color, texture, and shape, and high-level semantic concepts, such as emotions, events, or scenes [1][2]. Color is considered a powerful cue for content-based image retrieval (CBIR) [3][4] and is also an effective feature in image analysis because color is robust with noise, image orientation, and resolution [2][4–6]. Therefore, various color descriptors have been proposed in previous studies [7–9]. The color histogram [7] and its enhancements [10–12] are good attempts at creating color descriptors because of their ability to solve translation and rotation invariant problems. However, the color histogram and its enhancements lack the spatial correlation of colors that allows different images to be considered similar.

One of the most promising approaches in solving the problem is color correlogram [13][14], which preserves the spatial correlations of color information for accurate image retrieval. The color correlogram approach demonstrates high effectiveness compared with the color histogram and an earlier spatial-color approach called the color coherence vector (CCV) [9]. The color correlogram is a table indexed by color pairs (C_i, C_j) , where the k^{th} entry specifies the probability of finding a color C_i at a distance k from a color

C_j in the image; i, j are the indexes of colors within a range of m quantized colors; and k is a distance within the range of a maximum distance d .

The problem of the correlogram lies in the expensive cost of memory space and computation time, with the correlogram requiring $O(m^2d)$ complexity. This cost is an infeasibility problem for use in a huge database, especially regarding memory space. Several gigabytes are required for a large database, which may not be available in the main memory of a computer. Therefore, the *Autocorrelogram* [13] is proposed to reduce the time and space complexity into $O(md)$ by finding the spatial correlation of each color with only itself. The accuracy of the *Autocorrelogram* is certainly lower than the original correlogram because the correlations of a particular color with other colors are ignored, and the only correlation with the same color is kept.

In this paper, a compact and generalized representation of the color correlogram is proposed, which reduces the complexity of the correlogram from $O(m^2d)$ to $O(m^2/2 + m/2)$. The proposed representation is slightly more complex than the *Autocorrelogram* (or less complex than the *Autocorrelogram* in some cases when d is large). The proposed method outperforms the *Autocorrelogram* and achieves the same (or slightly lower) accuracy than the original correlogram. The satisfactory performance is caused by the preservation of the spatial correlations among all the colors in the image, which reduces the memory space of the colors (m^2) to approximately half ($m^2/2 + m/2$). The proposed representation also generalizes all d distances into one distance value by taking the average of all distances. In this case, keeping many distances that refer to separate spatial correlations becomes expensive. Instead, averaging the distance significantly reduces the complexity of the descriptor with very little degradation in accuracy.

Color descriptors, including the color correlogram, are weak in image recognition or discrimination because the naive rules that the color descriptors are based on do not simulate the human visual system [3][4]. Therefore, more improvements can be done in this field. Much research has been conducted on human color perception (e.g., [15][16]), which show that humans use only a few of the prominent or dominant colors of the image to judge similarity. Two rules on the model human visual and color perception exist. The first rule states that two images are considered similar if the images have the same dominant colors (DCs). The second rule indicates that the two images are perceived to be similar if the images have the same distribution of DCs irrespective

of content [4]. Humans consider the DCs and the spatial distributions of images to judge color similarity. Therefore, dominant color descriptors have been introduced in many studies (e.g., [17–20]) instead of descriptors that use a large number of colors, such as the color histogram and color moments [7][10][12].

The color correlogram has perceptual and infeasibility problems in large databases. Thus, a perceptual correlogram was introduced [4], which applies *ColGrm* concepts on a few DCs instead of a large number of colors. However, the perceptual correlogram has some deficiencies in simulating the original correlogram through the imperfect similarity measure, which will be explained in Section IV. In the present paper, an adaptation of the perceptual correlogram [4] is proposed. The adapted descriptor is called a DC-based color correlogram (DCBC), which adapts the dissimilarity measure of the perceptual correlogram by correctly simulating the original correlogram on a few dominant colors.

The rest of this paper is organized as follows: Section II presents the related works on the color-based CBIR. Section III introduces the compact-generalized correlogram (CGC), and Section IV presents the adapted perceptual DC-based correlogram. The experimental results of the proposed and adapted descriptors and their integration are compared with some candidate descriptors in Section V. Section VI concludes the paper.

II. RELATED WORK

Many studies have been conducted on content-based image retrieval (CBIR), such as Visual-SEEk [21], QBIC [22], Photobook [23], and Image-Rover [24]. In these studies, several visual (low-level) features, such as color, texture, and shape, were used. Color is one of the most commonly used features in CBIR [3][4]. Therefore, this study focuses solely on color to retrieve images in the CBIR domain.

Previous studies vary in their usage of color descriptors [3]. Some use global color descriptors (GCDs), and others used spatial color descriptors (SCDs). The former is used to measure the similarity between two images by taking the colors and their percentages in the images into account, such as the color histogram [7][10] and the dominant colors [17–19][28]. The latter measures the similarity between the two images by considering the existing colors and their distributions or arrangements in the image, such as in the CCV [9] and color correlogram [13][14].

The color histogram in GCDs proposed in [7] has been extensively used as a GCD to solve translation and rotation invariant problems. This color histogram is characterized by easy implementation and accuracy particularly in small databases. Many enhancements in histogram-based approaches have been achieved, as reported in [10][11][25]. The original representation of the RGB color is 24-bits 16-million colors that are assigned to each pixel in the image, leading to an infeasibility problem for both time and memory space. Therefore, static quantization [39] is used to reduce color space to make storage and time more reasonable. However, histogram-based approaches have several

drawbacks. The first one is the dependence on a static quantization method, which suffers from low discrimination power. The lack of discrimination is caused by a large number of similar colors set to different bins, making the similarity measure (i.e., L1, L2 or histogram intersection) between the two histograms inefficient. The second drawback is the mismatch of the methods with human color perception [19][26][28]. Humans cannot perceive more than eight colors [27] or can only perceive a few prominent colors in an image [4][15][16]. Therefore, extracting DCs from the image becomes the best solution because DCs require less time and storage consumption.

Although DCs are not that effective in color-based image retrieval, it is still considered a GCD. The basic problem of these descriptors is the lack of spatial correlations of colors within the image. This absence leads to considering different images in terms of color distribution as similar because of the same color percentages. The complement part (spatial relationship of colors) of the similarity of images is “where the colors are located” [3][4]. GCDs work without the complement part; thus, the results are not satisfactorily presented in the CBIR field. Many methods have been proposed to include the complement part, such as the CCV [9]. A pixel is considered coherent if its color is similar to the color of the region to which it belongs; otherwise, the pixel is considered incoherent. Many approaches have also been proposed to prove the high effectiveness of the spatial relationship among image colors such as [29], which used the concept of color boundaries, and [5][8], which used the color adjacency concept. These approaches lead to a simple conclusion that the relative distance (inter-distance) of the colors of an image can capture the true or real composition of the colors in the image.

These SCDs have two important properties: translation and rotation invariant. One of the most active approaches among all the SCDs is the color correlogram (*ColGrm*) [13][14]. The *ColGrm* is a table indexed by color pairs (C_i, C_j), where the k^{th} entry specifies the probability of finding a color C_i at a distance k from a color C_j in the image; i, j are indexes of the colors within the range of m quantized colors; and k is the distance within the maximum distance d .

The *ColGrm* complexity is $O(m^2d)$, which consumes high CPU time and memory space, especially in a large database. For example, the image of width ($W=500$) and height ($H=400$) is assumed. In such dimensions, a suitable value for d would be 40–200, corresponding to the following formula: $d \approx 10\%$ to 50% of the smaller dimensions in the image [2][3]. Any value of d less than this range will not be suitable in capturing the true spatial color distributions of the image because only colors within a small range will be described. The complexity of correlogram algorithm remains too high and requires several processing hours per image on a computer even with a lower bound selection of the range of distance d ($d=40$). Moreover, an infeasible memory space for the feature vector will be required even for a small image database. Several possible solutions can be applied to solve this infeasibility problem. The first solution is the reduction of the range of distance d (e.g., let $d \approx 10$), which will reduce the complexity by only fourfold (an insignificant reduction)

and will be unable to precisely identify the true spatial color distribution. Another solution is the reduction of the color space using the quantization algorithm. The typical quantization for RGB color space is 8 partitions for each band ($8 \times 8 \times 8$), which will equal to 512 colors and will speed up the ColGrm by about 30-fold. However, immense memory space (about 80 megabytes (MB) per image) will still be required, making the ColGrm applicable for only small databases. Therefore, a simplified version of the ColGrm called *AutoCorrelogram* is introduced [13]. The *Autocorrelogram* only characterizes the spatial color distribution of the same colors, i.e., each color with itself without identifying correlations with other colors. The latter case may cause the degradation of the color descriptor, which actually occurred in many studies [1–4][13][14] that reported the ColGrm to be better in retrieval accuracy than the *Autocorrelogram*.

Ma and Zhang [30] and recently [31][32] show (using extensive experiments) that the ColGrm and Autocorrelogram can achieve better performance than those of other global and spatial color descriptors, such as the color histogram, color moments, and CCV, despite the limitations of correlograms. Some extensions were made to both, such as the Markov Stationary Features [33], which is an extension of the Autocorrelogram, Wavelet Correlogram [34], and Gabor wavelet correlogram [35]. All these approaches perform only slightly better than the original ColGrm descriptor, which has more time complexity [4]. A recent method reduces the time complexity of ColGrm by the approximation of a descriptor [36], depending on the randomization of selecting the neighbors of the pixels. The method certainly decreases the accuracy compared with the original ColGrm but decreases time complexity to half. The drawbacks of this method are that the complexity of the memory space remains $O(m^2d)$ and that the accuracy of such an algorithm is not fixed because of the dependence on the randomization of selecting the candidate pixels to build the ColGrm feature vector. Therefore, the proposed method depends on the original ColGrm in adaptation and comparison. A compact representation of the ColGrm is proposed in this paper to solve the problem of infeasibility. The details of the proposed method are presented in Section III.

DC-based methods can be used to solve the perceptual and infeasibility problems of the color histogram. However, the DC-based methods remain as GCDs and lack spatial color correlations. Therefore, the DC concept can be integrated with the ColGrm to obtain better performance than when each is applied separately. A satisfactory attempt was recently conducted to integrate the two in [4], which used a penalty trio model to find the dissimilarity between the two images by joining the global (from DCs) and the spatial (from ColGrm) information. Kiranyaz et al. [4] changed the dissimilarity equation of ColGrm, which is claimed to be inefficient, but such change would lead to serious performance degradation. This problem will be discussed in more detail in Section IV. A duo-model instead of a trio-model is proposed to solve the problem of the later model.

III. THE PROPOSED CGC DESCRIPTOR

The color correlogram offers the best performance among the GCDs and SCDs, as mentioned in Section II. However, the massive consumption of time and memory space remains its major drawback. Some reduction can be achieved for its time and feature vector space through a critical analysis of the process of the correlogram.

ColGrm $\gamma_{ci,cj}^{(k)}$ is a table of probabilities for finding the spatial correlation of a certain color with the other colors within an image from a specific distance. The table is indexed by the triple (Ci, Cj, k) , where Ci and Cj represent the colors their neighboring probabilities need to know in a distance k . The indexes' values i, j are within the m quantized colors, and the k value is within the maximum distance d . The ColGrm maintains spatial correlation among the colors in the image. The ColGrm table is depicted in Figure 1.

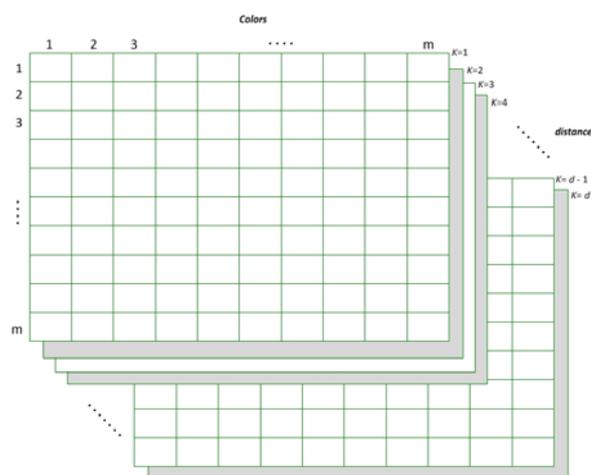


Figure 1. Original color correlogram feature vector representation with a complexity of $O(m^2d)$.

Figure 1 shows that the massive storage space of this representation lies in the colors and distances. Therefore, our proposed method is focused on the reduction of these two factors without a significant degradation of the performance of the original ColGrm.

A. Color Reduction

In the first factor (colors), the square matrix of colors in Figure 1 contains the probabilities of finding color i at the distance k from color j . A repetition of information is noticeable through a proper logical analysis of this color representation. The probability of finding color i with a specific distance from color j is located in the two positions in the matrix: locations (i, j) and (j, i) . Intuitively, the existence of a white color beside a black one, for example, has the same meaning as a black color being beside a white one. Black is on the right of white; thus, we can find white on the left of black. The co-occurrence matrix of the colors in the original representation increases in the two locations co-occurrence (black, white) and co-occurrence (white, black), as shown in Figure 2.

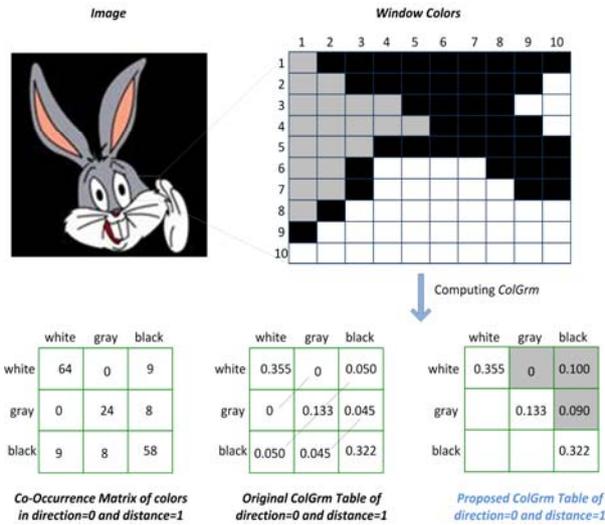


Figure 2. An example of computing the ColGrm table in the image window in a simple setting (direction = 0 and distance = 1), which shows the similarity of the lower rectangular matrix with the upper one.

Figure 2 is an example of the image window that has three colors, namely, white (w), gray (g), and black (b). A simple setting (direction = 0 and distance = 1) is considered as a simple explanation to the reader. The direction equal to zero and the distance equal to one indicate that only the horizontal (left and right) neighbors of the pixels are considered during the extraction of the ColGrm table. The elements of the co-occurrence matrix are shown in Figure 2. For example, the co-occurrence (white, black) = 9 means that 9 horizontal black neighbors to the white color exist. The ColGrm table holds the probability instead of the number of the occurrence of colors; thus, the co-occurrence matrix is simply divided by the number of all neighbors in the 10 × 10 window, which is 180. The ColGrm table shows that ColGrm (w, g) = ColGrm (g, w) and all the other elements in the lower triangular matrix are similar to the elements of the upper matrix. Therefore, repeating these elements is useless because one element is sufficient for each of the two colors instead of two elements. Keeping the upper triangular matrix with the main diagonal is enough to maintain the whole matrix. The upper triangular matrix in the new proposed representation is duplicated to substitute the absence of the lower matrix. The ColGrm complexity can be reduced to approximately half, i.e., $O(m^2/2+m/2)$ instead of $O(m^2)$, as shown in the shaded cells of Figure 3. Therefore, only the upper triangular matrix and the main diagonal must be computed and saved.

The dissimilarity measure equation remains the same as the original correlogram, as depicted in (1) [13][14].

$$ColGrmDisSimilarity(Q,I) = \sum_i \sum_j \sum_k^d \frac{|\gamma_{ci,cj}^{(k)}(Q) - \gamma_{ci,cj}^{(k)}(I)|}{1 + \gamma_{ci,cj}^{(k)}(Q) + \gamma_{ci,cj}^{(k)}(I)} \quad (1)$$

B. Distance Reduction

Distance is the second specified factor in reducing the complexity of the ColGrm feature vector. The number of distances required for the ColGrm to capture the true spatial correlations of the colors ranges from 10% to 50% of the smaller dimension in the image. This process consumes CPU time and memory space, as depicted in Figure 1.

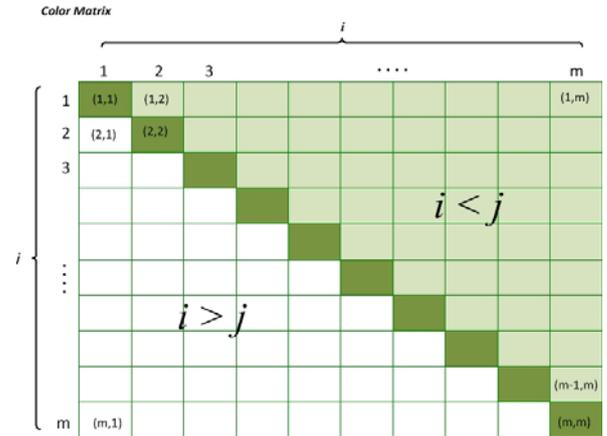


Figure 3. New ColGrm matrix representations. The shaded cells represent the actual required storage space for the colors (approximately half the size of the colors in the ColGrm).

The reduction of time and space is required to apply the ColGrm in a large database. The proposed solution to reduce these distances is generalization. The proposed generalization scheme that can be applied for distances is to average all distances. Distances are very important in measuring how many pixels exist between a certain color and other colors. For example, the image with three colors in Figure 2, where $d = 1$ and ColGrm (white, gray, d) = 0.15, indicates that the probability of finding a white color far from a gray color by one pixel is 0.15. When $d = 2$ and ColGrm (white, gray, d) = 0.149, the probability of finding a white color from a gray color by two pixels is 0.149. This pixel-based structure is unfortunately one of the main drawbacks of the ColGrm [3][4] because the color vicinity is characterized at a pixel level, which is unfeasible (in terms of time and space) in high resolution images and has no meaning with regard to the human visual system. The individual pixels cannot be perceived by the human eye. The average of all distances can be computed to eliminate this effect from the ColGrm and generalize the distance. The layers of the distance shown in Figure 1 can be abbreviated to one layer that contains the probabilities of generally finding the colors in the image I. In Figure 1, the size of the image is 10 × 10; thus, a distance = 5 is selected as 50% of the smaller dimension in the image. When distance = 5 and the generalization of ColGrm is applied, one layer is produced. For example, ColGrm (white, gray) = 0.145 means that the probability of finding the white color far from the gray color by 3 pixels (average of 5 distances is 3) is generally 0.145. This process ensures the generality of the descriptor and eliminates pixel-level dependency (especially in high-resolution images). The spatial correlations of the

colors in general for all the images in the database are also described. A general vision of the image contents (colors) is drawn instead of depending on the individual distances of the colors that lack feasibility and human perception. The complexity of the distance becomes 1 (instead of d in original *ColGrm*). The total complexity of the proposed *ColGrm* after color and distance reduction becomes $O(m^2/2+m/2)$ instead of $O(m^2d)$, as shown in Figure 3. The complexity of the proposed compact *ColGrm* is $O(m^2/2+m/2)$ during the image retrieval process and storage space but is $O(m^2d)$ during the feature extraction process because all distances must first be computed prior to the average. Feature extraction is an offline process, which means that extraction is performed once and the feature vectors can be saved in a database ready for the online retrieval process. Therefore, the speed of the interactive process with the user is not significantly affected.

IV. THE ADAPTED DC-BASED CORRELOGRAM DESCRIPTOR

DC-based approaches are introduced to solve the perceptual problem of the conventional color-based approaches by simulating human color perception. One of the most promising DC-based approaches is the method proposed by S. Kiranyaz et al. [4], which integrates DCs with *ColGrm* to solve the problems of both methods. These problems include the lack of the spatial colors information problem of the DC-based approaches and the infeasibility problem of the original *ColGrm* descriptor, especially in large databases. The method is called perceptual correlogram. The DCs are extracted from an image through a method similar to [18], which simulates human color perception. Then, these DCs are back-projected on the image to extract the color correlogram that depends on the DCs. This method proposes a trio-model to measure the dissimilarity of the two images, as depicted in (2) [4].

$$P_{trio}(Q, I) = P_{\phi}(Q, I) + (\alpha P_G(Q, I) + (1 - \alpha)P_{Corr}(Q, I)). \quad (2)$$

The trio-model has three measuring metrics: P_{ϕ} , P_G , and P_{Corr} . The first metric (P_{ϕ}) measures the mismatching colors and their percentages in the two compared images, as depicted in (3) [4]. W_i and C_i represent the percentages and the colors values in the mismatching color list (S^{ϕ}). The other two metrics (P_G and P_{Corr}) measure the difference between the matched colors of the two images. P_G measures the global difference between the two images, as expressed in (4) [4]. Nm represents the number of matching colors of the two images; T_s represents color similarity threshold, and β is the value between 0 and 1, which represents the adjustment between the two terms of (4). P_{Corr} measures the spatial (or *ColGrm*) difference between the two images, as shown in (5) [4], where MC represents a list of similar (matched) colors between the two images Q and I. $\gamma_{ci,cj}^{(k)}$ is the probability of finding DC C_i at a distance k from DC C_j .

$$P_{\phi}(Q, I) = \frac{\sum (W_i | C_i \in S^{\phi})}{2} \leq 1 \quad (3)$$

$$P_G(Q, I) = \beta \sum_{i=1}^{Nm} |W_i^Q - W_i^I| + (1 - \beta) \sqrt{\frac{\sum_{i=1}^{Nm} (C_i^Q - C_i^I)^2}{T_s N_m}} \leq 1 \quad (4)$$

$$P_{Corr}(Q, I) = \sum_{i,j \in MC} \sum_{k=1}^d \begin{cases} 0 & \text{if } \gamma_{ci,cj}^{(k)}(Q) = \gamma_{ci,cj}^{(k)}(I) = 0 \\ \frac{|\gamma_{ci,cj}^{(k)}(Q) - \gamma_{ci,cj}^{(k)}(I)|}{\gamma_{ci,cj}^{(k)}(Q) + \gamma_{ci,cj}^{(k)}(I)} & \text{else} \end{cases} \quad (5)$$

In other words, P_{ϕ} and P_G measure the global differences, and P_{Corr} measures the spatial differences between the compared images. A proper critical analysis of the trio-model reveals serious drawbacks. The first drawback occurs in computing the *ColGrm* dissimilarity metric (P_{Corr}), and the second drawback lies in existence of the P_G and P_{ϕ} that compute general dissimilarity and represent a different perspective from *ColGrm* dissimilarity. The limitation of P_{Corr} in (5) is identified through a comparison with the dissimilarity measure of the original *ColGrm*, as shown in (2). The results of both dissimilarity measures are compared in Table I.

The similarity measure of the method proposed [4] has a serious problem, which is the lack of discrimination of the dissimilarity measure between large and small differences of the probability values because the dissimilarity values of large differences are equal to those of small differences (as depicted in the fourth column of Table I). This matter is contrary to human visual perception because the human eye cannot recognize small differences. The original *ColGrm* dissimilarity keeps these differences linear. If the difference is large, the dissimilar value is also large; and if it is small, the result is also small.

TABLE I. COMPARISON BETWEEN TWO DISSIMILARITY MEASURE METHODS FOR THE ORIGINAL COLGRM (THIRD COLUMN) AND FOR THE PERCEPTUAL COLGRM 4 (FOURTH COLUMN) TO SHOW THE DRAWBACKS OF THE LATTER.

x	y	$\frac{ x - y }{(1+x+y)}$	$\frac{ x - y }{(x + y)}$	Difference Amount
0	0	0	0	Zero (Equal)
0.5	0.5	0	0	Zero (Equal)
0.5	0	0.333	1	Large
0.005	0	0.005	1	Small
0.5	0.1	0.25	0.66	Large
0.005	0.001	0.004	0.66	Small
0.5	0.4	0.05	0.11	Large
0.005	0.004	0.001	0.11	Small

The dissimilar value of the perceptual descriptor is illogical because even a small difference obtains a large dissimilar value (reaching to 1), and an image may have many small colors. The other metrics (i.e., P_{ϕ} and P_G) in the dissimilarity measure also have values based on the percentages of colors (from 0 to 1), which conflicts with P_{Corr} that has a value fixed in both the large and small percentages

of color. Therefore, the dissimilarity measure of the original *ColGrm* is better than that of the perceptual *ColGrm* [4]. The original *ColGrm* has a fixed color space, whereas the perceptual *ColGrm* has a dynamic and variable number of colors.

In other words, computing P_\emptyset and P_G (global differences) with P_{Corr} (spatial difference) is unsuitable because of the difference in values and perspective. In the perceptual *ColGrm*, Kiranyaz et al. was forced to use the two together because the P_{Corr} metric computes the dissimilarity of the matched colors only, whereas the original *ColGrm* dissimilarity measure equation computes the dissimilar values for matched and mismatched colors together. Therefore, the concept of original *ColGrm* can be applied to the adapted DC-based *ColGrm*, which computes the matched and mismatched colors in the same metric. The probability values of the matched colors between the two images in the adapted method will be directly compared because the mismatched colors for each of the two images will be compared with zeros, as in (8).

The corresponding probability values of the mismatching colors in the adapted *ColGrm* can be considered as zeros similar to those in the original *ColGrm*. The original *ColGrm* is simulated and can be considered the second term aside from P_{Corr} . The proposed duo-model of the adapted DCBC is expressed as follows:

$$P_{duo}(Q, I) = P_{match}(Q, I) + P_{mismatch}(Q, I) \quad (6)$$

$$P_{match}(Q, I) = \begin{cases} \sum_{i,j \in MC} \sum_{k=1}^d \frac{|\gamma_{ci,cj}^{(k)}(Q) - \gamma_{ci,cj}^{(k)}(I)|}{1 + \gamma_{ci,cj}^{(k)}(Q) + \gamma_{ci,cj}^{(k)}(I)} * a_{i,j} & \text{if } i=j \\ \sum_{i,j \in MC} \sum_{k=1}^d \frac{|\gamma_{ci,cj}^{(k)}(Q) - \gamma_{ci,cj}^{(k)}(I)|}{1 + \gamma_{ci,cj}^{(k)}(Q) + \gamma_{ci,cj}^{(k)}(I)} & \text{if } i \neq j \end{cases} \quad (7)$$

$$P_{mismatch}(Q, I) = \sum_{i,j \in Q_MMC} \sum_{k=1}^d \frac{|\gamma_{ci,cj}^{(k)}(Q) - 0|}{1 + \gamma_{ci,cj}^{(k)}(Q) + 0} + \sum_{i,j \in I_MMC} \sum_{k=1}^d \frac{|\gamma_{ci,cj}^{(k)}(I) - 0|}{1 + \gamma_{ci,cj}^{(k)}(I) + 0} \quad (8)$$

MC represents the list of the matched colors between the two images Q and I. Q_MMC and I_MMC represent the lists of the mismatched colors of images Q and I, respectively. Moreover, a_i, j represents the similarity ratio between the colors C_i and C_j , which can be computed using the following equation [17]:

$$a_{i,j} = \frac{d_{i,j}}{d_{max}} \text{ where } d_{i,j} \leq Tc \quad (9)$$

where $d_{i,j}$ represents the L1 distance between C_i and C_j , and the abbreviation C represents the 3D color values in the CIE-LUV color space, which can be computed as follows [17]:

$$d_{i,j} = |C_i^L - C_j^L| + |C_i^U - C_j^U| + |C_i^V - C_j^V| \quad (10)$$

The color threshold Tc represents the maximum distance, in which the two colors are considered similar, and is set to 20, and $d_{max} = \alpha Tc$, $\alpha = 1$, or 1.2. In (7), $a_{i,j}$ is multiplied to the *ColGrm* dissimilarity values when $d \leq 5$. The reasons behind multiplying only the main diagonal of the *ColGrm*

array by the color similarity ratio ($a_{i,j}$) is that the main diagonal values often represent the percentages of the colors in the image (especially when d is small) because it contains the probability of finding each color with itself, except for the colors that are too scattered in the image, and is rarely used in images that are converted into images of 8 DCs as the maximum. The other values in the *ColGrm* matrix represent the probabilities of finding a certain color with other colors (spatial correlations). Therefore, multiplying the color similarity ratio with the percentages of the DCs simulates the DC-based approaches to alleviate the problem of non-identical matched colors.

In sum, the differences between the adapted DC-based *ColGrm* descriptor and the perceptual *ColGrm* descriptor lie in two positions. The first difference is that the perceptual descriptor depends on the dissimilarity measure from the different perspectives, with the three metrics measuring the dissimilarity between two images. P_\emptyset and P_G are used to measure the global differences of colors. These metrics are produced from the approach perspective of DC. P_{Corr} measures the spatial correlations of the matched colors only between the two images. This metric represents the *ColGrm* perspective. Combining different perspective metrics may lead to the inconsistency of these metrics, which may produce an inaccurate dissimilarity value. Nevertheless, the adapted DC-based *ColGrm* depends on the *ColGrm* perspective only, which can measure global and spatial color differences together efficiently, making accuracy better than that of the perceptual descriptor (as shown in the experimental results in Section V). The second difference is the dissimilarity measure of the perceptual *ColGrm* descriptor (P_{Corr}), which is different from the original metric. The new metric has a serious limitation of being unable to differentiate between large and small probabilities of the correlations of color in the image, as shown in Table I.

V. EXPERIMENTAL RESULTS

The evaluation of the proposed compact-generalized *ColGrm* descriptor, adapted DC-based *ColGrm* descriptor, and the integration of both is conducted on two datasets: 1) the well-known Corel-10K dataset that contains 80 classes and 10,800 images, with 100 images existing for each class in the dataset; and 2) the Cartoon-11K dataset that contains 11,120 images collected from the web, with 146 classes (cartoon characters) existing, with each one having at least 35 images. The two datasets are used to show the superiority of the proposed color-based descriptors in large databases.

The descriptors selected to be compared with the proposed CGC, adapted DCBC, and the integration of the two are the original *ColGrm* [13][14] (whenever applicable), *AutoCorrelogram* [13], MPEG-7 Dominant Color Descriptor (DCD) [17], and Perceptual *ColGrm* [4]. The rationale for this selection is the representation of the first two descriptors of the original *ColGrm* descriptor, which are considered the base of the proposed descriptors. The third descriptor (DCD) is the base of any DC-based approach, which is used in DCBC. The last descriptor represents the original descriptor, which has been adapted to produce the DCBC descriptor.

A. Performance Measure Metrics

A quantitative performance measure metrics is utilized to measure the accuracy of the proposed descriptors with the other ColGrm descriptors chosen for comparison. Two of the metrics are the average retrieval rate (ARR) [37] and the average normalized modified retrieval rank (ANMRR) [17][37]. These metrics are used by the MPEG-7 committee to evaluate its work and are considered two of the most widely used metrics. These metrics combine many conventional metrics, including hit-miss counters, precision-recall, and ranking information, and they represent all-in-one value. The third metric is the mean average precision (MAP), which is one of the most widely used metrics in CBIR and is a compromise between precision and recall in a single metric [32][38]. This metric has become one of the leading performance evaluation metrics in ad hoc retrieval systems [38]. The fourth metric is P(10), which is a precision value of the first 10 retrieved images by a specific query. This metric is the most widely used metric for web-based image retrieval [32], as the user tends to see the result of his query in the first page or prefer to reformulate the query instead of checking the second page. The best value for the metrics ARR, P(10), and MAP is close to 1, indicating that the relevant images are retrieved in good standing. The best value of ANMRR is close to 0. MAP differs from ANMRR in that MAP measures the retrieval accuracy to all relevant images in the database to a particular query, whereas ANMRR measures the retrieval accuracy within a specific window (W) size. The window size is normally equal to twofold of the ground truth size of a specific image query.

The complexity of the proposed descriptors, in terms of time and memory space, is also urgently computed as the fifth metric to prove their applicability in large databases. The applicability of the proposed descriptors in large databases is the main aim of this study. The accuracy metrics are used to prove that the compactness of the proposed descriptors does not significantly degrade performance.

B. Retrieval Performance

The retrieval performance of the competing descriptors in the specified datasets can be measured using the accuracy and complexity metrics. The complexity metrics represent the computing time and memory space needed for the comparison of the proposed descriptors with the competing descriptors. Time is divided into feature extraction time (offline) and image retrieval time (online). Memory space is referred to as the main memory or disk space required by the descriptors. The diversity of queries is also important in ensuring fair and honest results [38]; thus, the evaluation queries are selected from all classes of the databases.

1) Retrieval performance of the Corel-10K dataset

An experiment is conducted on the Corel-10K dataset [40] with 111 queries. The results of the four evaluation metrics and the complexity of the memory space are given in Tables II and III, respectively, to show the accuracy and efficiency of the proposed methods compared with other descriptors. A single value in the "MPEG7 DCD" column in Table II indicates that this descriptor does not have a

different setting of distances to compute unlike other descriptors. The percentages of colors are depended upon rather than the distances among colors, which are used in spatial ColGrm methods. The left part of Table II (i.e., the first three columns) shows that the best accuracy values are those of the original ColGrm, which are better than proposed CGC. However, the values are applicable only for minimum settings ($3 \times 3 \times 3$ colors of each band and distances equal to 5 and 10), as shown in Tables II and V. The slight degradation of the accuracy of the proposed descriptor is caused by the generalization of the distances that loses the values of the accurate distances. A comparison is then made by increasing the setting, such as $4 \times 4 \times 4$ colors and 5, 10, and 40 distances. Only the *Autocorrelogram* and the proposed CGC can be applied in this case. The proposed descriptor also outperforms the *Autocorrelogram* because of the preservation of the spatial correlation of each image color with other colors, whereas the *Autocorrelogram* has a spatial correlation of each color with itself and ignores the others. In the ColGrm descriptors, the accuracy is decreased when the number of distances is increased because the unsuitable distances will have an effect on the suitable distances, which is a certain distance indicating that the actual distance between the specific color and the other colors in the image exists. The memory space and image retrieval time remain $O(m^2/2+m/2)$, which are online processes (performed when comparing the query image ColGrm with all database images of the ColGrms), despite the increase in the distances of the proposed descriptor. This increase in distances only affects the feature extraction process, which is an offline process (performed once only when creating the database away from an interaction with users), and the extraction query image ColGrm, in which the complexity of its computation and memory space is $O(m^2d)$ and is equal to the original correlogram.

The middle part of Table II (i.e., second three columns) shows that the adapted DCBC outperforms the three original descriptors (i.e., DCD, ColGrm, and the perceptual ColGr). The adapted descriptor is more accurate than the original version [4] because the latter has many drawbacks, as mentioned in Section IV. The complexity of the perceptual and proposed DCBC descriptors is $O(8^2d)$ as the maximum, where 8 represents the maximum DCs that can be extracted from the image. The significant degradation accuracy of the perceptual descriptor when increasing the distance is also noticeable because the incompatibility between the spatial dissimilarity (P_{Corr}) and global dissimilarity (P_\emptyset and P_G) when increasing the distance leads to the significant change in the P_{Corr} . The global dissimilarity values (i.e., P_\emptyset and P_G) remain unchanged. The dissimilarity measure of the perceptual ColGrm descriptor has a serious limitation. The integration of the proposed methods is achieved by applying the compactness and generalization concepts of the CGC (first proposed descriptor) on the DC-based ColGrm (second adapted descriptor). The combination outperforms

all three original descriptors (i.e., MPEG-7 DCD, *ColGrm*, and Perceptual *ColGrm*), with a maximum complexity of $O(8^2/2+8/2) = O(36)$.

The single value in an entire row in Table III indicates that either the descriptor does not have different distances in its computations (e.g., MPEG7 DCD) or that the descriptor produces the same memory space for all distances (e.g., the proposed CGC and the integration of CGC and DCBC). Tables II and III show that the integration of CGC and DCBC is a promising approach to the minimal consumption not only of memory space but also of image retrieval time.

Increasing the setting of *ColGrm* to four colors for each band ($4 \times 4 \times 4 = 64$ colors) leads to the proposed CGC outperforming all the other competing DC-based descriptors. This result is caused by the variety of colors (64 in CGC) being higher than that of the DC-based *ColGrm* approaches (8 colors maximum).

Table V clearly shows that the original color correlogram is inapplicable in a setting with 64 ($4 \times 4 \times 4$) colors. The original color correlogram has serious limitations, such as high computational complexity and memory storage (Table IV). Only the *Autocorrelogram* and all the compact descriptors can be applied. The proposed descriptors and their integration also outperform the *Autocorrelogram* and the perceptual *ColGrm*. The perceptual *ColGrm* appears worse than the *Autocorrelogram* because of the aforementioned limitations shown in Tables II and V. The key contribution of this paper is solving the feasibility problems (in computations and memory space) of the original *ColGrm*. Increasing the setting to more than four colors in each band is not shown in this paper because the results are similar to those of the setting of four colors.

2) Retrieval performance of the Cartoon-11K dataset

The four evaluation metrics are computed for the 158 queries on the Cartoon-11K dataset (this database is collected from Google and will be published soon) in Tables VI and VII, respectively, to show the accuracy of the proposed methods compared with other descriptors.

Table VI shows that the adapted DCBC descriptor outperforms all competing descriptors, including the perceptual descriptor. The proposed descriptor CGC shows the same accuracy as the original *ColGrm* but with a significant reduction in complexity $O(m^2d)$ to $O(m^2/2+m/2)$.

Table VII, with a setting of four colors, shows that the proposed CGC outperforms the adapted DCBC because the abundance of the colors can be expressed on the image content more efficiently than DCs. The storage space required for the Cartoon database is approximately equal to that in the Corel database, as depicted in Tables III and IV. These tables show that the compactness of the proposed, adapted, and integrated descriptors increases the speed of the image retrieval process.

VI. CONCLUSION

In this paper, two compact correlogram descriptors are proposed for large databases. The first descriptor, CGC, solves the inapplicability problems of the color correlogram in large databases. CGC reduces the colors of *ColGrm* approximately by half and performs a generalization of all the distances into a single representative distance. This descriptor also has less degradation accuracy than the original *ColGrm*, but the latter cannot be applied in a large setting with increased colors, distances, or database sizes. CGC also outperforms the *Autocorrelogram*, which can be applied in large settings, because the proposed method keeps the correlations of each color in the image with other colors, whereas the *Autocorrelogram* keeps the correlations of each color only with itself and ignores the relations with other colors. The second descriptor is the DC-based *ColGrm* adapted from the perceptual *ColGrm* [4], which suffers from serious limitations in its dissimilarity measure. DCs offer both perceptual and compact descriptions of colors. Therefore, the combination of DCs with *ColGrm* surpasses all the competing descriptors in terms of accuracy, time, and storage space. Integrating the proposed descriptors also shows promising results in significantly reducing complexity.

TABLE II. ANMRR, ARR, P(10), AND MAP VALUES FOR ALL COMPETING DESCRIPTORS ON COREL-10K DATABASE WITH 111 QUERIES (WITH NO. OF COLORS EQUALS $3 \times 3 \times 3 = 27$ COLORS AND DISTANCE = 5, 10, AND 40). BEST ACCURACY VALUES ARE IN BOLD.

Descriptor Metric	Original Correlogram	Auto-Correlogram	Proposed CGC	MPEG-7 DCD	Perceptual CG	Adapted DCBC	Integration CGC+DCBC
ANMRR	0.646/651/NA	0.705/714/739	0.648/659/680	0.710	0.688/779/935	0.591/595/605	0.600/612/632
RR	0.287/280/NA	0.240/233/208	0.285/278/258	0.235	0.250/188/052	0.337/334/326	0.328/325/301
P(10)	0.57/.55/NA	0.43/.42/.41	0.57/.55/.51	0.40	0.50/.37/.21	0.62/.62/.60	0.60/.60/.57
MAP	0.294/285/NA	0.232/225/200	0.293/283/257	0.206	0.241/166/042	0.328/324/317	0.317/311/290
Average	0.376/366/NA	0.299/291/269	0.375/363/269	0.282	0.325/236/092	0.423/420/409	0.411/406/382

TABLE III. SIZE OF THE FEATURES' DATABASE FOR COREL-10K AND COLORS IS (3 × 3 × 3) 27 FOR ALL COMPETING DESCRIPTORS.

ColGrm Method	Distance=5	Distance=10	Distance=40
Original ColGrm	278.1 M	556.2 M	2.17 G
AutoCorrelogram	10.3 M	20.6 M	82.4 M
Proposed CGC		28.8 M	
MPEG7 DCD		0.85 M	
Conceptual ColGrm	25.2 M	49.7 M	196.1 M
Proposed DCBC	25.2 M	49.7 M	196.1 M
Integration of CGC+DCBC		3.6 M	

TABLE V. ANMRR, ARR, P(10), AND MAP VALUES FOR ALL COMPETING DESCRIPTORS ON COREL-10K DATABASE WITH 111 QUERIES (WITH NO. OF COLORS EQUALS 4 × 4 × 4 = 64 COLORS AND DISTANCE = 5,10, AND 40). BEST ACCURACY VALUES ARE IN BOLD.

Descriptor Metric	Original Correlogram	Auto-Correlogram	Proposed CGC	MPEG-7 DCD	Perceptual CG	Adapted DCBC	Integration CGC+DCBC
ANMRR	N/A	0.619/650/661	0.552/559/580	0.710	0.688/779/935	0.591/595/605	0.600/612/632
RR	N/A	0.317/305/298	0.377/367/358	0.235	0.250/188/052	0.337/334/326	0.328/325/301
P(10)	N/A	0.53/.52/.51	0.64/.62/.61	0.40	0.50/.37/.21	0.62/.62/.60	0.60/.60/.57
MAP	N/A	0.323/315/300	0.387/380/357	0.206	0.241/166/042	0.328/324/317	0.317/311/290
Average	N/A	0.387/372/361	0.463/453/444	0.282	0.325/236/092	0.423/420/409	0.411/406/382

TABLE VI. ANMRR, ARR, P(10), AND MAP VALUES FOR ALL COMPETING DESCRIPTORS ON CARTOON-11K DATABASE WITH 158 QUERIES (WITH NO. OF COLORS EQUALS 3 × 3 × 3 = 27 COLORS AND DISTANCE = 5,10, AND 40). BEST ACCURACY VALUES ARE IN BOLD.

Descriptor Metric	Original Correlogram	Auto-Correlogram	Proposed CGC	MPEG-7 DCD	Perceptual CG	Adapted DCBC	Integration CGC+DCBC
ANMRR	0.853/0.853/NA	0.880/890/902	0.853/854/855	0.945	0.927/944/969	0.838/838/839	0.841/844/851
ARR	0.118/117/NA	0.094/088/077	0.117/117/116	0.041	0.057/041/023	0.130/130/130	0.126/123/117
P(10)	0.35/.35/NA	0.29/.26/.24	0.35/.35/.35	0.08	0.20/.17/.10	0.39/.38/.39	0.37/.37/.36
MAP	0.098/097/NA	0.075/069/060	0.097/097/097	0.029	0.045/038/023	0.105/105/104	0.102/098/094
Average	0.178/.177/NA	0.144/131/118	0.177/177/176	0.051	0.093/076/044	0.194/194/194	0.189/186/180

TABLE VII. ANMRR, ARR, P(10), AND MAP VALUES FOR ALL COMPETING DESCRIPTORS ON CARTOON-11K DATABASE WITH 158 QUERIES (WITH NO. OF COLORS EQUALS 4 × 4 × 4 = 64 COLORS AND DISTANCE=5,10, AND 40). BEST ACCURACY VALUES ARE IN BOLD.

Descriptor Metric	Original Correlogram	Auto-Correlogram	Proposed CGC	MPEG-7 DCD	Perceptual CG	Adapted DCBC	Integration CGC+DCBC
ANMRR	N/A	0.867/870/892	0.830/833/835	0.945	0.927/944/969	0.838/838/839	0.841/844/851
ARR	N/A	0.107/100/089	0.136/135/133	0.041	0.057/041/023	0.130/130/130	0.126/123/117
P(10)	N/A	0.32/.28/.25	0.41/.40/.38	0.08	0.20/.17/.10	0.39/38/39	0.37/.37/.36
MAP	N/A	0.083/079/070	0.114/110/108	0.029	0.045/038/023	0.105/105/104	0.102/098/094
Average	N/A	0.160/147/129	0.208/203/196	0.051	0.093/076/044	0.194/194/194	0.189/186/180

REFERENCES

[1] Tungkaathan, S. Intarasema, and W. Premchaiswadi, "Spatial color indexing using ACC algorithm," 7th International Conference on ICT and Knowledge Engineering, 1-2 Dec. 2009, pp. 113 – 117.

[2] W. Premchaiswadi and A. Tungkaathan, "A Compact Auto Color Correlation using Binary Coding Stream for Image Retrieval," Proceedings of the 15th WSEAS international conference on Computers, 2011, pp. 430-436.

[3] S. Kiranyaz, M. Birinci, and M. Gabbouj, "Perceptual color descriptor based on spatial distribution: A top-down approach," Journal of Image and Vision Computing vol. 28 (8), 2010, pp. 1309–1326.

[4] S. Kiranyaz, M. Birinci, and M. Gabbouj, Perceptual Color Descriptors. Foveon, Inc. / Sigma Corp., San Jose, California, USA: Boca Raton, FL, CRC Press, 2012.

[5] Y. H. Lee, K. H. Lee, and H. Y. Ha, "Spatial Color Descriptor for Image Retrieval and Video Segmentation," IEEE Trans. Multimedia, vol. 5(3), 2003, pp. 358–367.

[6] R. Schettini, G. Ciocca, and S. Zuffi, "A Survey Of Methods For Colour Image indexing And Retrieval In Image Databases," Proc. Schettini A survey, 2001, pp. 312-322.

[7] M. Swain and D. Ballard, "Color Indexing," International Journal of Computer Vision, vol. 7(1), 1991, pp. 11–32.

[8] M. A. Stricker and M. Orengo, "Similarity of color images," Proc. SPIE, Storage Retrieval Still Image Video Databases IV, vol. 2420, 1996, pp. 381–392.

[9] G. Pass, R. Zabih, and J. Miler, "Comparing Images Using Color Coherence Vectors," Proc. ACM on Multimedia, 1997, pp. 65-73.

[10] J. Hafner, H. S. Sawhney, W. Esquitz, M. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 17, 1995, pp. 729-736.

[11] J. Cox, M. L. Miller, S. O. Omohundro, and O. N. Yianilos, "PicHunter: Bayesian relevance feedback for image retrieval," Proceedings of Int'l Conference on Pattern Recognition, 1996, pp. 361–369.

TABLE IV. SIZE OF THE FEATURES' DATABASE FOR COREL-10K AND COLORS IS (4 × 4 × 4) 64 FOR ALL COMPETING DESCRIPTORS.

ColGrm Method	Distance=5	Distance=10	Distance=40
Original ColGrm	1.52 G	3.1 G	12.2 G
AutoCorrelogram	24.4 M	48.8 M	195.3 M
Proposed CGC		158.7 M	
MPEG7 DCD		0.85 M	
Conceptual ColGrm	25.2 M	49.7 M	196.1 M
Proposed DCBC	25.2 M	49.7 M	196.1 M
Integration of CGC+DCBC		3.6 M	

- [12] Y. Gong, C. H. Chuan, and G. Xiaoyi, "Image indexing and retrieval using color histograms," *Multimedia Tools and Applications*, vol. 2(1996), 1996, pp. 133–156.
- [13] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Image indexing using color correlograms," *Proceedings of Computer Vision and Pattern Recognition*, 1997, pp. 762–768.
- [14] Kunttu, L. Lepistö, J. Rauhamaa, and A. Visa, "Image correlogram in image database indexing and retrieval," *Proceedings of 4th European Workshop on Image Analysis for Multimedia Interactive Services*, London, UK, 2003, pp. 88–91.
- [15] E. V. D. Broek, P. Kisters, and L. Vuurpijl, "The utilization of human color categorization for content-based image retrieval," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 5292, 2004, pp. 351–362.
- [16] Mojsilovic, J. Kovacevic, J. Hu, R. Safranek, and K. Ganapathy, "Matching and retrieval based on the vocabulary and grammar of color patterns," *IEEE Transactions on Image Processing*, vol. 9(1), 2000, pp. 38–54.
- [17] Yamada, M. Pickering, S. Jeannin, and L. C. Jens, "MPEG-7 Visual Part of Experimentation Model Version 9.0-Part 3 Dominant Color," *ISO/IEC JTC1/SC29/WG11/N3914*, Pisa, 2001.
- [18] Y. Deng, C. Kenney, M. S. Moore, and B. S. Manjunath, "Peer group filtering and perceptual color quantization," *IEEE Int. Symp. Circuits Syst. VLSI (ISCAS'99)*, vol. 4, 1999, pp. 21–24.
- [19] N. -C. Yang, W.-H. Chang, C.-M. Kuo, and T.-H. Li, "A fast MPEG-7 dominant color extraction with new similarity measure for image retrieval," *Journal of Visual Communication and Image Representation*, vol. 19, 2008, pp. 92–105.
- [20] K. M. Wong, L. M. Po, and K. W. Cheung, "A compact and efficient color descriptor for image retrieval," *Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2007)*, Beijing, China, 2007, pp. 611–614.
- [21] J. R. Smith and S. F. Chang, "VisualSEEK: a fully automated content-based image query system," *Proceedings of ACM Multimedia*, Boston, 1996, pp. 87–98.
- [22] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafine, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The QBIC system," *IEEE Computer*, 1995.
- [23] Pentland, R. W. Picard, and S. Sclaroff, "Photobook: tools for content based manipulation of image databases," *Proceedings of SPIE (Storage and Retrieval for Image and Video Databases II)*, vol. 2185, 1994, pp. 34–37.
- [24] S. Sclaroff, L. Taycher, and M. L. Cascia, "Image-Rover: a content-based image browser for the world wide web," *Proceedings of IEEE Workshop on Content-based Access Image and Video Libraries*, Puerto Rico, 1997, pp. 2–9.
- [25] V. Ogle and M. Stonebraker, "Chabot: retrieval from a relational database of images," *IEEE Computer*, vol. 28 (9), 1995, pp. 40–48.
- [26] L. M. Po and K. M. Wong, "A new palette histogram similarity measure for MPEG-7 dominant color descriptor," *IEEE International Conference Image Processing (ICIP'04)*, vol. 3, 2004, pp. 1533–1536.
- [27] Mojsilovic, J. Hu, and E. Soljanin, "Extraction of perceptually important colors and similarity measurement for image matching, retrieval, and analysis," *Transaction of Image Processing*, vol. 11 (11), 2002, pp. 1238–1248.
- [28] A. Talib, M. Mahmuddin, H. Husni, and L. E. George, "A weighted dominant color descriptor for content-based image retrieval," *Journal of Visual Communication and Image Representation*, vol. 24, 2013, pp. 345–360.
- [29] Nagasaka and Y. Tanaka, "Automatic video indexing and full video search for objects," *Visual Database Systems II, IFIP*, 1992, pp. 113–127.
- [30] W. Ma and H. Zhang, "Benchmarking of image features for content-based retrieval," vol. 1, 1998, pp. 253–256.
- [31] Y. Chun, N. Kim, and I. Jang, "Content-based image retrieval using multi-resolution color and texture features," *IEEE Transactions on Multimedia*, vol. 6(10), 2008, pp. 1073–1084.
- [32] O. A. B. Penatti, E. Valle, and R. d. S. Torres, "Comparative Study of Global Color and Texture Descriptors for Web Image Retrieval," *Journal of Visual Communication and Image Representation (Elsevier)*, 2012, pp. 359–380.
- [33] J. Li, W. Wu, T. Wang, and Y. Zhang, "One step beyond histograms: Image representation using markov stationary features," *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.
- [34] S.-J. Lee, Y.-H. Lee, H. Ahn, and S.-B. Rhee, "Color image descriptor using wavelet correlogram," *The 23rd International Technology Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, 2008, pp. 1613–1616.
- [35] H. Moghaddam and M. Saadatmand-Tarzan, "Gabor wavelet correlogram algorithm for image indexing and retrieval," *18th International Conference on Pattern Recognition*, vol. 2, 2006, pp. 925–928.
- [36] Taranto, N. D. Mauro, S. Ferilli, and F. Esposito, "Approximate image color correlograms," *Proceedings of the international conference on Multimedia*, Firenze, Italy, 2010, pp. 1127–1130.
- [37] S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 11, 2001, pp. 703–715.
- [38] M. Grubinger, "Analysis and Evaluation of Visual Information Systems Performance," in *PhD Thesis*, School of Computer Science and Mathematics, Faculty of Health, Engineering and Science, Victoria University, Melbourne, Australia, 2007.
- [39] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, 1998, pp. 2325–2383.
- [40] Corel 10000 image database [retrieved: September, 2012]. Available: <http://wang.ist.psu.edu/docs/related.shtml>.

Automatic Aerial Image Alignment for GeoMemories

Giuseppe Amato, Fabrizio Falchi, Fausto Rabitti
Information Science and Technology Institute (ISTI)
National Research Council (CNR)
Pisa, Italy
name.surname@isti.cnr.it

Andrea Marchetti, Maurizio Tesconi
Institute of Informatics and Telematics (IIT)
National Research Council (CNR)
Pisa, Italy
name.surname@iit.cnr.it

Abstract—In the last few years, aerial and satellite photographs have become more and more important for historical records. The availability of Geographical Information Systems and the increasing number of photos made per year allows very advanced fruition of large number of contents. In this paper we illustrate the GeoMemories approach and we focus on its automatic image alignment architecture. The approach leverages on a set of georeferenced images used as knowledge base. Local features are used in combination with compact codes and space transformation to achieve high level of efficiency.

Keywords—*image alignment; BoW; VLAD; local features; aerial photos;*

I. INTRODUCTION

GeoMemories¹ is a joint project between the AeroFototeca Nazionale (AFN) and the Institute of Informatics and Telematics (IIT) of the CNR. AFN has an extensive set of aerial photographs that constitute a historical record of the Italian territory from the end of the XIX century up to the end of the XX century. The project has created a web platform that allows browsing aerial photos traveling along the spatial-temporal dimensions. For efficient and effective managing such a huge archive, automatic alignment algorithms are crucial for placing historical aerial photos on top of nowadays maps.

During the last decade, local features have emerged, as an effective approach for image alignment, copy detection, landmark recognition, etc. A drawback of the use of local features is that a single image is represented by a possible large set of descriptors that should be individually matched and processed in order to compare the visual content of two images. In order to improve the efficiency of image matching on a large scale, a very popular method is the Bag of Features or Bag of Words (BoW) approach. This approach describes each image as a subset of predefined (visual) words. Thus, techniques successfully applied to text retrieval, can be easily applied also in the context of content based image retrieval.

Recently, another promising direction has emerged to simplify the representation of local features, based on the use of compact codes. The Vector of Locally Aggregated Descriptors (VLAD) [1] and its probabilistic counterpart (i.e., Fisher vectors) [2] compactly represent the local features of an image as a single fixed size vector. Image pairs can be then compared by using similarity (or distance) functions applied to

the compact vectors. In this way, all the techniques extensively studied for efficient similarity search can be applied. In fact, an image query corresponds to a single fixed size vector [3]. For instance, in [4] it was shown that Euclidean Locality-Sensitive Hashing (LSH) [5] techniques can be efficiently and effectively applied with VLAD.

In this paper we propose a comprehensive approach for efficient alignment of aerial images available in the GeoMemories project. The approach consists of 3 stages with increasingly cost of the analysis but with decreasing number of candidate images. While the whole dataset is searched for finding similar images to the one to be georeferenced, in the subsequent steps more robust approach are considered only between the candidate set selected at the step before. In this way we are able to merge the high efficiency of the recognition algorithms developed in the Multimedia Information Retrieval field (i.e., VLAD and BoW) with the high effectiveness of Computer Vision approaches relying on local features and Random Sample Consensus (RANSAC) [6].

II. BACKGROUND

A. Local Features

A local feature is an image pattern, which differs from its immediate neighborhood. It is usually associated with a change of an image property or several properties simultaneously, although it is not necessarily localized exactly on this change. Local features describe interesting regions in an image. Interesting regions differ from their immediate neighborhood and should be consistently identified on any two images representing the same visual content [7]. The description of each interesting region [8] have to be robust to region transformation such as scale, rotation, affine, homography, etc...

The Scale Invariant Feature Transform (SIFT) [9], the most famous local feature, is a representation of the low level image content that is based on a transformation of the image data into scale-invariant coordinates relative to local features extracted from keypoints in an image. Keypoints are invariant to scale and orientation, selected by choosing the most stable points from a set of candidate location. Each keypoint in an image is associated with one or more orientations, based on local image gradients. Image matching is performed by comparing, typically using the Euclidean distance, the descriptions. Even if many other local features have been proposed in the last few year (e.g., SURF [10]) SIFT is still the most widely adopted.

¹<http://www.geomemories.org>

B. Bag of Words (BoW)

The goal of the BoW approach [11] is to substitute each description of the region around an interest point (i.e., each local feature) of the images with visual words obtained from a predefined vocabulary in order to apply traditional text retrieval techniques to content-based image retrieval. The visual vocabulary is typically built selecting the centroids of clusters identified using as *k-means*. The second step is to assign each local feature of the image to the identifier of the first nearest word in the vocabulary. At the end of the process, each image is described as a set of visual words. The retrieval phase is then performed using text retrieval techniques considering a query image as a disjunctive text-query. Typically, the *cosine* similarity measure in conjunction with a term weighting scheme is adopted for evaluating the similarity between any two images.

In this paper we will not use BoW for indexing images but for finding candidate matching pairs between two images. In fact, for indexing, we will make use of the more efficiency and effective VLAD approach described below. Given a set of candidate matches, for effective image alignment it necessary to perform RANSAC [6] on candidate matching pairs of points between the query image and the referenced ones. To speed up this process we make use of the BoW approach considering any region of interest described with the same word as matching.

C. Compact codes

Fisher kernels [12] describe how the set of descriptors deviates from an average distribution modeled by a parametric generative model. Fisher kernels have been applied in the context of image classification [13] and large scale image search [2]. In [4] it has been proved that Fisher vectors (FVs) extend the BoW. While the BoW approach counts the number of descriptors assigned to each region in the space, FV encodes the proximate location of the descriptors in each region and has a normalization that can be interpreted as an IDF term [14]. The FV image representation proposed by [13] assumes that the samples are distributed according to a Gaussian Mixture Model (GMM) estimated on a training set.

The VLAD representation was proposed in [1]. As for the BoW, a codebook $\{\mu_1, \dots, \mu_K\}$ is first learned using a cluster algorithm (e.g. *k-means*). Each local descriptor x_t is then associated to its nearest visual word $NN(x_t)$ in the codebook. For each codeword the differences between the vectors x_t assigned to μ_i are accumulated:

$$v_i = \sum_{x_t: NN(x_t)=i} x_t - \mu_i \quad (1)$$

The VLAD is the concatenation of the accumulated vectors, i.e. $V = [v_1^T \dots v_K^T]$.

In order to compare two VLAD image representation with the inner product similarity function two normalization are performed: first, a power normalization with power 0.5; second, a L2 normalization. The observation that VLAD descriptions are relatively sparse and very structured suggests a principal component analysis (PCA) that is usually performed to reduce the size of the Kd -dimensional VLAD vectors.

In [4], it has been proved that VLAD is a simplified non-probabilistic version of FV: VLAD is to FV what *k-means* is to GMM clustering. The *k-means* clustering can be viewed as a non-probabilistic limit case of GMM clustering. In [4] Euclidean Locality-Sensitive Hashing and its variant have been proposed to efficiently search VLAD descriptions.

D. Image alignment

Image alignment is the task of discovering the correspondence relationships among images with varying degrees of overlap. The survey given in [15] groups the approaches in the following categories: motion models [16], direct pixel-to-pixel comparisons, features based. In this work, we make use of a feature based approach. Our proposal is inspired by state-of-the-art image registration (i.e., the process of transforming different sets of data into one coordinate system) in combination with high-efficiency techniques developed mainly for landmark recognition.

III. GEOMEMORIES

GeoMemories is a joint project between the AeroFototeca Nazionale (AFN) of the Italian Ministry of Cultural Heritage in Rome and the Institute of Informatics and Telematics (IIT) of the Consiglio Nazionale delle Ricerche (CNR). The project is funded by the Italian Internet Domains Registry. AFN has an extensive set of aerial photographs that constitute a large archive of memory creating a historical record of the Italian territory between XIX and XX centuries. This huge archive of some millions of photos consists of several collections among which the Royal Air Force and the USA Air Force ones represent a view of Italy as it was 70 years ago. This landscape no longer exists and it has been transformed by the post-war reconstruction, the economic boom, the modernization and some natural disasters.

The project has developed a web platform for browsing aerial photos traveling along the spatial-temporal dimensions with the opportunity to also integrate multimedia data from other open archives or from social contributions. Since Version 5.0 Google Earth has added a timeline to display historical imagery but this new feature is very basic and has few functionality. For instance, it is not possible to create fading effects between two historical maps of the same area. In addition, aerial photos provided by Google are relatively recent and the biggest limitation is the lack of relevant historical aerial photos. Italy, for example, has a significant coverage only from 2003, moreover the only samples of actual historical imagery (1943) concerns some cities (Rome, Florence, Naples, Turin, Trieste and Venice) and the image resolution is very low. Our aim is to rebuild a virtual globe as similar to Google Earth oriented to the management of the time providing historical and spatial information.

The aerial photos made available to the project are digitized and stored to form a parallel virtual archive: this is an important measure for the protection of the originals, all on paper or film, which over time can thus be withdrawn from the direct manipulation, and preserved in the best conservative way. Digital images are then subjected to different steps, as described by Fig. 2, to create historical maps. Each photo, after being digitized, cropped and eventually equalized to

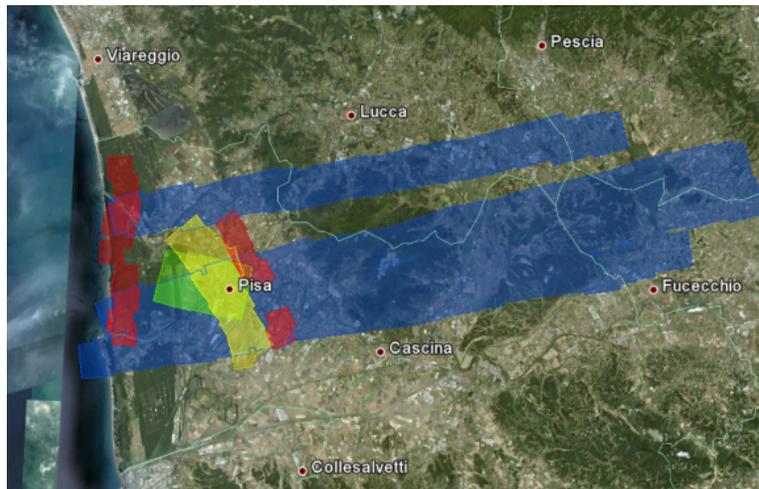


Fig. 1. Some aerial strip photos corresponding to 4 historical maps (20-08-1943 blue map, 18-02-1944 green map, 13-04-1944 yellow map and 14-05-1944 red map)

eliminate some exposure differences, are orthorectified and georeferenced. Google Imagery is used as reference map. Finally the georeferenced photos are joined by using mosaicing techniques.

Initially we have started to process 200 photos covering the northern part of Tuscany, and the years from 1943 to 1945, then from the second part of 2011 the set has been increased to 1000/2000 photos. This small set has been useful to develop and test several procedures to process the aerial strip photographs. At the end of the process we have realized four historical maps of the province of Pisa covering a time range from 20 August 1943 to 14 May 1944.

It is worth noting that the work flow for creating historical maps is really onerous and even though we can reduce the human component by developing automatic procedures, the huge size of the archive requires to find solutions based on social contributions so an important future activity of the project will concern the development of collaborative web applications to realize some steps such as georeferencing exploiting web volunteers (see the online tool <http://www.georeferencer.org/>).

The historical maps, each referring to a specific historical period, are browsable in the 4 spatio-temporal dimensions by using a web application based on Google Earth plugin and some javascript libraries. Fig. 3 shows a screenshot of the application. The user filter region and time of interest can select the corresponding historical maps and play with a fade in/out slide bar to display the evolution of the current area.

Google Earth integrates on its map different layers such as video, pictures, web cam by using the geographic reference (geotagging), we would like to use the same mechanism adding the time value (timetagging). The geo-historical data layers will come from the web obtained through web mining techniques, or filtered from open archives as wikipedia, youtube, flicker, and finally by raising social contributions related to initiatives for preserving historic memories. The result will be a sort of Historical Geographical Atlas where it will be possible to build tour to travel in the space-time dimension to tell stories.

A. Automatic Image Alignment

The image alignment task in GeoMemories consists in identifying overlapping of a non-georeferenced images with one of the georeferenced images already inserted in Geomemories. Our approach is features based and consists in retrieving first the most similar images in the knowledge based that are considered the best candidate for having an overlap with the query image.

As reported in [17] many geometric transformation can be found using the RANSAC algorithm [6]. Given the peculiar characteristics of our scenario, where the image dataset consists of aerial photos taken mostly from the vertical, we used the rotational and scale transformation, which provided us with the right compromise between simplicity and precision of results. In fact, even if the aerial photos represent a ground which is not strictly rigid and flat, more general transformations, such as affine and homography transformation, typically results in more noisy results and more difficult computation.

Given a point in a georeferenced image and a point in a non georeferenced image, we search for a rotational and scale transformation able to map the referenced point on the query image. Once such a transformation is found we are able to georeference the query image.

Unfortunately, comparing each aerial photos with subregions of the reference maps or another aerial photos using local features matching and RANSAC not only does not scale, but it is unfeasible for the scale of the problem we had. A single comparison takes seconds and the whole process would take hours. In GeoMemories, to speed-up the process, BoW and VLAD approaches are adopted for finding a set of candidates. Each image is processed as follows:

- 1) the SIFT local features are extracted using state of the art open source software (we used OpenMAJ but also OpenCV could be used). In principle any other local feature such as SURF or ORB could be used and are under investigation.
- 2) word assignment for BoW approach is achieved comparing each local features with the vocabulary defined

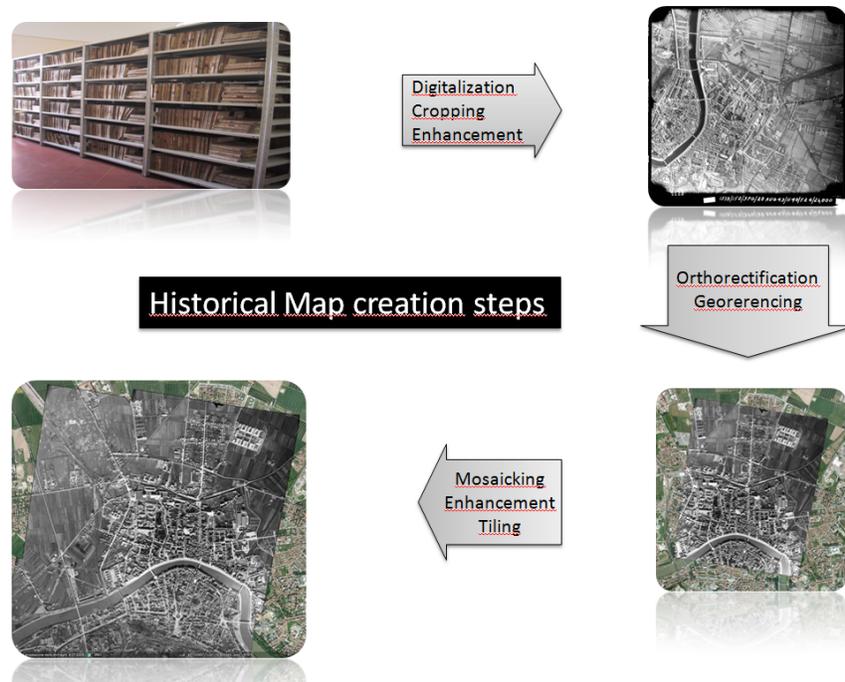


Fig. 2. The process steps for creating historical maps from aerial strip photography

- using the k-means on a sample set of images. To speed-up the assignment kd-tree has been adopted.
- 3) the VLAD descriptions are obtained by comparing each local features with the small vocabulary (64 reference) that the VLAD requires. The VLAD description is then composed as described in Sec. II-C.

For efficiently finding candidate overlapping images, VLAD descriptions of each image was inserted in a similarity search index, namely the MI-File [18].

Given a not georeferenced image, the process of automatic alignment is performed as following:

- 1) local features, BoW and VLAD description are extracted as previously described
- 2) the VLAD description is used as a query for finding the 100 most similar georeferenced images in the index
- 3) the BoW description of the query image are compared with the 100 most similar images using RANSAC and searching for rotational and scale transformation
- 4) the 3 most promising images are then compared with the query using the SIFT local features and searching for a rotational and scale transformation using RANSAC

After each step the candidate set becomes smaller and more costly algorithms are used. While the second step consider all the images in the dataset but only relies in the VLAD descriptions without geometric consistency checks, at the third step RANSAC is used to check for overlaps. However at this stage local features are considered only for their label (the visual word assigned). Eventually, in the last step, the actual SIFTs are used for effective matching.

As mentioned before, both VLAD and BoW approaches require a set of reference local features to be selected between a knowledge base. In our experiments we considered a first set of 1000 aerial images and selected 128 local features for VLAD and 10.000 words to be used for BoW.

IV. CONCLUSION

In this work, we presented the GeoMemories project focusing on the automatic image alignment approach developed for georeferencing aerial photographs given a set of manually aligned images. The approach rely on a increasing knowledge based given that both automatic and manually georeferenced images inserted in GeoMemories are used for georeferencing subsequent images. Our approach efficiently and effectively aligns aerial photographs combing techniques from the Multimedia Information Retrieval field based on VLAD and BoW with high effective Computer Vision approach relying on local features and RANSAC. The approach consists of 3 stages with increasingly cost of the analysis but with decreasing number of candidate images. While the whole dataset is searched at first, in the subsequent steps more robust approach are considered only for the candidate set selected at the steps before. In this way we are able to merge the high efficiency of recognition approaches from the Multimedia Information Retrieval community with the high effectiveness of the algorithms developed in the Computer Vision field..

The proposed approach is actually under experimentation on the GeoMemories infrastructure. The low percentage of manually aligned photos available at this time, did not allow us to report experimental results in this preliminary work. However, subjective results are promising and we plan to report objective results in future works.

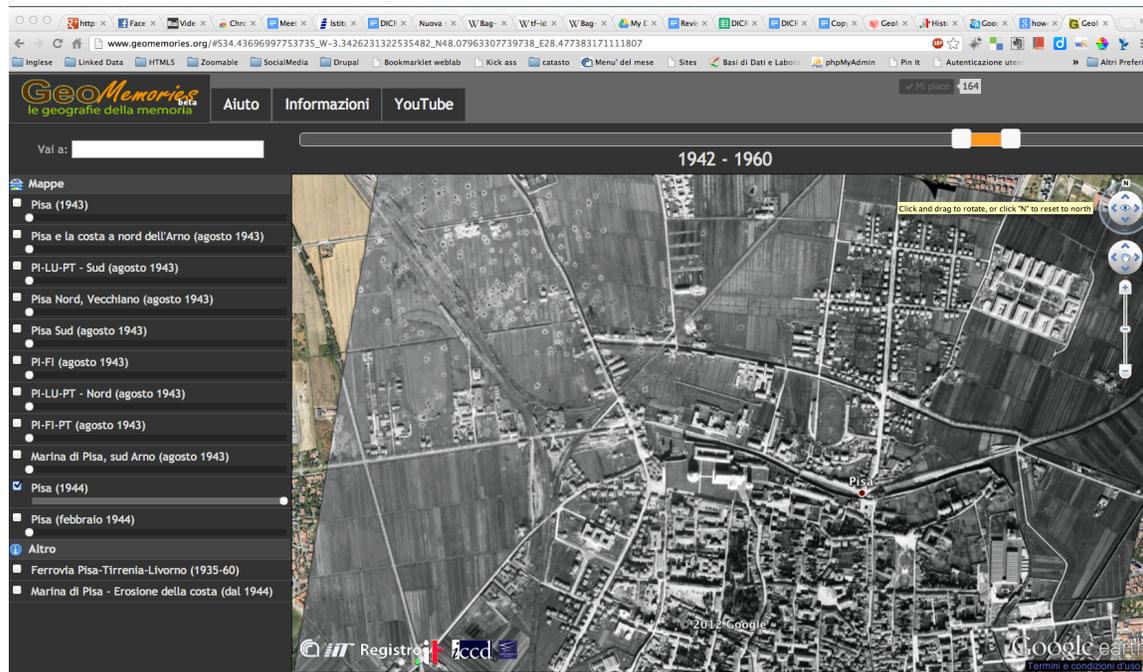


Fig. 3. A screenshot of the first prototype for browsing the historical maps

REFERENCES

- [1] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *Int. J. Comput. Vision*, vol. 87, pp. 316–336, May 2010. [Online]. Available: <http://dx.doi.org/10.1007/s11263-009-0285-2>
- [2] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June 2010, pp. 3384–3391.
- [3] P. Zezula, G. Amato, V. Dohnal, and M. Batko, *Similarity Search: The Metric Space Approach*, ser. Advances in Database Systems. Springer-Verlag, 2006, vol. 32.
- [4] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Sep. 2012, qUAERO. [Online]. Available: <http://hal.inria.fr/inria-00633013>
- [5] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proceedings of the twentieth annual symposium on Computational geometry*, ser. SCG '04. New York, NY, USA: ACM, 2004, pp. 253–262.
- [6] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [7] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Found. Trends. Comput. Graph. Vis.*, vol. 3, no. 3, pp. 177–280, 2008.
- [8] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, 2005.
- [9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [10] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *In ECCV*, 2006, pp. 404–417.
- [11] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ser. ICCV '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 1470–.
- [12] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *In Advances in Neural Information Processing Systems 11*. MIT Press, 1998, pp. 487–493.
- [13] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, June 2007, pp. 1–8.
- [14] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.
- [15] R. Szeliski, "Image alignment and stitching: a tutorial," *Found. Trends. Comput. Graph. Vis.*, vol. 2, no. 1, pp. 1–104, Jan. 2006.
- [16] r. Szeliski, "Video mosaics for virtual environments," *IEEE Comput. Graph. Appl.*, vol. 16, no. 2, pp. 22–30, Mar. 1996.
- [17] G. Amato, F. Falchi, and C. Gennaro, "Geometric consistency checks for knn based image classification relying on local features," in *SISAP '11: Fourth International Conference on Similarity Search and Applications, SISAP 2011, Lipari Island, Italy, June 30 - July 01, 2011*. ACM, 2011, pp. 81–88.
- [18] G. Amato and P. Savino, "Approximate similarity search in metric spaces using inverted files," in *Proceedings of the 3rd international conference on Scalable information systems*, ser. InfoScale '08. ICST, Brussels, Belgium, Belgium: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2008, pp. 28:1–28:10. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1459693.1459731>

Impact of Packet Loss on H.264 Scalable Video Coding

Siyu Tang
Bell Labs, ALCATEL-LUCENT
Antwerp, Belgium
siyu.tang@alcatel-lucent.com

Patrice Rondao Alfance
Bell Labs, ALCATEL-LUCENT
Antwerp, Belgium
Patrice.Rondao_Alfance@alcatel-lucent.com

Abstract—This paper presents an exact study of the impact of packet loss on H.264 scalable video coding (SVC). A Markov Chain (MC) with 2^N states is developed to describe the error propagation process inside a group of pictures (GOP). The model is extended to estimate the number of frames affected by transmission errors for a video sequence composed of multiple GOPs. By analyzing the inter-frame prediction rules, we examine the performance of different GOP structures against transmission errors. From the exact analysis, several metrics are analytically determined. Based on the proposed metrics, the performance of the SVC hierarchical B-frame structure and the advance video coding (AVC) IPPP structure (compatible base layer in SVC) are evaluated and compared under the assumption of random packet loss with rate p .

Keywords—QoE; SVC; AVC; Markov Chain

I. INTRODUCTION

In order to transport compressed video over a packet-based network (e.g., IP-network), the encoded bit-stream needs to be fragmented according to the *maximum transfer unit* (MTU). In an error-prone environment, packet loss may occur and cause distortion on the perceived video quality at the receiver. The resulting video distortion, however, varies according to the inter-frame predicting rules being used. Hence, it is important to understand the impact of transmission errors on different encoding schemes.

The scalable video coding (SVC) amendment [1] for the H.264/AVC (Advance Video Coding) standard provides scalable video streams in the temporal, spatial and quality dimensions with graceful adaptation between different scalability layers. It is considered to be a promising approach to offer quality adaptation to heterogeneous receivers with varying bandwidth constraints. Video distortion caused by packet loss in SVC is of great importance, as it determines the level of quality degradation, and provides insights on efficient quality adaptation.

Predicting transmission distortion in SVC is challenging, due to the three scalable dimensions and the hierarchical inter-frame prediction structures (the terms of *inter-frame prediction* and *frame dependency* are used interchangeably in this paper.). As we will discuss in Section II, an exact model of the transmission distortion in SVC seems missing. In this paper, we focus on an inter-frame prediction model for the hierarchical temporal prediction structure. In particular, we aim to investigate the error propagation process in

SVC hierarchical prediction structure and the robustness of different prediction mechanisms against transmission errors. Encoding schemes studied in this paper are: 1) SVC hierarchical B-frame structure (efficient compression, applicable for non-real-time video application), 2) AVC IPPP mode (SVC base layer compatible, applicable for real-time video delivery with stringent delay requirements).

The contributions of the paper are two-fold. First of all, to our knowledge, it is the first exact analysis that studies error propagation in the SVC hierarchical prediction structure. Instead of relying on extensive simulations, the exact model allows us to evaluate the performance of difference GOP (group of pictures) structures under packet loss accurately. Secondly, results obtained from this work can be used as a guideline in choosing the preferred codec (or GOP structure) that is more robust against transmission errors.

The rest of the paper is organized as follows. In Section II, related work is discussed. Section III presents preliminary definitions and the description of the error propagation problem. Section IV describes the proposed analytic model, along with the performance metrics. In Section V, we present the analytic results and key findings. Section VI concludes the paper.

II. RELATED WORK

There exists a rich number of studies focusing on the transmission distortion problem on the baseline profile of H.264/AVC. To simplify the loss-model, most of the studies assumed an additive model when consecutive packet losses occur, e.g., [6], [7] and [8]. In [9], the non-linearity of transmission distortion was considered. The proposed algorithm has shown superior performance over the linear models.

Existing transmission distortion models in AVC, however, cannot be directly applied to SVC due to the hierarchical prediction structure being employed. Most rate-distortion models regarding SVC aimed to optimize the perceived video quality as a function of different encoding parameters, e.g., [13], [14]. Studies about transmission distortion in SVC are either performed by experimental measures, or via approximations. Monteiro *et al.* [10] quantified the impact of packet loss on the SVC video stream with extensive simulations. Ghareeb *et al.* [11] have shown that the effect of packet loss can be reduced when delivering SVC layers

through multiple paths based on experiments conducted in the ns-2 simulator. In [12], a loss-distortion approximation model was developed but did not provide an exact analysis. In order to accurately predict the impact of packet loss on the SVC hierarchical prediction structure, an exact analysis is required, which is the focus of this paper.

III. PRELIMINARY DEFINITIONS AND PROBLEM DESCRIPTION

We consider the delivery of N encoded frames over a lossy network, where a unique *identification number* (ID) k ($0 \leq k \leq N-1$) is assigned to each frame. Let d ($0 \leq d \leq n_{gop} - 1$) be the unique ID of each GOP, where n_{gop} is the total number of GOPs in the video sequence. Let t ($0 \leq t \leq m$) be the unique identifier of each temporal sub-layer (or simply temporal layer) where m is the number of temporal layers in each GOP.

Employing variable bit-rate (VBR) encoding at the encoder leads to variable frame sizes after encoding. Let s_k denote such a random variable (r.v.), where k is the frame ID. In IP-networking, the *maximum service unit* (MST) is fixed to 1460 bytes excluding the 40-byte header. Hence, the number of packets n_k consisting of frame k after IP fragmentation is also a random variable. The total number of fragmented packets of the N frames is therefore obtained by $M = \sum_{k=0}^{N-1} n_k$.

Losing a packet in frame k will not only affect the current frame, but also propagates the initial error to subsequent frames due to the hierarchical inter-frame coding. In this paper, we identify the reason for a frame becoming erroneous as follows. If at least one packet is lost in a frame, we say that the frame is *erroneous due to packet loss*. Error propagation within a GOP takes place step by step, which is defined as *error propagation steps* r . We consider a worst case where a corrupted Macroblock will be used for inter-frame prediction and further propagates the error to successive frames. Any frame that is predicted from an erroneous frame due to the inter-prediction structure is considered as *being erroneous due to frame dependency*. Note that a frame impacted by packet loss can be again affected by frame dependency. We do not apply any advanced error concealment technique to the video sequence, so that error propagation is only evaluated under the influence of GOP structure. Pixels containing errors are simply concealed with zeros.

We measure the error propagation process within a GOP by the total number of erroneous frames, $Y[r]$, after each propagation step r . Let $\{Y[r], r \geq 0\}$ describe such a stochastic error propagation process. In the successive steps, all erroneous frames in the previous step keep disseminating the error to their neighbouring frames according to the inter-prediction rule, resulting in $Y[r]$ erroneous frames after step r . The total number of erroneous frames, $Y[r]$, is non-decreasing with r . The process of $Y[r]$ varies with respect

State index i	$f_{N-1}f_{N-2}\dots f_3f_2f_1f_0$
0	00.....0000
1	00.....0001
2	00.....0010
.....
$2^N - 1$	11.....1111

Table I
STATE SPACE OF THE ERROR PROPAGATION PROCESS WITH N FRAMES.

to the prediction structure being used.

Given the above definitions, we formulate our problem as follows: given 1) different GOP structures; 2) M encoded frames with variable frame size and number of fragmented IP packets; and 3) random packet loss over the N fragmented packets with probability p , we want to find out: The *probability density function* (pdf) that there are exactly y frames affected by packet loss and by frame dependency respectively.

IV. MODELING ERROR PROPAGATION IN SVC

A. A Markov Chain with 2^N states

In this section, we develop an exact analysis of the error propagation process in a GOP. The notion N is confined as the total number of frames in a GOP. The analysis is extended to predict the number of erroneous frames in a video sequence with multiple GOPs in Section IV-C2. At each discrete propagation step r , an arbitrary frame k can enter two states: 1) affected by errors, denoted by $F_k[r] = 1$; and 2) not affected by errors, denoted by $F_k[r] = 0$.

The state $Y[r]$ of the GOP at step r is defined by all possible combinations of the states, in which the N frames can be at step r

$$Y[r] = [Y_0[r], Y_1[r], \dots, Y_{2^N-1}[r]]^T \quad (1)$$

where $Y_i[r] = 1$ if $i = \sum_{k=0}^{N-1} F_k[r] \cdot 2^k$, and $Y_i[r] = 0$ otherwise. The total number of states is $\sum_{k=0}^N \binom{N}{k} = 2^N$, and the state space of the error propagation process is organized with $f_k \in \{0, 1\}$ as shown in Table I.

The error propagation process can be described exactly as a discrete Markov Chain (MC) since the current erroneous frames $Y_j[r]$ at step r only depends on those from the previous step $Y_i[r-1]$. The number of states with i erroneous frames is $\binom{N}{i}$ out of the 2^N ones.

Let P be an $(2^N) \times (2^N)$ transition probability matrix. Each entry in P , $P_{ij} = \Pr [Y_{r+1} = j | Y_r = i]$, denotes the probability that the MC moves from state i to state j in one step. The *probability state vector* $s[r]$ in step r is denoted by $s[r] = [s_0[r], s_1[r], \dots, s_{2^N-1}[r]]$, with $\sum_{i=0}^{2^N-1} s_i[r] = 1$ and $s_i[r] = \Pr [Y_r = i] = \Pr [F_0[r] = f_0, F_1[r] = f_1, \dots, F_{N-1}[r] = f_{N-1}]$.

The probability state vector can be calculated in terms of the initial state vector $s[0]$ and the transition probability matrix P from

$$s[r] = s[0] \cdot P^r \quad (2)$$

In the above proposed discrete MC, the maximal number of steps until entering an *absorbing state* (defined by $P_{ii} = 1$) is bounded by $r_{max} = 2^N - 1$. In other words, we take $2^N - 1$ as the upper bound of r for numerical calculations. Consequently, the steady-state vector is given by

$$\pi = s[r_{max}] = s[0] \cdot P^{r_{max}} \quad (3)$$

with $r_{max} = 2^N - 1$.

In order to solve (3), the initial state vector $s[0]$ and the transition probability matrix P need to be determined. In Section IV-C1, we present our approach of determining $s[0]$. Notice that both $s[0]$ and the state space description of the MC are independent of the GOP structure. The transition probability P_{ij} , however, is highly dependent on the inter-prediction structure being used. In Section IV-B, we discuss the calculation of the transition probabilities of the MC in different GOP structures.

B. The transition probabilities P_{ij}

In this section, we first explore the principles of frame ID assignment and inter-frame prediction. Afterwards, a so-called *dependency matrix* is developed to describe the dependency between frames. Finally, the transition probabilities P_{ij} are computed based on the dependency matrix. Two GOP structures are studied in the sequel.

1) *The hierarchical B-frame structure*: Following the conventions in [3], inter-prediction in the hierarchical B-frame structure is jointly initiated from I-frames in the current and preceding GOP, see Fig. 1. Hence, in our model, the number of frames in a GOP is defined as $N = 2^m + 1$, including the I-frame preceding the current GOP. Frame dependency is indicated by the inter-prediction arrows in Fig. 1. For example, frame 2 is predicted from 1 if there exists a outgoing arrow from 1 to 2.

Denote $v^d[t]$ a *frame ID vector* in each temporal layer t of GOP d . An entry in $v^d[t]$, $v_i^d[t]$, refers to the i -th frame ID in layer t . As shown in Fig. 1, frames in the base layer are always assigned by $v^d[0] = [d \cdot 2^m, (d+1) \cdot 2^m]$, with $0 \leq d \leq n_{gop} - 1$. The assignment of frame IDs in layer t ($t \geq 1$) of GOP d obeys a general rule, that is, a frame ID in layer t is iteratively computed by adding or subtracting

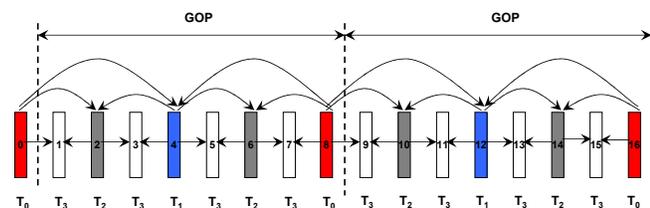


Figure 1. Hierarchical B-frame prediction structure of SVC temporal scalability. Frames are presented in display order, while the number below each frame indicates its corresponding temporal layer. Frames in the base layer T_0 are coded as I-frames. Those in the enhancement layers T_1, T_2, \dots are coded as B-frames.

2^{m-t} from the frame it depends on in layer $t-1$. By using such a rule, frames are ranked according to the displaying order. For instance, each frame in layer $t-1$, $v_i^d[t-1]$, determines two frame IDs in layer t on its left and right side by

$$\begin{cases} v_j^d[t] = v_i^d[t-1] - 2^{m-t} & \text{left side} \\ v_{j+1}^d[t] = v_i^d[t-1] + 2^{m-t} & \text{right side} \end{cases} \quad (4)$$

Notice that both I-frames determine a single frame in layer $t=1$ of the current GOP by $d \cdot 2^m + 2^{m-1}$ or $d \cdot 2^m - 2^{m-1}$. The number of frames in layer t , $w[t]$, is given by

$$w[t] = \begin{cases} 2 & \text{if } t = 0 \\ 1 & \text{if } t = 1 \\ 2^{t-1} & \text{if } 2 \leq t \leq m \end{cases} \quad (5)$$

Next, we investigate the frame prediction rule of the hierarchical B-frames as a function of frame IDs. Let k_{t_1} be an arbitrary frame in layer t_1 ($k_{t_1} \in v^d[t_1]$), and k_{t_2} an arbitrary frame in layer t_2 , ($k_{t_2} \in v^d[t_2]$). Four observations are revealed from Fig. 1: 1) k_{t_2} is predicted from k_{t_1} if $t_2 > t_1$. 2) Bi-directional prediction applies to frames in $t \geq 1$. Frame k_{t_1} predicts one single frame in each succeeding layer along one direction. 3) Inter prediction of the two I-frames ($t=0$) are bounded on their right- and left- side respectively. 4) Frames in layer $t_2 = m$ are not used to predict other frames. To summarize, a frame, k_{t_2} , is predicted from k_{t_1} if and only if the following relationships are satisfied:

$$k_{t_2|t_1} = k_{t_1} + \frac{2^m}{2^{t_2}} \quad \text{or} \quad k_{t_2|t_1} = k_{t_1} - \frac{2^m}{2^{t_2}} \quad (6)$$

where $0 \leq t_1 < t_2 \leq m$, and $t_2|t_1$ denotes the dependency of frame k_{t_2} on k_{t_1} . The number of frames, n_{dep} , that are predicted from an arbitrary frame in layer t is

$$n_{dep} = \begin{cases} m-t & \text{if } t = 0 \\ 2 \cdot (m-t) & \text{if } 0 < t < m \\ 0 & \text{if } t = m \end{cases} \quad (7)$$

To better illustrate the dependency between frames in (6), we develop a so-called $N \times N$ *dependency matrix* M . Each entry e_{xy} in M defines the dependency of frame y on frame x ($0 \leq x \leq 2^m + 1$ and $0 \leq y \leq 2^m + 1$). If the dependency of frame y on x is true, we have $e_{xy} = 1$. Otherwise, it is $e_{xy} = 0$. Combining (6) and (7), the entries in M are organized by

$$e_{xy} = \begin{cases} 1 & \text{if } x = y \\ 1 & \text{if } y = x - \frac{2^m}{2^{t_2}} \\ 1 & \text{if } y = x + \frac{2^m}{2^{t_2}} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

with $0 \leq t_1 < t_2 \leq m$, where t_1 is the temporal sub-layer that frame x belongs to, and t_2 is the temporal sub-layer that frame y belongs to. The first condition $x = y$ assures that a frame stays erroneous once it is contaminated. The second

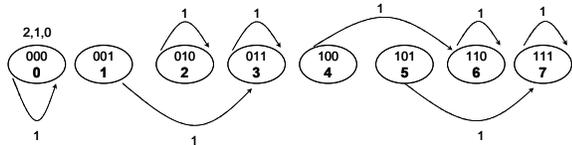


Figure 2. State diagram of the hierarchical B-frame structure with $m = 1$, $N = 3$ and $2^N = 8$ states. The number above state 0 indicates the frame ID with the right-to-left order. The arrows refer to the transition between states.

and third conditions includes all frames that are dependent on frame x .

The transition probability P_{ij} is derived together with the dependency matrix as

$$P_{ij} = \begin{cases} 1 & \text{if } \forall F_y \in Y_j: F_y = \bigvee_{x=0}^{N-1} F_x \cdot e_{xy} \text{ with } \forall F_x \in Y_i \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where Y_i and Y_j are the i -th and j -th states in Table I, and $F_y = \bigvee_{x=0}^{N-1} F_x \cdot e_{xy} = F_0 \cdot e_{0y} \vee F_1 \cdot e_{1y} \vee \dots \vee F_{N-1} \cdot e_{N-1,y}$. The operation $A \vee B$ is 1 if A or B , or both are 1. If both are 0, $A \vee B$ is zero. Fig. 2 shows an example of the Markov state diagram with $N = 3$ frames and $2^N = 8$ states. As we can see from Fig 2, it is only possible to transit from state i to state j if $j \geq i$.

2) *The IPPP prediction mode:* The first frame in a GOP of the IPPP prediction mode is always an I-frame. All succeeding frames are encoded as P-frames, see Fig. 3. Note that there is no such concept of “temporal layer” in AVC, N is simply the number of frames in a GOP (including one I-frames and all succeeding P-frames).

The $N \times N$ dependency matrix of the IPPP mode is

$$e_{xy} = \begin{cases} 1 & \text{if } y = x \\ 1 & \text{if } y = x + 1 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

as a frame x only affect one single frame on its right side. Substitute (10) to (9), P_{ij} of the IPPP mode is computed. State diagram of the IPPP mode is not presented due to space limitations.

C. Performance evaluation

As shown in (3), the performance of error propagation is determined together by the initial state vector $s[0]$ and the transition probability matrix P . In this section, we discuss our approach to obtain $s[0]$ and the methodology to predict

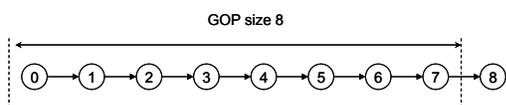


Figure 3. The structure of IPPP mode with $N = 8$ frames.

the number of erroneous frames affected by transmission errors both inside a GOP and within a video sequence consisting of multiple GOPs.

1) Estimating number of erroneous frames in a GOP:

Given random packet loss with probability p over M packets, the number of frames affected by packet loss, denoted by Y_{pkt} , is a r.v. depending on n_k and p . Let Y_{dep} be the number of erroneous frames contaminated by frame dependency after error propagation. The sum of Y_{pkt} and Y_{dep} equals $Y[r_{max}]$ at step r_{max} .

The initial state vector $s[0]$ defines the probability that each state in Table IV-A may occur, where frame k is affected by packet loss with probability $1 - q^{n_k}$, and not affected with probability q^{n_k} . Given n_k packets in frame k and $q = 1 - p$, $s[0]$ is computed by

$$s_i[0] = \prod_{k=0}^{N-1} \{1_{f_{k;i}=1} \cdot (1 - q^{n_k}) + 1_{f_{k;i}=0} \cdot q^{n_k}\} \quad (11)$$

where $f_{k;i}$ denotes the status of frame k in state Y_i . The indicator function 1_y is defined as 1 if y is true, otherwise 1_y is zero [2, pp. 30].

Consequently, the pdf of Y_{pkt} is derived as

$$\Pr[Y_{pkt} = y] = \sum_{i=0}^{2^N-1} 1_{y_i=y} \cdot s_i[0] \quad (12)$$

where $y_i = \sum_{k=0}^{N-1} f_{k;i}$ is the number of frames with status 1 in state Y_i . The pdf of Y_{tot} is

$$\Pr[Y_{tot} = y] = \sum_{i=0}^{2^N-1} 1_{y_i=y} \cdot s_i[r_{max}] \quad (13)$$

with $y_i = \sum_{k=0}^{N-1} f_{k;i}$ and $s_i[r_{max}]$ is computed from (3). Let $i \rightarrow s$ be a realization (i.e., a sample path) of the error propagation process from an initial state i to the absorbing state s . The number of erroneous frames caused by inter-frame dependency for a single sample path from state i to state s is

$$y_{dep;i \rightarrow s} = y_s - y_i = \sum_{k=0}^{N-1} f_{k;s} - \sum_{k=0}^{N-1} f_{k;i} \quad (14)$$

Pdf of the number of frames influenced by frame dependency is therefore, determined by

$$\Pr[Y_{dep} = y] = \sum_{i=0}^{2^N-1} \sum_{j=0}^{2^N-1} 1_{y_j-y_i=y} \cdot 1_{P_{ij}^{r_{max}}=1} \cdot s_i[0] \quad (15)$$

where $y_i = \sum_{k=0}^{N-1} f_{k;i}$, $y_j = \sum_{k=0}^{N-1} f_{k;j}$. Elements in the n -step transition probability matrix, $P_{ij}^{r_{max}} = \Pr[Y_{r_{max}} = j | Y_0 = i]$, defines the probabilities to move from initial state i to the steady-state j .

The mean and variance of Y_{tot} are $E[Y_{tot}] = \sum_{i=0}^N i \cdot \Pr[Y_{tot} = i]$ and $\text{Var}[Y_{tot}] =$

$\sum_{i=0}^N (i - E[Y_{tot}])^2 \cdot \Pr[Y_{tot} = i]$ respectively. The mean and variance of Y_{pkt} and Y_{dep} can be obtained in an analogous way.

2) *Estimating number of erroneous frames in a video sequence:* Performing an exact analysis about the erroneous frame estimation problem in a video sequence with multiple GOPs is very difficult. At step r , an arbitrary GOP d may enter N state: $G_d[r] = i$ with $0 \leq i \leq N$ and N is the total number of frames in a GOP.

The major challenge is to describe the entire system exactly. That is, to find all possible combinations the states that the n_{gop} GOPs can be at step r . As discussed in Section IV-A, we use 2^N states to describe N frames with two possible states. Given N states instead of 2 for each GOP, the exact analysis becomes more complex and requires a huge state space. Hence, we resort to a simple approximation to compute the number of erroneous frames in a video sequence.

Let Y_{pkt}^* , Y_{dep}^* and Y_{tot}^* be the number of frame affected by packet loss, frame dependency and the sum of the above two r.v. in the entire video sequence. Instead of seeking for the pdf of Y_{pkt}^* , Y_{dep}^* and Y_{tot}^* , we derive their expectations as

$$E[Y_{tot}^*] = \sum_{d=0}^{n_{gop}-1} E[Y_{tot;d}] \quad (16)$$

where $E[Y_{tot;d}]$ is calculated in Section IV-C1 for a single GOP d . $E[Y_{pkt}^*]$ and $E[Y_{dep}^*]$ are computed analogously.

V. RESULTS AND DISCUSSIONS

In this section, a high definition (HD) video sequence *old_town_cross* with the resolution of 1980x1080 and frame rate of 50fps is encoded with the SVC reference encoding software JSVM (Joint Scalable Video Model) [4] to generate the hierarchical B-frame and IPPP mode encoded frames (compatible with AVC). The version of JSVM under use was 9.18. We assign $m=3$ in the B-frame structure, which results in 9 frames in each GOP. To have fair comparison, the GOP size of the P-frame mode is set to 9 as well. Encoding parameters such as the quantization parameter are the same for both schemes.

We also developed a simulation program by using the C language to simulate IP fragmentation of the encoded bit-stream. The program also simulates packet loss over the fragmented packets. Since JSVM cannot decode frames containing errors, an extra script was developed to filter out damaged (hence undecodable) frames before decoding. The JSVM software, therefore, only decodes those frames that are error-free. Results obtained from the analysis are compared with the results derived from the simulated program. For each simulated result, 10^4 iterations are carried out.

A. The hierarchical B-frame structure

First of all, we compare the analytic $\Pr[Y_{pkt} > y]$ and $\Pr[Y_{dep} > y]$ with simulated results in Fig. 4(a) and (b). As

we can see from Fig. 4, both curves match the simulated results very well. The tail probability $\Pr[Y_{pkt} > y]$ (or $\Pr[Y_{dep} > y]$) defines the probability that more than y frames are affected by packet loss (or frame dependency). Given the packet loss rate of $p = 0.1$, the probability that more than 3 frames are impacted by packet loss is approximately 0.32, shown in Fig. 4(a). The probability that more than 6 frames being affected by packet loss decreases to the order of 10^{-3} and 10^{-6} for $y = 8$ frames. Frame dependency seems to play an influential role, as illustrated in 4(b). In 90% of the cases, more than 4 frames are affected by frame dependency. In 28% of the cases, more than 6 frames are contaminated. With the exact analysis, we are able to evaluate the performance of $\Pr[Y_{pkt} > y]$ (or $\Pr[Y_{dep} > y]$) with high accuracy, which is normally very difficult to achieve with simulations. For instance, in order to have an accuracy of 10^{-6} for $y = 8$, as in Fig. 4(a), the simulation needs to be performed 10^{12} times, which is time consuming. In the following, only analytic results will be discussed except for Fig. 5(b) where the approximation in (16) is verified with simulated results for a video sequence with multiple GOPs.

In Fig. 5(a), the analytic results of $E[Y_{pkt}]$, $E[Y_{dep}]$ and $E[Y_{tot}]$ are presented respectively. When packet loss rate is smaller than 0.1, frame dependency is a dominant factor in propagating errors. However, with $p > 0.2$, $E[Y_{dep}]$ started to decrease and $E[Y_{pkt}]$ begins to grow more drastically. This is because, given a fixed amount of frames in a GOP, more frames being affected by packet lost naturally leads to less frames being influenced by dependency. The inset of Fig. 5(a) plots $E[Y_{tot}]$ together with its associated upper and lower bounds. The upper and lower bounds are computed by $E[Y_{tot}] \pm \sigma$ respectively, where $\sigma = \sqrt{\text{Var}[Y_{tot}]}$ is the standard deviation. The two bounds for $E[Y_{pkt}]$ and $E[Y_{dep}]$ are not presented due to space limit. Notice that the upper and lower bounds indicate the worst and optimal performance. For instance, we see that with $p = 10^{-2}$, the maximal number of erroneous frames reaches 9 and the minimal number of contaminated frames is around 6. From

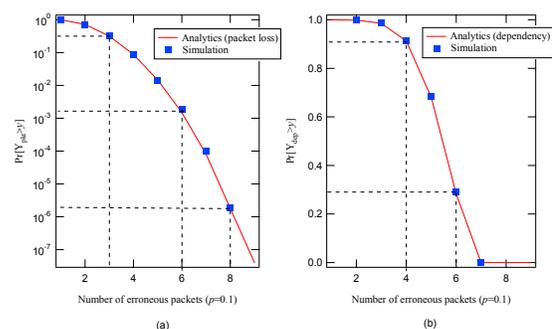


Figure 4. The tail behavior of $\Pr[Y_{pkt} > y]$ (a) and $\Pr[Y_{dep} > y]$ (b) versus number of erroneous frames y with $p = 0.1$.

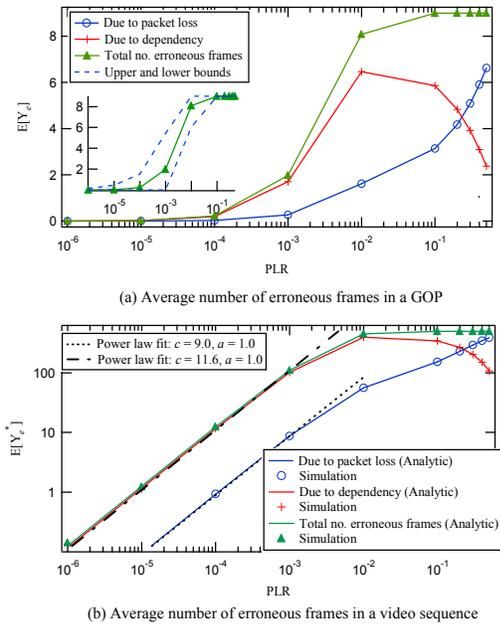


Figure 5. (a) Average number of erroneous frames in a single GOP as a function of p (lin-log scale). The inset plots the upper and lower bounds of the total number of erroneous frames. (b) Average number of erroneous frames in a video sequence with 62 GOPs as a function of p (log-log scale). Both figures present the analytic results. The approximation of (16) is verified with simulations in (b).

a statistic point of view, all results of individual packet loss events and the resulting erroneous frames are bounded by the two dotted curves.

In Fig. 5(b), we plot $E[Y_{pkt}^*]$, $E[Y_{dep}^*]$ and $E[Y_{tot}^*]$ for a video sequence with 62 GOPs (approximately 9.3 seconds). As revealed from Fig. 5(b), equation (16) approximates simulated results very well. On a log-log scale, $E[Y_{pkt}^*]$, $E[Y_{dep}^*]$ and $E[Y_{tot}^*]$ exhibit straight lines until the point of $p = 10^{-3}$, conforming to the power law distribution (defined as $y = c \cdot x^a$, where c is a normalization constant). The fitting parameter of a defines the slope of a curve. The fitting curves in Fig. 5(b) allow us to predict the average number of erroneous frames based on a and c without employing (16). However, the power law distribution fails to approximate the curves if $p > 10^{-3}$. We see clearly that, less than 20% of the frames ($E[Y_{tot}^*]$) are affected by transmission errors up to $p = 10^{-3}$. With $p = 10^{-2}$, around 80% of the frames are affected. If $p \geq 10^{-1}$, almost all frames are contaminated.

B. Comparing the B-frame structure with the IPPP mode

In this section, the performance of the hierarchical B-frame structure is compared with the IPPP mode. As shown in Fig. 6, the IPPP mode is more sensitive to packet loss than the B-frame structure. This is because the IPPP mode has lower compression efficiency compared to the B-frame structure. A P-frame generally consists of more packets than a B-frame. According to (11), larger n_k incurs higher

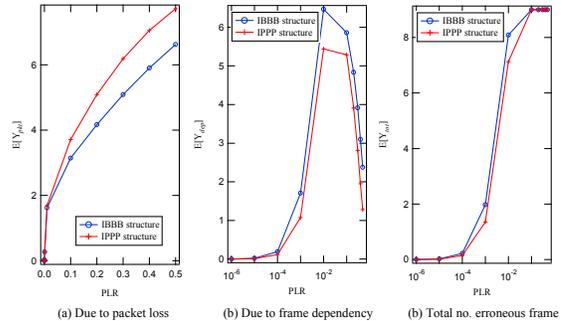


Figure 6. Average number of erroneous frames caused by packet loss (a), frame dependency (b) and both (c) of the IBBB and IPPP structure. Figure (b) and (c) are plotted on lin-log scale for easier reading.

frame error probability. Therefore, the IPPP mode is more vulnerable to packet loss, especially with larger p .

The frame dependency of the B-structure is, by nature, higher than the IPPP mode, as plotted in Fig. 6(b). The maximal absolute difference between the two curves occurs at $p = 10^{-2}$, where the inter-frame prediction in the B-frame mode leads to one more erroneous frame than in the P-frame mode. The total number of erroneous frames in Fig. 6(c) is comparable between the IBBB and the IPPP structure. The average absolute difference between the B-frame and the P-frame mode is around 0.5, 0.7 and 0.2 frames in Fig. 6(a), (b) and (c) respectively. Considering the higher coding efficiency of the IBBB structure, and the marginal difference between the B- and P- frame mode against transmission errors, the SVC IBBB hierarchical coding structure appears to be a good candidate for video transmission in an error-prone environment with random packet loss.

VI. CONCLUSION

This paper presents an exact analysis to examine the impact of packet loss on the H.264 scalable video coding. With a simple approximation, the model is extended to predict the performance of a video sequence consisting of multiple GOPs. Major conclusions from the performance analysis are: 1) Frame dependency in SVC B-frame structure is a dominant factor in propagating transmission errors with packet loss rate $p < 0.1$. 2) The upper and lower bounds obtained from the analysis suggests the worst and optimal performance of individual loss events. In order to satisfy users with the worst performance, it is important to look at the upper bounds and thereafter enhance the robustness of the bit-stream to be delivered under packet loss. 3) When $p \leq 10^{-3}$, the average number of frames affected by transmission errors can be approximated by the power law distribution. To avoid drastic increment in the number of erroneous frame, the packet loss rate should be controlled as $p \leq 10^{-3}$ (depending on system requirements). 4) Despite the hierarchical frame dependency, the overall performance

of the B-frame structure is, in fact, comparable with the IPPP mode under random packet loss.

Performance evaluation presented in this paper is based on the number of frames affected by transmission errors. Examining other metrics, such as the PSNR, that can properly reflect the pixel-level quality degradation is the focus of our future work. The error propagation process investigated in this paper can be directly employed to describe the inter-frame motion compensation in the future model. Besides, adapting the initial state vector $s[0]$, the burst packet loss process can be incorporated.

VII. ACKNOWLEDGMENT

The authors would like to thank W. Acke for proofreading this paper.

REFERENCES

- [1] ITU-T and ISO/IEC JTC1, JVT-W201, "JointDraft 10 of SVC Amendment", Apr. 2007.
- [2] P. Van Mieghem, "Performance Analysis of Communications Networks and Systems", Cambridge University Press, 2006.
- [3] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H. 264/AVC standard", *IEEE TCSVT*, 2007. 17(9): pp. 1103–1120.
- [4] Team, J.V., H. 264/SVC reference software (JSVM 9.18) and manual. CVS sever at garcon. ient. rwth-aachen. de, 2009.
- [5] S. Tang, E. Jaho, I. Stavrakakis, I. Koukoutsidis, and P. Van Mieghem, "Modeling gossip-based content dissemination and search in distributed networking", *Computer Communications*, 2011. 34(6): pp. 765–779.
- [6] K. Stuhlmuller, N. Farber, M. Link, and B. Girod, "Analysis of video transmission over lossy channels". *IEEE Journal on Sel. Areas in Communications*, 2000. 18(6): pp. 1012–1032.
- [7] Y.J. Liang, J.G. Apostolopoulos, and B. Girod, "Analysis of packet loss for compressed video: Does burst-length matter?", *Proc. IEEE ICASSP*, Apr. 2003. vol. 5: pp. 684–687.
- [8] S. Tao, J. Apostolopoulos, and R. Guérin, "Real-time monitoring of video quality in IP networks", *IEEE/ACM Trans. on Networking (TON)*, 2008. 16(5): pp. 1052–1065.
- [9] Z. Chen and D. Wu, "Prediction of Transmission Distortion for Wireless Video Communication: Analysis", *IEEE Trans. on Image Processing*, 2012. 21(3): pp. 1123–1137.
- [10] J.M. Monteiro, C.T. Calafate, and M.S. Nunes, "Evaluation of the H. 264 scalable video coding in error prone IP networks", *IEEE Trans. on Broadcasting*, 2008. 54(3): pp. 652–659.
- [11] M. Ghareeb, A. Ksentini, and C. Viho, "Scalable Video Coding (SVC) for multipath video streaming over Video Distribution Networks (VDN)", *Int. Conf. on Information Networking (ICOIN)*, 2011: pp. 206–211.
- [12] H. Mansour, P. Nasiopoulos, and V. Krishnamurthy, "Modeling of loss-distortion in hierarchical prediction codecs", *IEEE Int. Symp. on Signal Processing and Information Technology*, 2006: pp. 536–540.
- [13] Y. Wang, Z. Ma, and Y.F. Ou, "Modeling rate and perceptual quality of scalable video as functions of quantization and frame rate and its application in scalable video adaptation", *IEEE Int. Packet Video Workshop*, 2009: pp. 1–9.
- [14] M.R. Ardestani, A. Shirazi, and M.R. Hashemi, "Rate-distortion modeling for scalable video coding", *IEEE 17th Int. Conf. on Telecommunications (ICT)*, 2010: pp. 923–928.

News Video Semantic Topic Mining Based on Multi-wing Harmoniums Model

Xin Wen Xu and Yu Bo Shen

Department of traffic and transportation engineering
National University of Defense Technology
Changsha, China
{xinwen_xu, shenyubo}@126.com

Guo Hui Li

Department of System Engineering
National University of Defense Technology
Changsha, China
guohli@nudt.edu.cn

Abstract—Two-layer undirected graphical model, Harmoniums, is a new approach to mine latent semantic topics from observed data. For the multi-modal heterogeneous features of news video, this paper proposes multi-wing Harmoniums (MWH) model that represents news video stories as latent semantic topics derived by jointly modeling the transcript text, color histogram and edge histogram of the video. This model includes a multivariate Poisson distribution and two multivariate Gaussian distributions. It extends and improves earlier models based on two-layer random fields, which capture bidirectional dependencies between hidden topic aspects and observed inputs. The model especially facilitates efficient inference and robust topic mixing, and provides high flexibilities in modeling the latent topic spaces. The variational algorithm efficiently reduces the difficulty of model learning. The experiments results on the CCTV news video collections and an extensive comparison with various extant models show the efficiency of MWH on news video semantic mining.

Keywords—news video multi-wing Harmoniums; video mining; semantic mining; news video

I. INTRODUCTION

Along with the rapid development of processor speed and the Internet as well as the availability of inexpensive massive digital storages, there have been strong demands for the modeling and mining of the multiple media sources, like text, image, audio and video. Numerous researchers have shown great interests to the news video as a mass medium for its easiness to acquire and richness in information. To fill the semantic gap between the low-level features and the high-level semantic topic of news video, it's necessary to make full use of the rich information provided by the multi-modal heterogeneous data so that we can effectively achieve the data mining tasks on news video like classification, cluster, retrieval and image annotation. The multi-modal heterogeneous data refers to the data of the objects to be described collected from different approaches or perspectives. And we refer to each of the approaches or perspectives as a modality. For example, in the multi-modal face detection, the multi-modal data can consist of the 2d face images and 3d face shape model; in the mining of multi-modal videos, videos can be decomposed into subtitles, audios, images, etc. Therefore, the key issues of video semantic mining researches are to model the associated data from multiple sources jointly and explore appropriate lower dimensional

latent representations of the originally high-dimensional features. The fusion of multi-modal data like key frame image, audio and transcript text, has been a widely used technique in video processing. The fusion strategy includes the feature-level fusion in earlier stage and the later decision-level fusion. It is an open question as to which fusion strategy is more proper for a task. Snoek et al. [1] compares the two strategies in the classification of videos.

There are many approaches to obtain low-dimensional intermediate representations of video data. Principal component analysis (PCA) [2] has been the most popular method, which projects the raw features into a lower-dimensional feature space where the data variances are well preserved. Independent Component Analysis (ICA) [3] and Fisher Linear Discriminant (FLD) [4] are also widely used in dimensionality reduction. Recently, there are also many proposals on modeling the latent semantic topics of the text and multimedia data. For example, Latent Semantic Indexing (LSI) [5] finds a linear transform of word counts into a latent eigenspace of document semantics. Though LSI can roughly obtain the latent semantic and work well in automatic index application, it generates overfitting for failing to meet the statistics principles. Later, LSI is extended to probabilistic Latent Semantic Indexing (pLSI) [6], which models the latent topic into the probability distribution of words and the documents into the probability distribution of the latent topic. The pLSI is based on the principle of probability and defines proper generative model, so it can be applied to model composition and can control the complexity. The Latent Dirichlet Allocation (LDA) [7] is a directed graphical model that provides generative semantics of text documents, where each document is associated with a topic-mixing vector and each word is independently sampled according to a topic drawn from this topic-mixing. LDA is later extended to Gaussian-Mixture LDA and Correspondence LDA [8], both of which are used to model annotated data such as captioned images or video with transcript text.

In fact, the methods mentioned above are mainly used to transform high-dimension of raw features to low-dimensional presentation and presumably gain the latent semantics of data. However, these methods are mainly applied to single modal data and can't or can hardly be applied to the multi-modal heterogeneous data. Two-layer undirected graphical model, Harmoniums [9], is a new approach to mine latent semantic topics from observed data [16]. Based on Harmoniums models, we present news video

multi-wing harmoniums (NVMWH) model that represents story unit as latent semantic topics derived by jointly modeling the transcript keywords, color histogram and edge histogram features of news video data for news video semantic topic mining.

The rest of the paper is structured as follows. In Section 2, we present the latent semantic topic model of news video-based on multi-wing harmoniums model and the Learning and Inference of its parameters. Section 3 presents the experiments and discussions. The paper concludes in Section 4.

II. THE LATENT SEMANTIC TOPIC MODEL OF NEWS VIDEO BASED ON MULTI-WING HARMONIUMS MODEL

A. The basic Harmoniums model

The basic Harmoniums model, which was originally studied by Smolensky (1986) [9] in his Harmony theory, defines a complete bipartite undirected graphical model containing two layers of nodes (Fig. 1). Let $H = \{h_i\}$ denotes the set of hidden units in such a graph, and let $X = \{x_i\}$ denotes the set of input units. The Harmoniums model creates a random field for the undirected graph model.

$$p(x, h | \theta) = \frac{1}{Z(\theta)} \exp\{\sum_i \theta_i \phi_i(x_i) + \sum_j \theta_j \phi_j(h_j) + \sum_{ij} \theta_{ij} \phi_{ij}(x_i, h_j)\} \quad (1)$$

where $\phi_e(\cdot)$ denotes the potential function defined on either a singleton or a connected pair of units (indexed by e) in the model, θ_e denotes the weight of the corresponding potential, and $Z(\theta)$ stands for the log-partition function.

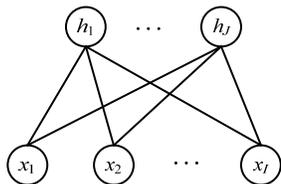


Figure 1. The basic Harmoniums Model

The bipartite topology of the harmoniums graph suggests that if the nodes within the same layer are given then the nodes of the other opposite layer are conditionally independent. This makes possible a feasible definition of the Harmoniums distribution based on the conditional distribution function $p(x|h)$ and $p(h|x)$ between the two layers: $p(x|h) = \prod_i p(x_i | h)$, $p(h|x) = \prod_j p(h_j | x)$. Hence, it is semantically simple and easy to design. For simplicity, all conditional probabilities considered here adopt exponential forms.

$$p(x_i | h) = \exp\{\sum_a \hat{\theta}_{ia} f_{ia}(x_i) - A_i(\{\hat{\theta}_{ia}\})\} \quad (2)$$

$$p(h_j | x) = \exp\{\sum_a \hat{\lambda}_{ja} g_{ja}(h_j) - B_j(\{\hat{\lambda}_{ja}\})\} \quad (3)$$

where $\{f_{ia}(\cdot)\}$ and $\{g_{ja}(\cdot)\}$ respectively denote the sufficient statistics of variable x_i and h_j . $A_i(\cdot)$ and $B_j(\cdot)$ denote

respective log-partition functions. And the shifted parameters $\hat{\theta}_{ia}$ and $\hat{\lambda}_{ja}$ are defined as $\hat{\theta}_{ia} = \theta_{ia} + \sum_b W_{ib}^a g_{jb}(h_j)$ and $\hat{\lambda}_{ja} = \lambda_{ja} + \sum_b W_{ja}^b f_{ib}(x_i)$. It's produced by the input and all the matching pairs in hidden layers. Welling et al. [10] showed that these easily comprehensible and manipulable local conditionals precisely map to the Harmoniums random fields:

$$p(x | h) \propto \exp\{\sum_{ia} \theta_{ia} f_{ia}(x_i) + \sum_{jb} \lambda_{jb} g_{jb}(h_j) + \sum_{ijab} W_{ia}^{jb} f_{ia}(x_i) g_{jb}(h_j)\} \quad (4)$$

θ_{ia} , λ_{ja} and W_{ia}^{jb} are the set of parameters associated with their corresponding potential functions. It's very difficult to make parameter estimation for the appearance of the log-partition function with joint probability, so ratio symbol rather than precise equal symbol is used in the formula. Such a model was referred to as the Exponential Family Harmoniums (EFH) [10]. In the sequel, we will take advantage of this bottom-up strategy to construct specific Harmoniums from the easily comprehensible local conditionals. It can be shown that there is no marginal independence for either input or hidden variables in a Harmoniums model. However, an EFH enjoys the advantages of conditional independence between hidden variables, which is generally violated in the directed models. This property greatly reduces reasoning difficulty. But typically, learning harmonium is more difficult due to the presence of a global partition function.

B. The multi-wing Harmoniums model

The hidden and input units in a Harmoniums model are symmetrical, which cannot contribute to explain their causal relationship in semanteme. However, the definition mentioned above of the local condition independence based on two layers can provide explanations for the bidirectional causality of Harmoniums structure. Essentially, the hidden unit H can be considered as latent topic which defines the production of input. Conversely, H can also be seen as the predictors produced by a discriminative model of the input unit.

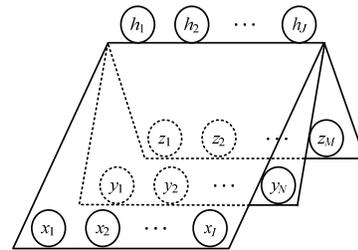


Figure 2. Multi-wing Harmoniums Model

In many applications, the input to the model does not have to be from a single source and/or of a homogeneous data type. For example, in the analysis of typical multi-media application video streaming, the input from video clips contains much relevant information, such as transcript texts, pictures, sounds and motion vectors. Assuming that all these inputs are combined together to present one topic center, it

will be natural to model the shared topic center using a set of hidden units, and to group observations from all sources into multiple homogeneous arrays of input units, each corresponding to a single source. Thus a multi-wing Harmoniums model is constructed, as shown in Fig. 2. This model consists of three canonical Harmoniums joint by a shared array of hidden units. It's straightforward to construct a multi-wing Harmoniums model from a canonical Harmoniums model. For example, a three-wing Harmoniums model added by two input sets $Y=\{y_n\}$ and $Z=\{z_m\}$ can be related to H via $p(y|h)=\prod_n p(y_n|h)$ and $p(z|h)=\prod_m p(z_m|h)$, where

$$p(y_n | h) = \exp \left\{ \sum_c \hat{\gamma}_{nc} e_{nc}(y_n) - C_n(\hat{\gamma}_{nc}) \right\} \quad (5)$$

$$p(z_m | h) = \exp \left\{ \sum_d \hat{\eta}_{md} s_{kd}(z_m) - D_m(\hat{\eta}_{md}) \right\} \quad (6)$$

$$\hat{\gamma}_{nc} = \gamma_{nc} + \sum_{jb} U_{nc}^{jb} g_{jb}(h_j) \quad (7)$$

$$\hat{\eta}_{md} = \eta_{md} + \sum_{jb} V_{md}^{jb} g_{jb}(h_j) \quad (8)$$

Together with (2), and slightly modified (3) that takes into account the influences from X , Y , and Z by loading the parameter $\hat{\lambda}$ with additional shift

$$\hat{\lambda}_{jb} = \lambda_{jb} + \sum_{ia} W_{ia}^{jb} f_{ia}(x_i) + \sum_{nc} U_{nc}^{jb} e_{nc}(y_n) + \sum_{md} V_{md}^{jb} s_{md}(z_m),$$

where $\{U_{nc}^{jb}\}$ and $\{V_{md}^{jb}\}$ stand for the matching parameters between the hidden unit and the input sets Y and Z . Thus, we can get the random field of three-wing exponential family Harmoniums.

$$p(x, y, z, h) \propto \exp \left\{ \sum_{ia} \theta_{ia} f_{ia}(x_i) + \sum_{nc} \gamma_{nc} e_{nc}(y_n) + \sum_{md} \eta_{md} s_{md}(z_m) + \sum_{jb} \lambda_{jb} g_{jb}(h_k) + \sum_{ijab} W_{ia}^{jb} f_{ia}(x_i) g_{jb}(h_k) + \sum_{njcb} U_{nc}^{jb} e_{nc}(y_n) g_{jb}(h_k) + \sum_{mjdb} V_{md}^{jb} s_{md}(z_m) g_{jb}(h_k) \right\} \quad (9)$$

where $\{f_{ia}(\cdot)\}$, $\{e_{nc}(\cdot)\}$, $\{s_{md}(\cdot)\}$, $\{g_{jb}(\cdot)\}$ stands for the fully statistical characteristics of x_i , y_n , z_m , h_k . This construction maintains the conditional independence between hidden variables given inputs and hence ensures the efficiency of inference (once the model is parameterized). Note that x , y , z are marginally dependent to each other, as can be quickly verified from the bipartite graph structure. This enables the application of inferring the features of one source from another source, for example, automatic image annotation which attempts to infer related words from a given image.

C. The multi-wing Harmoniums model of news video

To model video streams, which contain both text and image information, in the following, we outline a news video multi-wing harmonium (MWH) model based on a text submodel and two image submodels using the modular constructive technique described above.

The parameters of this model are defined as follows:

- The story unit of news video s denotes a four-dimensional tuple of (x, y, z, h) , which respectively denotes the keywords, color features and texture features of key-frame, and latent semantic topics.
- The vector $x=(x_1, x_2, \dots, x_I)$ denotes the keyword feature extracted from the transcript associated with the shot. Here I is the size of the key words dictionary, and $x_i \in \{0,1\}$ indicates the absence or presence of the keyword i in the story of news video.
- The vector $y=(y_1, y_2, \dots, y_N)$ denotes the boundary histogram feature of the key frame in the story of news video. Similar to the color feature, each key frame is divided into rectangular regions of size N in fixed size. $y_n \in R^c$ presents the color histogram feature of the region of size N denoted by C -dimensional vector. So y is a stack vector whose length is CN .
- The vector $z=(z_1, z_2, \dots, z_N)$ denotes the color histogram feature of the key frame in the story of news video. Each key frame is evenly divided into rectangular regions of size N in fixed size. $z_n \in R^c$ presents the color histogram feature of the region of size N denoted by D -dimensional vector. So z is also a stack vector whose length is DN .
- The vector $h=(h_1, h_2, \dots, h_J)$ denotes the latent semantic topics of the story of the news video. J is the number of latent topic, $h_j \in R$ denotes the degree of correlation between the news story and the latent topic of the size j .

1) Text feature model

Following the traditional bag-of-word model [11] for texts, we model the video transcript texts by adopting the word-count Poisson distribution model of the document. Instead of using a continuous surrogate of the discrete counts (as done in a mixture of Gaussian setting), or assuming that the counts of words are accumulated from independent draws from multinomial distributions (as done in LDA), the text Poisson model is based on the hypothesis that the rate of the word in a document can be described as the word's accumulation of the Poisson distribution in the dictionary, namely the latent topic features associated with each document directly determine the expected rate of each word in a document. In this way, a Poisson distribution is assigned to the observation counts of each word in a document. This text model has key differences from a multinomial model, which is achieved directly by topic mixing in the distribution of word rates in the documents combining specific topic feature, rather than via an additive effect of multiple single topic draws of the same word or via marginalization of the latent topic of each word. In the conditional Poisson model for word counts, topic mixing is still stable and robust even when a word appears only once or a few times in a document, which is typical in video captions though. Whereas in the multinomial word model, single word i can only come from a single topic and is thus unable to satisfy the topic mixing directly. The text Poisson submodel is defined as follows: For each word $i=\{1, \dots, I\}$, its rate x_i is distributed as:

$$p(x_i | h) = \text{Poisson}(x_i | \exp(\alpha_i + \sum_j h_j W_{ij})) \quad (10)$$

This model shows that, each key word of the transcript text in the story of news video relies on a Poisson distribution of the latent semantic topic h . In other words, the probability of a key word's appearance is determined by the weighted array of latent topic feature h . The parameters α_i and w_{ij} are scalar variables. $\alpha = (\alpha_1, \dots, \alpha_i)$ is a I-dimension vector. $W = [W_{ij}]$ is a matrix of $I \times J$. Because of the conditional independence between x_i and h , there is $p(x|h) = \prod_i p(x_i|h)$.

2) Image feature model

As to the image feature of the story in news video, we denote the feature by adopting the 72-dimension feature of color histogram and the 80-dimension of edge histogram in the partition HSV of key frame image in video stories. The color histogram feature y_n of the zone N in key frame image and the edge histogram feature z_n separately satisfies the distribution of conditional multivariate Gaussian distribution.

$$p(y_n | h) = N(y_n | \sigma_n^2 (\beta_n + \sum_j h_j U_{nj}), \sigma_n^2) \quad (11)$$

$$p(z_n | h) = N(z_n | \mu_n^2 (\tau_n + \sum_j h_j V_{nj}), \mu_n^2) \quad (12)$$

where β_n and U_{nj} are 72-dimension vectors. $\beta = (\beta_1, \dots, \beta_n)$ is a stack vector of $72 \times N$. $U = [U_{nj}]$ is a matrix of $72 \times N \times J$, and σ_n^2 is a 72×72 -dimension covariance matrix. τ_n and V_{nj} are 80-dimension vectors, $\tau = (\tau_1, \dots, \tau_n)$ is a stack vector of $80 \times N$. $V = [V_{nj}]$ is a matrix of $80 \times N \times J$, and μ_n^2 is a 80×80 -dimension covariance matrix. We adopt unit matrix to simplify operation. For the conditional independence between y_n , z_n and h , there are $p(y|h) = \prod_n p(y_n|h)$ and $p(z|h) = \prod_n p(z_n|h)$.

3) Hidden semantic topic model

As to the hidden unit h of latent topic feature in the news video story unit, we assume that each feature is a conditional unit-variance Gaussian distribution, whose mean is determined by a weighted combination of the key word feature in the video news story unit's transcript text, the color histogram and the edge histogram.

$$p(h_j | x, y, z) = N\left(h_j | \sum_i W_{ij} x_i + \sum_n U_{nj} y_n + \sum_m V_{mj} z_m, 1\right) \quad (13)$$

where W_{ij} , U_{nj} and V_{mj} share the same parameters with (10), (11) and (12). Similarly there is $p(x|h) = \prod_i p(x_i|h)$ from the conditional independence. This model denotes the topic vector as a random point in Euclidean space, while other feature models based on polynomial denote their topic joint vector as a point in the space of single feature. The condition distribution of all vectors in the model is shown above. These local conditional distributions can map into the random filed of Harmoniums shown as follows.

$$p(x, y, z, h) \propto \exp\left\{\sum_i \alpha_i x_i + \sum_n \beta_n y_n + \sum_n \tau_n z_n - \sum_n \frac{y_n^2}{2} - \sum_n \frac{z_n^2}{2} - \sum_j \frac{h_j^2}{2} + \sum_{ij} W_{ij} x_i h_j + \sum_{nj} U_{nj} y_n h_j + \sum_{nj} V_{nj} z_n h_j\right\} \quad (14)$$

By integrating out the hidden variables h in (14), we obtain the marginal distribution over the observed keyword and color features in the stories of news video.

$$p(x, y, z) \propto \exp\left\{\sum_i \alpha_i x_i + \sum_n \beta_n y_n + \sum_n \tau_n z_n - \sum_n \frac{y_n^2}{2} - \sum_n \frac{z_n^2}{2} + \frac{1}{2} \sum_j (\sum_i W_{ij} x_i + \sum_n U_{nj} y_n + \sum_n V_{nj} z_n)^2\right\} \quad (15)$$

which also contains a hidden partition function in this distribution.

The parameter of the NVMWH model, $s = (\alpha, \beta, \tau, W, U, V)$, is learnt by the maximum likelihood of a news video story set, where the likelihood function is defined by (15). Due to the presence of the global partition function, the learning process requires approximate inference methods, which will be discussed in the next section. We define the variance of the latent variables given the input variables to one in order to simplify the parameter estimation. Introducing a covariance matrix Σ can offer additional freedom for joint distribution $p(h_i | x_i, y_n, z_n)$, but it would not lead to more general representations in terms of probability $p(x, y, z)$ [10].

From the analyses above, we can find that the NVMWH model can denote the latent semantic topic of the story units in news video. In other words, the NVMWH model can be used to infer the latent semantic topic h if given the text feature x , the visual feature y and z of key frame image of the story unit in the news video.

D. The learning of the model's parameter

By the analysis in the section above, we can find that the NVMWH model can be used to gain the hidden semantic topic of the story unit in news video but the parameters of the model have to be determined before using this model. There are many methods to estimate the model parameters. We adopt the maximum likelihood method according to the NVMWH model defined in the section above. Assuming that the training set contain N independent identically distributed (IID) story units, namely $\{x, y, z\} = \{x_n, y_n, z_n, n=1, \dots, N\}$, the parameter of the NVMWH model $s = (\alpha, \beta, \tau, W, U, V)$ is estimated by maximizing the log-likelihood of the data defined by (15). Due to the complexity of this model, there is no closed-form solution to the maximization problem and we have to resort to an iterative method like gradient ascent. The learning rules (i.e., the gradients) can be obtained by setting the derivatives of (15) with respect to model parameters:

$$\begin{aligned} \delta \alpha_i &= \langle x_i \rangle_{\tilde{p}} - \langle x_i \rangle_p, \delta \beta_n = \langle y_n \rangle_{\tilde{p}} - \langle y_n \rangle_p, \\ \delta z_n &= \langle z_n \rangle_{\tilde{p}} - \langle z_n \rangle_p, \delta W_{ij} = \langle x_i h_j' \rangle_{\tilde{p}} - \langle x_i h_j' \rangle_p, \\ \delta U_{nj} &= \langle y_n h_j' \rangle_{\tilde{p}} - \langle y_n h_j' \rangle_p, \delta V_{nj} = \langle z_n h_j' \rangle_{\tilde{p}} - \langle z_n h_j' \rangle_p \end{aligned} \quad (16)$$

where $h_j' = \sum_i W_{ij} x_i + \sum_n U_{nj} y_n + \sum_n V_{nj} z_n$, $\langle \cdot \rangle_{\tilde{p}}$ and $\langle \cdot \rangle_p$ denote expectation under empirical distribution (i.e., data average) or model distribution of the harmonium, respectively. Like other undirected graph models, there is global normalizer (a.k.a partition function) in the likelihood function of the

NVMWH model, so it's very difficult to directly calculate $\langle \cdot \rangle_p$. Instead, we need approximate inference methods to estimate these model expectations $\langle \cdot \rangle_p$. We explored three approximate inference methods in our work, which are briefly discussed below.

1) The mean field approximation

Mean field (MF) is a variational method that approximates the model distribution p through a factorized form as a product of marginals over clusters of variables [12]. We use the naive version of MF, where the joint probability p is approximated by a surrogate distribution q as a product of singleton marginals over the variables:

$$q(x, y, z, h) = \prod_i q(x_i | t_i) \prod_n q(y_n | \zeta_n, \sigma_n) \prod_n q(z_n | \zeta_n, \mu_n) \prod_j q(h_j | \xi_j) \quad (17)$$

where the singleton marginals are defined as $q(x_i) \sim \text{Poisson}(t_i)$, $q(y_n) \sim N(\zeta_n, \sigma_n)$, $q(z_n) \sim N(\zeta_n, \mu_n)$ and $q(h_j) \sim N(\xi_j, 1)$, and $\{t_i, \zeta_n, \mu_n, \xi_j\}$ is variational parameter.

The variation parameters can be computed by minimizing the KL-divergence between p and q , which results in the following fixed-point updating equations:

$$t_i = \exp(\alpha_i + \sum_j W_{ij} \xi_j) \quad (18)$$

$$\zeta_n = \sigma_n^2 (\beta_n + \sum_j U_{nj} \xi_j) \quad (19)$$

$$\zeta_n = \mu_n^2 (\tau_n + \sum_j V_{nj} \xi_j) \quad (20)$$

$$\xi_j = \sum_i W_{ij} t_i + \sum_n U_{nj} \zeta_n + \sum_n V_{nj} \zeta_n \quad (21)$$

We iteratively update the variational parameters using the above fixed-point equations until they converge, and then the surrogate distribution q is fully specified. We replace the intractable model expectations $\langle \cdot \rangle_p$ with $\langle \cdot \rangle_q$ in (16), which are easy to compute from the fully factorized surrogate distribution q . Then, we can update the model parameters using the learning rules defined in (16). Besides, when using the gradient ascent method in the mean field, the learning process includes two nested loops: the outer loop iteratively updates the model parameters using the learning rules (16), while the inner loop iteratively updates the variational parameters in order to approximate the model expectations in the learning rules. Whenever the model parameters are updated (and so are the model distribution p), the whole inner loop needs to be executed to recompute the surrogate distribution q to approximate the updated model distribution p .

2) Gibbs sampling

Gibbs sampling, as a special form of the Markov chain Monte Carlo (MCMC) method, has been used widely for approximate inference in complex graphical models [13] [14]. This method repeatedly samples variables in a particular order, with one variable at a time and conditioned on the current values of the other variables. If the iteration number is big enough, the sampling of joint distribution and the boundary distribution is gained successively. For example, in the Poisson text submodel mentioned in this

paper, the sampling order is defined as $x_1, \dots, x_j, h_1, \dots, h_j$ and other variables are defined as inputs. First, $\{h_j\}$ is set as the current value and each x_i is sampled from the condition distribution defined in (10). Then, x_i is set as the current value and each h_j is sampled from the condition distribution defined in (13), and repeat this process iteratively. After a large number of iterations ("burn-in" period), this procedure guarantees to reach an equilibrium distribution that in theory is equal to the model distribution p . Therefore, we use the empirical expectation computed using the samples collected after the burn-in period to approximate the true expectation $\langle \cdot \rangle_p$. The number of "burn-in" iterations and samples is at least thousands and typically around tens of thousands.

3) Contrastive divergence

An alternative to exact gradient ascent search based on the learning rules in (16) is the contrastive divergence (CD) algorithm [15] that approximates the gradient learning rules. In each step of the gradient update, instead of computing the model expectation $\langle \cdot \rangle_p$, CD starts from the empirical values as the initial samples, runs the Gibbs sampling for up to only a few iterations and uses the resulting distribution q to approximate the model distribution p . It has been proved that the final values of the parameters by this kind of updating will converge to the maximum likelihood estimation. In our implementation, we compute $\langle \cdot \rangle_q$ from a large number of samples obtained by running only one step of Gibbs sampling with different initializations. Straightforwardly, CD is significantly more efficient than the Gibbs sampling method since the "burn-in" process is skipped.

III. EXPERIMENTS AND DISCUSSIONS

To verify the effectiveness of the NVMWH model, our experiments mainly include two parts. First, we show some illustrative examples of the latent semantic topics derived by the proposed models and discuss the insights they provide about the structure and relationships of video categories. In the second part, we evaluate the performance of our models in video classification in comparison with some of the existing approaches.

A. Experimental data and features selection

The experimental data adopted in our experiment come from the news video stories from CCTV news. We collect news programs of two years and four months, from May of 2006 to July of 2008, which contain 25776 news story units. Each story unit of the news video is considered as a document or a training test example. Our experiment adopts 4214 news story units which belong to 18 categories, namely fire disaster, flood, earthquake, storm (typhoon and hurricane), Olympic games, bird flu, Taiwan, the Korean nuclear issue, the United Nations, the United States, Russian, Japan, Iran, Iraq, terrorist attack, country, oil price and football. Each story unit is related to a category. Because the CCTV news shows only includes various important news, the distribution is uneven of the story unit in each category. The number of every kind of story unit in this experiment ranges from 26 to 828. 30% of the samples of each category

are randomly selected as training samples and the rest are considered as the test samples. Table 1 describes the training and the test samples of the experiment in detail.

TABLE I. THE SAMPLE DISTRIBUTION OF TRAINING SET AND TEST SET

Serial number	Category name	Total number of samples	Number of training samples	Number of test samples
1	fire disaster	80	24	56
2	flood	61	18	43
3	earthquake	627	188	439
4	storm	106	32	74
5	Olympic games	828	248	580
6	bird flu	26	8	18
7	Taiwan	113	34	79
8	the Korean nuclear issue	38	11	27
9	the United Nations	180	54	126
10	the United States	363	109	254
11	Russian	259	78	181
12	Japan	276	83	193
13	Iran	244	73	171
14	Iraq	142	43	99
15	terrorist attack	85	26	60
16	country	643	193	450
17	oil price	114	34	80
18	football	29	9	20
Total		4214	1264	2950

As to the text feature, we download all the transcript texts of all the news story units from the CCTV website. We adopt the word segmentation software of Beijing Language and Culture University to conduct word segmentation, and leave out all stopwords, and merger synonyms and near-synonyms. About nine thousand key words are gained from the 18 topics above. To further reduce the complexity of the model operation, we ignore the low-frequency words whose frequencies are less than 6 and extract 3182 keywords as the text feature. Hence, the text feature of each story unit in the news video is denoted as a 3182-dimension binary feature vector where 1 stands for the appearance of a certain keyword in the story unit and 0 stands for no appearance. As to the visual feature, we adopt the 72-dimension HSV color histogram feature and the 80-dimension edge histogram feature of the key frame image in MPEG-7.

By default, NVMWH is trained via contrastive divergence with up to 1000 steps of gradient ascent. We adopt the mean field approximation and the Gibbs sampling to conduct the model training.

To mitigate the issue of “identifiability” [10] that allows multiple parameters to share the same marginal likelihood, the initial estimations of parameters W , U and V in the NVMWH were determined by a SVD on the design matrix of text/images features over shots. We do not strongly emphasize this issue because in our analysis NVMWH is not mainly used to directly capture the exact semantics of the latent factors underlying the data space. In order to achieve semantically more accurate and informative latent factor representations, we can apply a subsequent clustering

procedure on the lower-dimensional representations provided by NVMWH.

The parameters of GM-Mix and GM-LDA were obtained using EM. We infer the latent topic captured by GM-Mix using the conditional probabilities of hidden variables $p(h|x,z)$ and those by GM-LDA based on variational Dirichlet posteriors of the topic weights.

B. The results and analysis of the latent topic mining experiment



Figure 3. Examples of latent topics

The NVMWH model can automatically discover low-dimension meaningful latent topics from the high-dimension multi-modal features of the texts and images in the news videos. We illustrate 8 latent topics out of the 18 topics learned by NVMWH in Fig 3. Each topic is described by the first 8 keywords and the first 5 key frame images related to the video shots. These keywords and key frame images possess highest condition probability in the latent topics. It is shown that the first 6 topics correspond to the scenes of oil price, forest fire, flood, bird flu, the Korean nuclear and the earthquake. They are a cluster based on the transcript texts and images. The last two topics illustrate some interesting patterns discovered by NVMWH. At the first sight, these key frames of the shots show great differences in visual features.

It seems to present several totally different topics for the different scenes like helicopter, stadium, meeting, shore and meteorological chart. However, by examining the transcript texts of the news story units, we find that their semantic topics share some common aspects. Several news stories in topic 7 all mention the similar or same words, such as typhoon, attack and weather which are words related to the topic typhoon. Similarly, several news stories in topic 8 also mention some related words like Olympic Games, stadium and security work. Obviously, NVMWH model discovers the last two topics mainly based on the similarity between the key words of the transcript texts in the videos, while the visual features functions a lot in discovering other topics.

C. The results and analysis of the Classification performance experiment of NVMWH model

In order to show the predictive power of the low-dimension latent semantic topic produced by NVMWH, we adopt classification, which is the most important task in multimedia analysis and application, to evaluate the performance of the NVMWH. In the experiment, the dimension of the latent semantic topic is set as less than 50 and its original feature dimension is set as 3182-dimension. The color histogram feature of the key frame image is set as 72-dimension and the edge histogram feature is set as 80-dimension.

First, we evaluate the performance of NVMWH on classifying testing examples into one of the predefined categories, and compare this method with LSI, GM-Mix and GM-LDA. For each algorithm, the parameters are estimated using all data, ignoring their true class labels. Once the models are learned, we use them to project every example into a lower-dimensional latent topic space. Then we split the data into a training set and a testing set as shown in Table 1, use the SVM^{Light} package to learn a support vector machine (SVM) on the training data, and predict on the testing data.

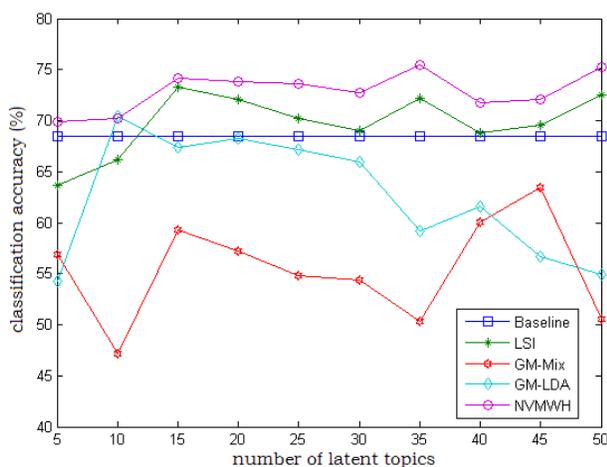


Figure 4. The classification accuracy of every model at different topic dimension

Fig. 4 shows the classification accuracy of five different models at different latent topic feature dimension ranging

from 5 to 50, and the dimension is the number of the latent semantic topics. The Baseline method retains the available feature classification results of all the variables in training and test. We find that the NVMWH model can always achieve higher classification accuracy than the Baseline even with a large dimensionality reduction. Under the same topic dimension, it also outperforms LSI with a good margin. We believe that this may be partially explained by the arguably better assumptions adopted by NVMWH on modeling text/image features. Surprisingly, GM-Mix produces a considerably worse performance than the baseline because the modeling power of GM-Mix is too limited to capture multiple latent topics for each text/image pair. Too much information is eliminated in GM-Mix's representations because the posterior distribution is usually peaked at one latent topic. Compared to GM-Mix, GM-LDA offers more flexibilities in modeling associated text/images and indeed it is (slightly) superior to all other models when latent aspect dimension is set to be 10. But it appears that GM-LDA may have suffered from overfitting or a low-dimensionality bias as its error curve rises significantly in higher-dimensional latent space. In contrast, we observe that the performances of LSI and NVMWH are relatively stable over a wide range of dimensions, which may reflect the robustness and expressiveness of their representation schemes for the latent aspects (i.e., as Gaussian variables rather Dirichlet variables).

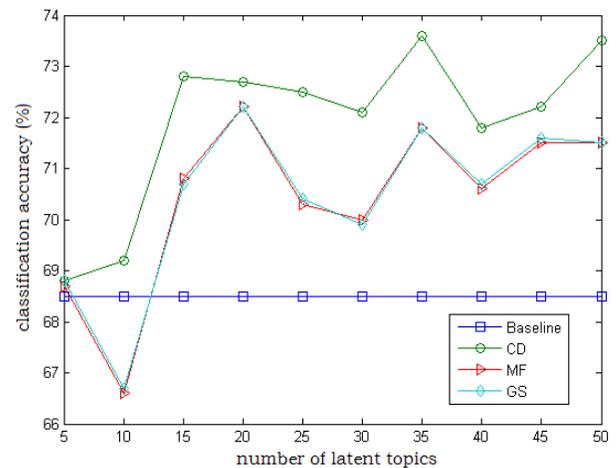


Figure 5. The classification accuracy of different learning algorithms in the NVMWH model

Fig. 5 presents the comparison of the performance between mean field, Gibbs sampling and contrastive divergence. We discover that mean field and the Gibbs sampling are similar in performance, while the accuracy of contrastive divergence is slightly better than the two. It's mainly because the latter approach uses a fully factorized distribution to approximate the true distribution whereas the former uses a Monte Carlo approximation. Meanwhile, we make a study of the efficiency of the three methods by examining the time taken to reach the convergence of the learning methods during the training. The results show that

the mean filed possess the highest efficiency at 2 minutes, the contrastive divergence method ranks second at 10 minutes and the Gibbs sampling has the lowest efficiency for about 50 minutes. Therefore, it is better to adopt the contrastive divergence method to conduct the learning and inference of the model parameters for the comprehensive consideration of accuracy and efficiency.

On the same news subject, why does not the classification performance increase steadily corresponding to the number of the latent topics? That is because the news subject is determined by the latent topics with high probability distribution, and the latent topics after certain dimension cannot help express the content of the subject of news. That is to say, the latent topics with low probability distribution have few associations, or even no associations, with the news subject. Thus, the classification performance would not increase with increasing number of latent topics.

IV. CONCLUSION AND FUTURE WORK

We make a thorough study of the latent semantic topic mining of news video by adopting the multi-wing two-layer undirected graphical model. First, we construct a news video multi-wing Harmoniums model of multi-modal heterogeneous features based on the basic Harmoniums model, the transcript texts and the key frame images in the news video. In this model, the multivariate Gaussian variables denoted the latent topics. The condition distribution of specific features models on the inputs of various kinds of data resources, namely a multiple Poisson distribution is used to model the text features and two multivariate Gaussian models respectively denote features of color histogram and of the edge histogram in the key frame image of news video story. The probability distributions are determined by all the topics so that a better topic mixing is achieved. It expands and improves the previous random field model, which is based on two layers by the bidirectional dependence relationship between the latent topics and the observed input data, whose performance is especially shown in the aspects of promoting effective reasoning, robust topic mixing and flexible latent topic modeling. The experiments of the latent semantic topic extraction and the prediction performance based on the NVMWH model prove the strong presentation ability and robustness of NVMWH model on latent semantic topic mining in news video stories.

In the future work, we will adopt more audio-visual features like facial feature, voiceprint feature, etc. in NVMWH model, and study a more effective learning algorithm for model's parameters, so as to improve the mining precision and efficiency of video semantic topics and apply semantic features to the mining and analysis of intelligence in public news videos.

REFERENCES

- [1] C. Snoek, M. Worring, and A. Smeulders. "Early versus late fusion in semantic video analysis," Proceedings of 13th ACM International Conference on Multimedia (MM 2005) in Singapore, ACM Press, Nov. 2005, pp. 399-402.
- [2] I. T.Jolliffe. "Principal component analysis," 2nd ed., Springer Press, 2002.
- [3] A. Hyvarinen, J. Karhunen, and E. Oja. "Independent component analysis," New York, USA: wiley, 2001.
- [4] L. Devroye, L. Györfi, and G. Lugosi, "A probabilistic theory of pattern recognition," Springer Press, 1996, pp. 46-47.
- [5] S. C. Deerwester, S. T. Dumais, and T. K. Landauer. "Indexing by latent semantic analysis," Journal of the American Society of Information Science, Sep. 1990, 41(6) pp. 391-407.
- [6] T. Hofmann. "Probabilistic latent semantic analysis," Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, Jul. 1999, pp. 289-296.
- [7] D. M. Blei, A. Y. Ng, and M. I. Jordan. "Latent dirichlet allocation," Journal of Machine Learning Research, MIT Press, Mar. 2003, pp. 993-1022.
- [8] D. M. Blei and M. I. Jordan. "Modeling annotated data," Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, Jan. 2003, pp. 127-134.
- [9] P. Smolensky. "Information processing in dynamical system: foundations of harmony theory. Parallel distributed processing: explorations in the microstructure of cognition foundations", Cambridge: MIT Press. 1986, pp. 194-281.
- [10] M. Welling, M. Rosen-Zvi, and G. Hinton. "Exponential family Harmoniums with an application to information retrieval," In Advance Neural Information Processing Systems, Dec. 2005, pp. 1481-1488.
- [11] D. Metzler. "Beyond bags of words: effectively modeling dependence and features in information retrieval," SIGIR Forum, Dec. 2008, 42(1), pp. 77-77.
- [12] E. Xing, M. Jordan, and S. Russell. "A generalized mean field algorithm for variational inference in exponential families," In uncertainty in artificial intelligence (UAI2003). Morgan Kaufmann Publishers, Aug. 2003, pp. 583-591.
- [13] W. R. Gilks, S. Ripley, and D. J. Spiegelhalter. "Markov chain Monte Carlo in practice," Cambridge, UK: Chapman & Hall/CRC Press, 1996.
- [14] J. R. Smith and S. F. Chang. "Visually searching the web for content," IEEE Multimedia Magazine, Jul.-Sep. 1997, 4(3), pp. 12-20.
- [15] M. Welling and G. E. Hinton. "A new learning algorithm for mean field Boltzmann machines," Proceedings of the International Conference on Artificial Neural Networks, London, UK, Springer-Verlag, Aug. 2002, pp. 351-357.
- [16] E. Xing, R. Yan, and A. Hauptmann. "Mining associated text and images with dual-wing harmoniums," Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence, Jul. 2005, pp. 633-641.

Video Object Detection by Classification Using String Kernels

Wan-Hsuan Yu, Chi-Han Chuang

Department of Computer Science and Engineering
National Taiwan Ocean University
Keelung, Taiwan
e-mail: yuwanhsuan@gmail.com

Shyi-Chyi Cheng

Department of Computer Science and Engineering
National Taiwan Ocean University
Keelung, Taiwan
e-mail: csc@mail.ntou.edu.tw

Abstract—Video object detection is one of the most important research problems for video event detection, indexing, and retrieval. For a variety of applications such as video surveillance and event annotation, the spatial-temporal boundaries between video objects are required for annotating visual content with high-level semantics. In this paper, we define spatial-temporal sampling as a unified process of extracting video objects and computing their spatial-temporal boundaries using a learnt video object model. We first provide a learning approach to build a class-specific video object model from a set of training video clips. Then the learnt model is used to locate the video objects with precise spatial-temporal boundaries from a test video clip using graph kernels. A frame sorting process as a preprocessing is also proposed to transform the graph, modeling the shot configuration of a video clip, into a string of shots. Thus, the computation of graph kernels is simplified to be string kernels. The string kernels for support vector machine (SVM) classification are finally adopted to train the SVM classifiers from a set of training samples and detect the video objects in a test video clip by classification. A human action detection and recognition system is finally constructed to verify the performance of the proposed method. Experimental results show that the proposed method gives good performance on several publicly available datasets in terms of detection accuracy and recognition rate.

Keywords—video objects; string kernels; dynamic programming; video object modeling; SVM classification.

I. INTRODUCTION

Video object detection (VOD) is the primary step to semantically annotate a video sequence in semantic video database indexing and retrieval, intelligent video surveillance, and advanced man-machine interfaces [1,2]. Early works in video object detection focused on detecting and recognizing the scene and objects shown in a representative key-frame of a video shot, thus the temporal information of video objects is lost [3,4]. Recently, semantic-based video analysis tended to model a video clip as a graph whose nodes are high-level video objects performing a specific action individually [5]. Techniques of graph matching are then applied to annotate the event type of the input video clip [6]. Detection and classification of video objects from video clips help bridge the semantic gap between high-level features and low-level features and the construction of modern semantic-based video analysis.

Conventional VOD algorithms, which characterize objects as spatially cohesive with locally smooth trajectories, use these techniques for tracking or body pose estimation to extract spatial-temporal tubes from the input video clip [7-9]. However, using a tracking or body pose estimation in real world videos is generally

not reliable due to object occlusion, distortion and changes in lighting. Instead, we formulate the tracking process for VOD [7, 10] as a classification problem because objects are, in general, spatially and temporally cohesive. Also, by assuming relatively slow camera motions, the shape and location of these objects vary slowly from frame to frame. Thus, the size of the search space to track an object across many frames is reduced significantly by exploiting this coherence. By considering a parameter set in the feasible search space as a class, the object tracking for VOD casts into a classification framework [11].

A primary motivation for the work presented here is to question the benefits of tracking object boundaries across frames for video-based applications such as activity analysis. In practice, the accuracy of any boundary estimate is limited by a number of systemic factors such as image resolution, noise, motion skew, and the accuracy of the model. For example, formulating VOD as motion segmentation using optical flow rests on the assumption of brightness constancy, which is violated at moving boundaries, resulting in poor estimates of object contours [12]. For some applications, the object detection at each frame only needs to be known up to a limited precision, as long as good shape and trajectories are maintained.

In addition to segmentation, conventional VODs also try to detect and segment the observed motions into semantic meaningful instances of particular activities from videos [13,14]. To reach this goal, recent approaches consider the detection and recognition of the video object as an extension of 2D object detection [15,16] with higher dimensionality. Some well-known approaches include space-time interest-point detectors [17] and bag-of-words models [18]. These techniques aim at employing a combination of local space-time features and global 3D shape features to estimate the space-time boundaries of a given video object. Two issues which are therefore of particular importance are dealing with local patch sampling and exploring the rich relationships among spatial-temporal "words" inherited from objects [19].

Video object classification is the key step in high-level video-based applications. Conventional machine learning techniques are applied to train the state-of-the-art methods using a large, diverse set of manually annotated images. The typical level of annotation needed is a bounding-box for each object instance [15]. To ensure the performance of a detector, a large amount of annotated instances is generally needed [20]. Recently, object classification approaches borrowed from unlabeled or weakly annotated data have attracted much attention to reduce tedious manual annotation to a minimum [21]. However, training a detector without location annotation is very difficult and performance is still below fully supervised methods [22].

Recent approaches for video object detection follow the following steps: a target object is initialized by human annotation or with a preexisting detector in one frame, then a classifier is trained on-line to redetect the object in each frame [23]. A

significant limitation to these approaches is the trained classifier is that a video-specific detector but not a generic class detector. In contrast, Ali et al. [24] proposed a semi-supervised boosting variant that exploits temporal consistency of video frames to learn a complex appearance model from a subset of fully annotated frames in each training video for video object detection. Testing is performed on videos of the same scene, but at different time instances.

An image object often consists of several parts arranged in a deformable configuration [15]. The use of visual patterns of local patches in shape modeling is related to several ideas including the approach of local appearance codebooks [16] and the generalized Hough transform (GHT) [25] for object detection. At training time, these methods learn a model of the spatial occurrence distributions of local patches with respect to object centers. At testing time, based on the trained object class classifiers, the appearances of interest points in images or video are matched in the visual codebooks to detect a specific object using the voting framework of GHT. The effectiveness of visual pattern grouping by Hough voting is thus well verified.

In this paper, we formulate a video object as a graph of postures (key-objects) to model the temporally relationship between key-objects. The graph edit distance (GED) can then be used to measure the spatial-temporal content difference between two video objects. A frame sorting process [26] as a preprocessing is also used to transform the graph, modeling the shot configuration of a video clip, into a string of shots. Thus, the computation of graph kernels is simplified to be string kernels. The string kernels for support vector machine (SVM) classification are finally adopted to train the SVM classifier from a set of training samples and detect the video objects in a test video clip by classification. We also create a template video object for each class to achieve the goal of speeding up the VOD process. A human action detection and recognition system is finally constructed to verify the performance of the proposed method. Experimental results show that the proposed method gives good performance on several publicly available datasets in terms of detection accuracy and recognition rate.

II. PROBLEM DEFINITION

The proposed video object detection by classification using string kernels is inspired from the work of [4] but of very different implementation. Let $V = \{F_t\}_{t=1}^n$ and $\bar{O} = \{O_t\}_{t=1}^n$ be a video clip of n frames and the corresponding video object consisting of n 2D target objects, respectively. Suppose $\bar{x} = \{x_t(s_t)\}_{t=1}^n \in R^{D \times n}$ be the feature vectors for every location s_t to locate O_t in F_t , we want to build a classifier

$$\varphi: R^{D \times n} \rightarrow R \quad (1)$$

such that the set of locations $\bar{s} = \{s_t\}_{t=1}^n$

$$\{\bar{s} : \varphi(\bar{x}(\bar{s})) \geq 0\} \quad (2)$$

detects a visible target video object from video frames. Given a training data set that comprises N input vectors $\bar{x}_1, \dots, \bar{x}_N$, with corresponding target values y_1, \dots, y_N where $y_i \in \{-1, 1\}, i=1, \dots, N$. The support vector machines (SVMs) approach [27] finds the linear decision boundary $\varphi(\bar{x})$ as:

$$\varphi(\bar{x}) = w^T \phi(\bar{x}) + b \quad (3)$$

where ϕ denotes a fixed feature-space transformation, b is a bias parameter, so that, if the training data set is linearly separable,

$y_i \varphi(\bar{x}_i) > 0$ for all points. The maximum marginal solution of SVMs is found by solving for the optimal weight vector $\bar{\alpha} = (\alpha_1, \dots, \alpha_N)$ by maximizing

$$\tilde{L}(\bar{\alpha}) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \langle \phi(\bar{x}_i), \phi(\bar{x}_j) \rangle \quad (4)$$

with respect to $\bar{\alpha}$, that is subject to the constraints:

$$\sum_{i=1}^N \alpha_i y_i = 0, \quad \alpha_i \geq 0, \text{ for } i=1, \dots, N \quad (5)$$

$\langle \phi(\bar{x}_i), \phi(\bar{x}_j) \rangle$ is the inner product of \bar{x}_i and \bar{x}_j in the feature space.

The parameters w and b are then derived from the optimal $\bar{\alpha}$.

The computational cost of the inner product could be much reduced by introducing kernel functions to avoid explicitly perform the transformation $\langle \phi(\bar{x}_i), \phi(\bar{x}_j) \rangle$. If the kernel function k satisfies the Mercer condition, then there exists a feature space and a mapping function ϕ such that k acts as an inner product in the feature space [28]. In this work, we propose to use the string kernel, starting from a Gaussian kernel, which has been proved to be effective for event recognition [1]. The string kernel is defined as

$$k(\bar{x}, \bar{x}') = \exp(-d(\bar{x}, \bar{x}')) \quad (6)$$

where $d(\bar{x}, \bar{x}')$ is the distance between \bar{x} and \bar{x}' using the dynamic programming process retaining the spatial-temporal consistency of the targets.

A challenge of the problem is it might require a large training set which results in tedious human-labeled effort in training the classifier φ . In this work, we tackle this problem by using an initial hand-labeled training video, and by going back-and-forth between the optimization of the labels of non-labeled videos. Many approaches train object detectors from images without location annotation [8]. Although the outputs of these operators are not precise, they can provide the initial training video object for learning the classifier φ . In this case, the proposed learning algorithm can be performed automatically without any human-labeled effort.

Another challenge is the performance of the string-kernel approach degrades greatly when the input video clip contains repetitive behaviors. In this case, we first represent a video clip as a set of shots and then lexicographically sort these shots to obtain a compact and normalized string of postures. The complexity of string-kernel computation is thus reduced by representing a video clip as a shot sequence.

III. THE PROPOSED APPROACH

Figure 1 shows the block diagram of the system. A preprocessing to lexicographically sort video frames is first applied to temporally normalize the video frames. Then, a key-frame detection procedure is applied to detect key-frames from a normalized video sequence to achieve the goal of eliminating redundant frames. The system is divided into the training and detection phases, where both of them are based on the proposed SVM classification with string kernels.

A. The Training

Many various image analysis tasks have verified the effectiveness of presenting video frames using bag-of-words (BoW) [1]. A common BoW approach to model video class is to extract features from all video patches in all training video clips of a video class to learn the appearance variability of the class, which is modeled as a local appearance codebook consisting of multiple codewords,

where each of them is determined by the mean features of a video patch cluster. Based on this codebook, we could compute a histogram of codeword frequencies to represent a video frame by mapping every patch of the frame to a codeword. Thus, each frame is represented as a BoW histogram.

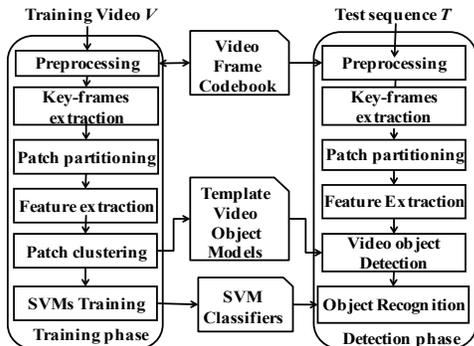


Figure 1. Block diagram of the proposed video object detection by classification using the Hough-voting approach.

In the preprocessing, the first step to temporally normalize the video frames in a video clip is to generate a video shot codebook through vector quantization of large sets of BoW histograms extracted from a collection of training video frames. The video frame codebook is generated by clustering the video frames in the feature space using k -means clustering algorithm and Euclidian distance as the clustering metric. The center of each resulting cluster is defined as a frame codeword. Let the video frame codebook FC have m cluster centers. Our approach uses FC to temporally normalize a video clip by grouping similar frames in which the temporal information is preserved. Given a video clip V of n BoW histograms, $h_i \in V, i = 1, \dots, n$, there is a collection of cluster assignment: $A = \{c_1, c_2, \dots, c_n\}$ where c_i is the cluster label indicating that cluster center i is the nearest neighbor of h_i in FC . By sorting A in lexicographical order, we can obtain $\tilde{A} = \{\tilde{c}_1^{\pi(1)}, \tilde{c}_2^{\pi(2)}, \dots, \tilde{c}_n^{\pi(n)}\}$ where $\tilde{c}_i^{\pi(i)}$ is the cluster label of the i -th video shot in \tilde{A} and $\pi(i)$ returns the index of frame i in V . The pair $(\tilde{c}_i^{\pi(i)}, \tilde{c}_j^{\pi(j)})$, $i < j$, belongs to \tilde{A} if and only if either $\tilde{c}_i^{\pi(i)} < \tilde{c}_j^{\pi(j)}$ or $(\tilde{c}_i^{\pi(i)} = \tilde{c}_j^{\pi(j)}) \wedge (\pi(i) < \pi(j))$. We finally permute the frames of V using \tilde{A} . The preprocessing step brings the system two obvious advantages: (1) similar frames are clustered to transform the repetitive activities into a single activity implicitly performed by the corresponding video object; (2) all video objects in the same class are starting from a common posture when we represent an activity as a sequence of postures.

A video shot detection procedure is then followed to separate a normalized video clip into multiple video shots, where each of them is represented as a key-frame. Finally, a video clip is represented as a sequence of key-frames. Let $\bar{O}_1 = \{\bar{o}_{1,j}\}_{j=1}^m$ denote the initial video object of m key-objects detected from corresponding key-frames of a training video clip in a class. For each key-object $\bar{o}_{1,i} \in \bar{O}_1$, we partition it into a set S_i of patches $P_j = (f, \vec{d}, s_p)$ where f is the feature vector characterized by a histogram of orientations (HOG) [29]; \vec{d} is the displacement vector defined from the patch center to key-object center; s_p is the size of the patch. As shown in Figure 2, the patch set S_i forms a GHT model and implicitly describes the

structure of $\bar{O}_{1,i}$ which can be used to detect similar objects from another image using the Hough-voting technique [15,16].

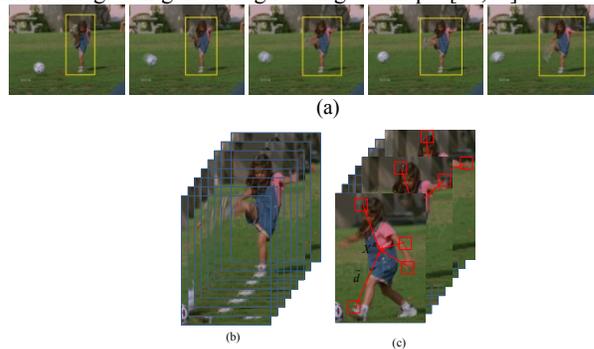


Figure 2. Representing a video clip by a sequence of key-frames: (a) detecting the key-frames and key-objects from a video clip; (b) piling up the normalized key-objects to form a 3D video object; (c) modeling (b) using a sequence of GHT models.

To achieve the goal of detecting the target object from an image I using the patch set S of key-object \bar{O} , we look after similar patches $P' \in S$ for each patch $P \in I$ located at (x_p, y_p) using the following distance function:

$$d(P, P') = 1 - \sum_i \sqrt{h_p(i)h_{p'}(i)} \quad (7)$$

where $h_p(i)$ and $h_{p'}(i)$ are the fractions of the i -th bin of the HOGs of P and P' , respectively. The local distance measurement for (P, P') should be added to the entry of the Hough-voting volume $H_I(x, y, s)$ at the image I :

$$H_I(x, y, s)^{(new)} = H_I(x, y, s)^{(old)} + (1 - d(P, P')) \quad (8)$$

where $s = s_p / s_{p'}$ is the ratio of sizes of P to P' and $(x, y) = (x_p, y_p) - s \times \vec{d}_{p'}$. Furthermore, a match pair of its similarity value less than a pre-defined threshold, i.e., 0.8, is excluded from casting a vote on the Hough-voting volume to avoid generating spurious peaks. Obviously, the peaks in H_I group patches in I into meaningful objects. The member patches to constitute a key-object can be found through performing the inverse Hough transform on the corresponding peak. Also, multiple peaks can be detected from H_I to locate multiple similar objects for the target object \bar{O} .

We also propose a parameter verification process to fine tune the location $\Lambda = (x, y)$ of the detected object in I . For each object, including the target and detected objects, we also construct a global HOG to characterize the shape of the object [15]. The distance between the detected and target objects can then be obtained by (7). The object O^* located at Λ^* is thus defined as

$$\Lambda^* = \arg \max_{\Lambda^* \in N(\Lambda)} [1 - d(o_{\Lambda^*}, \bar{o})] \quad (9)$$

where $N(\Lambda)$ returns all significant peaks from the neighborhood of Λ in H_I . Based on Λ^* , the system fine tunes the location of the detected object o_{Λ^*} in I . Moreover, the similarity between the detected and target objects is obtained. Although this process results in additional time for fine tuning the geometric transformation parameters, our experimental results show that it significantly improves the accuracy of object locations.

The core idea of our approach is to automatically compute labels for non-labeled samples belonging to the same class by minimizing the video object detection errors using dynamic

programming. The dynamic programming process (DPP) optimally aligns the initial (seed) video object $\bar{O}_1 = \{\bar{o}_{1,i}\}_{i=1}^m$ with the frames of the input video clip $V = \{F_i\}_{i=1}^n$ with the shortest distance. Let $A[i,j]$ denote the distance of the optimal alignment of $\bar{O}_1^{(i)} = (\bar{o}_{1,1}, \bar{o}_{1,2}, \dots, \bar{o}_{1,i})$ and $V_j = (F_1, F_2, \dots, F_j)$. The recurrent equation used to align $\bar{O}_1^{(i)}$ and V_j with the shortest distance in a bottom-up fashion by dynamic programming is

$$A[i, j] = \min(A[i-1, j-1], A[i, j-1], A[i-1, j]) + d(\bar{o}_{1,i}, o_{s_j}) \quad (10)$$

where o_{s_j} is the object detected from F_j at location s_j with the distance measurement $d(\bar{o}_{1,i}, o_{s_j})$ using (7). The goal of the recurrent equation is to find out the value of $A[m][n]$ which denotes the error to detect the video object from the input video clip using the seed video object. The initial condition for $A[i,j]$ is

$$A[i, j] = \begin{cases} 0 & \text{if } i = j = 0 \\ \infty & \text{if } i \neq 0 \wedge j = 0. \\ \infty & \text{if } i = 0 \wedge j \neq 0 \end{cases} \quad (11)$$

Given the set of detected video objects of a class, the SVM classifier ϕ with the string kernel defined in (6) is then trained to generate a new seed video object for further improving the detection and classification accuracy by an optimization loop. Let $C = \{V_i\}_{i=1}^N$ be set of training video clips and K be the maximal number of iterative loop. Given that initial video object $\bar{O}_1 = \{\bar{o}_{1,i}\}_{i=1}^m$, we define the proposed class-specific training (CST) algorithm as follows.

```

CST( $C, \bar{O}_1, K$ ) {
   $O_t \leftarrow \bar{O}_1$ 
  for  $k = 1$  to  $K$  do {
    for  $i = 1$  to  $|C|$  do  $\tilde{O}_i \leftarrow DPP(O_s, V_i)$ ;
     $\phi \leftarrow SVM\_Training(\{\tilde{O}_i\}_{i=1}^{|C|})$ ;
     $O_t \leftarrow \arg \max_{\tilde{O}_i, i=1, \dots, |C|} [\phi(\tilde{O}_i)]$ ;
  }
  return ( $O_t, \phi$ );
}

```

Finally, each class is represented as a template video object O_t and a SVM classifier ϕ . The former is used to detect a candidate video object from an input video clip using the proposed dynamic programming process. The classifier is then used to verify the correctness of the detected object.

B. The Detection

Given a test video clip, we first perform the same preprocessing procedure to temporally normalize the input video. The normalized video is also represented as a set of key-frames using the same key-frames detection in the training phase to reduce the time complexity of the successive video object detection using the Hough voting and dynamic programming. The detected video objects are then verified by the classifier ϕ .

The video object detection actually consists of two major steps: (1) detect the target video object O_V from the input video clip V based on the template video object of a class obtained in the training phase using the Hough voting and dynamic programming; (2) the class label of O_V is then defined to be

$$c(O_V) = \arg \max_{c \in C} \phi_c(O_V) \quad (12)$$

where $c(O_V)$ is the class label of the video object O_V ; C is the set of classes; ϕ_c is the SVM classifier of the class c .

Let $T = \{1, \dots, T\}$ be the set of time steps, and $\Omega = \{1, \dots, W\} \times \{1, \dots, H\}$ the set of locations, where W and H are the width and heights of the video frames. Given a classifier, the complexity of the video object detection by classification would be $O(W^T H^T)$ if we check all candidate video objects in a brute-force fashion. The time complexity of the proposed video object detection by dynamic programming is $O(T^2 W^2 H^2)$ which is much faster than the brute-forth approach.

IV. EXPERIMENTAL RESULTS

A series of experiments was conducted on an Intel PENTIUM Dual Processor 3.0GHz PC and three video datasets, the KTH dataset [30], the Weizmann dataset [31], and the UCF sports [32] are constructed to evaluate the performance of the human action detection and recognition system. The KTH video sequences have been used in many human action recognition studies. It contains six types of human actions: walking, jogging, running, boxing, hand waving and hand clapping. Each action is performed several times by 25 actors in four different scenarios: outdoor, outdoor with camera zooming, outdoor wearing different clothes, indoor. In total, there are 599 videos. The Weizmann dataset provides 90 video sequences of 9 actors performing 10 different actions. The UCF sports dataset is a collection of 150 broadcast sports sequences from network news videos and features ten different events: diving, golfing, kicking, weight-lifting, horseback-riding, running, skateboarding, swinging 1 (gymnastics, on the pommel horse and floor), swinging 2 (gymnastics, on the high and uneven bars), and walking. It is a very challenging dataset due to the camera motion and background clutter. These datasets have been used in many human action recognition studies.

The class labels, as the ground truth, for video sequences in the test datasets are used to determine the relevant matches in the test dataset to the query templates. Evaluations were done with a leave-one-out cross-validation. Classification results are shown in Table I and compared with state-of-the-art recognition systems [13, 18, 26, 43, 48-52]. The classification results provided in [13] include three variations of training and testing data: (A) training and testing on tracks generated from ground-truth annotations; (B) training on tracks from ground truth and testing on automatically extracted tracks; and (C) training and testing on automatically extracted tracks. The data variation C is used to construct the system. Table I shows that the classification accuracy of the method has better performance using the detected video objects as the input to class-specific SVM classifiers.

We follow the same localization evaluation rules in [13]: a detection is considered correct if, (1) the action object was correctly classified, and (2) the intersection-union ratio of the detection and ground truth bounding box is greater than 0.5. For the KTH and UCF datasets, selected frames were hand-annotated with bounding boxes, and the bounding boxes for the frames in between were generated by linear interpolation. For the UCF dataset, bounding boxes were provided as part of the ground truth annotation released with the data. Tables II and III show the performance comparison in localization accuracy using datasets KTH and UCF, respectively. All the compared methods perform well in action object detection and the proposed approach has the best performance in average detection accuracy. This illustrates the effectiveness of the GHT-based method in video action object detection.

TABLE I. CLASSIFICATION COMPARISON OF KTH, WEIZMANN, AND UCF WITH OTHER METHODS. ‘-’ MEANS THE DATA IS NOT PROVIDED IN THE ORIGINAL PAPERS.

Method	Weizmann	KTH	UCF
Proposed	100 %	95.2 %	83.4 %
Hough forest (A) [13]	97.8 %	93.5 %	86.6 %
Hough forest (B) [13]	95.6 %	92.0 %	81.6 %
Hough forest (C) [13]	92.2 %	93.0 %	79.0 %
Rodriguez et al. [32]	-	85.66%	69.2%
Wang et al. [33]	-	90.1%	81.6%
Yeffet & Wolf [34]	100%	90.1%	79.2%
Niebles et al. [18]	90 %	83.3 %	-
Schindler et al. [35]	90 %	92.7 %	-
Laptev et al. [36]	100%	91.8 %	-
Ommer et al. [21]	97.2 %	87.9 %	-

TABLE II. KTH LOCALIZATION RESULTS.

Method	Proposed	Hough Forest [13]	voc. Forest [37]
Precision			
Boxing	0.97	0.88	0.98
Hand Clapping	0.98	0.96	0.97
Jogging	0.90	0.84	0.79
Running	0.80	0.72	0.78
Walking	0.95	0.95	0.86
Hand Waving	0.98	0.98	0.96
Average	0.93	0.89	0.89

TABLE III. UCF LOCALIZATION RESULTS.

Classes	Precision	
	Proposed	Hough Forest [13]
Diving	0.62	0.52
Weight Lifting	1	1
Walking	0.70	0.67
Golfing	0.79	0.77
Skateboarding	0.41	0.39
Kicking	0.41	0.28
Running	0.43	0.37
Horseback Riding	0.78	0.66
Swing 1	0.46	0.44
Swing 2	0.32	0.26
Average	0.59	0.48

Figure 3 shows a result of human action detection and recognition using the proposed method. The system correctly detects and classifies the video object in a test video clip belonging to the class ‘‘Hand Waving’’ using the template video object and classifier of ‘‘Hand Waving’’. On the contrary, the voting results of matching the sampled patches of the test video clips to other template video objects on the Hough voting volume H will generate low responses. The peaks of H are obvious and easy to detect using a simple thresholding technique. As compared with conventional VOD methods, the system detects video objects belonging to a specific class. Non-meaningful video objects are discarded by the system.

V. CONCLUSION

In this paper we have presented a method for video object detection and recognition based on the fusion of template video object modeling and dynamic programming. The proposed template video object modeling encodes each class-specific template video object as a Hough model sequence. The dynamic programming framework is then used to optimally align the frames of an input test video sequence with the model sequences. The alignment results determine the positions of the corresponding video object in the test video sequence. The trained SVM classifiers are then used to annotate the type of the detected video object. An application to human action detection and recognition is

also constructed to verify the performance of the system. As compared with related GHT-based human action detection and recognition methods, the proposed method has the following contributions. First of all, this paper models the process of video object detection by the fusion of Hough voting and dynamic programming which is optimally retain the spatial-temporal information of a video object. Secondly, taking the detected video objects as the input, a training procedure with effort of human-made labeling to learn SVM classifiers with string kernels is discussed. The SVM classifiers estimate the possibility of a specific video object which performs a certain activity. In the test phase, the system detects and recognizes video objects from the

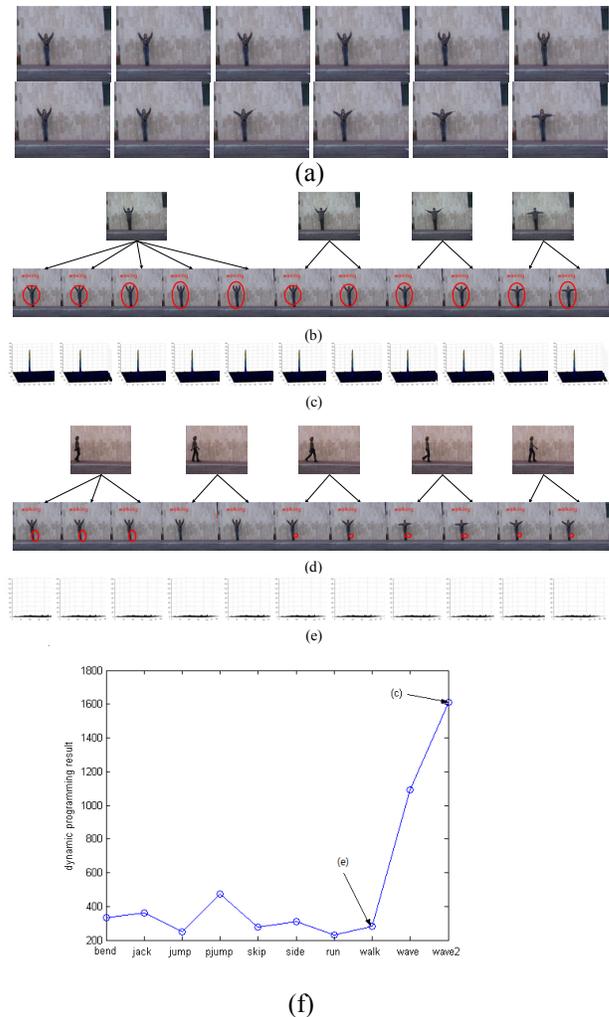


Figure 3. An example of human action detection and recognition using the proposed method on the dataset ‘‘Weizmann’’: (a) partial frames of a test video sequence belonging to the class ‘‘hand waving’’; (b) detection results of (a) using the template video object of ‘‘hand waving’’; (c) Hough voting results on the each frame of the test video sequence in (b); (d) detection results of (a) using the template video object of the class ‘‘walking’’; (e) Hough voting results on the each frame of the test video sequence in (d); (f) Hough voting results of (a) for classification.

input video clip automatically. Finally, the key-object representation is robust to temporal scaling in video object detection and recognition. Experimental results show that the proposed method gives good performance on several publicly

available datasets in terms of detection accuracy and recognition rate.

The proposed method suffers from the following limitations. The computational complexity of the approach using class-specific model matching by dynamic programming and GHT is essentially high. To implement the system on a parallel architecture, e.g., a GPU machine can solve the problem. Basically, GHT-based approaches can detect multiple objects from images or videos. However, the system based on its current implementation does not deal with the problem. Future work will deal with adding the detection of multiple video objects in a scene to the system, and increasing the database size.

ACKNOWLEDGMENT

This work was supported in part by National Science Council, Taiwan under Grant Numbers NSC 100-2221-E-019-054-MY3 and NSC 101-2918-1-019-003.

REFERENCES

- [1] L. Ballan, M. Bertini, A. D. Bimbo, L. Seidenari, and G. Serra, "Event detection and recognition for semantic annotation of video," *Multimedia Tools and Applications*, Vol. 51, No. 1, 2011, pp. 279-302.
- [2] G. Lavee, E. Rivlin, and M. Rudzsky, "Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video," *IEEE Trans. Systems, Man, and Cybernetics* Vol. 39, No. 5, 2009, pp. 485-504.
- [3] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, 2010, pp. 976-990.
- [4] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 32, No. 2, 2010, 288-302.
- [5] C. Xu, J. Cheng, Y. Zhang, Y. Zhang, H. Lu, "Sports video analysis: Semantics extraction, editorial content creation and adaptation," *Journal of Multimedia*, Vol. 4, No. 2, APRIL 2009, pp. 69-79.
- [6] T. Zhang, C. Xu, G. Zhu, S. Liu, H. Lu, "A generic framework for event detection in various video domains," in *Proc. ACM Multimedia*, 2010, pp. 103-112.
- [7] W. Brendel and S. Todorovic, "Video object segmentation by tracking regions," in *Proc. Int'l Conf. Computer Vision (ICCV)*, 2009.
- [8] T. Brox and J. Malik, "Large displacement optical flow: descriptor matching in variational motion estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (to appear)
- [9] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *Proc. ECCV*, 2010.
- [10] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Analysis & Machine Intelligence*, vol. 29, no. 12, 2007, pp. 2247-2253.
- [11] L. Shang, P. Jasiobedzki, and M. Greenspan, "Model-based tracking by classification in a tiny discrete pose space," *IEEE Trans. Pattern Analysis & Machine Intelligence*, vol. 29, no. 6, 2007, pp. 976-989.
- [12] M. Nicolescu and G. Medioni, "A voting-based computational framework for visual motion analysis and interpretation," *IEEE TPAMI*, vol. 27, no. 5, 2005, pp.739-752.
- [13] A. Yao, J. Gall, and L. V. Gool, "A Hough transform-based voting framework for action recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.
- [14] A. Oikonomopoulos, I. Patras, and M. Pantic, "An implicit spatiotemporal shape model for human activity localization and recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.
- [15] P. Felzenszwalb, R. Girshick, and J. McAllester, "Cascade object detection with deformable part models," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.
- [16] B. Leibe, A. Lenardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *International Journal of Computer Vision*, vol. 77, no. 1-3, 2008, pp. 259-289.
- [17] P. Scovanner, A. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition," in *Proc. ACM Int'l Conf. Multimedia*, 2007.
- [18] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," in *Proc. Int'l Computer Vision*, vol. 79, no. 3, 2008, pp. 299-318.
- [19] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: A statistical framework," *International Journal of Machine Learning and Cybernetics*, 1, 1, 2010, pp. 43-52.
- [20] S. Vijayanarasimhan and K. Grauman, "Large-scale live active learning: Training object detectors with crawled data and crowds," in *Proc. CVPR*, 2011.
- [21] B. Ommer, T. Mader, and J. M. Buhmann, "Seeing the objects behind the dots: Recognition in videos from a moving camera," *International Journal of Computer Vision*, 83, 1, 2009, pp. 57-71, 2009.
- [22] A. Prest, C. Leistner, J. Civera, C. Schindl, V. Ferrari, "Learning object class detectors from weakly annotated video," in *Proc. CVPR* 2012.
- [23] Y. Kalal, J. Matas, and K. Mikolajczyk, "P-N learning: Bootstrapping binary classifiers from unlabeled data by structural constraints," in *Proc. CVPR*, 2010.
- [24] K. Ali, D. Hasler, and F. Fleuret, "Flowboost - appearance learning from sparsely labeled video," in *Proc. CVPR*, 2011.
- [25] D. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognition*, vol. 13, 1981, pp. 111-122.
- [26] Y.-L. Chen, S.-C. Cheng, and Y.-P. Phoebe Chen, "Reordering Video Shots for Event Classification Using Bag-of-Words Models and String Kernels," in *Proc. Intl. Conf. Image and Vision Computing (IVCNZ '12)* 2012.
- [27] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," In *Proc. of ACM Int'l Workshop on Computational Learning Theory*, 1992.
- [28] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge University Press, New York, 2004.
- [29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 60(2) (2004): 91-110.
- [30] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild'," in *Proc. Int'l Conf. IEEE Computer Vision and Pattern Recognition*, 2009.
- [31] J. Gall and V. Lempitsky, "Class-specific Hough forests for object detection," in *Int'l Conf. Computer Vision and Pattern Recognition*, 2009, 1022-1029.
- [32] M. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2008.
- [33] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. 20th British Machine Vision Conference*, 2009.
- [34] L. Yeffet and L. Wolf, "Local Trinary Patterns for human action recognition," in *Proceedings of International Conf. Computer Vision*, 492-497, 2009.
- [35] K. Schindler and L. J. V. Gool, "Action snippets: How many frames does human action recognition require?," in *Proc. ICCVPR*, 2008.
- [36] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. CVPR*, 2008.
- [37] K. Mikolajczyk and H. Uemura, "Action recognition with motion-appearance vocabulary forest," In *Proceedings of IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

Robust TV Stream Labelling with Conditional Random Fields

Abir Ncibi*, Emmanuelle Martienne†, Vincent Claveau‡, Guillaume Gravier‡ and Patrick Gros*

*INRIA †IRISA-Univ. of Rennes 2 ‡IRISA-CNRS

Campus de Beaulieu, F-35042 Rennes, France

Email: *firstname.lastname@irisa.fr*

Abstract—Multi-label video annotation is a challenging task and a necessary first step for further processing. In this paper, we investigate the task of labelling TV stream segments into programs or several types of breaks through machine learning. Our contribution is twofold: 1) we propose to use simple yet efficient descriptors for this labelling task, 2) we show that Conditional Random Fields (CRF) are especially suited for this task. In particular, through several experiments, we show that CRF out-perform other machine learning techniques, while requiring few training data thanks to its ability to handle the different types of sequential information lying in our data.

Keywords-Conditional Random Fields; video-stream labelling; TV segmentation; robust descriptors; sequentiality.

I. INTRODUCTION

Many digital TV channels have emerged in the recent years, making large amounts of video streams available. Yet, any new service based on these streams, such as video retrieval, information extraction, repurposing, requires, as a first step, to be able to structure the video flow into meaningful elementary units. In practice, the process of video stream structuring requires two main tasks: (1) a segmentation task that consists in detecting program boundaries, and (2), a labelling task that consists in giving each program a label describing its type or content.

Several studies [1]–[4] have already dealt with this issue. However, their main drawback is that their labelling processes chiefly rely on program information provided by the channels, on some reference databases, or on TV program guides. They all have underlined the limits of using such an external knowledge which is sometimes inaccurate or incomplete. In particular, TV guides don't contain information for small programs like commercials. Moreover, such TV guides are not always available. In this paper, to avoid this pitfall, we explore a different approach based on supervised machine learning. Of course, such an approach also requires some expert knowledge to build a training set, but we assume that this supervision is more easily available than a complete program information. More precisely, our goal is to investigate the use of a specific machine learning technique for the labelling task, namely the Conditional Random Fields (CRF), which are known to be suited to handle sequences. In that respect, our objective is manifold; we show that:

1 – CRF are efficient to induce programs labels, and out-perform other standard machine learning techniques;

2 – these good results can be obtained with few data and simple but robust descriptors;

3 – this good performance can be theoretically explained by the CRF's capability to use contextual relationships among a sequence of programs.

The remainder of this paper is organized as follows: next section is an overview of related work. In Section III, basic information about CRF and their learning algorithms are presented. Then, in Section IV, we detail our experimental settings, including the datasets used, the features and the evaluation measures. Sections V, VI and VII report the experiments we performed. Finally, we conclude in Section VIII.

II. RELATED WORK

To our knowledge, [1] are the first who proposed a complete solution for the video stream structuring problem. Their approach requires a reference database containing different kinds of breaks that are manually annotated. Breaks that repeat in the video stream are detected by matching the video stream with the breaks included into the reference database. If the video stream contains a new break that is not in the database, this new break is added to the database to update it. The main drawback of this method is its dependency to the reference database. Indeed, the latter has to be created for each channel and updated periodically to take into account all the breaks broadcasted by this channel. Another approach is proposed by [3] and consists in modelling program schedules by contextual hidden Markov models, that are able to predict all the possible schedules for a particular day. This approach gives good results in terms of precision of the prediction, but requires many annotated learning data. Another method developed by [4] uses an Inductive Logic Programming (ILP) tool to identify two classes of broadcasts: programs and breaks. The drawbacks of this approach are twofold: (1) it requires at least seven days of manually annotated programs and (2) it is not able to identify different kinds of breaks.

In this paper, we focus mainly on the labelling task. We suppose that the video stream has been divided, manually or automatically, into sequences of video segments. We propose a robust approach that uses CRF to label all the resulting

video segments. The highlights of our method are: (1) each segment is described with robust descriptors that are very easy to compute, and (2) the use of CRF allows for building an efficient model that predicts the types of the segments by taking into account the sequentiality of the data and the relationships between neighbouring segments.

CRF have been successfully applied to text processing, such as part-of-speech tagging [5], [6] or shallow text parsing [7]. They have also been used in video processing for detecting semantic events [8], [9] or identifying players in sports videos [10]. For all these tasks, CRF proved high efficiency and outperformed other probabilistic models, especially generative models like Hidden Markov Models (HMM) [11] and other discriminant models like Maximum Entropy Markov Model (MEMM) [12].

III. CONDITIONAL RANDOM FIELDS: BASIC CONCEPTS AND RELEVANT ALGORITHMS

A. Basic concepts

Conditional random fields [12] are undirected graphical models which aim at modeling a probability distribution of annotations y conditioned on known observations x based on labelled examples. CRF are defined as follows: we assume $G(V, E)$ an undirected graph (graph of independence) where V are vertices of the graph and E are edges of the graph. X and Y are two random fields over respectively the set of observations and the associated set of labels. For each vertex $v \in V$, it exists a random variable Y_v in Y . (X, Y) is called a conditional random field when each random variable Y_v depends only on observations X and its neighbours in the graph G . Based on this condition and according to the fundamental theorem of random fields (Hammersley and Clifford, 1971), the conditional probability of a sequence of annotations y given a sequence of observations x is written in terms of potential functions ψ_c over all cliques of the graph G :

$$p(y|x) = \frac{1}{Z(x)} \prod_{c \in \mathcal{C}} \psi_c(y_c, x) \quad (1)$$

where:

- \mathcal{C} is the set of all cliques of the graph G (completely connected subgraphs).
- y_c are configurations of random variables over vertices of the clique c .
- $Z(x)$ is a normalization factor.

B. Linear-chain CRF

The main use of CRF in the literature, mainly in natural language processing, is labelling sequences. In this case, the graph of independence G is a first-order linear chain (as the one shown on Fig. 1).

In this graph:

- cliques are adjacent edges and vertices of the graph;

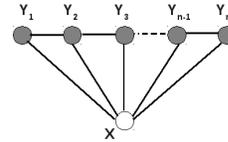


Figure 1. Graphical representation of a sequential CRF

- each label depends only on the previous and the next labels and the entire observations sequence x .

For linear CRF [12], the potential function ψ_c can be written as an exponential of weighted functions over the two types of cliques of the graph G as follows:

$$P(y|x) = \frac{1}{Z(x)} \exp \left(\sum_{k=1}^{k_1} \sum_i^n \lambda_k f_k(y_i, x) + \sum_{k=1}^{k_2} \sum_i^n \mu_k g_k(y_{i-1}, y_i, x) \right) \quad (2)$$

where:

- $Z(x) = \sum_y \exp \left(\sum_{k=1}^{k_1} \sum_i^n \lambda_k f_k(y_i, x) + \sum_{k=1}^{k_2} \sum_i^n \mu_k g_k(y_{i-1}, y_i, x) \right)$ (3)
- f and g are called *features functions*. The f functions characterize local relations in terms of labels and link the current label at position i to the sequence of observations x ; the g functions describe transitions between the graph vertices (states) and are defined for each pair of labels (or states) at position i and $i - 1$ and the sequence of observations.
- k_1 , k_2 and n are respectively: number of features functions f , number of features functions g and the size of the sequence of labels to be predicted.

Functions in f and g are generally binary functions which show the occurrences of particular combinations of label(s) and observation(s). These functions are fixed by the user, they reflect the knowledge of the user on the application field. Each function is applied to all the positions of the sequence. For instance, let's define $f(x_i, y_i)$ which relates the current observation x_i to its current label y_i . If we apply this function to all couples of labels and observations in the sequence, it will generate $|x| \times |y|$ functions features. Let's consider the sequence of observations $x = (15s, 10m, 10s, 1h)$, where each observation is the duration of the corresponding program, and its associated sequence of labels $y = (commercial, trailer, commercial, program)$. We also suppose that we fix two functions $f(x_i, y_i)$ and $g(y_i, y_{i-1})$. At position $i = 3$, the following feature functions are generated:

$$f(x_i, y_i) = \begin{cases} 1 & \text{if } x_i = 10s \text{ and } y_i = \text{commercial} \\ 0 & \text{else} \end{cases}$$

$$g(y_i, y_{i-1}) = \begin{cases} 1 & \text{if } y_i = \text{commercial} \text{ and } y_{i-1} = \text{trailer} \\ 0 & \text{else} \end{cases}$$

Features functions are associated with weights λ_k and μ_k that estimate the importance of information given by each feature function.

The conditional nature of CRF allows for relaxing the assumption of the conditional independence of observations fixed in the HMM, and allows for neighbourhood interactions among the observed data. CRF also avoid the *label bias* problem met with the HMM (or extensions like MEMM). This problem is caused by the fact that the probability mass received by y_{t-1} *must be* transmitted to y_t (at time t) regardless the corresponding observation x_t (for the interested reader, a good illustration of the label bias problem is presented by [12]). CRF are not impacted by such considerations since the way adjacent pairs y_t and y_{t-1} influence each other is not directed and is determined by input features x .

C. Learning and inference with CRF

Learning CRF models consists in estimating the vector of parameters $\theta = (\lambda_1, \lambda_2, \dots, \lambda_{k_1}, \mu_1, \mu_2, \dots, \mu_{k_2})$ given a training set $D = (x^{(i)}, y^{(i)})_{i=1}^N$ which maximizes the log-likelihood of the model:

$$L_\theta = \sum_{i=1}^N \log(p_\theta(y^{(i)}|x^{(i)})) \quad (4)$$

This function is concave, guaranteeing convergence to the global maximum. This optimization can be resolved by traditional iterative scaling learning algorithms such as the Improved Iterative Scaling (IIS) algorithm [13] but it has been proved that the Limited memory BFGS (L-BFGS) quasi-Newton method [14] converges much faster to estimate the parameters θ . The advantage of L-BFGS is that it avoids the explicit estimation of the Hessian matrix of the log-likelihood by building up an approximation of it, using successive evaluations of the gradient.

After this step of training, applying the CRF consists in finding the most probable sequence of labels y^* given a sequence of observations x :

$$y^* = \arg \max_y p_\theta(y|x) \quad (5)$$

As for other stochastic methods, y^* is generally obtained with a Viterbi algorithm, which calculates the marginal probability of states at each position of the sequence, using a dynamic programming procedure.

IV. EXPERIMENTAL SETTINGS

In this section, we present the data we used for the experiments that were conducted to evaluate CRF, as well as other machine learning algorithms, on video stream labelling.

A. Data

In our experiments, we used a TV stream containing three weeks of broadcasts. Within this stream, each segment was manually identified and given a label corresponding to its type: program or break. Four additional types have also been used to distinguish between different kinds of breaks: trailer, commercial, sponsorships and jingle. Two datasets were produced from this stream, each dataset resulting from the application of a particular segmentation method:

- A manual segmentation method which identifies precisely the beginning and the end of each broadcast. This segmentation will be useful for evaluating the relevance of CRF on the labelling of a perfectly segmented video stream. In this case, a video segment is equivalent to a program (movie, TV serie, talk-show, etc) or a break (commercial, trailer...). 7,591 video segments were extracted using this manual segmentation method.
- An automatic segmentation method that aims at evaluating CRF for the labelling in a more realistic setting. The automatic segmentation method we used is based on the detection of repeated segments [15]. Applied to TV streams, this method tends to over-segment the stream: 48,544 video segments were detected. By examining the result of the segmentation, we observe that each broadcast (corresponding to a unique segment with the manual segmentation method) is divided into several segments, each segment having a short duration.

The distribution of the segments over the different types, inside both datasets, is shown on Table I.

Table I
DISTRIBUTION OF THE SEGMENTS OVER THE DIFFERENT TYPES

Label	Manual segmentation	Automatic segmentation
Program	1,506	22,557
Trailer	1,290	4,075
Commercial	1,050	18,089
Sponsorship	1,714	2,201
Jingle	2,031	1,622
Total	7,591	48,544

B. Descriptors

Within both datasets, each video segment is described by three descriptors:

- its duration: we distinguish between ten possible values, each value being an interval: [0-15s], [15s-30s], [30s-45s], [45s-1min], [1m-15m], [15m-30m], [30m-1h], [1h-2h], [2h-4h].
- the moment in the week it was broadcasted: business day, off-day or weekend.

- the period in the day it was broadcasted: morning, noon, afternoon, evening, night.

These features are robust since they are very easy to compute, and they do not depend on the quality of image or sound signal of the stream. Some examples of segments with these features and their labels are shown in Table II.

Table II
EXAMPLES OF SEGMENTS WITHIN THE DATASETS, WITH THEIR
FEATURES AND LABELS

Segment	Moment in the week	Period in the day	Duration	Label (class)
<i>Seg₁₅</i>	Business day	morning	[10s,15s[commercial
<i>Seg₁₆</i>	Business day	morning	[0s,10s[trailer
<i>Seg₁₇</i>	Business day	morning	[0s,10s[jingle
<i>Seg₁₈</i>	Business day	afternoon	[15min,30min[program
<i>Seg₁₉</i>	Business day	afternoon	[10s,15s[commercial

C. Labelling tasks and evaluation measures

For all the experiments, the first two weeks of a dataset were used to train and construct the labelling model, and the last third week to test and evaluate this model. Two kinds of labelling tasks have been considered:

- a binary labelling task in which a segment is either a program or a break, i.e., there is no distinction between different kinds of breaks;
- a multiple labelling task in which a distinction is made between different kinds of breaks. Consequently, five labels are used: program, commercial, sponsorship, trailer or jingle.

In the experiments reported in the next sections, the performance is evaluated on the test sequences by comparing the labels produced by the technique with those from the ground-truth. Different evaluation measures are used. As a global measure, we compute the accuracy rate, that is, the proportion of correctly labelled segments in the test streams. For each label, we also evaluate the recall, precision and f-score, and we then compute a weighted average over the labels (weighted according to the amount of segments of each class). Note that the weighted average recall is equivalent to the accuracy rate that we use as a global measure.

D. Video stream labelling with CRFs

To use efficiently CRF, we consider that a sequence groups together all the segments of a day of broadcasting. We have 15 sequences, i.e., the first two weeks for learning the labelling model and 8 sequences, i.e., the last third week, for testing the model. In a sequence, observations are the vectors of descriptors and labels are the types of the segments.

To learn the labelling model, appropriate features functions must be chosen to express the dependencies that may exist in a sequence between observations or labels. In our

experiments, we used the tool CRF++¹. In this tool, feature functions are defined in a *template file* where:

- feature functions f are equivalent to *unigram templates* that describe relationships between the current label and the observations in the sequence (see Section III-A);
- feature functions g are equivalent to *bigram templates* that describe only the relationships between two successive labels (see Section III-A).

For each dataset, an appropriate template file which defines the f and g functions was created:

- *Manual segmented stream dataset*: for this dataset, we used feature functions which run over a window of eight neighbours; four before and four after the current observation.
- *Automatic segmented stream dataset*: for this dataset, feature functions are more complicated. We used also a window of eight neighbors but more combinations of previous, current and following observations were taken into account. For this dataset, we used only local feature functions f (unigrams) to avoid the over-fitting.

In Section VII, we presents some results obtained with different templates. Each experiment was performed on both datasets. To highlight the efficiency of CRFs in sequential data labelling, we compare the results obtained by CRFs with the results obtained by different non sequential classification methods (SVM, Naive bayes, Random Forest) for the same labelling task. For each method, two settings are presented. The first one uses a naive description in which only the current observation is considered (noted as *simple* hereafter). The second one (noted as *contextual* hereafter) takes into account the context of the observations by adding the descriptors of the surrounding observations in the description of the current segment. Different sizes of contexts have been tested; here, we report the ones yielding the best results, that is when considering the two previous and the two next segments. We also compare CRFs to HMMs to study the impact of the label context which is also taken in account by CRFs. To the contrary of CRF, HMM only take into account the current observation of the segment to be labelled. To complete this comparison, we also indicate the results of two baselines:

- *Baseline1* where only the most frequent label is predicted;
- *Baseline2* which uses a features function which considers only the duration of the current segment to be labelled. We choose this baseline because duration is the most discriminant descriptor.

Again, the experiments are performed on both manually and automatically segmented datasets.

¹<http://crfpp.googlecode.com/svn/trunk/doc/index.html>

V. RESULTS ON THE MANUAL DATASET

This section is dedicated to the results obtained with the manually segmented stream. In the first part, we summarize the results obtained by CRF and other classification methods. In the second part, we focus on the detailed results obtained by CRF on binary and multiple labelling.

A. Global results on the manual dataset

Results are presented on Tables III and IV. The best results for SVM, reported hereafter, were obtained with a RBF (Radial Basic Function) kernel where γ is set to 0.1. For the results, we attribute to each label a weight that is proportional to its frequency in the learning dataset. Then, we calculate weighted averages of recalls, precisions and F-scores for each predicted label.

Table III
PERFORMANCE FOR THE BINARY LABELLING TASK WITH THE MANUAL DATASET, USING CRF AND VARIOUS CLASSIFICATION METHODS

	Accuracy - Recall (%)	Precision (%)	F-score (%)
CRF	95.66	95.63	95.63
HMM	88.79	90.2	89.2
SVM simple	86.3	85.4	85.2
N.Bayes simple	86.6	87	86.8
R.Forest simple	87.1	88.3	87.5
SVM contextual	94.9	94.8	94.8
N.Bayes contextual	89.4	91.0	89.9
R.Forest contextual	94.9	94.8	94.8
Baseline1	79.48	63.16	70.39
Baseline2	87.37	86.66	86.53

From these results, several points are noteworthy. First, the task of binary labelling seems easy enough to yield high score with the baseline techniques. Secondly, the difference between the simple and contextual settings of the usual machine learning algorithms underlines the importance of taking the context of the current observation into account as it is naturally done with CRF. Thirdly, the label context, naturally taken into account by HMM and CRF, seems also beneficial for the performance.

Table IV
PERFORMANCE FOR THE MULTIPLE LABELLING TASK WITH THE MANUAL DATASET, USING CRF AND VARIOUS CLASSIFICATION METHODS

	Accuracy - Recall (%)	Precision (%)	F-score (%)
CRF	84.08	85.13	84.54
HMM	74.52	53.82	49.22
SVM	59.5	64.6	56.5
N.Bayes	59.6	62.8	56.5
R.Forest	58.7	61.3	55
SVM contextual	76.1	76.8	75.8
N.Bayes contextual	68.6	69.6	68.8
R.Forest contextual	74.9	75.8	74.6
Baseline1	27.36	7.49	11.76
Baseline2	64.3	66.2	64.8

The multiple labelling task is more difficult, resulting in lower performance. Here again, the importance of taking the context of the current observation into account appears clearly. As it is suggested by the HMM results, the context of the label is not enough to cope with these more complex data. Yet, the CRF model, which takes both contexts into account yields the better results and outperforms any other technique. Finally, these results highlight the two following interesting points:

- using features functions, CRF are the most competitive method for the task of video sequence labelling;
- robust descriptors are discriminant enough to label the manual dataset.

B. CRF results on the manual dataset

In order to analyse the errors, we report detailed results for the CRF in Tables V and VI.

Table V
DETAILED PERFORMANCE FOR THE BINARY LABELLING TASK WITH THE MANUAL DATASET USING CRF

	Number of segments	Recall (%)	Precision (%)	F-score (%)
Inter-Program	1.184	97.52	97.03	97.27
Program	564	88.47	90.23	89.34
Weighted average		95.66	95.63	95.65

CRF have interesting results in binary labelling: both programs and breaks are identified by the learned model. Breaks are better identified than programs because they are more numerous in the learning dataset and are characterized by their short duration.

Table VI
DETAILED PERFORMANCE FOR THE MULTIPLE LABELLING TASK WITH THE MANUAL DATASET USING CRF

	Number of segments	Recall (%)	Precision (%)	F-score (%)
Program	564	88	92.31	90.10
Trailer	381	88.47	90.23	89.34
Jingle	752	85.37	81.87	83.53
Commercial	309	87.37	82.56	84.9
Sponsorship	742	76.17	81.43	78.71
Weighted average		84.08	85.13	84.54

In multiple labelling, CRF are still able to predict labels with a F-score equal to 84.54%. Programs are the best predicted class with 92.31% precision and 90.10% F-score.

VI. RESULTS ON THE AUTOMATIC DATASET

The automatic segmentation technique used to produce what we refer as the automatic dataset tends to over-segment the stream, as programs or breaks are generally divided into several segments. For the labelling task, these multiple segments belonging to one broadcast, have to get the same label. This section follows the same structure than the previous one: we start by presenting the global results of

the different machine learning methods, before giving more detailed results about the CRF. For all these experiments, the best results with SVM were obtained with a linear kernel.

A. Global results on the automatic dataset

Results are shown in Tables VII and VIII.

Table VII
PERFORMANCE FOR THE BINARY LABELLING TASK WITH THE AUTOMATIC DATASET, USING CRF AND VARIOUS CLASSIFICATION METHODS

	Accuracy - Recall (%)	Precision (%)	F-score (%)
CRF	69.54	72.94	67.9
HMM	62.7	73.37	56.8
SVM simple	57.9	57.8	57.7
N. Bayes simple	57.7	57.7	57.7
R. Forest simple	58.7	58.7	58.7
SVM contextual	64.7	68.4	61.3
N. Bayes contextual	63.9	65.4	61.7
R. Forest contextual	63.4	65.7	61.7
Baseline 1	52.19	27.24	35.79
Baseline 2	54.29	55.16	53.77

Table VIII
PERFORMANCE FOR THE MULTIPLE LABELLING TASK WITH THE AUTOMATIC DATASET, USING CRF AND WITH VARIOUS CLASSIFICATION METHODS

	Accuracy - Recall (%)	Precision (%)	F-score (%)
CRF	57	51.25	52.45
HMM	37.65	55.4	37.32
SVM simple	46.4	36.9	40.2
N. Bayes simple	46.7	39.5	41.7
R. Forest simple	47.5	40.8	42.5
SVM contextual	50.6	45.6	45.9
N. Bayes contextual	51.0	46.3	48.2
R. Forest contextual	51.7	47.8	47.8
Baseline 1	47.8	22.85	30.92
Baseline 2	47.32	38	41.26

Several facts are worth noting. First, in the binary labelling task as in the multiple labelling one, CRF still provide the best performance (in terms of Accuracy and F-scores) compared to other methods. One other interesting point is that all the methods yield lower results than for the manual dataset, with about a 30% F-score loss. This unsurprising result can be explained by the over-segmentation that resulted from the use of an automatic segmentation tool.

B. CRF detailed results on the automatic dataset

Table IX
DETAILED PERFORMANCE ON THE BINARY LABELLING TASK WITH THE AUTOMATIC DATASET USING CRF

	Number of segments	Recall (%)	Precision (%)	F-score (%)
Break	9,198	64.97	90.34	75.58
Program	8,426	81.63	46.83	59.52
Weighted average		72.94	69.54	67.9

Table X
DETAILED PERFORMANCE ON THE MULTIPLE LABELLING WITH THE AUTOMATIC DATASET USING CRF

	Number of segments	Recall (%)	Precision (%)	F-score (%)
Program	8426	64.61	67.17	65.87
Trailer	1459	1.23	10.17	2.19
Jingle	635	0.00	0.00	0.00
Commercial	6149	74.37	49.26	59.27
Sponsorship	945	0.94	20.45	1.80
Weighted average		51.25	56.99	52.45

As for the manual dataset, we provide detailed results of the CRF performance in Tables IX, X and XI. In multiple labelling of the automatic dataset, CRF provide the highest F-score in average (52.45%), even if there is a high confusion between labels as shown on the confusion matrix (see Table XI). Commercials and programs are the best recognized by CRFs. Other broadcasts are difficult to be correctly labelled, especially jingles and sponsorships.

Table XI
MULTIPLE LABELLING OF THE AUTOMATIC DATASET USING CRF - CONFUSION MATRIX

Predicted \ Real	Program	Trailer	Jingle	Commercial	Sponsorship
	Program	5444	115	8	2836
Trailer	562	18	0	876	4
Jingle	204	4	0	426	2
Commercial	1529	38	4	4573	6
Sponsorship	370	2	1	573	9

We note a high confusion between the following labels (see Table XI):

- jingle and commercial: more than 50% of jingles are labelled as commercials and the remainder as programs;
- trailer and commercial: more than 50% of trailers are labelled as commercials and the remainder as programs;
- sponsorship and commercial: more than 50% of sponsorships are labelled as commercials and the remainder as programs;
- commercial and program: almost 25% of commercials are labelled as programs and almost 30% of programs are labelled as commercials.

These results highlight the fact that the descriptors used to describe the segments are not discriminant enough to separate many successive segments.

VII. EXPLORING THE EFFICIENCY OF CRF

Results obtained in the previous experiments show that CRFs are better suited to our labelling tasks than other usual machine learning techniques. In this section, two related issues regarding this good performance are explored. We first shed light on the importance of taking into account

the sequential nature of our data, and how this is done, at different levels in CRFs. As the supervision task is tedious and costly, we then examine how CRF deal with different training set sizes.

A. About sequentiality in CRF

Four experiments were conducted in order to shed light on the ability of CRF to take into account the sequential nature of the data. This is simply done by using different template files defining the model. Here are the different settings used for these experiments:

- *CRF-all*: this template indicates that the CRF uses a) information about the current observation as well as the four before and the four next ones (corresponding to features function $f(y_i, x_{i-4}, \dots, x_i, \dots, x_{i+4})$), and b) information about the neighbouring labels, called bigram template, corresponding to feature functions $g(y_i, y_{i-1})$.
- *CRF-CO*: here, we use a) an unigram template which considers only the current observation and its associated label (corresponding to features function $f(y_i, x_i)$) and b) a bigram template; so finally:

$$P(y|x) = \frac{1}{Z(x)} \exp \left(\sum_{k=1}^{k_1} \sum_i \lambda_k f_k(y_i, x_i) + \sum_{k=1}^{k_2} \sum_i \mu_k g_k(y_{i-1}, y_i) \right) \quad (6)$$

In terms of information taken into account, this formulation can be compared to the HMM one.

- *CRF-nonB*: this template is similar to CRF-all, without the bigram template.

$$P(y|x) = \frac{1}{Z(x)} \exp \left(\sum_{k=1}^{k_1} \sum_i \lambda_k f_k(y_i, x_{i-4}, \dots, x_{i+4}) \right) \quad (7)$$

This formulation can be compared to the *contextual* setting of the standard machine learning algorithms.

- *CRF-CO-nonB*: this last template only includes an unigram template which considers the current observation (corresponding to features function $f(y_i, x_i)$). This formulation can be compared to the *simple* setting of the standard machine learning algorithms.

These different CRF versions are tested on the manual dataset on the multiple label task. Table XII presents the results they obtain (report to Table IV for other methods' performance). From these experiments, one can assess the importance of the two types of sequential information taken into account in CRF. Indeed, both the label dependency and the neighbouring observations help to yield the best results. On this particular task, the latter has a greater impact than the former. It is interesting to compare the results of

Table XII
PERFORMANCE FOR THE MULTIPLE LABELLING TASK WITH THE MANUAL DATASET, USING CRF AND VARIOUS CLASSIFICATION METHODS

	Accuracy - Recall (%)	Precision (%)	F-score (%)
CRF-all	84.08	85.13	84.54
CRF-nonB	78	78.43	78.33
CRF-CO	66.27	69	66.79
CRF-CO-nonB	58.7	61.27	55.09

the CRF-CO and HMM since they both exploit the same information. Yet, the CRF clearly outperforms HMM thanks to its undirected representation of the label dependency preventing any label bias problem (cf. section III-B). As expected, CRF-nonB yields similar results to those of the *contextual* setting of the SVM, Naive Bayes or Random Forests. Similarly, the CRF-CO-nonB, whose prediction only relies on the current observation, is comparable to standard machine learning techniques such as SVM, or Random Forests with the simple setting and thus also obtains similar results.

B. Training set size

As it has been said before, due to the cost of supervision, it is interesting to examine how the performance of the labelling techniques are dependent on the training set size. For this experiment, we adopt the most difficult setting: multiple labelling with the automatic dataset. At every learning step, we add a new sequence of segments broadcasted on the same day to the training set, learn the CRF parameters, and apply this CRF on the test set. To prevent any bias, the sequences are randomly selected, and the results are averaged over several runs. The results are shown in Fig. 2.

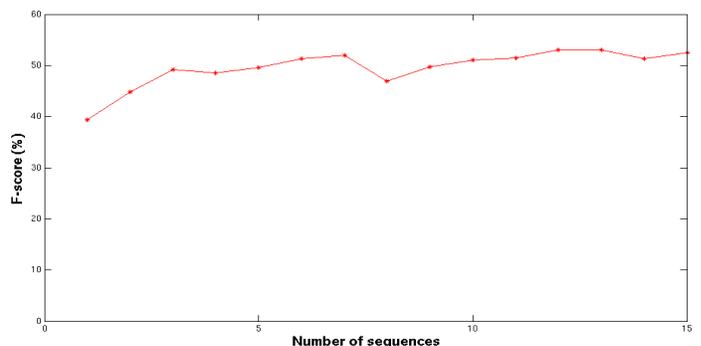


Figure 2. Results of CRF on multiple labelling of the automatic dataset, using different sizes of the learning dataset

We note that even with a small number of sequences (3 sequences), CRF have a F-score near of its optimum. This efficiency of CRF could be explained by the advantage of CRF compared to other probabilistic graphical models that do not account for local conditional probabilities, so

being free of biased estimations of these quantities when too few labeled data are available. Instead, feature functions weights account for positive as well as negative contributions of the observations to the labels. This result is interesting when compared with existing methods requiring extensive annotated data such as [4].

VIII. CONCLUSION AND FURTHER WORK

In this paper, we applied CRF to the labelling of a segmented TV stream where video segments are described with robust descriptors. These descriptors can be computed easily because they do not depend on the signal quality and do not require knowledge in the field of signal or image. In addition, these descriptors don't depend on a specific channel nor on the period in the year. The TV stream was segmented with two different segmentation processes, each process leading to a specific dataset: manual and automatic. Our goal was to identify five kinds of broadcasts in each dataset. We obtained interesting results on the manual dataset where the precision and the recall were up to 90%. Results are lower on the automatic dataset, especially in multiple labelling where we noticed many confusions between labels. Nevertheless, CRF's results exceed those of other classification methods such as Hidden Markov Models, which is also a probabilistic graph-based model. Indeed, the CRF's capability to handle the sequential context between video segments makes it possible to separate different kinds of programs and breaks, even when they are described with very simple features. Of course, this approach chiefly relies on the quality of the stream pre-processing steps. Dealing with the automatically segmented data is thus more challenging, especially for the multiple labelling task, which leads to high confusion between certain labels (commercial vs. jingle, commercial vs. and sponsorship...). This weakness can be explained by the over-segmentation of the automatic dataset: broadcasts are divided into many consecutive segments that features are not informative enough to discriminate.

Different perspectives are foreseen for this work. To improve our results on the multiple labelling task, especially for the automatically segmented dataset, we plan to investigate the use of content-based features, namely audio features that are specific to some kinds of breaks (for instance, there is no music and no speech in jingles). Another challenge is also to reduce the need for already labelled data in the building of the model. To achieve this goal, we plan to introduce unlabelled data and to explore using active learning strategies to induce CRF.

REFERENCES

- [1] X. Naturel, G. Gravier, and P. Gros, "Fast structuring of large television streams using program guides," in *4th International Workshop on Adaptive Multimedia Retrieval, AMR'06*, 2006.
- [2] X. Naturel and P. Gros, "Detecting repeats for video structuring," *Multimedia Tools and Applications*, vol. 38, no. 2, pp. 233–252, 2008.
- [3] J.-P. Poli, "An automatic television stream structuring system for television archives holders," *Multimedia systems*, vol. 38, no. 2, pp. 255–275, November 2008.
- [4] G. Manson and S.-A. Berrani, "An inductive logic programming-based approach for TV stream segment classification," in *IEEE International Symposium on Multimedia, ISM'08*, Berkeley, California, USA, December 2008.
- [5] A. Pranjali, R. Delip, and R. Balaraman, "Part of speech tagging and chunking with hmm and crf," in *NLP Association of India (NLPAI) Machine Learning Contest*, 2006.
- [6] M. Constant, I. Tellier, D. Duchier, Y. Dupont, A. Sigogne, and S. Billot, "Intégrer des connaissances linguistiques dans un CRF : Application à l'apprentissage d'un segmenteur-étiqueteur du français," in *Traitement Automatique du Langage Naturel (TALN'11)*, 2011.
- [7] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, ser. NAACL '03, 2003.
- [8] T. Wang, J. Li, Q. Diao, Y. Z. Wei Hu, and C. Dulong, "Semantic event detection using conditional random fields," in *Computer Vision and Pattern Recognition Workshop (CVPRW '06)*, 2006.
- [9] N. Zhang, L.-Y. Duan, Q. Huang, L. Li, W. Gao, and L. Guan, "Automatic video genre categorization and event detection techniques on large-scale sports data," in *Conference of the Center for Advanced Studies on Collaborative Research (CASCON'10)*, 2010.
- [10] W.-L. Lu, J.-A. Ting, K. P. Murphy, and J. J. Little, "Identifying players in broadcast sports videos using conditional random fields," in *Computer Vision and Pattern Recognition (CVPR '11)*, 2011.
- [11] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," in *IEEE ASSP Magazine*, 1986.
- [12] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *International Conference on Machine Learning (ICML)*, 2001.
- [13] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Inducing features of random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 380–393, 1997.
- [14] N. N. Schraudolph, J. Yu, and S. Günter, "A stochastic quasi-Newton method for online convex optimization," in *11th International Conference on Artificial Intelligence and Statistics*, ser. Workshop and Conference Proceedings, vol. 2, San Juan, Puerto Rico, 2007, pp. 436–443.
- [15] Z. A. A. Ibrahim and P. Gros, "TV stream structuring," *ISRN Journal on Signal Processing*, vol. 2011, no. 1, April 2011.

Determinants of Behavioral Intention to Mobile Banking

Case From Yemen

Rashed, Abdullah, Santos, Henrique

Algoritmi Centre,
University of Minho,
Guimarães, Portugal
rashed, hsantos@dsi.uminho.pt

Al-Eryani, Arwa

IT Faculty,
Saba University
Sana'a, Yemen
arwa_y@hotmail.com

Abstract— Nowadays, new tools and technologies are emerging rapidly. They are often used cross-culturally before being tested for suitability and validity. However, they must be validated to ensure that they work with all users, not just part of them. Mobile banking (as a new technology tool) has been introduced assuming that it performs well concerning authentication, among all members of the society. Our research aimed to evaluate authentication mobile banking user acceptance, through Technology Acceptance Model (TAM), in Arabic countries, namely Yemen. The results confirm the previous studies that have shown the importance of perceived ease of use and perceived usefulness. Furthermore, perceived ease of use plays a determinant role.

Keywords- *Technology acceptance models; Mobile Banking; Arabic culture.*

I. INTRODUCTION

Technologies make our lives easy but not secure [19] especially for financial issues. Most organizations already provide the services via the Internet and mobile appliances [18]. Furthermore, during the last ten years, the improvement of mobile communication technologies has changed the banking industry, as users are able to conduct banking services at anyplace and at any time [5] via mobile phones. Mobile Banking provides many services to the customers such as: requesting the balance and the latest transactions; transferring funds between accounts; buying and selling orders, for the stock exchange; and receiving portfolio and price information [2]. For individuals it would be difficult to remember their user names and PINs [14]. For that reason, many users select easy to remember passwords [3], which are considered a security trade-off. Security specialists are looking for more advanced techniques that would improve its performance [13].

Mobile Banking is still in a development phase in most countries especially middle-east, where small markets with few users have been reported. This is due to lack of customer acceptance and poor time response services [2]. In the other hand, mobile payments are mainly used with popular mobile services since there are few alternative payment solutions available [10].

There are three types of authentication [15]:

1) Something you know: a PIN, a password, or a passphrase.

2) Something you have: a passport, key, ATM card or cell-phone [6].

3) Something you are (Biometrics): fingerprints, signature, ear shape, keystroke, voice, finger geometry, iris, retina, DNA, hand geometry [11] and odour [16].

Acceptance of technology is a milestone [20]. It is very important to predict users' intention to use mobile banking [5] so various alternative approaches have been used to analyze customer's acceptance phenomenon. Within this context, TAM is one of the most widely accepted tools among information systems researchers [2].

In this paper we investigate the acceptance of mobile appliances, focusing in authentication effectiveness, in Arabic countries. The rest of the paper is organized as following: in Section 2 we overview the previous studies, as literature review; in Section 3 we describe our methodology and discuss results. We conclude and present future work in Section 4.

II. LITERATURE REVIEW

Khanfar et. al. [8] conducted the customer satisfaction with internet banking web site for a bank. Their covered factors were: customer support, security, ease of use, digital products/services, transaction and payment, information content, and innovation. The results found a narrow-based satisfaction with internet banking in all factors. They found that all factors have a positive impact on the customer satisfaction. Moreover, they found that there was no relation between all demographics data and customer satisfaction due to the high computer literacy among customers.

Gaurav et. al. [4] discussed Automatic Teller Machine (ATM) authentication techniques. They aimed to propose solution that uses the personal mobile devices to interact with the service outlets. They used public key Infrastructure for mutual authentication of the service and the personal device in their model. Their idea depends on the following policy:

- After users' registration, their mobile carries the public key whereas their smart card contains the private key.
- Mobile phone authenticates itself to ATM.
- Mobile phone establishes a session key using standard key exchange protocols such as Diffie-Hellman key

exchange along with an integrated authentication to avoid man-in-middle attack.

- Users would access the service of the ATM using the signed application either loaded by the bank during registration or by the ATM.

So users need to carry only their personal devices to access various services. They did their simulation on different platforms.

AlZomai et. al. [1] discussed the authentication problems of security in online banking of using SMS for transactions. Their experiment aimed to simulate the online bank using website to do the transactions. They suggest enhancing online banking security by focusing on usability more than security technical and mechanisms. They suggested SMS authorization scheme. They attacked their approach to make sure that it would work properly. Their attack succeeded in 21%. They justified that as user should have more experience.

Gu et. al. [5] examined and validated the determinants of users' intention to mobile banking. They used a structural equation modeling (SEM) to test the causalities in the proposed model. They verified the effect of perceived usefulness, trust and perceived ease-of-use on behavioral intention in mobile banking. The results indicated strong support for the validity of proposed model with 72.2% of the variance in behavioral intention to mobile banking. The study also found that self-efficiency was the strongest antecedent of perceived ease-of-use, which directly and indirectly affected behavioral intention through perceived usefulness in mobile banking. In addition, they found that structural assurances were the strongest antecedent of trust, which could increase behavioral intention of mobile banking.

Hua et. al. [7] investigated the factors affect mobile commerce adoption in China and the United States. They conducted a survey on 190 individual mobile commerce users in China and USA. Results showed that there are several significant cultural differences on consumer intention to use mobile commerce.

Yaseen et. al. [21] used TAM model to study the m-commerce technology deployment in Jordan. They distributed 210 questionnaires to mobile commerce users in Stock Exchange for Brokers and Investors. Their factors were trust, perceived usefulness, perceived ease of use, social and cultural values and economic issues that influence a decision maker intention to adopt this type of technology in doing business. Their results showed that perceived trust, perceived usefulness, perceived ease of use, social and cultural values had significant association with intention to deploy mobile commerce technology while economical issue is not significant.

Maiyaki et. al. [9] studied determinants of consumer behavioral intention in Nigerian commercial banks. They investigated the influence of perceived service quality, perceived value, corporate image and switching cost on the consumer behavioral intention in the context of commercial banks in Nigeria. They found that the service of quality, customer perceived value and image of the corporate had significant influence on customer behavioral intention.

Barati et. al. [2] studied the factors that affect acceptance of mobile banking. They presented a set of factors that could potentially positively affect the success of mobile banking and should be taken into account by banks while adopting mobile technology as shown in Figure 1. They found that perceived usefulness and perceived ease of use are significant. Moreover they found that role of facilitating conditions in acceptance of mobile services is very important.

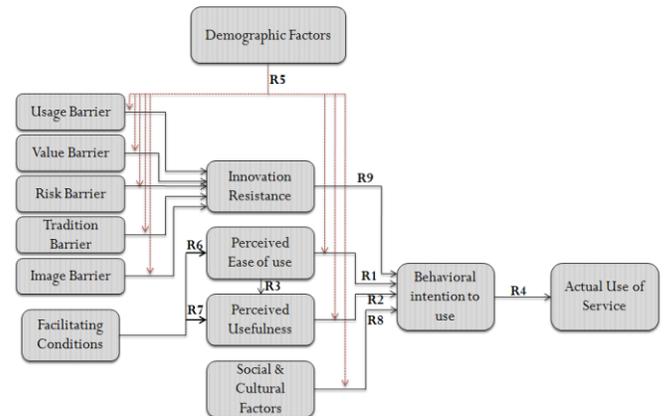


Figure 1: Acceptance model for Mobile Banking [2]

Ramayah et. al. [12] studied and examined the intention to use an online bill payment among part time MBA students in University Sciences Malaysia, Penang. They developed and modified the extended TAM and Social Cognitive Theory to identify factors that would determine and influence the intention to use an online bill payment system. They conducted a survey that involved 120 students. They found that perceived ease of use and perceived usefulness are the significant drivers of intention to use the online bill payment system. In addition to that, they found that subjective norm, image, result demonstrability and perceived ease of use were to be the key determinants of perceived usefulness whereas perceived risk was found to be negatively related to usefulness. Moreover, computer self-efficacy played a significant role in influencing the perceived ease of use of the online bill payment system.

III. METHODOLOGY AND DISCUSSION

TAM has two pillars that determine the users' acceptance of a new technology: perceived ease of use and perceived usefulness. Perceived ease of use is defined as the degree to which the users expect that the target system would require a low effort to learn to use, while perceived usefulness is defined as "the individuals' subjective probability of using a specific application system, will increase their job performance within an organizational context" [17].

Table 1 shows the research variables required by TAM and its characterization. Perceived ease of use and perceived usefulness act as independent and dependent variables at the same time. Besides, the demographic factor is considered as independent, while intention to use acts as dependent as it

depends on perceived ease of use and perceived of usefulness.

The research hypotheses are:

H1: Perceived ease of use will have a positive effect on intention to use Mobile Banking.

H2: Perceived usefulness will have a positive effect on intention to use Mobile Banking.

H3: Demographic factor will have a positive effect on intention to use Mobile Banking. Perceived performance is defined as the degree to which users expect that the target system would support the performance perceive. Saving time and effort is defined as the degree to which the users expect that the target system would save the time and effort when comparing with the old method. Social and cultural factors are defined as the degree to which the users expect that the social and cultural factors will affect its decide to use the target system. We directly asked the respondents about the mentioned factors to measure their intention and behaviour.

TABLE 1: RESEARCH VARIABLES

Variable	Type	Scale
Technology acceptance	Dependent	Discrete (1-5) 1: Extremely Likely 5: Extremely Dislikely
Perceive ease of use	Independent/Dependent	
Perceived usefulness	Independent/Dependent	
Demographic factor	Independent	

The factors affecting acceptance of Mobile Banking as a new technology in financial payments and transactions are presented in Table 2. The model expands TAM with innovation resistance, performance perceive saving time and effort, and social and cultural factors. Moreover, proposed model includes experience that represents the familiarity of the mobile device and ATM, technology use skills, etc.

Descriptive Analysis

As shown in Table 2, our sample consisted of 76% males and 24% females. The majority of the sample were young (48%) in the interval [21-30], 6% were less than 21. 33% were within the interval [31-40]. 47% of the respondents have post graduate degrees, 41% bachelor degree and most of them (54%) are proficient IT users. 89% of the respondents use ATM machines and 98% preferred to use it rather than dealing with a human being clerk. 78% of the respondents liked the idea 20% did not decide. 74% of the respondents considered using mobile banking easy 10% considered it as difficult and 16% did not decide.

87% of the respondents considered using mobile banking as a brilliant idea and 15% did not decide whether they considered mobile banking as good or bad idea. 5% considered it as a stupid idea. 45% intended to use mobile banking and 40% did not decide. 83% perceived the usefulness of using mobile banking and 15% did not decide. 90% of the respondents think that using Mobile Banking will

improve the performance in their lives. 89% considered the idea would help in exploiting the time.

TABLE 2: SAMPLE PROFILE

Variable		Frequency
Gender	Female	76
	Male	24
Race	Yemenis	79
	Arab	21
Age	15-20	6
	21-30	48
	31-40	33
	More than 41	13
Specializations	IT	54
	Finance	5
	Administration	10
	Medicine	7
	Engineering	13
Jobs type	Others	11
	Public Sector	23
	International organizations	5
	Private Organization	41
	Family business	3
	Other	28

Figure 2 shows our proposed model for Mobile Banking acceptance. This model expands TAM adding factors such as experience, Innovation, performance, social factors, saving time. The experience of using mobile would affect the responses and similar technologies would help users to perceive both ease of use and usefulness.

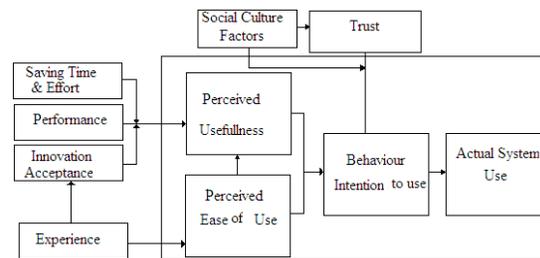


Figure 2: Proposed framework

The review showed that the demographic characteristics have an impact on the adoption of mobile technology. However, we find that age has no effect on intension to use mobile in financial transactions. Furthermore, we found that gender has significant effect as males have strong intention to use the new technologies more than females. Experience factor is significant. Social and cultural factors are important in acceptance of mobile banking. Mobile services are innovation and each innovation comes with resistance of consumers.

IV. CONCLUSION

Our sample consisted of young educated individuals. Moreover, most of them were frequent users of ATM machines and preferred to use technologies rather than the old methods.

Our results confirm the previous studies results. However, they conflict with some of them. Our results find that both perceived ease of use and perceived of usefulness are significant factors. However, the results show that perceived ease of use plays the most significant role.

From our results, it can be concluded that many of participants accept mobile banking for reasons such as saving time and improvement of their daily life.

Most of participants think it is easy to use mobile banking and some think they need some help. The result shows the gap between accepts the new technology as an idea and the actual use of it. We recommend awareness campaign that leads to user perceptions for new technologies.

ACKNOWLEDGEMENT

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

REFERENCES

- [1] AlZomai M., AlFayyadh B., Audun Jøsang A. and cCullagh A.(2008), An Experimental Investigation of the Usability of Transaction Authorization in Online Bank Security Systems, Proceedings of the sixth Australasian conference on Information security - Volume 81, Wollongong, NSW, Australia , pp:65-73, ISBN ~ ISSN:1445-1336 , 978-1-920682-62-0
- [2] Barati S. and Mohammadi S. (2009), An Efficient Model to Improve Customer Acceptance of Mobile Banking, Proceedings of the World Congress on Engineering and Computer Science 2009 Vol. II WCECS 2009, October 20-22, San Francisco, USA.
- [3] Coventry L., De Angeli A. and Johnson G. (2003), Usability and Biometric Verification at the ATM Interface, Proceedings of the SIGCHI conference on Human factors in computing systems, Ft. Lauderdale, Florida, USA, ISBN:1-58113-630-7, pp: 153 - 160.
- [4] Gaurav A., Sharma A., Gelara V. and Moona R.(2008) Using Personal Electronic Device for Authentication-based Service Access, IEEE International Conference on Communications (ICC2008), Beijing, 19-23 May 2008.
- [5] Gu J., Lee S. and Suh Y. (2009), Determinants of Behavioral Intention to Mobile Banking, Expert Systems with Applications: An International Journal, Vol. 36 , Issue 9 (November 2009): 11605-11616.
- [6] Herzberg A. (2003), Payments and Banking with Mobile Personal Devices, Communications of the AC, Volume 46, Issue 5, 2003, ISSN: 0001-0782, pp: 53 - 58.
- [7] Hua D. and Prashant P. (2009), Mobile Commerce Adoption in China and The United States: A Cross-Cultural Study, ACM SIGMIS Database, Vol.40, Issue 4 (November 2009): 43-61
- [8] Khanfar K., Rashed A., Elzamy, A. and Elmasri, A. (2005) Customer Satisfaction with Internet Banking Web Site (Case study on the Arab Bank), the 4th International Multiconference on Computer Science and Information Technology CSIT 2006, Amman, Jordan. ISBN: 9957 - 8592 - 0 -X, National Number: 2129/9/2005
- [9] Maiyaki A. and Mokhtar S.(2010), Determinants of Consumer Behavioural Intention in Nigerian Commercial Banks, International Conference on Business and Economic Research (ICBER 2010), Malaysia (15 - 16 March 2010)
- [10] Mallat N., Rossi M. and Tuunainen V. (2004): Mobile banking services. Commun. ACM 47(5): 42-46.
- [11] Prashanth C. , Ganavi S., Mahalakshmi T. ,Raja K.,Venugopal K. and Patnaik L. (2009), Iris Feature Extraction Using Directional Filter Bank, for Personal Identification, Proceedings of the 2nd Bangalore Annual Compute Conference on 2nd Bangalore Annual Compute, Article No. 6, ISBN:978-1-60558-476-8
- [12] Ramayah T., Chin Y.L., Norazah, M. and Amlus, I. (2005), Determinants of Intention to Use an Online Bill Payment System among MBA Students, E-Business, Issue 9, pp. 80-91.
- [13] Rashed A. and Santos H. (2010a), Odour User Interface for Authentication: Possibility and Acceptance: Case Study, The International MultiConference of Engineers and Computer Scientists 2010 (IMECS2010), (The 2010 IAENG International Conference on Bioinformatics), Hong Kong.
- [14] Rashed A. and Santos H. (2010b), Multimodal Biometrics and Multilayered IDM for Secure Authentication, accepted, ICGS3 6th International Conference for on Global Security, Safety and Sustainability, 1-3 September 2010, Braga, Portugal.
- [15] Rashed A. and Santos H. (2010c), OTM Machine Acceptance: in the Arab Culture, accepted, ICGS3 6th International Conference for on Global Security, Safety and Sustainability, 1-3 September 2010, Braga, Portugal.
- [16] Rashed A. and Santos H.(2010d), Validating TAM with Odour Interface in ATM Machines, Global Journal of Computer Science and Technology GJCST Vol. 10 Issue 7: July/August,2010.
- [17] Röcker C. (2009), Perceived Usefulness and Perceived Ease-of-Use of Ambient Intelligence Applications in Office Environments, HCD 09 Proceedings of the 1st International Conference on Human Centered Design: Held as Part of HCI International.
- [18] Segev A., Porra J. and Roldan M. (1998), Internet Security And the Case of Bank of America, Internet security and the case of Bank of America, Volume 41, Issue 10, 1998,ISSN:0001-0782, pp: 81 - 87.
- [19] Sukhai N.(1998), Access Control & Biometrics, Proceedings of the 1st annual conference on Information security curriculum development, Kennesaw, Georgia, ISBN:1-59593-048-5, pp: 124 - 127
- [20] Szajna B. (1996), Empirical Evaluation of the Revised Technology Acceptance Model, Management Science, INFORMS, 42(1): 85-92.
- [21] Yaseen S. and Zayed S. (2010), Exploring Critical Determinants in Deploying Mobile Commerce Technology, American Journal of Applied Sciences 7 (1): 120-126.

MobiStream: Live Multimedia Streaming in Mobile Devices

Chrysa Papadaki

Department of Informatics
Athens University of Economics and Business
Athens, Greece
chrpapa@intracom.gr

Vana Kalogeraki

Department of Informatics
Athens University of Economics and Business
Athens, Greece
vana@aueb.gr

Abstract— In recent years, many techniques have been proposed so as to enable resource-constrained devices to consume or deliver live multimedia streams. The majority of the existing techniques use distributed multimedia services and powerful servers to handle streams on behalf of clients. This is due to the fact that, multimedia streaming, when smartphones act as both clients and servers, can generate many challenges due to the heterogeneity of the multimedia streaming protocols, the media formats and codecs supported by today's smartphones. In addition, multimedia processing is resource consuming and, in many cases, unsuitable for a plethora of resource-constrained devices. To overcome these challenges, we present **MobiStream** a device-to-device multimedia streaming system for resource-constrained devices that achieves efficient handling of live multimedia streams. The design of **MobiStream** architecture provides solutions to several issues including resource constraints, streaming among heterogeneous operating systems and platforms, generation, synchronization and presentation of multimedia streams. We have developed the **MobiStream** prototype system on Java 2 SE and Android platforms; this paper presents the implementation details and the experimental evaluation of our system.

Keywords-live multimedia streaming; Android platform; streaming protocol; resource-constrained devices.

I. INTRODUCTION

In recent years, the demand for real-time multimedia services, including voice over IP (Internet Protocol), audio and video streaming, has been growing rapidly so that multimedia streaming applications have become dominant in present communications systems. Furthermore, the explosive development of mobile networks and the availability of mobile devices in the hands of the masses, have made real-time multimedia delivery popular on mobile devices, such as smartphones and tablets, which have now become a major part of everyday life. It is an indisputable fact that cellular traffic is growing tremendously, with a share of video traffic increasing from 50% now to an expected 66% by 2015 [2]. Consequently, the demand for innovative smartphone applications that allow users to receive and deliver live or on-demand rich content has increased dramatically.

Today's smartphones are equipped with significant processing, storage and sensing capabilities, as well as wireless connectivity through cellular, Wi-Fi and Bluetooth.

They provide ubiquitous Internet access, primarily through their cellular connection and secondarily through Wi-Fi, and enable a plethora of distributed multimedia applications. However, the acquisition and transmission of large amounts of video data even on modern mobile devices create important challenges. Issues like resource allocation, energy consumption, CPU, memory and bandwidth constraints, as well as the software development platform must all be taken into consideration. It is, therefore, essential to address these challenges by efficiently managing the resources and employing effective streaming techniques.

Current solutions for mobile multimedia streaming assume a centralized architecture where a resource-powerful server can support heterogeneous sets of media encoders, decoders and streaming protocols and is able to adapt content on behalf of clients to provide multimedia streams to resource-constrained mobile devices [6][12]. On the other hand, solutions for multimedia streaming over ad hoc networks assume the existence of distributed multimedia services and require cooperation between mobile devices for content dissemination; however, these either do not consider the scenario of content adaptation [7] or are cross-layered [8]. Din and Bulterman [11] demonstrate the use of synchronization techniques for distributed multimedia, but without addressing the issue of energy reduction. Recently, lightweight middleware targeting mobile multimedia applications have been proposed to address the issues of heterogeneity on modern smartphones. One of the latest efforts is the **Ambistream** middleware [9], which provides an additional layer as an intermediate protocol and the associated container format for multimedia streaming among heterogeneous nodes. For the generation and presentation of the multimedia streams, **PacketVideo OpenCore** [13] and **Stagefright** [14] multimedia frameworks are used, respectively. Moreover, these multimedia frameworks are based on cross-platform solutions. One of them is **FFmpeg** (Fast Forward MPEG) [15], which is an Open Source lightweight multimedia framework that allows encoding, multiplexing and streaming of videos in different formats. However, **FFmpeg** has several limitations; it does not support a wide range of audio/video codecs, especially for Android devices and is better suited for streaming from a single source.

Multimedia streaming is a challenging problem when smartphones act as both clients and servers. This is due to

the fact that, the framework needs to be integrated into multiple mobile platforms to provide live streaming among multiple smartphones because of the variability of the supported media formats, codecs and streaming protocols. In addition, multimedia processing, especially in the case of handling streams of high-quality content, is resource-consuming and needs to be carefully handled in the case of mobile devices. To address the above challenges, in this paper, we present MobiStream, a mobile-to-mobile live multimedia streaming system that enables mobile devices to easily handle live multimedia streams leveraging the available multimedia software stack of the applied platform. We assume the scenario of a mobile device that requests to deliver a live multimedia stream to one or more peers. In fact, MobiStream enables mobile devices to act as both clients and servers and allows clients to process and deliver live multimedia streams to mobile devices or desktop servers, while considering resource constraints. An important feature of MobiStream is that it can also materialize the scenario of live multimedia streaming over an ad hoc network. For example, the Android Ice Cream Sandwich devices provide peer-to-peer (P2P) connectivity using WiFi Direct [10], so, either a laptop or an Android device can easily act as a virtual access point (AP). Thus, the system using nodes that act simultaneously as servers and clients can support this kind of scenarios. The streaming client in our approach does not act as a relay client for other phones. Taking all the above into account, we envision a system that provides sustainable solutions to a wide range of applications, such as streaming a live event directly to other devices reachable on the network, voice and video call applications, private audio-visual communication between peers without involving third party servers, sharing live multimedia content in cases of unavailable infrastructure, etc. We have implemented our prototype system that is running on both Android and Java 2 SE platforms to demonstrate the feasibility of our approach.

The rest of the paper is structured as follows. In Section II, we describe the system design in detail and discuss several design issues concerning the generation, transmission, synchronization and presentation of the live multimedia streams and the choices we made to address them. Section III demonstrates our approach on the synchronization of the streams. In Section IV, we present the prototype system we have implemented and discuss implementation details, including challenges specific to Android phones. In Section V, we present the system performance evaluation results of our testbed for a range of scenarios and conclude the paper in Section VI.

II. SYSTEM DESIGN

A. System Overview

MobiStream is structured in a client-server model, where devices are able to act as servers and clients simultaneously. These can communicate over cellular or WiFi. Each device can assume both roles, as it can be a client, when uploads content to a server, or a server, when it receives one or multiple media streams from the clients. The Client consists of the Dispatcher component, the Synchronization Module and the Media Recorders. The Dispatcher is responsible for communicating with the Server and packaging and transmitting the generated Media Units (MUs). The MUs are produced by the Audio and Video Recorders which are independent sub-applications of the Client. The Synchronization Module is responsible for synchronizing the generated media units before the final stage of transmission. The Server is designed to run on mobile devices as well as desktop computers. It comprises the Receiver component, the Sync Manager and the Media Players. The Receiver component is used to listen for incoming client requests, using a built-in TCP Server which is running independently in the background, and depackages and separates the received MU packets. The Sync Manager is responsible for the synchronization of the received MUs, while the Audio and Video Players are in charge of the presentation of the final synchronized multimedia stream. Both clients and servers are multithreaded so as to enable the server to receive multimedia streams from many clients and the client to transmit to multiple destinations. Fig. 1 illustrates the overall system architecture.

In the remaining of this section, we give an overview of the building blocks and the interaction between them.

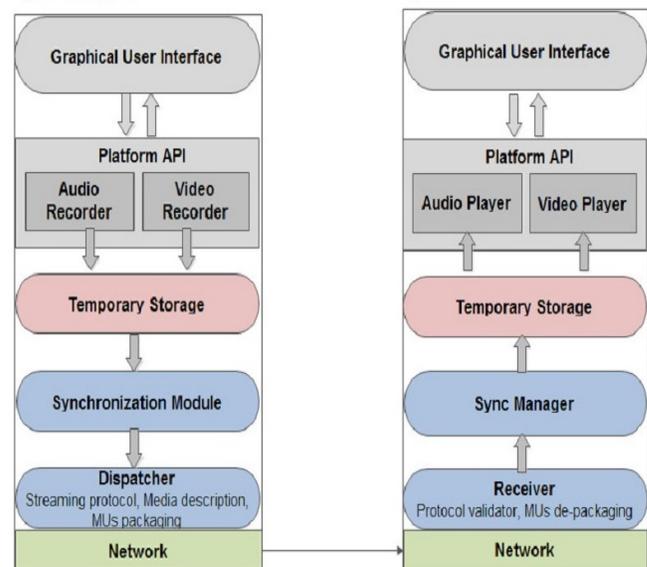


Figure 1. MobiStream Framework Architecture. Streaming Client (left) – Server (right)

B. Streaming Client

The Streaming Client is in charge of generating multimedia streams and transmitting them to the Server. More specifically, it comprises the following main components:

Media Recorders: one of the first design challenges we faced was the design of media components for Android devices that would enable the generation of live video streams. Currently, the available APIs (Application Programming Interface) of the latest Android SDK do not include specifications to allow developers to capture fragments of live video streams. To circumvent these problems, we designed and developed our media components, which are able to produce and consume media units of specific formats. Thus, for the audio recording, we designed the Audio Recorder, a component that records uncompressed PCM (Pulse-code modulation) frames of fixed size from the input hardware device and stores them in a concurrent data structure used in parallel with the Synchronization module (discussed below). For the video recording, we designed the Video Recorder, a component that obtains an instance of the hardware input camera, sets camera parameters, frame rate and preview resolution, starts updating the preview surface and simultaneously capture the preview frames and stores them. The module that captures preview surface frames actually captures a sampling of the video, consequently a lower-quality video than the expected is being produced and second, during the video recording, the FPS (frames per second) vary, and that would significantly affect the smoothness of the video play out.

Synchronization module: we designed the Synchronization Module in order to eliminate the variability of video capture rate and synchronize the audio and video streams. The Synchronization Module is responsible for monitoring video and audio in order to capture the rate, based on the formulas we discuss in the next section, and propagate the frames and the samples to the packaging stage at the Dispatcher application.

Dispatcher: the main responsibility for the Dispatcher is to establish a connection, setup a multimedia session and packetize the media units, that polls from the local buffers. The co-operation of Dispatcher and Synchronization module results in the transmission of the synchronized multimedia streams. The overall technique for the synchronization at the client side is described in details in Section III.

C. Server

A significant feature of our proposed Server design is that it is modular and platform-independent. The Server is multi-threaded in order to be able to present more than one multimedia streams from different sources. This component is responsible for handling client requests, configuring the

requested multimedia sessions, receiving and reconstructing the multimedia streams, and displaying feedback during the experiments.

Receiver: the Receiver is in charge of handling incoming connection requests, de-packetizing the incoming RTP packets using a packet Validator module, and separating the streams by drawing information from the header. Then, the receiver provides Sync Manager with the received MUs in order to proceed to synchronization stage.

Sync Manager: the Sync Manager is one of the most significant components of our system as it is used to address several major problems related to synchronization of the media units and the presentation of the final stream. It consists of a multimodal functionality as described below. In case of an unreliable link for the uploading of the multimedia streams, the Receiver enables the entire functionality of the Sync Manager in order to execute the audio/video synchronization algorithm we discuss in Section III, so as to prevent the out-of-order presentation of the MUs and the lack of synchronization between audio and video. Given the video frame rate, the Sync Manager is able to compute the video and audio playback time in order to achieve the same temporal correlation of MUs as at the transmission point and synchronize them in order to be played by the Media Players without letting network delays affect the video presentation. In the case of a reliable connection, the Sync Manager assumes that the packets arrive in order, as the underlying protocol is TCP, so, it decides not to use the synchronization algorithm and only adopts a buffering technique in order to synchronize the media streams and provides them to Media Players in a constant rate which represents the playback rate of the multimedia stream at the origin. The proposed buffering technique is presented in the next section in detail.

Media Players: we designed these components in order to enable the presentation of multimedia streams of PCM and JPEG units at the receiver end. For both players we followed a producer-consumer design, using concurrent data structures. The Sync Manager is the producer that produces the MUs in order and the players are the consumers that consume the available media units. For video presentation, we created a user interface handler that updates the video screen when a new video frame is available. For audio play out, we designed an Audio Player that is able to play audio samples in a specific frame format and sampling frequency (discussed in details in the next section).

III. PROPOSED APPROACH

The system follows a client-server model of two independent audio and video decoders. Using multi-

threaded software, we managed to accelerate the process of video reconstruction by separating the multimedia streams, synchronizing them whenever required, at negligible CPU overhead, as we show in our experimental evaluation, and executing parallel decoding of each stream. This way, an application based on this system is able to run efficiently on resource-constrained devices minimizing the processing overhead and reduce processing delays, which are critical for real-time multimedia applications. Apart from software architecture and computer performance, another significant contributory factor to live multimedia streaming is the network availability. The Internet, like other packet networks, occasionally loses and reorders packets and delays them by variable amounts of time. To overcome these impairments, we designed a protocol for real-time communication following the Real-Time Transport Protocol (RTP) specifications [1] that provides end-to-end delivery services for data with real-time characteristics, such as interactive audio and video.

A. Proposed Real Time Protocol

One important feature of our real-time protocol was to provide a way to reconstruct audio and video streams with a controlled delay for play out. To achieve this goal, we use the RTP header to packetize MUs in order to provide the receiver with payload identification, timing information and a sequence number, the last two allow receivers to calculate packet losses and jitter as well. Although the proposed protocol follows the general design of RTP, it does vary in several major ways.

First, RTP does not provide any mechanism to ensure timely delivery or provide other Quality-of-Service guarantees i.e. prevention from out-of-order delivery. It actually uses underlying protocols, usually UDP, for transport and multiplexing functionality. In an audio/video session [3] as opposed to [5] where an algorithm is proposed for synchronizing of streams carried in separated sessions. This type of streaming is acceptable over low bandwidth communication channels. Thus, to begin live streaming, the establishment of one end-to-end connection over either TCP or UDP is required. In addition, each device is able to start multiple sessions to transmit video to different destinations. To achieve multimedia streaming in one session, we had to keep the payload type constant and allocate different values to the synchronization source identifier (SSRC) field regarding the media type of the payload. In comparison to RTP specifications where if both audio and video media are used in a conference, they are transmitted as separate RTP sessions, therefore SSRC identifier is a randomly chosen value meant to be globally unique within a particular RTP session. In Table I, we describe the attributes of the header we use to packetize the media units. Our goal in the streaming protocol is to support live multimedia services either over TCP or UDP.

TABLE I. PACKAGING ATTRIBUTES

Name	Size	Description
payload type	1 byte	This field identifies the format of the RTP payload and determines its interpretation by the application. It holds the same value for all packets regardless of the media payload type, because all packets represent one multimedia stream.
sequence number	2 bytes	The sequence number increments by one for each data packet sent, and may be used by the receiver to detect packet loss and to restore packet sequence
time stamp	4 bytes	The timestamp reflects the sampling instant of the first octet in the RTP data packet. The sampling instant MUST be derived from a clock that increments monotonically and linearly in time to allows synchronization and jitter calculations
SSRC	4 bytes	The SSRC field identifies the synchronization source. This identifier should be chosen randomly, with the intent that no two synchronization sources within the same RTP session will have the same SSRC identifier
Payload	N bytes	Data

We implement a buffering technique that we discuss in the next section, consisting of two major parts. The first part refers to a dispatcher-side buffering in order to facilitate the synchronization of the generated MUs and the second part concerns the adoption of a receiver-side buffer to accommodate initial throughput variability and inter-packet jitter. The experimental results we conducted shown that the proposed buffering technique can be integrated into applications using TCP-Friendly transmission of multimedia streams, and benefit from TCP mechanisms as it is reliable and guarantees delivery of packets in order. However, using TCP as transport layer may induce long delays because of the TCP retransmission mechanism. This actually leads to long video pauses at the receiver-end, which highly degrade the real-time communication. To cope with this issue, we monitor the transmission delay between successive incoming packets and drop those that are late with respect to specific thresholds, we discuss later, related to the actual time user conceives. As far as the scenario of using a UDP-based streaming protocol is concerned, we adopt the proposed streaming protocol over UDP using the buffering technique, we discuss in the next session, and the time-oriented audio and video synchronization algorithm that we present in Section C.

B. Buffering Technique

One of our major design challenges was how to create a synchronized multimedia stream with a constant playback rate produced by two different media sources, as the capturing and coding rate on each source is different and induces variable delays. To address this problem, we first synchronized the camera and microphone capture rates by setting up our system's audio recorder appropriately so as to

capture audio samples depending on the capture frame rate of smartphone's camera. Moreover, we provide a client-side buffering so as to adjust multimedia stream capture rate by prefetching multimedia data into a buffer in a controlled rate, which represents the playback rate at the receiver. This assures the elimination of the variable delays induced by sources. Thus, media streams have well-defined temporal relations among themselves and can be sent synchronized to the server. More precisely, the relation among the audio samples, video frames and playback time is given by the following formulas:

$$VP_i = V_i / VR \quad (1)$$

where VP_i is the video playback time of the i^{th} video frame in seconds, V_i is the i^{th} video frame number which is an integer that increases by one representing the i^{th} generated video frame and VR represents the video frame rate (Frames per second) of the source. In practice, applying the (1), the system is able to accurately calculate the playback time of a particular video frame in seconds. To calculate the audio playback time, AP_j , of an audio frame, we use (2), where the $num_samples$ represent the number of the encoded audio samples of 16-bit each of the produced PCM frame. In our approach, in stereo mode, a PCM audio frame contains 512 samples and, in mono mode, a frame contains 1024 mono samples, thus, it follows that each audio frame consists of 2048 bytes minimum. This size applies to all fragments of the audio stream. Note that using Android Media package, data should be read from the audio hardware in chunks of sizes subject to the total recording buffer size. In (2), A_j is the j^{th} audio frame number which is an integer that increases by one representing the j^{th} generated audio frame and sampling frequency corresponds to the produced samples per second (Hz).

$$AP_j = num_samples \times (1 / \text{sampling frequency}) \times A_j \quad (2)$$

Taking the above-mentioned into account, we conclude to (3), which calculates the audio frame that must be presented in the VP_i^{th} second in order to achieve synchronization.

$$A_j = VP_i \times \text{sampling frequency} / num_samples \quad (3)$$

Using the above formulas, the Synchronization module of the Client application is able to estimate the correlation among the produced MUs and provide the Dispatcher with a synchronized multimedia stream so as to transmit the MUs in the right order so as to be presented in sync at the Receiver, in case of transmission under ideal circumstances, no further processing would be required at the Server in order to present a synchronized multimedia stream. Nevertheless, a critical aspect lies in the lack of synchronization that may exist between audio and video streams at the receiver-end due to the fact that the characteristics of IP-based network, delay and jitter, affect the temporal relations present in multimedia streams. To

circumvent these problems, we use a receiver buffer for the temporary storage of incoming media units comparing (1) to (2). In practice, the Sync Manager of the Server checks whether the playback time of a newer video frame is the same with the playback time of the corresponding audio frame. If this is the case, it follows the presentation of MUs at the proper time. The use of a MU buffer introduces some delay in the application, which is directly proportional to the size of this buffer. The objective of the process is to provide a presentation that resembles as much as possible the temporal relations that were created during the encoding and multiplexing process at the Client.

C. Audio/Video Synchronization Algorithm

In our system, the real-time delivery of the packets can be accomplished by using either TCP or UDP as the transport layer. Taking for granted that the media streams are synchronized at the origin, we need to achieve the same temporal correlation for playback at the receiver. This can be a quite difficult issue when the system performs transmission over UDP, which is unreliable and does not provide Quality-of-Service mechanisms, such as prevention from out-of-order delivery of packets. To cope with this challenge, we propose the following synchronization algorithm which imposes negligible CPU overhead, as shown in experimental results below, which is important as we have to deal with resource-constrained devices and real-time communication. In order to ensure a better quality of the reconstructed material, priority is given to audio information. We chose audio stream to be played regardless of the state of the video because human perception is more sensitive to degradation in audio quality than in video [4]. This means that audio would be played upon arrival as long as it is in order, regardless of the state of the video stream. In practice, if the audio stream anticipates the video stream, the receiver simply discards the video packets.

In the case of receiving a video packet, first, the audio/video synchronization algorithm checks the SSRC field of the packet header in order to determine whether the payload contains audio or video data. Then, it checks if the received video frame is newer than the displayed one by comparing the new timestamp with the old one. If this is the case, it calculates the video and audio playback times, using (1) and (2), respectively. If the audio is ahead of the video, the algorithm calculates the difference between their playback times, $AP_i - VP_i$. In the case of $AP_j - VP_i > \text{threshold}$, where threshold is the maximum level at which humans detect frames as being in sync, the video is considered too old to be displayed and it is dropped, otherwise it is rendered. The threshold is tuned based on the application characteristics. In [4], a detailed study of the end user capability to detect harmful impacts of de-synchronization on QoE (Quality of Experience) is provided. The author indicates that an absolute skew smaller than 160 ms is harmless and greater than 320ms is harmful for QoE. The author identifies a double temporal area [-160,-80] and [80,160] called transient, in which the impact

of the skew heavily depends on the experimental conditions.

IV. IMPLEMENTATION

Our software architecture was motivated by the need to have a simple and platform-independent implementation. We chose Java as the development language. The object oriented features of Java and its simplicity enables our system to be simple and modular. Thus, MobiStream can run on any platform that supports Java and requires a real-time streaming protocol for multimedia services. The software for the smartphones is an Android application that enables the device to act simultaneously as client and server and runs efficiently on Android v2.3 or later versions. For the laptop server, we used in some experiments, the software runs on Java 2 SE. We have also developed a graphical user interface (GUI) and the code for the media components.

In this section, we describe the implementation details, the major challenges we faced specifically on Android phones, and the design choices we made to address them.

A. The Streaming Process

The phases required to complete the streaming process between two devices are media capture, media transmission and media presentation. In this section, we describe the implementation details of each phase and the technical problems we encountered.

1) Media Capture

Media content originates from hardware input devices, that is, camera and microphone. In most multimedia applications, the media capture phase is implemented using available APIs that provide access to built-in Multimedia Recorders that supports several media formats, encoders and streaming protocols in order to provide playable stream formats to Media Players. Developing on Android platform, we faced two major issues. First, the lack of hardware accelerated codec APIs when we implemented the prototype system and, secondly the fact that the exposed APIs do not provide the ability to stream live multimedia content from the built-in Media Recorder in a format playable from the built-in Media Player. To overcome these issues, we have implemented two independent Media Recorders. Each one is able to draw input from a different hardware device and use media formats and encoders supported by all platforms.

For the video recording, we used the Camera APIs to set image capture settings, start/stop preview and retrieve frames for encoding for video. An instance of the camera is actually a client for the Camera service, which manages the actual camera hardware. We install a callback to be invoked for every preview frame, using pre-allocated buffers, in addition to displaying them on the screen. The callback will be repeatedly called for as long as preview is active and buffers are available. The purpose of this method is to improve preview efficiency and frame rate by allowing preview frame memory reuse. The image format for preview pictures is either NV21 or YV12, since they are supported

by all camera devices. To reduce the size of the video images, we use a JPEG encoder. The video frame size depends on the video resolution and the quality of the compressed data.

For the audio recording, we used the AudioRecord class of the Android SDK which manages the audio resources for Java applications to record audio from the audio input hardware of the platform. This is achieved by reading the data from the AudioRecord object. Upon creation, an AudioRecord object initializes its associated audio buffer that it will fill with the new audio data. The size of this buffer, specified during the construction, determines how long an AudioRecord can record before "over-running" data that has not been read yet. Data should be read from the audio hardware in chunks of sizes inferior to the total recording buffer size. Thus, the Audio Recorder captures uncompressed PCM samples of a specific sampling rate and size. In our prototype system, we set the sampling rate and the size of the recorded samples accordingly to the video frame rate in order to facilitate the synchronization process, as described previously. The captured MUs are stored in concurrent data structures so as to enable the co-operation of the modules involved in capture and transmission phases.

2) Media Transmission

At the end of the capture phase, since the MUs cannot be directly transmitted over IP-based networks, they are wrapped within media containers that provide the necessary meta-information to facilitate the decoding and correct presentation at the receiver end. This task is assigned to the module that packages the media units following the specifications of the real-time streaming protocol we discussed previously. At the server side, the receiver performs the de-multiplexing and de-packaging process and provides the separated media streams to the Sync Manager in order to synchronize them before the presentation phase. A contributory factor to the efficiency of the collaboration among the modules of the different phases is the use of Android Services, which are independent application components that host the main processes of our system and execute long-running operations while not interacting with the user.

3) Media Presentation

Using the above-mentioned Media Recorders, the proposed real-time streaming protocol and the synchronization algorithms we discussed previously, the system is able to reproduce the initial media streams and proceed to the presentation phase. In order to present the MUs, we developed two independent Media Players. For the video playback, first the decoding of the compressed data from the playback buffer takes place and then the User Interface Handler which extends the Handler class of Android SDK updates the video view. This process is executed as soon as there is a new video frame in the playback buffer. For the audio playback, we developed an Audio Player, using the AudioTrack class of the Android SDK which manages and plays a single audio resource for

Java applications. It actually allows streaming PCM audio buffers to the audio hardware for playback.

B. Streaming Protocol

We used the java.net library to implement a library that provides a streaming protocol for real-time applications, based on Real-time Transport Protocol, for multimedia services and can be ported to any platform supports Java and its network libraries. Using this library, the system is able to set up, start and handle multiple unicast sessions using UDP or TCP as the transport layer, and transmit multimedia data supporting a wide range of media formats for the packaging and de-packaging stages, even though in the prototype system we used specific formats in order to facilitate the porting of the live multimedia streaming process to different platforms.

V. EXPERIMENTAL RESULTS

A. Experimental Setup

We have conducted a set of experiments in order to evaluate the efficiency and performance of MobiStream. The testbed of the experiments is presented in Table II. Additionally, we provide screenshots of the android application in Fig. 2. This setup can be used in various scenarios, for example, in streaming video, in mobile video, e-health, assistive technologies. First, we assume a Streaming Client running on an Android-powered device that uploads live multimedia streams to a Server. We conducted a series of experiments using different levels of signal strength - weak, medium and strong using TCP, monitoring the five bar scale of the smartphones which basically measures radio signal levels maintained by the wireless network adapter, in decibels (dB) on a more linear scale. For each experiment, we report the averaged results of five runs. We also repeated the process using UDP and compared the results. To run this test we used the Xperia Neo V Android smartphone as Streaming Client and the Samsung Windows 7 laptop as Server. The second scenario concerns a Streaming Client delivering a live multimedia stream to multiple receivers of the network. The network comprises a wireless access point (*i.e.*, router), a streaming client (Xperia Neo V) and 5 to 20 receivers. We executed the experiment using a different number of receivers so as to record the end-to-end delay and the jitter, in order to investigate how these measurements affect the quality of the video at the receiver. In all cases, no external peers injected traffic in the network the server allows a few seconds (3s to 5s regarding the signal strength) startup delay, which is a common practice in commercial streaming products. All packets arriving earlier than their playback times are stored in the server's local buffer. In comparison to Ambistream in which a 30s start-up delay is introduced by the middleware layer to allow protocol translation. This aspect restricts the

TABLE II. TEST DEVICES

Test Devices	Sony Ericsson Xperia Neo V	HTC Explorer	Samsung NP300V5A-S05
Role	Client/Server	Client/Server	Server
Platform	Android 4.0.4	Android 2.3.5	Windows 7
CPU	1 GHz	600MHz	I5-2450M 2.5GHz
Memory	420MB	256MB	4GB

use of the middleware for real-time applications. The multimedia stream has a QCIF (176 by 144) frame size in 200kbs and 400kbs video bitrates, whereas in 600kpbs, 800kpbs and 1000kpbs we apply a CIF (352 by 288). The stream duration is 180 seconds and the video capture rate varies accordingly to the video bitrate presented in the experimental results; in total, more than 12 hours of streaming required among the testing devices.

D. Experimental Results

1) System Evaluation

We first present the experimental results of the mobile-to-server scenario. We focus on the following Quality of service metrics: end-to-end delay (*i.e.*, the time taken for a packet to be transmitted from the client to the server), the jitter (*i.e.*, packet delay variation measured at the server) and the download rate (*i.e.*, the transmission bitrate measured at the server). In Fig. 3 and Fig. 4, we present the download rate of the desktop server using TCP and UDP respectively. We chose a high video bitrate of approximately 1100kpbs, in order to evaluate network throughput. In case of using TCP, Fig. 3 clearly depicts the behavior of the transport protocol in the weak signal strength case, as it shows intense variability of the download rate induced by the retransmission mechanism of TCP. In the medium and weak signal strength cases, the download rates recorded were 4,96% and 17,97% lower than the rates observed in strong signal strength case. In case of using UDP, we observe from Fig. 4 that the download rates in medium and weak signal



Figure 2. (a) The mobile screen while recording a live event and the video window of the server running on a desktop, (b) Server Configuration screen, (c) Session Configuration screen, (d) Statistics screen.

download rate is higher regardless of the signal state due to the client-side buffering employed in the framework. Fig. 5 illustrates the jitter for different packets using TCP. In weak signal strength we recorded high values of jitter, e.g., 786ms, at 387th packet. This fact entails long pauses at the video presentation. Nevertheless, our proposed approach discussed in Section III accomplishes a good quality of the video stream without degrading the real-time communication. In medium signal strength, the highest absolute values of jitter are smaller than the values recorded in weak signal state. In strong signal strength, the highest positive value of jitter recorded was 40ms. Regarding the second scenario of the use of multiple server applications running on the network, we measured the end-to-end delay in case of 5, 10, 15, 20 receivers using TCP. Fig. 6 presents the mean end-to-end delay for different numbers of servers running in the network. The end to end delay remains within acceptable bounds in terms of video quality and Quality-of-Experience and increases proportionally to the number of receivers, approximately 28% from 5 to 10 receivers, 30% from 10 to 15 receivers and 48% from 15 to 20 receivers.

2) Evaluation of Memory and CPU usage

We also measured the resource usage of our approach. We run the experiments using the HTC Explorer smartphone described in Table II. Fig. 8 illustrates that the memory usage at the Server side remains constant. For higher data-rates, the memory usage may increase slightly because of the higher buffer sizes required. In the case of the Client application, the memory usage increases proportionally to video capture rate (including only JPEG data). In both applications, the framework re-uses the pre-allocated space in RAM in order for the multimedia application to be able to run under memory constraints, as in this scenario we run the experiments using a smartphone with 256MB RAM. Fig. 9 depicts the CPU overhead on both client and server mobile applications versus the video bitrate. In all experiments we observed slightly higher percentage of CPU overhead in Client application, this is due to the use of the hardware input camera and the YUV compression module. Nevertheless, in both applications when the video bitrate is greater than 700kbps the CPU overhead tends to be the same. In order to accurately estimate the CPU usage of the framework during the live streaming process, we divided the CPU monitoring into three phases; (I) initialization of media components, (II) streaming process, (III) media components finalization. In both client and server applications the CPU usage during the first phase were 67% and 55%, respectively. The second phase is illustrated by Fig. 9 and includes, from the client point of view, the recording, storing, packaging and transmission of the media units. Regarding the server application, the second phase includes the de-packaging, the synchronization, the storing and the presentation of the received media units. For the third phase, the server and client required approximately 55% and 68% CPU usage, respectively.

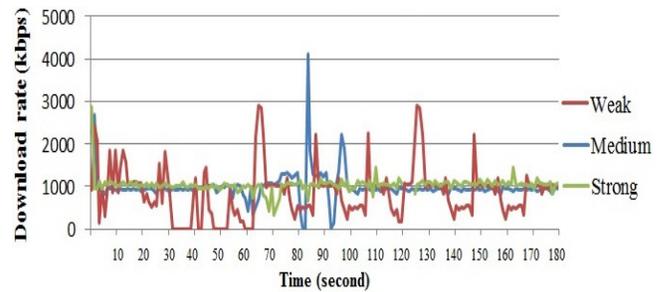


Figure 3. Download rate (kbps) - Signal Strength, using TCP.

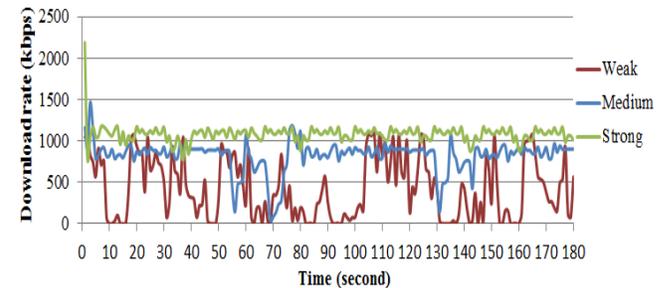


Figure 4. Download rate (kbps) - Signal Strength, using UDP.

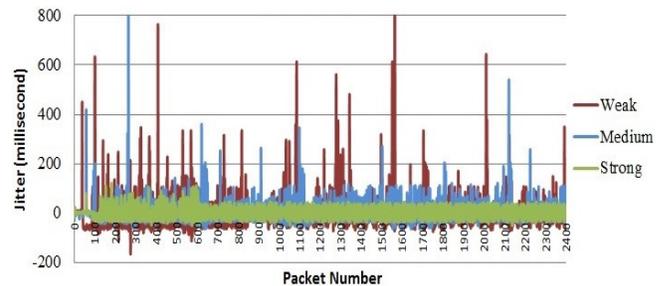


Figure 5. Jitter(ms) - Signal Strength, using TCP.

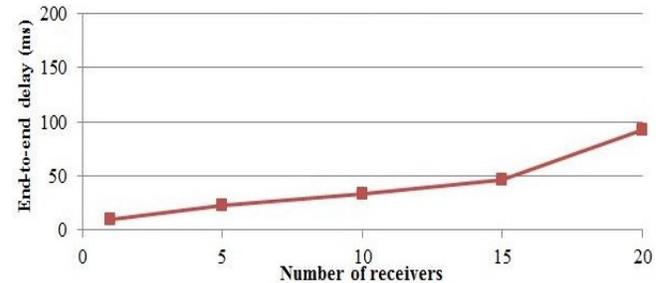


Figure 6. End to end delay (ms) – Number of Receivers, using UDP.

3) Evaluation of Energy Consumption

In the last set of experiments, we measured the energy consumption of our approach. We executed the scenario of mobile-to-mobile server running on smartphones and before the experiment both smartphones were fully charged. During the experiment, the battery states are recorded every 10 seconds. Fig. 7 presents the battery state as a function of time. The 100% percent corresponds to the fully charged battery. We chose a high video bitrate of 1100kbps and run

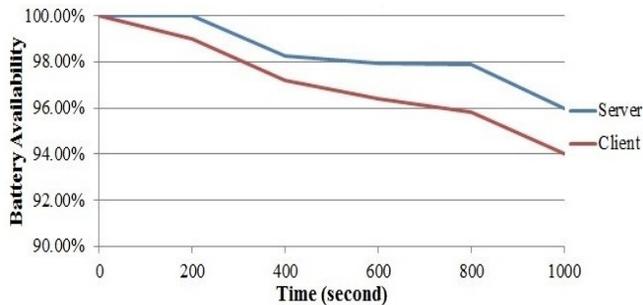


Figure 7. Battery Level (%) – Video bitrate (kbps)

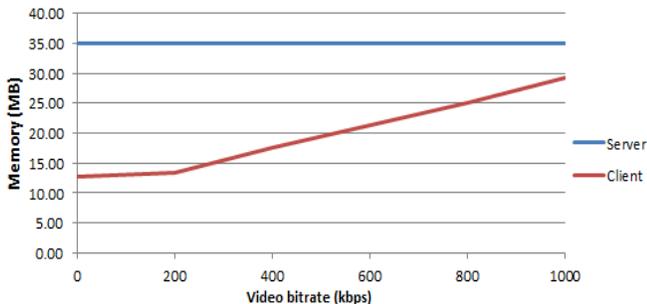


Figure 8. Memory (MB) – Video bitrate (kbps)

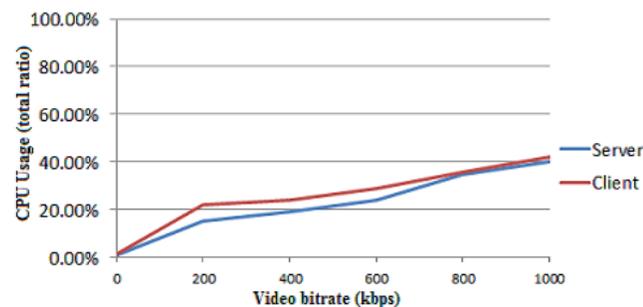


Figure 9. CPU Usage (total ratio) – Video bitrate (kbps)

each experiment for 16.6 minutes. Fig. 7 depicts that the Server hardware input Camera and framework's Audio Recorder compared to the Server application in which the main energy consuming component is the Audio Player.

II. CONCLUSION AND FUTURE WORK

In this paper, we designed, implemented, and evaluated a mobile multimedia system, MobiStream that enables resource-constrained devices to handle real-time multimedia streams. We designed a platform-independent framework so that we can support live multimedia streaming among heterogeneous mobile devices. We present our approach on

the synchronization of the media streams and the streaming process we employed. Our experimental results demonstrate significant performance benefits in terms of the usage of the mobile devices' resources and video quality. For our future work, we plan to evaluate the working of our approach using a larger number of heterogeneous mobile devices.

ACKNOWLEDGMENT

This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) – Research Funding Program: Thalys-DISFER, Investing in knowledge society through the European Social Fund.

REFERENCES

- [1] H. Schulzrinne, S.L. Casner, R. Frederick, and V. Jacobson. "RTP: A Transport Protocol for Real-Time Applications", IETF Request for Comments: RFC 3550, Jul. 2003.
- [2] Cisco Systems, "Cisco visual networking index: Global mobile data traffic forecast update", 2011-2016. <http://www.cisco.com>.
- [3] M. Westerlund and C. Perkins "Multiple RTP Sessions over a single Transport flow", Ericsson, University of Glasgow, Nov. 2011.
- [4] R. Steinmetz, "Human perception of jitter and media Synchronization", IEEE Journal on Selected Areas in Communications, vol. 14, no. 1, 1996, pp. 61–72.
- [5] R. Bertoglio, R. Leonardi, and P. Migliorati, "Intermedia Synchronization for videoconference over IP", Signal Processing: Image Communication, Sept. 1999, pp. 149-164.
- [6] K. Curran and G. Parr, "A Middleware Architecture for Streaming Media over IP Networks to Mobile Devices", IEEE Int. Conf. Wireless Communications and Networking, Mar. 2003.
- [7] N.M. Do, C.H. Hsu, J.P. Singh, and N. Venkatasubramanian, "Massive live video distribution using hybrid cellular and ad hoc networks", IEEE International Symposium on World of Wireless, Mobile and Multimedia Networks, Jun. 2011.
- [8] J. Domingo-Pascual, P. Manzoni, S. Palazzo, A. Pont, and C. Scoglio "On the Forwarding Capability of Mobile Handhelds for Video Streaming over MANETs", 10th International IFIP TC 6 Networking Conference, May 2011.
- [9] E. Andriescu, R. Cardoso, and V. Issarny, "Ambistream: A Middleware for Multimedia Streaming on Heterogeneous Mobile Devices", Middleware, volume 7049 of Lecture Notes in Computer Science, pp. 249-268, 2011.
- [10] C. Fragouli and E. Soljanin. "Network Coding Fundamentals." Now Publishers Inc, Delft, The Netherlands, Jun. 2007.
- [11] S.U. Din and D. Bulterman, "Synchronization Techniques in Distributed Multimedia Presentation", IARIA MMEDIA 2012, Apr. 2012, pp. 1-9.
- [12] T.E. Truman, T. Pering, R. Doering and R.W. Brodersen. "The InfoPad multimedia terminal: a portable device for wireless information access", IEEE transactions on computers, Oct. 1998, pp. 1073-1087.
- [13] PacketVideo Corporation, "PacketVideo OpenCORE Multimedia Framework", <http://www.opencore.net/>.
- [14] D. Hobson-Garcia, K. Matsubara, T. Hayama and H. Munakata. "Integrating a Hardware Video Codec into Android Stagefright using OpenMAX IL", http://elinux.org/images/5/52/Elc2011_garcia.pdf.
- [15] FFMPEG, "Developer Documentation", <http://www.ffmpeg.org>.

H.264 Parallel Optimization on Graphics Processors

Elias Baaklini*, Hassan Sbeity† and Smail Niar*

* University of Valenciennes, 59313, Valenciennes, Cedex 9, France

{elias.baaklini,smail.niar}@univ-valenciennes.fr

† Arab Open University, Beirut 2058 4518, Lebanon

hsbeity@aou.edu.lb

Abstract—Multimedia applications are present in most mobile hand-held devices. The H.264 standard is currently dominating the video compression world. H.264 has high computational complexity requiring large amount of processing resources. Many techniques emerged that optimize H.264 using parallelization on multicore systems ranging from groups of pictures until the smallest block of pixels. We propose a parallelization technique based on rows of macroblocks with a light dependency detection algorithm that optimizes data parallelization and minimizes dependency synchronization stall time. The parallel H.264 implementation is tested on 2, 4, 8, and 16 cores processors using CIF and HD video resolutions benchmarks. The experimental results show that, in terms of execution time and parallel scalability, CIF video sequences peak at 4 cores with a speedup of 3.1 and HD video sequences peak at 8 cores with a speedup of 6.2. The H.264 parallel implementation is then tested on a graphics processor simulator of the Evergreen family of AMD GPUs reaching a speedup up to 12.1 times without communications overhead. Our results shall aid to find the best parallel configuration of the H.264 standard with the most suitable multicore platform to use in terms of time complexity and parallel efficiency.

Keywords—Multimedia; H.264/AVC decoder; Video Compression; Optimization; Parallel Computing; Graphics Processors

I. INTRODUCTION

Multimedia hand-held devices are nowadays becoming more and more pervasive in many of modern world societies. Smart phones and tablet devices are equipped with high screen resolution and with relatively fast multicore embedded processors. DVD and blu-ray players, digital cameras, and LCD TVs support high resolutions like HD and Full-HD. However, few multimedia applications benefit from the computational potentials that multicore processors offer in these emerging powerful embedded devices. Furthermore, video coding standards like H.264/AVC [2] and HEVC [3] are adopting complex algorithms like context-adaptive binary arithmetic coding (CABAC) and variable length coding (CAVLC) in order to achieve better compression and thus lower transmission bitrates for high resolution video sequences. The additional complexity of these algorithms has a major impact by increasing execution time and energy consumption.

In our research, we intend to solve the problem of high complexity of the H.264 decoder using parallelization on multicore embedded processors and on graphical processors. Even with new cutting-edge processors, video resolutions are increasing rapidly, which require more processing time and consequently more energy consumption. Many solutions based

on parallel execution exist ranging from macroblocks (fine-grain) till groups of pictures (coarse-grain) parallel decoding. Macroblock parallel decoding is highly scalable since many macroblocks can be processed in parallel. However, dependencies and huge overheads are created as a result of communication and synchronization between macroblocks. Parallel decoding of groups of pictures require large memories for high definition video sequences. In addition, they have a lower scalability than macroblock decoding because of the limited number of groups of frames that can be decoded in parallel. Our solution is to decode macroblock rows in parallel. This level of parallel execution is considered between the coarse-grain and the fine-grain parallelization approaches. It also offers a balance between large overheads and high scalability of previous solutions.

Our main contribution in this paper is the design of a new approach for the parallelization of the macroblock rows of the H.264 decoder with an algorithm that detects dependencies on-the-fly based on isolating intra-prediction macroblocks (I-MBs). Experiments are conducted using simulations on 2, 4, 8, and 16 cores processors. We further experiment our parallel implementation on a graphical processor simulator of the Evergreen AMD GPU. We compare CPU and GPU experimental results. Our results define the best multicore processor with the highest speedup and the best parallel efficiency. For CIF resolutions, video sequences benchmarks reach their maximum throughput using 4 cores with a speedup of 3.1. For HD video sequences, 8 cores processors offer the best time and energy efficiency combined with a speedup of 6.2. On a GPU with 16 parallel computational units, the speedup reaches 12.1 for HD resolutions and 7.4 for CIF resolutions.

In our H.264 parallel implementation, the motion compensation (MC) stage for each row of inter-prediction macroblocks (P-MB) is executed in parallel on different cores. We experiment the parallel version using low and high definition resolutions, CIF and HD respectively, on multicore processors. Macroblock dependencies in the same picture slice are avoided by decoding intra-prediction macroblocks (I-MBs) when all other macroblocks are decoded. Overheads emerged as a result of shared memory communications and synchronization between cores. We simulated the parallel execution on multicore processors and graphical processors using a multicore simulator, Multi2Sim [8]. We further investigate the scalability of the multiple cores implementation, which shows the existence of a virtual threshold depending on the resolution when large numbers of cores are used.

The remainder of the paper is organized as follows. In Section 2, we present the related work concerning H.264 parallel optimizations. In Section 3, we describe our approach for parallel execution of macroblock rows of the H.264 decoder. In Section 4, we present the experimental results for execution time on CPUs and GPUs using a simulator for multicore processors. Final conclusion and future work are given in Section 5.

II. RELATED WORKS

Ever since the H.264/AVC standard [2] was published in 2003, researchers started to solve the high complexity issue of the new standard mainly using parallelism. Several modifications were suggested for the H.264 encoders and decoders in order to improve the performance in terms of execution time and memory usage. Parallel decoding techniques of H.264 exist from the highest level, which is the group of frames or pictures (GOP), the coarse-grain level, till the lowest level, which is the block inside a macroblock, the fine-grain level. Kannangara [12] reduced the complexity of the H.264 decoder (19-65%) by predicting the SKIP macroblocks using an estimation based on a Lagrangian rate-distortion cost function. Gurhanli [14] suggested a parallel approach by decoding independent groups of frames on different cores. The speedup is conditioned with the modification of the encoder in order to omit the start-code scanner process. Any modification to the encoder will require a long process for modifying the H.264 specification in order to be compliant with the standard. The exclusion of previously encoded video sequences is also an effect for modifying the H.264 encoder. Nishihara [18] proposed a load balancing mechanism among cores where partitions sizes are adjusted during runtime. He also reduced the memory access contention based on execution time prediction. Among frame-level and MB-level parallelization, the 3D-wave technique proposed by Azevedo [15] decodes independent MBs in parallel on different cores. A good scalability is proved for HD resolutions where macroblocks are scanned in zigzag mode and decode independent macroblocks in parallel. Chong [16] added a pre-parsing stage in order to resolve control dependencies for MB-level parallelization. Van Der Tol [20] mapped video sequences data over multiple processors providing better performance over functional parallelization. He groups macroblocks in a way that minimal dependency between cores is required. Horowitz [17] compared different H.264 implementations including FFmpeg [4] and the H.264 reference software JM [1]. He also analyzed the complexity of the H.264 decoder subsystems. Sihm [19] proposed a multicore pipeline for the deblocking filter based on the group of pictures data level partitioning. He also suggested software memory throttling and fair load balancing techniques in order to improve multicore processors performance when several cores are used. In our research we optimize the H.264 decoder knowing that our approach can be also applied to the H.264 encoder. We focus on improving the efficiency of the H.264 decoder using multicore processors. We decode rows of macroblock in parallel where rows are mapped to a number of cores. Dependencies between macroblocks are avoided by decoding intra-prediction macroblocks sequentially at the end of the decoding stage. We map our implementation on 2, 4, 8, and 16 cores. Speedup of the parallel implementation is calculated using simulated execution time. We further implement an

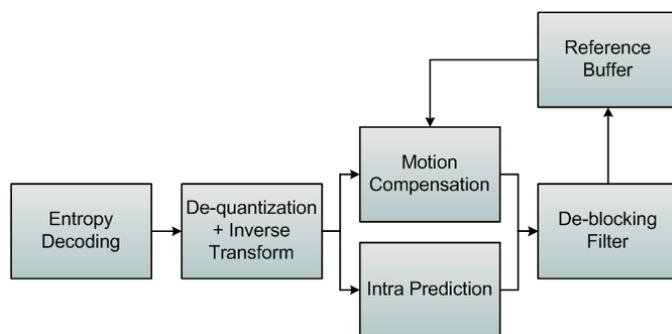


Figure 1. H.264 decoding process

OpenCL [5] version of our parallel H.264 implementation. Simulation experiments on graphics processors are conducted using a CPU-GPU simulation Multi2Sim [8].

In the following section, we describe in detail our parallel implementation of the H.264 decoder. In addition, we describe our environment configuration for the execution simulations on normal processors and graphics processors.

III. H.264 PARALLEL IMPLEMENTATION

In this Section, we describe our parallel implementation of the H.264 video decoder. We start with a brief overview of the decoder, then we explain how we parallelize the decoder, and finally we compare our approach to other similar parallel implementations.

A. Parallel Execution and Synchronization

Parallel execution is considered as a major potential solution for complex applications where sequential execution bounds the performance of these applications. Most processors that are currently available in the market have multiple cores and support many threads. Applications with low execution efficiency may benefit from a high potential speedup when data or functional parallelization is applicable. Even optimized implementations can still take advantage from parallelization techniques. In our research, we choose the H.264/AVC video decoder as our multimedia application benchmark for which we provide a parallel implementation using our approach. We further gather execution statistics and we compare results to other relatively similar implementations.

B. H.264 Standard

The Moving Picture Experts Group (MPEG) and the Video Coding Experts Group (VCEG) developed jointly in 2003 the "Advanced Video Coding" (AVC) standard published as ITU-T Recommendation H.264 and as part 10 of MPEG-4. Since the first commercial implementations, several multimedia device manufacturers adopted the new video codec. About 7 years after the first release of the final draft of the standard, H.264 is the mostly used video compression standard in most multimedia devices according the PCWorld.com [6]. Cameras, smart phones, PDAs, CCTV recorders, blu-ray disc players and many other devices use H.264 for encoding and decoding videos. H.264 achieves better compression and higher quality at the expense of more complex algorithms. Thus more computation

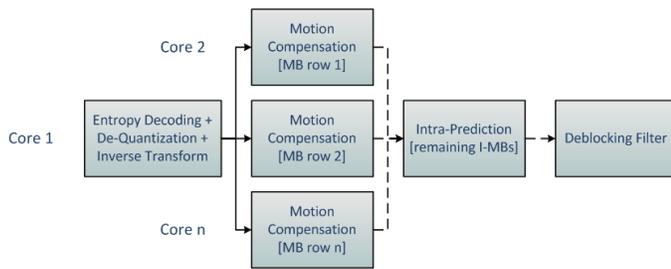


Figure 2. Decoding macroblock rows in parallel on n cores

resources are exploited and more energy is consumed in order to increase compression ratio of video files.

C. H.264 Decomposition

The H.264 decoder can be divided into five main functional parts: Entropy Decoder (ED), De-Quantization and Inverse Transform (IQT), Motion Compensation (MC) and Intra-Prediction (IP), and Deblocking Filter (DF). The H.264 decoder stages are illustrated in Figure 1. A slice of a picture is partitioned into blocks of 16 x 16 pixels called Macroblock (MB). The number of horizontal macroblocks and vertical macroblocks varies with the resolution of the frame picture that is being decoded. Entropy decoding is performed for all bits in a slice of a frame. Motion compensation or intra-prediction is applied for every macroblock of size 16 x 16 pixels. A macroblock can be also divided into sub-blocks of 16 x 8, 8 x 8, 8 x 4, and 4 x 4 pixels. The encoder chooses the sub-blocks sizes depending on the image complexity of the video sequences being decoded. The motion compensation stage uses a reference buffer in order to calculate the values of macroblocks in the current frame. The reference buffer contains a list of previously decoded frames. Macroblocks that are inter-predicted and motion compensated from previously decoded frames are either of type P or B (P-MBs and B-MBs). Macroblocks that depend on macroblocks in the current frame (called I-MBs) are intra-predicted. Deblocking filter is executed at the end of the decoding process in order to reduce the edging effect between macroblock borders.

D. Parallel Execution

The H.264 reference software, JM [1], is an open source implementation used as a reference implementation for the H.264 standards. In our research, we modified the JM [1] source code of the H.264 decoder in order to decode rows of macroblocks in parallel using the PThread library in C programming language. As shown in Figure 2, each core handles the motion compensation stage for macroblocks in a group of rows. Motion compensation and intra-prediction phases should be completed before applying the DF phase. Data parallelization is applied to the motion compensation phase of different macroblocks. The maximum numbers of parallel data execution is equal to the number of macroblock rows. One of the available cores is needed to coordinate the execution of the parallel decoding process on different cores. The coordinating core may be one of the cores that are used for parallel execution since parallel cores are only used for part of the decoding process. The level of parallel decoding of

```

input: list of macroblocks MBlst
        Number of cores cores

list of I-MBs DepMBlst in a slice

// main loop where each iteration is for 1 core
for each subMBlst of MBlst do // depending on cores
  for each element currMB in subMBlst do
    if currMB.TYPE == I16MB
      or currMB.TYPE == I8MB
      or currMB.TYPE == I4MB
    then
      DepMBlst.add(currMB);
      continue;
    else
      decode(currMB);
    endif
  endfor
for each element currMB in DepMBlst do
  decode(currMB);
endifor
endfor

```

Figure 3. Decoding rows of macroblock with intra-prediction dependency check algorithm

macroblock rows may be considered between coarse-grain and fine-grain approaches. High level approaches process multiple slices or frames in parallel. Low-level approaches decode macroblocks or blocks inside a macroblock in parallel. This balance between both approaches is also reflected between synchronization overheads and memory requirements. Coarse-grain methods need high memory usage in order to decode multiple frames in parallel. Fine-grain methods cause an enormous synchronization overhead affecting deeply the speedup. Our approach is aimed to benefit from the balance between both advantages and disadvantages. Macroblock rows require less memory than a frame and more than one macroblock. The number of rows is much less than the total number of macroblocks. For example, in HD resolution (1280 x 720), each frame has 3600 MBs, 80 horizontal MBs and 45 vertical MBs. Thus, the number of macroblock rows is less by a factor of 80 than the total number of macroblocks. As a result, the overhead for synchronization and communications between cores is also reduced by a factor of 80.

E. Dependencies between Macroblocks

In H.264, there are 3 types of macroblocks: I, P, and SKIP. I-MBs depend on other macroblocks in the same slice of a frame. P-MBs depend on macroblocks from previously decoded frames. SKIP-MB uses the same macroblock from a reference frame without transmitting the motion vector information. I-MBs require dependent macroblocks, which are in the same slice, to be previously decoded. So a dependency identification procedure is needed in order to satisfy I-MB dependencies. In order to overcome these dependencies, we start by decoding all P-MBs and SKIP-MBs rows in parallel. When this operation is completed, the remaining macroblocks, which are I-MBs in the current slice, are decoded sequentially. With this simple ordering, dependencies between macroblocks in the same slice are satisfied. The average number of I-MBs in P-Frames and B-Frames is 2.5% for CIF resolution and 4% for HD resolution. A video sequence always starts with an I-Frame (IDR), which is composed completely of I-MBs. This

1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4
4	4	4	4	4	4	4	4

Figure 4. Speedup of the motion compensation and intra-prediction stages on multicore processors

type of frames is available typically every 150 to 200 frames (3 to 8 seconds depending on frame rate) in a video sequences in order to overcome communication problems when some frame data are lost during communication transmission. We can increase or decrease the frequency of IDR frames in the encoder configuration. However, a high frequency of IDR frames, for example one I-frame every 30 or 50 frames, will significantly decrease the compression ratio of the decoder. The number of I-MBs in a P-Frame or a B-Frame depends on the complexity of the image and on the objects in the image and their rate of movements in the video sequences. P-Frames and B-Frames are mostly composed of P-MBs and SKIP-MBs with a small number of I-MBs. So the number of I-MBs does not significantly affect the overall speedup for the parallel decoding of MBs. Figure 3 shows the pseudocode for the macroblock dependency check algorithm in addition to the iteration over macroblock rows in a slice of a frame and the assignment of macroblock rows to different threads or cores of a processor. The list of all macroblocks and the number of cores are given as input data. A main loop iterates over groups of macroblocks assigned for each core. This loop is mapped onto the assigned cores in order to be executed in parallel. An inner loop checks every macroblock. I-MBs are added to an empty list. The remaining macroblocks are decoded. After the main loop, a second loop iterates over all I-MBs that are in the new list and decodes all the macroblocks in the list.

F. Macroblocks Partitioning

In a frame slice, while iterating over macroblocks, we skip intra-prediction macroblocks (I-MBs) and we decode P-MBs and SKIP-MBs in parallel on multiple cores as described above in the algorithm in Figure 3. Depending on the number of available cores, we group rows of macroblocks in order to be decoded in parallel. The slice is divided by the number of cores horizontally. In [7], 6 parallel representations are experimented in terms of stall time and core usage. Among the presented data partitioning approaches, our partition is similar to the slice-parallel splitting approach that is described in [7]. As shown by the author, this approach has a high stall time overhead. This stall time is caused by synchronization procedures in order to satisfy macroblock dependencies. However, with our approach for avoiding dependencies between macroblocks, the stall time overhead does not apply. We chose this method

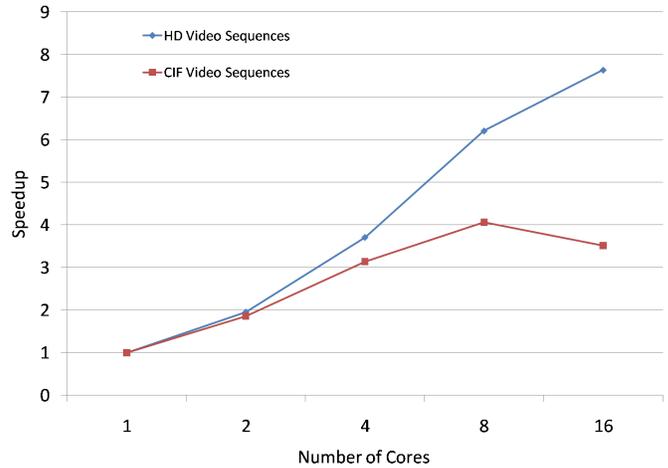


Figure 5. Speedup of the motion compensation and intra-prediction stages on multicore processors

TABLE I. Speedup of video sequences on multicore processors

Resolution	MB Rows	2	4	8	16
HD (1280 x 720)	45	1.959	3.710	6.207	7.636
CIF (352 x 288)	18	1.861	3.142	4.065	3.517

because of data locality and because of minimal data transfer initiation overhead. For example, in order to execute a slice of 80 rows of macroblocks on 4 cores processor, each core decode a chunk of 20 rows of macroblocks. Using this partition method, data is only transferred 4 times to the cores, which is the minimal number of transfers because it is equal to the number of available cores. In Figure 4, we show an example of a frame of size 64 x 64 pixels, 8 x 8 MBs, mapped onto 4 cores. The numbers inside the squares are the numbers of cores. Macroblocks in Figure 4 are assumed to be all P-MBs or B-MBs. I-MBs are not displayed for illustration purposes.

G. GPU Parallelization

The H.264 decoder parallel implementation is further modified to execute motion compensation process of P-MBs and B-MBs on graphics processors (GPUs). Part of the code is modified in order to comply with OpenCL [5] language, which is a unified framework for defining and controlling a GPU. Kernels, functions in C language, written in OpenCL are executed on a graphics device. Parallel execution of groups of macroblock rows are processed by work-groups. Slice data is first transferred to the graphics device and transferred back to the memory of the processor in order to complete the decoding process. Parallel execution of the motion compensation stage is performed as illustrated in Figure 2 where work-groups are considered as cores. Experimental results and comparisons with processor execution are discussed in the following section.

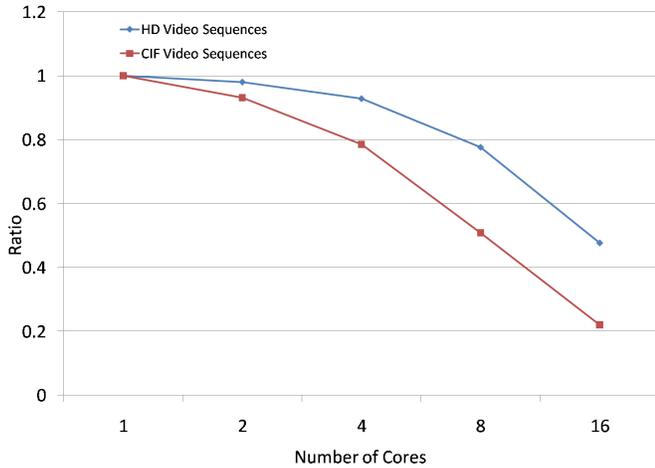


Figure 6. Speedup to number of cores ratio

IV. EXPERIMENTAL RESULTS

In this Section, we test our H.264 parallel implementation using multicore x86 processors and graphics processors. We gather simulation statistics and we compare our results with other results for parallel H.264 implementations.

A. Simulations

Our H.264 parallel implementation described in Section 3 is executed by Multi2Sim [8], a cycle-accurate simulator for multicore x86 and graphics processors. Cache and memory configurations comply with common x86 processors that are available nowadays in many Intel [11] or AMD [9] processor chips. Each core has a private L1 cache of 512 KB and All other cores have a shared L2 cache of 2 MB. We simulate the execution of our parallel H.264 decoder using 2, 4, 8, and 16 cores processors. We perform simulation experiments of the H.264 OpenCL version on the AMD Evergreen GPU family with the configurations of the AMD Radeon 5870 GPU [10]. We gather statistics using 3 video sequences with CIF resolution (bus, waterfall, and flowers) and 3 video sequences with HD resolution (Intotree, Parkrun, and Shields). Simulation is performed for the H.264 decoding process of 60 frames for each video sequence.

B. Results

Execution times with different number of cores using CIF and HD resolutions for the motion compensation stage are listed in Table I. The number of parallel rows of macroblocks increases with the video resolution. Thus HD resolution scales better than CIF resolution with the number of core. Experiments are conducted using simulations on 2, 4, 8, and 16 cores processors. Speedup results for the motion compensation stage are illustrated in Figure 5. For CIF resolutions, the maximum speedup of 4 is attained using 8 cores. With 16 cores, the speedup decreases to 3.5 due to large data communication overhead. For HD video sequences, a 7.6 speedup is reached with 16 cores processor. These optimization speedups are not efficient when compared to the number of cores used. Figure

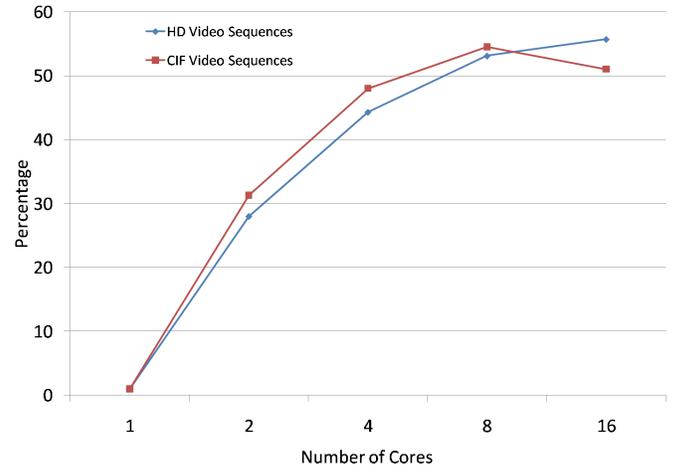


Figure 7. Percentage gain for the complete decoding process on multicore processors

6 shows that the ratio between the number of cores and the speedups is very high when using 16 cores. The best efficiency ratio is 4 cores with a speedup of 3.1 for CIF resolution and 8 cores with a speedup of 6.2 for HD resolution. The ratio of the speedup to the number of cores using 4 cores for CIF and 8 cores for HD is around 0.8. Doubling the number of cores drops the ratio to 0.5, which cannot be considered as efficient as we expect when running a parallel application on a multicore processor. In [13], the highest speedup is 5 on 8 cores and 8.1 on 16 cores. Our results have a better ratio, for less than 16 cores, related to the number of cores. For 16 cores and above, the results in [13] are better. However, decoder implementation, processor configurations and video resolutions vary between both approaches. Thus, exact comparisons are not applicable. The overall performance gain for all stages of the H.264 parallel decoder is illustrated in Figure 7. CIF resolutions reach 48% increase in performance using 4 cores and HD resolutions attain 53.1% using 8 cores.

C. Parallel Execution on Graphics Processor

We experiment our parallel implementation on a graphical processor simulator of the Evergreen AMD GPU. Figure 8 shows the speedups attained with the GPU devices. HD resolutions have a speedup of 12.1 and CIF resolutions a speedup of 7.4. These results exclude the data transfer time between the main processor and the graphics processor. This overhead limits the usability of the GPUs when the gain is low. In our case, the ratio of the speedup to the number of work-groups is around 0.75. Speedup simulation results for CIF and HD resolutions decoding on GPU are displayed in Table II. Graphics processor have high potential of parallel optimization. The number of work-groups and work-items is increasing significantly in new devices. Hundreds of work-groups and thousands of work-items can have a huge impact on applications with high parallel data processing.

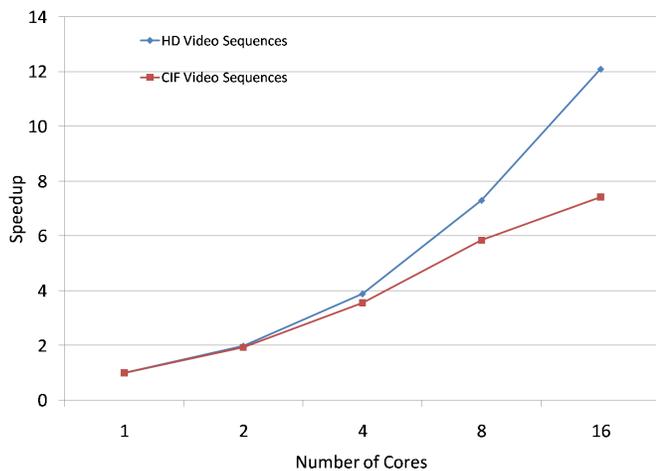


Figure 8. Speedup of H.264 parallel execution on Evergreen GPU

TABLE II. Speedup of video sequences on graphics processors

Resolution	MB Rows	2	4	8	16
HD (1280 x 720)	45	1.983	3.884	7.301	12.095
CIF (352 x 288)	18	1.928	3.560	5.839	7.417

V. CONCLUSION AND FUTURE WORKS

We have introduced a novel parallel technique for H.264 video decoder. Our approach decodes in parallel macroblock rows of the H.264 decoder with an algorithm that detects dependencies on-the-fly based on isolating intra-prediction macroblocks (I-MBs). Experiments using CIF and HD video sequences show that every resolution has a virtual threshold for the speedup when increasing the number of cores. This limit is due to the increase of data transfer between cores. The best speedup with the highest ratio to the number of cores is 3.1 for CIF resolutions using 4 cores and 6.2 for HD resolutions using 8 cores. A speedup of 12.1 is attained the H.264 parallel implementation is executed on a graphics processor. Additional research and experiments need to be conducted on the OpenCL implementation for GPUs. We plan to test our implementation on real boards and gather more statistics like memory usage and power consumption in addition to execution time and optimization efficiency in general.

REFERENCES

- [1] K. Suhring. H.264 reference software. <http://bs.hhi.de/~suehring/tml/>.
- [2] AISO/IEC. International standard. Part 10: Advanced video coding, 14496-10, 2003.
- [3] JCT-VC. High efficiency video coding (HEVC) text specification draft 8. 10th Meeting: Stockholm, SE, 1120 July 2012.
- [4] FFmpeg project. <http://www.ffmpeg.org/>.
- [5] OpenCL: The Open Standard for Parallel Programming of Heterogeneous Systems. <http://www.khronos.org/opencl>.
- [6] IDG Consumer & SMB. PCworld Magazine. <http://www.pcworld.com/>.
- [7] F. Seitner, M. Bleyer, M. Gelautz, R. Beuschel. Evaluation of data-parallel H.264 decoding approaches for strongly resource-restricted architectures. *Multimedia Tools and Applications*, 1(2010), S. 1 - 27.
- [8] R. Ubal, B. Jang, P. Mistry, D. Schaa, and D. Kaeli. Multi2Sim: A Simulation Framework for CPU-GPU Computing. *Proc. of the 21st International Conference on Parallel Architectures and Compilation Techniques*, Sep., 2012.
- [9] AMD Opteron Processor Family. <http://www.amd.com/>.
- [10] AMD Evergreen Family Instruction Set Arch. (v1.0d). <http://developer.amd.com/sdks/amdappsdk/documentation/>.
- [11] Intel Core Processor Family. <http://www.intel.com/>.
- [12] C. S. Kannangara and I. E. G. Richardson and M. Bystrom and J. Solera and Y. Zhao and A. Maclennan. Complexity reduction of H.264 using Lagrange Optimization Methods. *IEE VIE 2005*, Glasgow, UK, 2005.
- [13] M. A. Mesa, A. Ramirez, A. Azevedo, C. Meenderinck, B. Juurlink, and M. Valero. Scalability of Macroblock-level Parallelism for H.264 Decoding. *Proceedings of the 2009 15th International Conference on Parallel and Distributed Systems*, pages 236–243, ICPADS, 2009.
- [14] A. Gurhanli and S. Hung. Coarse grain parallelization of h.264 video decoder and memory bottleneck in multi-core architectures. *International Journal of Computer Theory and Engineering* vol. 3, no. 3, pages 375–381, 2011.
- [15] A. Azevedo, C. Meenderinck, B. Juurlink, A. Terechko, J. Hoogerbrugge, M. Alvarez, and A. Ramirez. Parallel h.264 decoding on an embedded multicore processor. *HiPEAC*, pages 404–418, 2009.
- [16] J. Chong, N. Satish, B. Catanzaro, K. Ravindran, and K. Keutzer. Efficient parallelization of h.264 decoding with macro block level scheduling. *ICME*, pages 1874–1877, 2007.
- [17] M. Horowitz, A. Joch, F. Kossentini, and A. Hallapuro. H.264/avc baseline profile decoder complexity analysis. *IEEE Trans. Circuits Syst. Video Techn.*, 13(7):704–716, 2003.
- [18] K. Nishihara, A. Hatabu, and T. Moriyoshi. Parallelization of h.264 video decoder for embedded multicore processor. *ICME*, pages 329–332, 2008.
- [19] K. Sihn, H. Baik, J. Kim, S. Bae, and H. Song. Novel approaches to parallel h.264 decoder on symmetric multicore systems. *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 09*, pages 2017–2020, Washington, DC, USA, 2009. IEEE Computer Society.
- [20] E. Van Der Tol, E. Jaspers, and R. Gelderblom. Mapping of h.264 decoding on a multiprocessor architecture. *Image and Video Communications and Processing*, pages 707–718, 2003.

Development of Context-Aware Real-Sense Services for Multi-Media and Multi-Device Environment

Hyunjeong Lee, Jaedoo Huh, Il-Woo Lee,

Energy IT Technology Research Section
Electronics and Telecommunications Research Institute
Daejeon, South Korea
{hjlee294, jdjuh, ilwoo}@etri.re.kr

Sang Ho Lee

Dept. of Software Engineering
Chungbuk National University
Cheongju, Chungbuk, South Korea
shlee@cbnu.ac.kr

Abstract- In this paper, we propose based real sense media services based on context-aware. Some real sense technologies provide the same services without considering the users' preferences and context. So, some users don't want the unilaterally and passively real sense services. For solving these problems, we consider the users' context, such as gender, age, preferences and provide the adequate real sense services for each user. First, we collect the context of users, environment and devices; then, we process and select the proper services considering the situation. Using this method, we can provide context-aware based real sense services for users more appropriately.

Keywords-real sense; context-aware; preference; ubiquitous.

I. INTRODUCTION

As many 3D contents and services are widely used, real-sense technologies are studied and developed to provide more realistic feeling for users. They focus on the real-sense media playback, multiple audio/video synchronization, sensory effects, and so on [1]. Most of real-sense services provide unilateral and passive effects. But, all users do not like the same real-sense effects, and some services are not adequate for some situations. For example, someone likes water spray effects, but the other does not. Therefore, if we consider context-awareness in real-sense service, then we can provide real-sense effects with considering users' preferences and context. Context-aware computing is the ability of a user's applications to discover and react to changes in the context in which they are situated [2][3]. If we think over these technologies for real-sense services, then users will feel more sympathy by considering contexts, such as users' age, gender, tastes, environmental situation, and so on [4]. So, we propose real-sense media services based on context-awareness for multi-media and multi-device environment. Multi-media means the contents are created and encoded using various media, such as several cameras, real-sense devices, in order to real-sense services for users. The contents are decoded as the original, and processed according to users' context. First, we collect the context of users, environment and devices from sensors and profiles. Additionally, we consider the environment and devices for users; then, we process and provide the proper services for users considering their situation. Using this method, we can

provide context-aware based real sense services more appropriately and effectively.

This paper is organized as follows. In Section 2, we explain the related works of context-awareness and real-sense media services. We describe the context-aware based real-sense media services in Section 3, and present our conclusions in Section 4.

II. RELATED WORK

As Mark Weiser introduced ubiquitous computing and its vision of people and environments augmented with computational resources, many works for context-aware computing have been studied to provide information and services wherever and whenever users desired [5]. But, it is not easy to recognize contexts about environment and people, and to provide automatic services according to the context without human repulsion [6]. To aware the situation of users and their environment, many technologies are required to collect data, such as sensors, identifiers, predefined context, and so on. These days, there are many researches of context-awareness and applications using it.

The EasyLiving project of Microsoft Research is to develop architecture and technologies for building intelligent environments that contains innumerable devices to provide intelligent environments [7]. Components in the environments include middleware, world modeling, and service description to provide users access to information and services. Also, the devices and systems in the environments have to understand the physical space in order to support richer interactions with users.

The context-aware middleware for ubiquitous robotic companion systems (CAMUS) is a middleware for context-aware applications with a development and execution methodology [8]. The CAMUS provides a context-aware framework for a ubiquitous robotic companion (URC), such as network-based hardware robots or software robots. To do this, CAMUS collects contextual information from various kinds of sensors and transfers the appropriate contextual information to variety of applications. Also, CAMUS provides autonomous service agents to recognize the context and to adapt themselves to different situations. In [9], content recommendation service agent (SA) and context-

aware task were developed based on the service framework and task development methodology of CAMUS.

Internet technologies are also considering context-awareness. As more and more users and services appear and utilize the Internet, it has some technological and operational limits imposed by its architecture in its attempt to give full support to the new requirements introduced by increasing services, applications, and content [10]. Also, users expect higher levels of performance, security, and reliability. Therefore context-aware service discovery is considered in future internet to provide adapted communications for new services, applications and content. This approach is expected to improve the satisfaction of users' expectations by matching offered service characteristics with requirements and preferences previously determined by them.

As we described above, context-aware computing aims for maximizing user's convenience and the utilization of environmental resource. This paper discusses technical issues regarding context-awareness for real-sense media service to consider the users' preference and context information.

III. CONTEXT-AWARE BASED REAL-SENSE SERVICES

A. Conceptual model for the Context-aware based real-sense services

In this paper, we describe a scenario for real-sense services based on context-awareness for the purpose of maximizing the users' comfort and convenience.

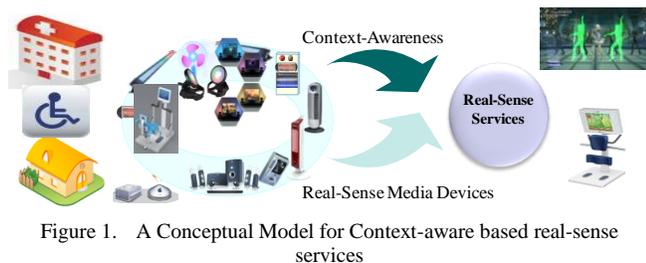


Figure 1. A Conceptual Model for Context-aware based real-sense services

The conceptual model for the Context-aware based real-sense services mentioned above is illustrated in Figure 1. There are many devices for real-sense effect in the home, hospital, education center, and so on. We recognize real-senses in the content, and determine which devices to apply for the service. In this stage, we have some issues on which devices are available and who is the user for the service. To solve the issue, we use some kinds of technologies to recognize the user and devices in the space.

In this paper, we use camera, infrared and 3D depth sensors in the device to detect users and users' motion. For real-sense rehabilitation service, we define some item to describe the users' context and preferences. Using the Real-

sense service based on context-awareness, users feel more comfortable and convenient in the future medical devices.

B. Technical Issues

There are some issues to discuss. Figure 2 shows technical issues for real-sense service based on context-awareness. First, sensor platform detects devices in the space or defines configuration information before hand. And we recognize users and environmental information. Temperature, humidity, weather is the environmental information, which is optional and is for further services. In our laboratory, we use devices with various sensors. They can detect the users' physical condition and motion. Secondly, a context-aware framework in a service server recognizes the situation that who is the user, what the service is needed, where the location the service is provided, and which adequate devices the real-sense service play. Then, the context-aware framework notifies this information to the service agent to provide the adequate real-sense service for the user.

Sensor Platform	Context-aware Framework	Service Agent
Device Detection	Location context	Rehabilitation Service
User Detection	User Preference	Therapy Service
Environment Detection	Adequate Devices	Broadcasting Service

Figure 2. Technical issues for Context-Aware based Real-Sense Media Services

C. System Architecture

A context-aware framework interfaces with a Sensor platform and an service agent, and arbitrates between them for feeling the physical effects according to the contexts. A sensor platform has sensor nodes, such as camera sensors, infrared sensors, 3D depth sensors, and so on, collect data of the device, environment, and users' situations, and send them to sensor coordinators, which manage several kinds of sensors and interfaces with a context-aware framework. A sensor networking sub-block adds, deletes and updates the information of sensor nodes and sensor coordinators. A sensor interface agent sends and receives messages to and from a context-aware framework in a service server. A context-aware framework consists of a semantic analyzer, a context ensembler, context manager and a context-aware interface agent. After a context-aware interface agent receives sensing data from a sensor platform agent, a semantic analyzer performs analysis of the received data's meaning. Then context ensembler configure and register the context information messages and send them to an service agent. In case of a locational context, a context-aware framework receives raw data related to location from a sensor platform, analyzes its meaning, forms context information message, saves them to a context repository as context schema, and sends them to an service agent. In an

service agent, real-sense services are modeled for users, and customized services are formed as the users' preferences. Then, a service interaction sub-block mediates the operation

of real-sense services as the predefined priority rule, where the rule is provided by users according to the their preference.

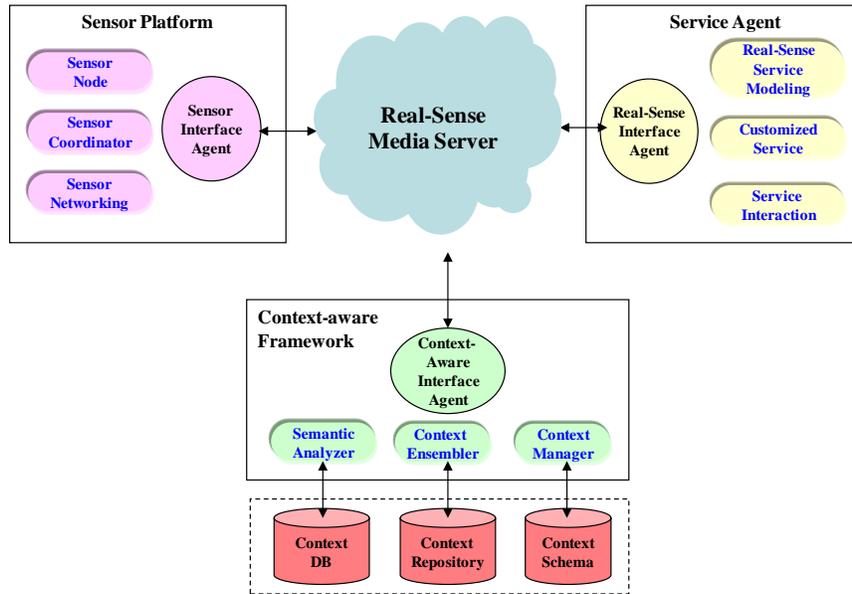


Figure 3. System Architecture for Context-Aware based Real-Sense Media Services

D. Service Flow

The proposed system consists of a sensor platform, a context-aware framework and an service agent, where a context-aware framework and an service agent is in a service server. The sequence flow for real-sense service based on context-awareness is shown in Figure 4. In the service scenario, the user, Harry, wants to play real-sense game. First, context-aware framework recognizes the user and his location using various sensors, such as camera, infrared and 3D depth sensors. Then, the context-aware framework

analyzes and recognizes his context and select the adequate real-sense service using his cotext and profile with preferences. He likes the wind and moving chair effects, and does not want water spray effects in the predefined his preference. Also, his preference can be changed using his actions in the effects. The service agent receives the context from the context-aware framework, and designs real-sense effects for Harry's game service. Finally, the service agent selects and triggers the devices according to the context received from context-aware framework.

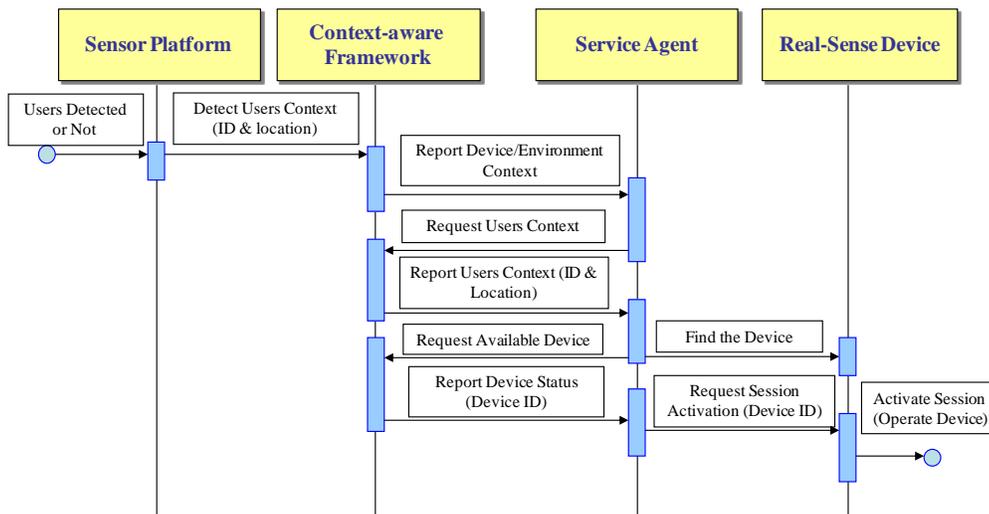


Figure 4. . A Sequence Flow of Context-Aware based Real-Sense Media Services

IV. CONCLUSION AND FUTURE WORK

In this paper, we proposed real-sense media services based on context-awareness. First, we collected the context of users, environment and devices from sensors and profiles. Then, we considered the environment and devices for the services and users. We processed context information to provide adequate services for users considering their situation. Using this method, we could provide context-aware based real sense services more appropriately and effectively.

In future work, the authors will focus on the interaction issues among several users in the same space. We will consider various scenarios to present service policy and interaction mechanism for solving the issues.

ACKNOWLEDGMENT

This work was supported by the MKE (Ministry of Knowledge Economy) [A004700008], Development of realistic sense transmission system with media gateway supporting multi-media and multi-device.

REFERENCES

- [1] J.K. Yun, J.H. Jang, K.R. Park, and D.W. Han, "Real-Sense Media Representation Technology Using Multiple Devices Synchronization," *Software Technologies for Embedded and Ubiquitous Systems, LNCS 2009*, vol. 5860/2009, 2009, pp. 343-353.
- [2] William Noah Schilit, "A System Architecture for Context-Aware Mobile Computing," PhD dissertation, Columbia University, New York, 1995.
- [3] S. A. N. Shafer, B. Brumitt, and J. Cadiz, "Interaction Issues in Context-Aware Intelligent Environment," *Journal of Human-Computer Interaction*, vol.16, no. 2, 2001, pp.363-378.
- [4] H. Lee, S. Lim, and J. Huh, "Design and Implementation of Intelligent Home Care System based on Context-awareness," 2004 international conference on computers, communications and systems, vol. 1, Daegu University, South Korea, 2004, pp. 169-173.
- [5] M. Weiser, "Some Computer Science Issues in Ubiquitous Computing," *Commun. ACM*, vol. 36, no. 7, July 1993, pp. 75-84.
- [6] W. N. Schilit, "A System Architecture for Context-Aware Mobile Computing," PhD dissertation, Columbia University, New York, 1995.
- [7] B. Brumitt, B. Meyers, J. Krumm, A. Kern, and S. Shafer, "EasyLiving: Technologies for Intelligent Environments," *LNCS 1927, Springer-Verlag*, September 2000, pp 12-29.
- [8] Y. Ha, J. Sohn, Y. Cho, and H. Yoon, "Towards Ubiquitous Robotic Companion: Design and Implementation of Ubiquitous Robotic Service Framework," *ETRI Journal*, vol. 27, no. 6, 2005, pp. 666-676..
- [9] A. Moon, H. Kim, H. Kim, and S. Lee, "Context-Aware Active Services in Ubiquitous Computing Environments," *ETRI Journal*, vol. 29, no. 2, April. 2007, pp. 169-178.
- [10] A. J. Gonzalez, R. M. de Pozuelo, M. German, J. Alcober, and F. Pinyol, "New Framework and Mechanisms of Context-Aware Service Composition in the Future Internet," *ETRI Journal*, vol. 35, no. 1, Feb. 2013, pp. 7-17.

Region of Interest Encoding in Video Conference Systems

Christopher Bulla and Christian Feldmann
 Institut für Nachrichtentechnik
 RWTH Aachen University
 Aachen, GERMANY
 {bulla,feldmann}@ient.rwth-aachen.de

Martin Schink
 MainConcept GmbH
 Aachen, GERMANY
 Martin.Schink@rovicorp.com

Abstract—In this paper, we present a region of interest encoding system for video conference applications. We will utilize the fact that the main focus in a typical video conference lies upon the participating persons in order to save bit-rate in less interesting parts of the video. A Viola-Jones face detector will be used to detect the regions of interest. Once a region of interest has been detected it will get tracked across consecutive frames. In order to represent the detected region of interests we use a quality map on the level of macro-blocks. This map allows the encoder to choose its quantization parameter individual for each macro-block. Furthermore, we propose a scene composition concept that is merely based upon the detected regions of interest. The visual quantization artifacts introduced by the encoder thus get irrelevant. Experiments on recorded conference sequences demonstrate the bitrate savings that can be achieved with the proposed system.

Keywords—region of interest coding; object detection; object tracking; scene composition; video-conferencing

I. INTRODUCTION

Video-conferencing greatly enhances traditional telephone-conferencing, with applications ranging from every day calls from friends and family to cutting management expenses by replacing business trips with video-conferences. This multi billion dollar market splits mostly into two segments: free applications with decent quality and expensive telepresence systems. Among the free applications Skype is probably the best known application offering decent video quality. When video-conferencing substitutes business trips, the costs for video-conferencing can be several million dollars. Telepresence systems, for example, outfit rooms at individual locations with exact replicas of furniture and life-size displays create an immersive environment which creates the impression of sitting at the same table with other conference participants. All solutions share operating costs for bandwidth as by far the most expensive part of the yearly budget. Naturally, the introduction of the H.264/AVC codec for current generation video-conference systems was a major advantage over legacy systems as it cut bit-rates in half, a pattern that is expected to repeat itself with the introduction of the upcoming HEVC codec.

This paper will present an approach that is also able to achieve a bit-rate reduction by around the same factor by

taking the context of the video application into account. Since the main focus of a video conference lies upon the participants our idea is to reduce bitrate in less interesting background areas. We will show how a combination of face detection, tracking, region of interest encoding and scene composition can be used to reduce bitrate while preserving a constant visual quality in the detected regions of interest.

The rest of the paper is organized as follows. In Chapter II we will in detail explain our region of interest encoding concept. Our achieved bitrate savings will be presented in Chapter III. Final conclusions as well as an outlook for future work in this area will be given in Chapter IV.

II. ROI VIDEO ENCODING

Our region of interest (ROI) video encoding system consists of four key components which interact with each other (see Fig. 1). In our system, the regions of interest correspond to the faces of all participating persons. The detection is done with the Viola Jones object detection framework. Once a face is detected a new tracker will be initialized. The tracker is necessary for two reasons: Our face detection algorithm may not provide a result in every frame, however, the encoder expect a result for each frame. Tracking of the detected persons across consecutive frames will provide the encoder with the necessary information even if the face detection is still active. A second motivation for the use of a tracker is given by the fact that persons may not look into the camera all the time. In this case, the face detector would also not be able to detect these persons which finally result in a classification of these areas as not of interest and thus in a bad visual quality.

The output of the tracker, which basically correspond to a quality value for each macro block will be forwarded to the encoder. The encoder is then able to encode the detected ROIs in a good and the background in bad quality.

Finally, the encoded video stream will be transmitted to all receiving clients which can then decode it, crop out the ROIs and render them in an arbitrary manner.

A detailed description of each component will be given in the following subsections.

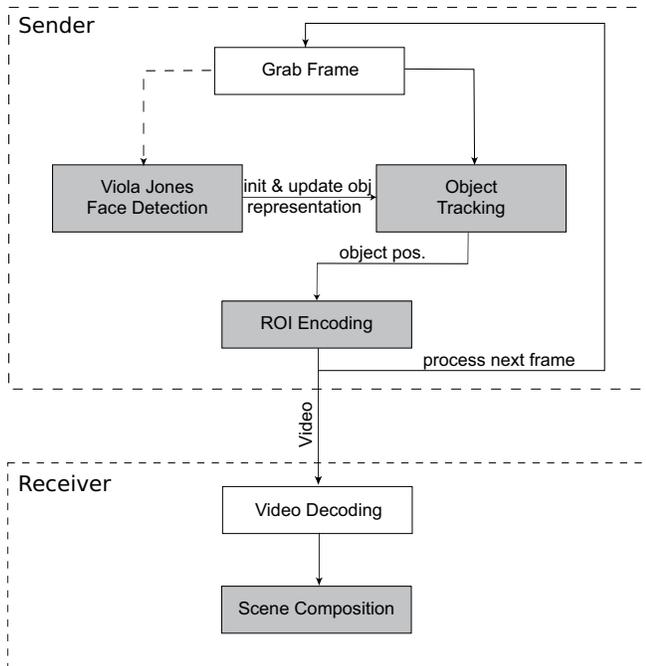


Figure 1. System overview. Interaction of face detection, tracking, video encoding and scene composition in sending and receiving client

A. Face detection

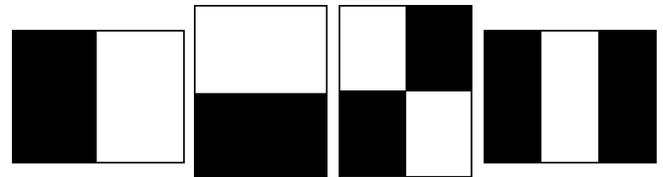
Our face detection algorithm is based on the Viola-Jones object detection framework [1]. It has three key components that will be briefly explained in the following. In a first step, a learning algorithm selects significant features in order to build efficient classifiers. The features used in this classifiers are Haar like and can be computed efficiently using an integral image representation. In order to speed up the classification process the single classifiers will be combined in a cascade.

The features that were used in the object detection system are exemplary depicted in Fig. 2a. The response of each feature is the sum of all pixel inside the black area subtracted from the sum of all pixel inside the white area. Using an alternative image representation, the integral image $II(x, y)$, these features can be computed very efficiently:

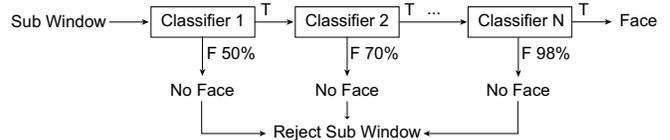
$$II(x, y) = \sum_{x' \leq x, y' \leq y} I(x', y'), \quad (1)$$

with $I(x', y')$ denoting the original image.

The integral image allows for the computation of the sum of all pixel inside a rectangle with only four memory access operations. The response of each feature can thus be computed very efficiently. The features are so called weak features, that means, that a classifier based on each single feature is only able to distinguish between a face and something else in a limited degree. However, a combination of these weak classifiers can yield to a strong classifier.



(a) Face detection features. Left to right: horizontal and vertical two-rectangle features, diagonal four-rectangle feature and horizontal three-rectangle feature.



(b) Cascaded classifier structure. Simple classifier reject many negative sub-windows while complex classifiers reduce the false positive rate

Figure 2. Rectangle features and cascaded classifier structure used in the face detection process

For a detection window of 24x24 pixel the entire set of possible rectangle features is about 45000. Since not all of them are necessary to detect faces in an image, a set of significant features have to be selected from all possible features what is done by AdaBoost [3].

Given a set of positive and negative training examples, the rectangle features that best separate the positive and negative examples need to be selected. The learning algorithm therefore determines the optimal threshold for a classification function such that the minimum number of examples are misclassified. The weak classifier $h_j(\mathbf{x})$ is then given by the function:

$$h_j(\mathbf{x}) = \begin{cases} 1, & \text{if } p_j f_j(\mathbf{x}) \leq p_j \theta_j \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

with f_j denoting the feature, θ_j a threshold, p_j a parity for the direction of the inequality and \mathbf{x} a patch of the image.

The final classifier $h(\mathbf{x})$ is then a linear combination of the selected weak classifiers:

$$h(\mathbf{x}) = \begin{cases} 1, & \text{if } \sum_{j=1}^J w_j h_j(\mathbf{x}) \leq \frac{1}{2} \sum_{j=1}^J w_j \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

with J denoting the total number of weak classifier and w_j a specific weight for each weak classifier. More information on the determination of the weights can be found in [1].

In order to reduce computation time and increase the detection performance the classifiers are arranged in a cascaded structure. An example of such a structure is depicted in Fig. 2b. Classifiers with relatively large false positive rates at the beginning of the cascade can be used to reject many negative sub-windows. Computationally more complex classifiers are then used at the remaining sub-windows to reduce the false positive rate. The idea is motivated by the fact that many sub-windows within an image won't contain a face.

B. Mean Shift Tracking

Since the face detection doesn't provide a detection result for each frame, a tracking of the face positions across consecutive frames is necessary. In the general case, given the object location and its representation in frame t we want to estimate the object location in frame $t + 1$. We will use a Mean Shift based tracking algorithm in order to fulfill this task. Mean Shift is an iterative technique for locating the mode of a density estimation based on sample observations $\{\mathbf{x}_n\}$ [2]. In the context of video object tracking, the samples $\{\mathbf{x}_n\}$ represent the pixel positions within the object region. In the following we will refer to the object that will be tracked as target, while possible locations of that object will be denoted as target candidates.

Let a kernel function G be given, the Mean Shift procedure estimates the new position of the target candidate \mathbf{y}_j based on a previous estimate of the target candidate position \mathbf{y}_{j-1} as follows:

$$\mathbf{y}_j = \frac{\sum_{n=1}^N w_n \mathbf{x}_n G\left(\frac{\mathbf{y}_{j-1} - \mathbf{x}_n}{h}\right)}{\sum_{n=1}^N w_n G\left(\frac{\mathbf{y}_{j-1} - \mathbf{x}_n}{h}\right)} \quad (4)$$

Here, N denotes the number of pixel within the object region, h the width of the kernel and w_n the weight at pixel position \mathbf{x}_n . The actual weight is given by:

$$w_n = \sum_{u=1}^M \sqrt{\frac{q_u}{p_u(\mathbf{y}_0)}} \delta(b(\mathbf{x}_n) - u), \quad (5)$$

with the normalized kernel-weighted M-bin target and candidate histograms $\mathbf{q} = \{q_u\}_{u=1, \dots, M}$ and $\mathbf{p}(\mathbf{y}) = \{p_u(\mathbf{y})\}_{u=1, \dots, M}$:

$$q_u = C \cdot \sum_{n=1}^N K(\mathbf{y}_0 - \mathbf{x}_n) \delta(b(\mathbf{x}_n) - u) \quad (6)$$

$$p_u(\mathbf{y}) = C_h \cdot \sum_{n=1}^N K\left(\frac{\mathbf{y} - \mathbf{x}_n}{h}\right) \delta(b(\mathbf{x}_n) - u). \quad (7)$$

Here, u denotes an index of a histogram bin, $b(\cdot)$ yields the bin index of the color at pixel location \mathbf{x}_n , $\delta(\cdot)$ is the Kronecker delta function and C and C_h are normalization constants.

The kernel functions $K(\mathbf{x})$ and $G(\mathbf{x})$ are connected through their individual profiles $k(x)$ and $g(x)$ for which $g(x) = -k'(x)$ holds [2].

Because the appearance of the target may change over time (eg. due to a change in the lighting or a change of the 3D object pose), we will update the target representation in each frame:

$$\mathbf{q}_t = \alpha \mathbf{q}_{t-1} + (1 - \alpha) \mathbf{p}(\mathbf{y}_{final})_t, \quad 0 \leq \alpha \leq 1. \quad (8)$$

Fig. 3 shows an example of the iterative Mean Shift procedure in a possible conference scenario. The target is

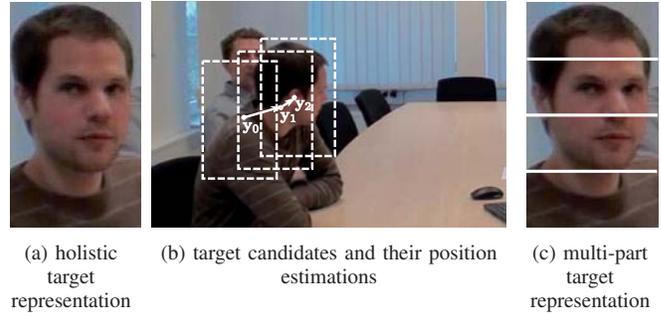


Figure 3. Target representation and new location estimation by iterative mean shift updates

depicted in Fig. 3a, the target candidates and the estimated locations as well as the final object location in Fig. 3b.

In order to get a more distinct object representation and thus an improved and robust tracking result, we divide our object representations according to [6] into parts which will be tracked separately. Fig. 3c shows an example of such a multi-part object representation. In contrast to the holistic representation illustrated in Fig. 3a, a multi-part representation provides information about the distribution of features for each subregion of the object.

C. ROI encoding

Implementing a region of interest algorithm alters the behavior of encoders and creates greatly different visual results. A traditional H.264/AVC encoder compresses a video stream, composed by a sequence of frames, by representing the content of these frames in a more efficient way; Although this compression is lossy, resulting in non-recoverable loss of image content, the effects are usually barely noticeable to the viewer. Rate distortion optimization makes sure that content with high importance to viewers perception of the videos quality, e.g., high frequency parts like the contours of a face or the pattern on a plant, is compressed less aggressively than content that contributes little to the viewers perception of the videos quality. Fig. 4a shows a scene with a person at a desk, and a bookshelf in the background; the scene is compressed with a standard H.264/AVC encoder and shows both the person and the bookshelves in about the same visual quality - the contours of both the person and the bookshelf are clearly identifiable, because both contribute equally to the overall visual quality. While this approach is very natural and pleasing to the human eye, it does not take the viewers attention into account: in a video-conference setting we are more interested in the person talking than in the books on the shelves. Taking the viewers attention into account means that the encoder should increase the quality of objects that are currently capturing the viewers attention, while paying for this increase in quality with lower quality on anything that is not important to the viewer; consequently, the goal of region of interest encoding is to redistribute bits for image

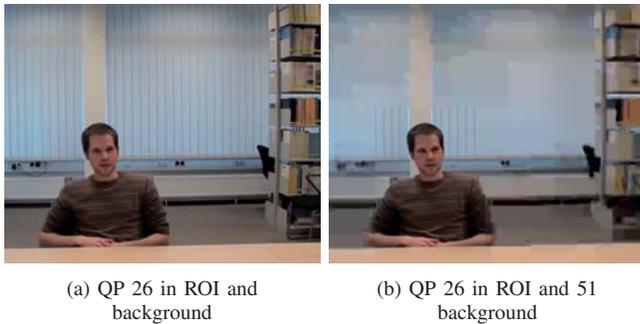


Figure 4. Comparison of image qualities within and outside of region of interest

compression from areas with little interest to areas with high interest. Fig. 4b shows a very extreme case of ROI encoding, where the bookshelf is now encoded in a much lower quality than the face of the person.

A region of interest is in its simplest form a rectangle containing the object of highest interest. In the case of video conferencing this is the face of the person currently speaking and the immediate area around it. However, the shape of the ROI is not limited to a rectangle but is flexible in shape as well as in the distribution of weights within the region.

A final thought should be given to H.264/AVC standard compliance. While it is possible to implement proprietary solutions that require an encoder and decoder pair capable of understanding the implemented region of interest algorithm, it is much preferred to make do without such requirements. Video-conferencing, just like telephone-conferencing, first and foremost requires interoperability. Consequently, a region of interest implementation may only modify the encoder, but must leave the decoder untouched, resulting in decodable content by every standard compliant vendor.

1) *ROI encoding using quantization parameters:* Taking all these conditions into account, we chose the modification of the quantization parameters for each individual macro-block (MB), similar to the approach by Ferreira et al. [4]. In H.264/AVC each frame is divided into MBs, each with a dimension of 16x16 pixels. These MBs are then transformed into the frequency domain using the discrete cosine transform (DCT), and are then quantized before entropy encoding [5]; the decoder performs the inverse steps to recover the final frame. Quantization is used to increase compression efficiency by mapping a large set of data to a smaller set of data. This operation is lossy and introduces a quantization error into the reconstructed values. By applying this technique to the transform coefficients the amount of coded data as well as the quality of the reconstructed picture can be controlled. In H.264/AVC, the quantization can be controlled by a quantization parameter ranging from 0 to 51, 0 being the finest quantization and 51 the coarsest.

We implemented ROI encoding in the MainConcept H.264/AVC encoder by quantizing the MBs within areas of

low interest very coarsely, e.g., with QPs in the range from 40 to 51, while quantizing MBs of interesting parts more finely to preserve as much of the original values as possible. Our approach generalizes the approach by Ferreira et al. [4] by allowing arbitrary values for the region of interest. As an example region of interest may include fading, e.g., values of 22 on the MBs covering the face of the active speaker, values of 28 in the MBs adjacent to the face and then QPs of 51 for the remaining background regions. Another reason for allowing a more flexible quantization of the MBs describing a region of interest are our two main use cases for video-conferencing: Without scene composition one will always view the entire frame in contrast to scene composition where parts of the frame are cropped, typically only showing the person and immediately adjacent content; since large parts of the frame aren't even seen during scene composition the quantization can easily be set to 51 for the background region that will be discarded during scene composition; likewise, without scene composition the less interesting MBs would probably not be quantized so harshly because they are clearly seen and are, while arguably less interesting, still negatively impacting the perception of quality due to the blocky nature of coarsely quantized MBs.

The quantization parameters for each MB are stored in an array which is the output of the face tracking algorithm. For convenience and to give extra to room rate distortion optimization and rate-control, we changed the values from 0 to 51 to 100 to 0, indicating the percentage of interest the viewer has in a MB - with a value of 0 resulting in the coarsest quantization and a value of 100 resulting in the finest quantization available. We choose to receive a QP array every frame, to allow for maximum flexibility for a region of interest, even though the region typically does not change rapidly due to the fact that people are rarely moving dramatically to warrant constant changes in the ROI.

The benefit of this approach is a very flexible region of interest, implemented in a H.264/AVC standard compliant manner. The downside of this approach is the MB based structure which can create blocky artifacts particularly with a very coarse quantization. Furthermore, a region of interest that resembles the exact contours of a face is also not possible due to the block based approach.

D. Scene Composition

The proposed region of interest concept offers at the receiving client the possibility to compose a video based on the detected persons. Inspired by the idea of a telepresence video conference systems, which creates the impression that all conference participants are sitting on the same table, and the fact that the focus of interest in a typical conference scenario is up to the participating persons, an alternative video composition could be achieved by showing only the detected persons. Each person is then scaled and placed side by side at the receiving client. This concept can be



Figure 5. Exemplary scene composition of four participants

extended in that way, that only the n most active speakers will be displayed at the receiving client. Determining the active speaker can be achieved through a combined audio and video analysis. The decision which person gets rendered at which client will be made by a central mixing component that analyzes an activity index of all participants.

Fig. 5 shows an example of the scene composition with four active participants. In addition to the advantage that our proposed scene composition depicts only relevant and active conference participants, the rough quantization artifacts depicted in Fig. 4 can be neglected. This kind of scene composition thus allows a very coarse quantization of the background.

III. EVALUATION

Our investigations focus on the bitrate savings achievable through region of interest (ROI) encoding in a video-conference. We thereby assume that the result of the detection and tracking algorithm is reliable. A separate evaluation of the performance of the face detection and tracking algorithm will not be the subject of this paper. Detailed information about the tracker performance for different object representation is given in [7].

Our goals for visual quality differ when scene composition is turned on or off: in case of scene composition, most of the video is cropped, so the ROI should achieve high bitrate reduction without regard to visual quality outside the ROI; without scene composition the effects of ROI encoding are directly visible to the viewer so our goal here was to find a sweet spot where bitrate savings and visual quality outside the ROI are in balance.

A. Test environment

In order to show the efficiency of our region of interest encoding approach we captured several videos with typical video-conferencing conditions. All of these videos have been recorded with a high-end consumer grade camera at a resolution of 720p and 50fps. All of the videos are 900 frames long. The videos *changing_lighting*, *disruption*, *next_to_each_other*, and *individual_spakers* show scenes with one, one, three and nine tracked people in them. Fig. 6 shows a typical frame from each of these videos. In addition to changing the number of tracked faces, we also

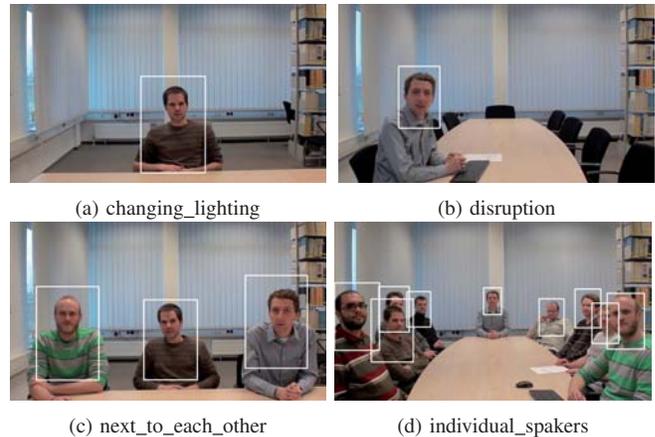


Figure 6. Sample Images of our test sequences with detected ROIs

included a change of light in the video *changing_lighting*: mid way through the video the light is turned off suddenly and gradually faded back in. Additionally, we included movement of a person in the video *next_to_each_other*. The area covered by the ROI box is 6% for *disruption*, 13% for *changing_lighting*, 23% for *next_to_each_other*, and 26% for the nine people video *individual_spakers*. For the quantization parameters of the ROI only two values have been chosen: all MBs inside the ROI have the same quantization value, just like anything outside has the same values.

The face tracker generates a box shaped region of interest sized with respect to the individual faces, showing head and shoulders. The region of interest encoding has been implemented in MainConcept's H.264/AVC encoder, based on MainConcept Codec SDK 9.5. The encoder itself has been configured to a low-delay setting suitable for video-conferences: no B-frames have been used, base profile, GOP length of 300, constant quantization instead of rate-control, and deblocking turned on. The long GOP of 300 allows some IDR frames to improve the robustness against network errors, but does not allow frequent joining of a conference; whenever a new user joins the video-conference a new IDR is requested. Deblocking helps improve the visual quality for highly compressed areas so it has been turned on for all videos. To further evaluate the efficiency of different profiles we also evaluated the *next_to_each_other* video with main profile (replacing CAVLC with CABAC) and high profile (enabling 8x8 transform and PCM prediction).

The quantization parameters inside the ROI ranged from 18 to 34; values below 18 no longer provide improved visual quality for the viewer, values above 34 produce artefacts that make reading facial expressions difficult. The outside of the ROI is quantized with a step size which is a multiple of six; The quantization parameters outside the ROI range from +0, to create a non-ROI reference, until they reach +18 for a very coarse quantization.

B. Results

In Fig. 7 the encoder performance for different quantization values for the ROI and the non ROI region are shown. Each graph represents a constant QP difference between the ROI and the non ROI area. For QP Difference 0 the ROI and the non ROI regions use the same quantization so this is the reference for encoding not using ROI information. With higher QP difference values the quality of the non ROI region decreases. The PSNR measure only takes the PSNR inside of the ROI into account.

We can see that especially at high bitrates the bandwidth savings using a coarser quantization for the background are enormous. For example, for the highest data point (ROI QP 22) we save about 77% using a QP of 28 for the background (QP difference 6) or 86% using a QP of 34 (QP difference 12). However, such high bitrates are unrealistic to be used in video conferencing applications. A more realistic QP range is between QP 26 and 30 where the conventional video coding approach uses a bitrate of about 1-2 Mbit/sec. In this area our ROI based encoding approach yields a coding gain of approximately 50%.

TABLE I. AVERAGE BD-RATE SAVINGS FOR THE TEST SET AT DIFFERENT QP DIFFERENCES.

QP Difference	Y	U	V
6	-43.51%	-48.75%	-48.45%
12	-46.41%	-52.70%	-52.21%
18	-44.90%	-51.25%	-52.28%

In Table I the average BD-savings in our test set are shown at different QP differences. In the table as well as in Fig. 7 one can see that the rate savings do not grow with the chosen QP difference. While a QP difference of 6 already gives great rate savings a difference of 12 or more does not further decrease the bitrate by the same magnitude. However, the perceived image quality of the not ROI regions suffers badly when the QP difference is increased to 12 or even 18.

IV. CONCLUSION AND FUTURE WORK

In this paper, we proposed a system that combines face detection, tracking and region of interest based encoding to improve the users video conferencing experience. By choosing a coarser quantization for the non ROI regions we can either save a significant amount of bandwidth or increase the quality of the video inside the ROI. When this system is combined with our proposed scene composition, the non ROI regions and their coding artifacts are removed which improves the quality of the video conference. However, also without scene composition the user experience is enhanced by shifting the encoder focus into the regions that are interesting to the user.

In future works, the accuracy of the face detection and tracking can be further improved to provide reliable information also in difficult environments. Additionally, the shape

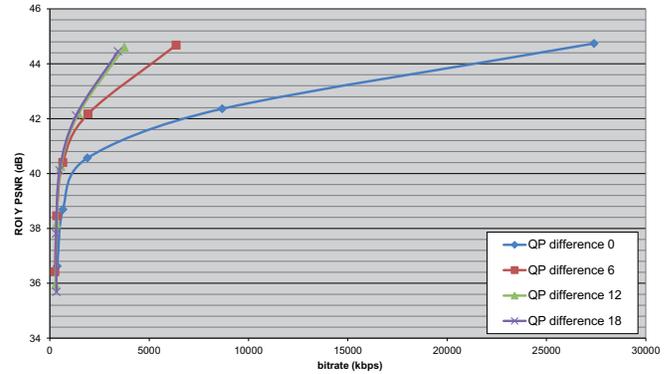


Figure 7. ROI Y-PSNR vs Bitrate for the sequence *changing_lighting* and different differences relations between the QP inside and outside of the ROI.

of the ROI region can be better adapted to the speaker (e.g. give a higher priority to the face) then choosing a constant QP in a rectangular region around the face.

ACKNOWLEDGMENT

This work was co-funded by the German federal state North Rhine Westphalia (NRW) and the European Union (European Regional Development Fund: Investing In Your Future).

REFERENCES

- [1] P. Viola and M.J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, 2004, pp. 137–154.
- [2] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, 1995, pp. 790–799.
- [3] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, 1997, pp. 119–139.
- [4] L. Ferreira, L. Cruz and P.A. Assunção, "H. 264/SVC ROI encoding with spatial scalability," in *Proc. of International Conference on Signal Processing and Multimedia Applications*, 2008, pp. 212–215.
- [5] T. Wiegand, G.J. Sullivan, G. Bjontegaard and A. Luthra, "Overview of the H. 264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, 2003, pp. 560–576.
- [6] D. Caulfield and K. Dawson-Howe, "Evaluation of multi-part models for mean-shift tracking," in *Proc. of International Machine Vision and Image Processing Conference*, 2008, pp. 77–82.
- [7] P. Hosten, A. Steiger, C. Feldmann and C. Bulla, "Performance evaluation of object representations in mean shift tracking," in *Proc. of International Conferences on Advances in Multimedia*, 2013.

Depth Map Compression with Diffusion Modes in 3D-HEVC

Yun Li, Mårten Sjöström, Ulf Jennehag and Roger Olsson

Dept. of Information Technology and Media

Mid Sweden University

Sundsvall, Sweden

yun.li@miun.se, marten.sjostrom@miun.se, ulf.jennehag@miun.se, and roger.olsson@miun.se

Abstract—For three-dimensional television, multiple views can be generated by using the Multi-view Video plus Depth (MVD) format. The depth maps of this format can be compressed efficiently by the 3D extension of High Efficiency Video Coding (3D-HEVC), which has explored the correlations between its two components, texture and associated depth map. In this paper, we introduce two modes for depth map coding into HEVC, where the modes use diffusion. The framework for inter-component prediction of Depth Modeling Modes (DMM) is utilized for the proposed modes. They detect edges from textures and then diffuse an entire block from known adjacent blocks by using Laplace equation constrained by the detected edges. The experimental results show that depth maps can be compressed more efficiently with the proposed diffusion modes, where the bit rate saving can reach 1.25 percentage of the total depth bit rate with a constant quality of synthesized views.

Keywords—Depth map coding; HEVC; diffusion modes

I. INTRODUCTION

The Multi-view Video plus Depth (MVD) video format consists of two components: texture and depth map, where a combination of these components enables a receiver to generate arbitrary virtual views. The 3D extension of High Efficiency Video Coding (3D-HEVC) [1] utilizes different prediction techniques to improve the compression efficiency for MVD data. We have previously devised an edge-based compression scheme by diffusion for depth images [2], as the depth image can be assumed to be piece-wise smooth bounded by sharp edges. The question is if better compression of depth maps can also be achieved by implementing diffusion modes block-wise in 3D-HEVC.

Three dimensional video representation using depth map reduces the number of views being transmitted, but coding of depth maps with the current techniques H.264/AVC [3] or its multi-view coding (MVC) extension [4] will introduce visible distortions in synthesized views. Therefore, they are not recommended for depth coding [5]. Various schemes have been developed to address problems of depth map coding. In paper [6], a comparative study showed that Block Truncation Coding (BTC) outperforms the Discrete Cosine Transform (DCT) and the Karhunen-løeve Transform KLT, and the adaptive BTC was devised that adaptively selects the block size for the BTC. Weighted mode filtering with depth dynamic range reduction [7] and Edge-weighted Optimization Concept (EWOC) with adaptive filtering [8] have

been proposed for depth compression. Model based intra coding approach using a depth lookup table and encoding the residuals in pixel domain in 3D-HEVC was devised in paper [9]. The edge-based depth image coding schemes [2][10] utilize diffusion to interpolate the smooth areas bounded by depth edges. There are still many other algorithms for depth map coding, but in this research work we focus on improving our previously proposed edge-based diffusion scheme [2].

The edge-based depth image compression scheme can preserve the transitions on the depth better than traditional video and image encoders [2]. However, such a scheme implies a very expensive encoding of edge contour information in terms of bit rate.

To solve this issue, edges can be extracted from the co-located texture image. Inter-component prediction for depth map coding has recently been implemented in 3D-HEVC [11]. It may employ an inter-component predicted wedgelet partitioning or a predicted contour partitioning for intra coding, the former separates a depth block into two parts by a straight line, the latter divides the block into parts of arbitrary shapes. Intensities of each part are then represented by constant values. The wedgelet partition for a depth block is found by searching for the best wedgelet pattern in the co-located texture luminance block. The contour partition is also detected from the texture luminance. This partition is selected depending on the pixel values in relation to their mean within the texture block, whereby the partition may be of arbitrary shape. The two depth values given to the different parts of the depth block are predicted from the partially reconstructed depth. Fig. 1 shows a depth block, where $P1$ and $P2$ are decoded values in the adjacent blocks. The edges in the current block are detected from the co-located texture luminance by thresholding, and the parts A and B are predicted by the mean of $P1$ and $P2$, respectively.

Another issue with our previously proposed edge-based scheme is the lack of rate distortion control for optimizing the compression ratio. Therefore, one of the solutions for this is to implement the diffusion process in a block-wise manner in HEVC.

A block-based diffusion method based on Laplace equation for H.264 was proposed in [12]. It detects an edge map from the depth map and encodes these edges by the

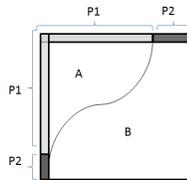


Figure 1. A depth block for inter-component prediction: the edge between part A and B are detected from the co-located texture block, and parts of the adjacent reconstructed blocks available for prediction are showed in dark and light grey.

bi-level image compression tool JBIG. The method uses these edges as constraints with the Laplace diffusion when reconstructing the depth map on blocks of a fixed size.

In this work, we propose two new modes based on block-wise Laplace diffusion. We replace two inter-component prediction modes in the 3D-HEVC by the proposed modes in order to save bits used for signaling. The novelties of this work are: (1) block-based diffusion modes are introduced into HEVC using inter-component prediction framework; (2) the block size is allowed to be further split in the same way as the original inter-component prediction modes; (3) the diffusion is conducted in two-step if isolated parts still exist in the block.

The overall aim of the work is to improve compression ratio for depth maps with a sustained 3D video quality. The work is limited to reusing the inter-component prediction framework, and the goals is to investigate the rate-distortion ratio for the new proposed diffusion modes, where quality is measured on synthesized views.

The sequel of the paper is organized as follows. We illustrate the proposed modes in Section 2, and test arrangements and evaluation criteria in Section 3. Section 4 presents the results and analysis, and Section 5 concludes the work.

II. PROPOSED METHOD

Fig. 2 illustrates all eight Depth Modeling Modes (DMM) in the 3D-HEVC software [13]. They are derived from the 3D-HEVC test model [1]. Among them, the modes (1), (2), (7) and (8) employ wedgelet partitions detected by a search on the depth block or predicted from the previously coded blocks, i.e., they are non-inter-component prediction modes. The inter-component prediction modes (3), (4), (5) and (6) derive, on the other hand, the partition information from the co-located texture block. Mode (2), (4), (6) and (8) employ so called delta constant partition value coding, i.e., they encode the difference between the mean of the original signal and the predicted constant value, which is the mean of the available adjacent prediction signal.

The original mode (5) in Fig. 2 is denoted as DMM-TEX-CONTOUR in the context of this paper. In this mode, the partition is detected from the co-located texture luminance by thresholding. As mentioned, the partition can also be detected by searching for the best Wedgetlet pattern in

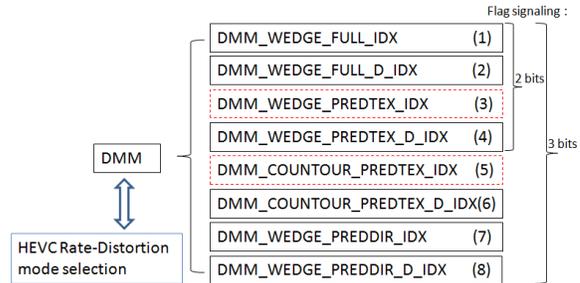


Figure 2. Depth modeling modes. (1) and (2) Explicit wedgelet signaling. (3) and (4) Inter-component predicted wedgelet partitioning. (5) and (6) Inter-component predicted contour partitioning. (7) and (8) Intra-predicted wedgelet partitioning.

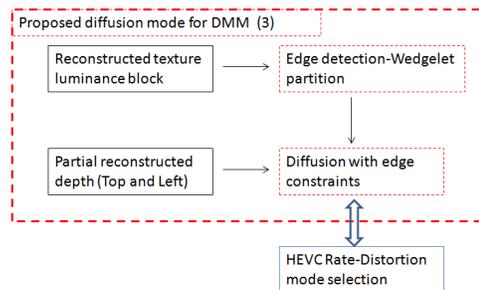


Figure 3. Proposed diffusion mode for DMM-TEX-WEDGE.

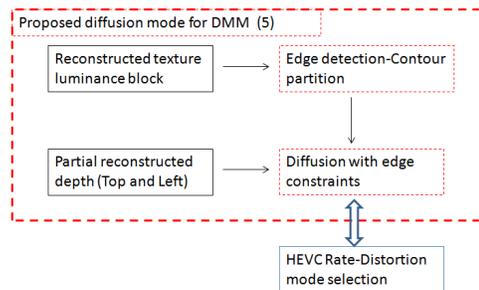


Figure 4. Proposed diffusion mode for DMM-TEX-CONTOUR.

the co-located texture block, i.e., in mode (3). This inter-component predicted wedgelet partition may avoid a possible inconsistency of the contour detection from the texture [11]. We also denote the original mode (3) in Fig. 2 as DMM-TEX-WEDGE.

The DMM-TEX-WEDGE and DMM-TEX-CONTOUR were replaced by the proposed diffusion modes for intra prediction. We replaced the existing modes instead of adding new modes in the inter-component prediction framework because no additional bits had been required for signaling the proposed modes. The original partitioning methods were kept and the obtained edges are used as constraints in the Laplace diffusion.

The proposed modes, illustrated in Fig. 3 and 4, thus include two processes: Edge detection and Diffusion with edge constraints, which are defined as follows:

Edge detection: The modes kept the original partitioning

methods using texture luminance for the edge detection. They require no extra bits for encoding the depth edges, whereas explicit coding of depth edges require substantial amount of coding bits. The methods for obtaining the edges are different for the contour and the wedgelet partitioning.

Edge detection-Wedgelet partition: The wedgelet partitioning is carried out by an efficient wedgelet search on the co-located texture block for the least distortion. Edge is the straight line that separates two parts.

Edge detection-Contour partition: The contour partition for the depth block is made by a thresholding process, in which parts from the partitioning are obtained based on if the value in the co-located texture block is above or below the mean value of this texture block. Edges are located at the transitions between the parts.

Diffusion with edge constrains: The new diffusion modes for intra prediction are also showed in Fig. 1. The parts A and B are diffused from $P1$ and $P2$ respectively.

Laplace equation

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0, \quad (1)$$

is employed for the diffusion. The unknowns of the equations are solved by a method described in [12]. The method refines the diffusion iteratively as:

$$f^{(n+1)}(x, y) = \frac{1}{N(x, y)} \sum_{(i, j) \in u_4(x, y)} C(i, j) f^n(i, j), \quad n = 1, 2, 3 \dots M \quad (2)$$

$$C(x, y) = \begin{cases} 1, & \text{if } f(x, y) \text{ is available} \\ 0, & \text{else,} \end{cases} \quad (3)$$

$$N(x, y) = \sum_{(i, j) \in u_4(x, y)} C(i, j). \quad (4)$$

The equations describe that the diffusion for a depth map block $f^{(n)}$ is refined iteratively with the number of iteration (n). $u_4(x, y)$ represents the four neighbors (up, right, down and left) around the current refined pixel with position (x, y) in the block. $C(i, j)$ denotes the availability of these neighbors (e.g., the pixels taken into calculation are available and belong to the same part), and $N(x, y)$ sums up the number of the available neighbors. The iteration stops with a convergence condition in Equation 5a. In addition to this condition, we also impose a time constrain for the diffusion, which is to limit the number of iterations M . Therefore, the diffusion process stops when either of the conditions a or b is satisfied:

$$\begin{cases} a. & |f^{(n+1)} - f^{(n)}| < 0.05 \\ b. & n \geq M. \end{cases} \quad (5)$$

A. Two-step diffusion

The contour partition may appear much more complex than the one showed in Fig. 1. The parts can be arbitrary shapes and even be isolated within a block. An example is depicted in Fig. 5. Our approach to fill these isolated parts is by using a two-step diffusion. In the first step, parts that are connected with the available prediction pixels are diffused, which is illustrated in Fig. 5(d). In the second step, the diffusion is carried out without the edge constraint for only those isolated parts. Fig. 5(e) shows the final diffused block. As to the maximum iterations for the diffusion in Equation 5, we set $M = 20$ for the diffusion step 1 and $M = 10$ for the step 2.

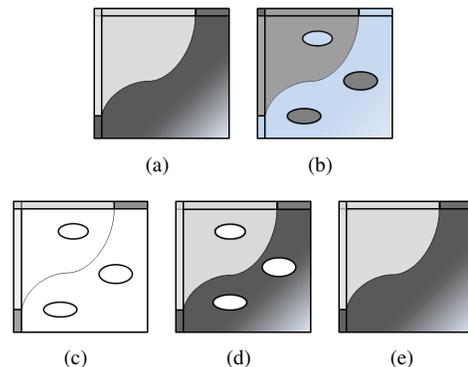


Figure 5. Two-step diffusion: (a) Original depth block, (b) co-located texture block, (c) detected edges on the texture, (d) diffusion after the first step, and (e) diffusion after the second step.

Such a diffusion process might produce erroneous depth values for the isolated parts, but these edges detected from the co-located textures might also not exist in the depth block. The HEVC rate-distortion process decides if the proposed modes are chosen.

III. TEST ARRANGEMENTS AND EVALUATION CRITERIA

The test arrangements and evaluation criteria are described as follows:

A. Implementation and Test setup

The proposed modes have been implemented in 3DV HEVC Test Model (3DV-HTM) software version 4.1 [13]. The evaluation partially followed the Call for Proposals on 3D Video Coding Technology [14]. However, we evaluated only the intra-frame coding to better understand the effectiveness of the proposed intra diffusion modes. Therefore, the bit rate anchors were not followed. We chose two-view configurations and four test sequences with fixed Quantization Parameter (QP) pairs for texture and depth. The MPEG test sequences [14]: Poznan Street [15], Poznan Hall [15], Undo Dancer, and Newspaper were selected. The first 50 frames from these sequences were evaluated.

We used Poznan Street view 3, Poznan Hall view 6, Undo Dancer view 2 and Newspaper view 4 for the evaluation

of depth. Virtual views were rendered at camera position 3.5 for Poznan Street, camera position 6.5 for Poznan Hall, position 3 for Undo dancer and position 5 for Newspaper for the assessment of synthesized views. The virtual views were synthesized from the decoded texture and the decoded depth, and compared to the virtual views synthesized from the original texture and the original depth. VSRS [16] version 3.5 was employed for the view synthesis.

The View Synthesized Optimization (VSO) [17] was turned off, i.e., in the HEVC rate-distortion optimization, the distortion is measured on the depth map instead of on the synthesized view when encoding of depth map. The QPs in (texture, depth) format were (20, 30), (25, 34), (30, 38) and (35, 42). These QPs were selected because the bit rate for the depth should be significantly lower than for the texture for an optimized bit rate allocation between texture and depth [17]. The results using the alternated 3D-HEVC with the proposed modes were compared to results using the original 3D-HEVC in the same testing conditions.

B. Evaluation criteria

The results were calculated using the BD-PSNR model [18]. In this model, a curve is fitted through the PSNR values of four bit-rate points. The difference between the integrals divided by their respective integration intervals is the average difference for two curves. In the evaluation, the bit rate change for depth was computed over the bit rates for the depth map versus PSNR of the decoded depth map, whereas the bit rate change for the synthesized views was calculated over the bit rates for the depth map versus PSNR of the synthesized view.

The complexity of the modes is presented as a ratio of total coding time between the proposed scheme and the 3D-HEVC.

IV. RESULTS AND ANALYSIS

The results are illustrated in Table I. The bit rate saving is around 0.64 percent for Poznan Hall, 0.47 for Poznan Street and 0.28 for Newspaper when only the depth quality is considered. When the evaluation of PSNR is on the synthesized views, around 0.49 percent bit rate savings were achieved for Poznan Hall and 0.31 percent for newspaper. Better bit rate savings were obtained for the synthetic sequence Undo Dancer, where 1.54 percent for the depth and 1.25 percent for the synthesized views were achieved. The results further show that there is no improvement for the Poznan Street sequence when considering the synthesized views.

Table II summarizes the complexity of the proposed modes. The complexity increases in average by 6.2 percent for encoding and 3.4 percent for decoding. An exception is for Undo Dancer sequence, where the decoding time is 4.2 percent less than for the 3D-HEVC. This implies that, in some cases, the proposed diffusion modes are more efficient

TABLE I. BD-PSNR FOR THE TESTED SEQUENCES (THE BIT RATE CHANGE IN PERCENTAGE OF THE TOTAL DEPTH BIT RATE)

Sequence	BD-rate(depth) (%)	BD-rate(virtual view) (%)
Undo Dancer	-1.536	-1.247
Newspaper	-0.282	-0.312
Poznan Street	-0.465	0.026
Poznan Hall	-0.642	-0.488
Average	-0.731	-0.505

TABLE II. CODING COMPLEXITY (TIME RATIO BETWEEN PROPOSED AND REFERENCE SCHEMES)

Sequence	Encoding	Decoding
Undo Dancer	1.041	0.958
Newspaper	1.055	1.096
Poznan Street	1.049	1.055
Poznan Hall	1.102	1.026
Average	1.062	1.034

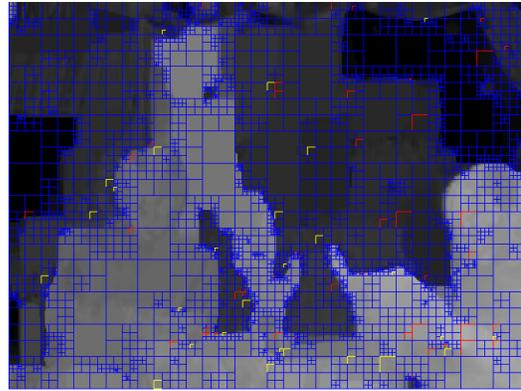


Figure 6. A depth image from the first frame of Newspaper: the blocks marked with red and yellow use the proposed modes that replaced the DMM-TEX-WEDGE and DMM-TEX-CONTOUR respectively.

in decoding than some of the other intra modes in the 3D-HEVC.

An example of block fragmentations and modes assignments are plotted in Fig. 6. Our proposed modes are marked with red and yellow color, which represent the two proposed modes that replaced DMM-TEX-WEDGE and DMM-TEX-CONTOUR, respectively. The total area covered by the proposed modes is 2.33 percent of the entire image among all intra prediction modes in Fig. 6.

The test results illustrate that better compression of depth maps can be achieved with the proposed modes in 3D-HEVC, and that the decoding complexity increases by less than 4 percent. The proposed modes target only inter-component prediction framework, and they cover a very small percentage of the entire depth map. Thus the effectiveness seems less significant. By replacing further intra-modes by diffusion modes, it is likely that further depth compression may be achieved.

The experimental results also demonstrate that the improvement for the quality of decoded depth is consistent. This implies that the Laplace diffusion process can better approximate the original depth signals than the constant

partition value coding in 3D-HEVC under the given testing conditions. The fast advancement in hardware processing power will likely make high computational complexity less of a problem in the future.

This work aimed at improving depth compression (reducing bandwidth consumption) for a better quality of synthesized views in 3D-HEVC, which is the state of the art in coding of 3D video contents. With the proposed diffusion modes, the proposed scheme outperforms the original 3D-HEVC. As coding of 3D video contents has been attracting many research attentions, we also aim at comparing our scheme with other novel methods and improving the proposed scheme further in the future research.

V. CONCLUSION

We have implemented two modes using diffusion in 3D-HEVC for coding of depth map and replaced two inter-component prediction modes by the proposed modes. They utilize edges from the associated texture and diffuse depth values in the block by using Laplace equation with texture edge constraints.

The experimental results illustrate that the proposed modes can improve the compression efficiency for depth map coding, and that the complexity increases by 3.4 percent in average for the decoding. When considering the quality of synthesized views, the bit rate saving can reach around 1.25 percentage of the total depth bit rate for the tested MVD sequences. The bit rate saving is efficient, considering that the proposed modes have been implemented in the inter-component prediction framework only and cover a very small percentage of the depth image among all intra prediction modes.

Future works consist of better edge detection schemes to reduce the partitioning errors for diffusion, investigating the possibility of introducing diffusion into further intra modes, optimizing the proposed modes with View Synthesized Optimization (VSO) enable and subjective quality oriented encoding by using the diffusion modes for a better view synthesis.

ACKNOWLEDGMENT

This work has been supported by grant 2009/0264 of the Knowledge Foundation, Sweden, by grant 00156702 of the EU European Regional Development Fund, Mellersta Norrland, Sweden, and by grant 00155148 of Länsstyrelsen Västernorrland, Sweden.

REFERENCES

- [1] "3D-HEVC Test Model Description Draft 1," ITU-T SG 16 WP 3 JCT3V-A1005_d0, July 2012.
- [2] Y. Li, M. Sjöström, U. Jennehag, and R. Olsson, "A scalable coding approach for high quality depth image compression," in 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2012, Oct. 2012, pp. 1–4.
- [3] "Advanced video coding for generic audiovisual services," ITU-T Recommendation H.264 and ISO/IEC 14496-10, Jan. 2012.
- [4] Y. Chen, Y.-K. Wang, K. Ugur, M. M. Hannuksela, J. Lainema, and M. Gabbouj, "The emerging mvc standard for 3d video services," EURASIP J. Appl. Signal Process., vol. 2009, Jan. 2008, pp. 8:1–8:13.
- [5] K. Muller, P. Merkle, and T. Wiegand, "3-d video representation using depth maps," Proceedings of the IEEE, vol. 99, no. 4, April 2011, pp. 643–656.
- [6] H. Nayyar and A. Wei, "A Comparative Study of Depth-Map Coding Schemes for 3D Video," Image and Video Compression, Stanford University, Mar. 2011.
- [7] V.-A. Nguyen, D. Min, and M. N. Do, "Efficient techniques for depth video compression using weighted mode filtering," Circuits and Systems for Video Technology, IEEE Transactions on, vol. 23, no. 2, Feb. 2013, pp. 189–202.
- [8] S. Schwarz, M. Sjöström, and R. Olsson, "Depth map up-scaling through edge-weighted optimization," in Proc. SPIE 8290, Three-Dimensional Image Processing (3DIP) and Applications II, Feb. 2012, pp. 829008–8.
- [9] F. Jager, M. Wien, and P. Kosse, "Model-based intra coding for depth maps in 3d video using a depth lookup table," in 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2012, Oct. 2012, pp. 1–4.
- [10] J. Gautier and O. Meur, "Efficient depth map compression based on lossless edge coding and diffusion," Picture Coding Symposium, 2012, pp. 81–84.
- [11] P. Merkle, C. Bartnik, and K. Muller, "3D video: Depth coding based on inter-component prediction of block partitions," in Proc. Picture Coding Symposium, 2012, pp. 149–152.
- [12] J. Chen, F. Ye, J. Di, C. Liu, and A. Men, "Depth map compression via edge-based inpainting," Picture Coding Symposium, 2012, pp. 57–60.
- [13] 3DV HEVC Test Model (3DV-HTM) version 4.1. Retrieved: 09, 2010. [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSsoftware/tags/HTM-4.1/
- [14] "Call for Proposals on 3D Video Coding Technology," ISO/IEC JTC1/SC29/WG11, Mar. 2011.
- [15] M. Domaski, T. Grajek, K. Klimaszewski, and M. Kurc, "Pozna Multiview Video Test Sequences and Camera Parameters," ISO/IEC JTC1/SC29/WG11, 2009.
- [16] "Report on experimental framework for 3D video coding," ISO/IEC JTC1/SC29/WG11 MPEG2010/N11631, Oct. 2010, Guangzhou, China.
- [17] G. Tech, H. Schwarz, K. Muller, and W. Thomas, "3D video coding using the synthesized view distortion change," Picture Coding Symposium, 2012, pp. 25–28.
- [18] G. Bjontegard, "Calculation of average PSNR differences between RD-curves," ITU-T VCEG-M33, Mar. 2001.

Efficient Stream-Reassembling for Video Conferencing Applications using Tiles in HEVC

Christian Feldmann
 Institut für Nachrichtentechnik
 RWTH Aachen University
 Aachen, Germany
 feldmann@ient.rwth-aachen.de

Christopher Bulla
 Institut für Nachrichtentechnik
 RWTH Aachen University
 Aachen, Germany
 bulla@ient.rwth-aachen.de

Bastian Cellarius
 RWTH Aachen University
 Aachen, Germany
 Bastian.Cellarius@rwth-aachen.de

Abstract—In a modern video conferencing application, the people participating at each client can be detected, tracked and placed in a virtual scene where all persons are of equal size and occupy a predefined rectangular space. This virtual scene can then be rendered on screen instead of a whole room with several people. As a result, a more immerse video conferencing impression is created. At a Multipoint Control Unit (MCU) it is beneficial to disassemble and reassemble the video streams so as to create a custom video stream for each client that only includes people that will be rendered by that client in order to use the available bandwidth to full capacity. In a conventional video coding approach, this video reassembling operation is only possible by decoding all incoming video streams, mixing in pixel domain and then encoding all outgoing video streams. However, this operation makes very high demands on the computational power of the MCU. In this paper, we demonstrate how in the upcoming video coding standard High Efficiency Video Coding (HEVC) the encoder can be modified to enable a reassembling operation that is HEVC compliant and works on a high syntax level in the bitstream. Hereby, no entropy en- or decoding is necessary which makes the operation very low complex.

Keywords- HEVC; video conferencing; video mixing; coded domain; tiles; slices;

I. INTRODUCTION

Current video conferencing systems have the ability to perform high quality, real time conferences between different parties all around the world. However, the demand for high video quality and a more immersive experience is often opposed by the available bandwidth and computational power at the central Multipoint Control Unit (MCU). In order to achieve these goals, an immerse conference scenario is considered in this paper that allows for a low complex video reassembling operation at the MCU.

In classical video conferencing approaches each connected endpoint has one camera. The captured video is then encoded into two video streams: One with a high resolution (e.g. 720p) and a second one with a lower resolution which is used as a thumbnail. Both streams are transmitted to the MCU, that decides which is the most active party and forwards the high resolution video stream of this party to all the other parties. The thumbnail views are always routed

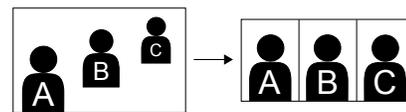


Figure 1. Each person in the scene is extracted from the captured video, scaled and placed side by side.

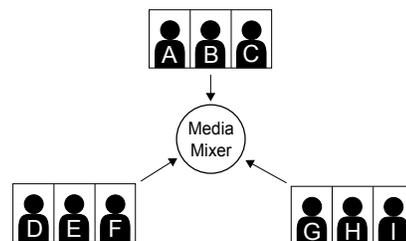


Figure 2. Three clients transmit their processed video stream to the MCU. The last most active people in the conference are E, B, C and I, descending in this order.

to all endpoints. Of course, the active party does not receive the high resolution video of itself but the high resolution video of the last active party. Hereby, each party can see a high resolution video of the active speaker and thumbnails of the other parties.

In this scenario, a combination of face detection, tracking and audio analysis is used in order to process the input video and extract persons from the captured video. Each person is then scaled and placed side by side (See Figure 1). This video is then only encoded in high resolution and transmitted to the MCU and from there to all clients (See Figure 2). Each client decodes the incoming video streams from the other clients and crops out only the last most active people to render them on screen. In Figure 3 an example is shown in which each client only renders the last most active speakers.

However, in this scenario the MCU transmits a lot of information to each client that the client discards after cutting out the people that it is going to render on screen. It is obvious that the required bandwidth can be significantly reduced, if the MCU supports a reassembling operation that allows outputting individual streams for each client containing only

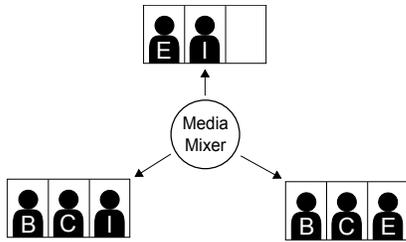


Figure 3. Each client receives and renders only the last most active and relevant people on screen. The last most active people in the conference are E, B, C and I, descending in this order.

relevant parts of the video. In the conventional approach the MCU would decode the incoming video streams, rearrange the video in the pixel domain and re-encode a new stream for each client. While this approach is simple, it has two disadvantages: Running decoders and encoders requires a lot of computing power at the MCU and has a negative impact on the overall compression performance [1] [2].

In the following Sections, we introduce a method that uses the upcoming video coding standard High Efficiency Video Coding (HEVC) [3] and the concept of Tiles [4] in order to logically split the video stream into sub-streams, with each sub-stream containing exactly one person. The video stream for each client can then be easily assembled in the MCU by only copying packets from the input video streams and altering some flags in the headers. After this operation, the output streams are still compliant to the HEVC standard and can be decoded by any device supporting HEVC.

Naturally, the proposed approach is not limited to video conferencing applications. It can be utilized in all applications where a video can be logically split into separate areas and only some of these areas need to be transmitted or the stream needs to be reassembled during transport.

In the following Section II, selected topics from HEVC will be presented that are used in the scope of the approach, before stream reassembling operations (Section III) and required encoder restrictions carried out in this approach are explained in Section III and IV. Afterwards, the arising compression loss due to the usage of Tiles and Slices and due to the encoder restrictions is measured in Section V. Finally, a conclusion about the approach is drawn in Section VI.

II. HEVC

In order to split the video stream into sub-streams, Slices and Tiles are combined in this approach to enable a high syntax level reassembling operations. This Section will give a brief overview of the HEVC tools and techniques that are used and/or modified in the proposed method.

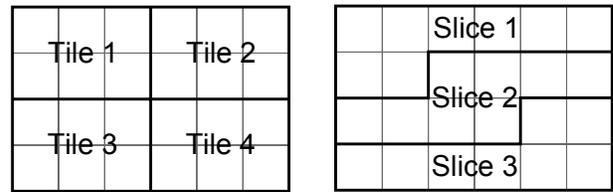


Figure 4. Subdivision of a picture with 24 CTUs into Tiles (left) and Slices (right).



Figure 5. The picture is divided into 24 CTUs and into 4 Slices as well as 4 Tiles of equal size. Each Tile and each Slice contains 6 CTUs.

A. Slices and Tiles

Both Slices and Tiles can subdivide a frame into logically separate parts that can be decoded independently. Tiles are defined via a number of Coding Tree Units (CTUs) for the width and the height of the Tile and are therefore always rectangular, while Slices simply contain a number of CTUs that are laid out in raster scan order (See Figure 4). Slices and Tiles can be used at the same time so that a Slice can contain Tiles or a Tile can contain Slices. A special situation can occur, when Slices and Tiles contain the same number of CTUs. In this case, each Tile contains exactly one Slice and the borders of Slices and Tiles match (See Figure 5).

Since Slices and Tiles do not allow prediction across Tile/Slice boundaries or entropy coding dependencies, they are independent with respect to the encoding and decoding process [4]. Thus, Slices and Tiles can be processed in parallel which can be utilized in parallel implementations and lower the latency of the en-/decoding process [5] [3].

B. Bitstream Syntax

As in H.264/AVC, in HEVC all coded content is embedded into Network Abstraction Layer (NAL) units, which are byte aligned and have a header identifying the kind of payload. A NAL unit can contain a Slice, but also different kinds of parameter sets. Several NAL units form an Access Unit (AU), where decoding an AU results in one decoded picture and must thus contain at least all Slices of that picture. Parameter sets contain information about the whole sequence or one picture and are not entropy coded. Each bitstream must contain at least one Sequence Parameter Set (SPS) and one Picture Parameter Set (PPS), which are valid until another parameter set of the same kind is referenced (example in Figure 6) [3].

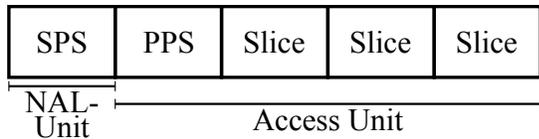


Figure 6. Bitstream containing a Sequence Parameter Set, a Picture Parameter Set and some Slices. More Access Units can follow of course.



Figure 7. Structure of the Bitstream when using one Slice in each Tile (colored parts are entropy coded).

In the bitstream each Slice is composed of a header and a payload. While the header contains general information about the Slice in high level syntax, the payload is entropy coded and may contain several Tiles, if used. The end of a slice is signaled using the *end_of_slice_flag* which is the last symbol coded for each CTU and is set if the current CTU is the last CTU in the Slice. If the Slice contains Tiles, all Tiles are either separated by fixed byte sequences in the Slice payload or the byte positions in the Slice payload are given in the corresponding Slice header. When using one Tile per Slice, the payload of each Slice NAL unit contains exactly one Tile, but no information about the Tile entry points in the Slice payload is necessary (see Figure 7). Thus, the bitstream appears to contain rectangular Slices that are laid out in raster scan order in the Frame. Still, the definition of column and row boundaries of the Tiles is present in the sequence and/or picture parameter set and the Slices are identified using the *slice_address* given in the Slice header, which is the raster scan index of the first CTU in the Slice [3].

C. Inter-Prediction in HEVC

In HEVC, each frame is subdivided into Coding Tree Units (CTUs) and each CTU can be subdivided into Coding Units (CUs) of different sizes. Each CU can then be split into one, two or four Prediction Units (PUs), which are predicted using either motion compensation or intra prediction. An example for the partitioning of a CTU into CUs and PUs is shown in Figure 8 [3].

There are two different ways of signaling motion information for inter prediction to the decoder:

- 1) A PU can use the so called merge mode where the reference frame index and the motion information for the current PU are inferred from a neighboring PU. In order to merge a PU, a candidate list is filled and only the index of this list is encoded into the bitstream. The order in which neighboring PUs are added is standardized so the decoder can build an identical merge candidate list. A candidate is only

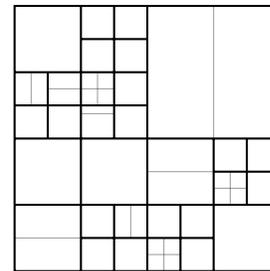


Figure 8. Example for a partitioning of a CTU in CUs (black) and PUs (grey).

added when the corresponding PU exists, has inter prediction related information and fulfills several more conditions (e.g. if it is located within the same CTU, Tile or Slice). There are two types of candidates (See Figure 9): The ones taken from a PU within the same frame (spatial candidates) and the ones taken from a PU in a collocated frame (temporal candidates). While the spatial candidates need to be in the same Slice/Tile as the current PU, temporal candidates do not have this restriction in general; the only restriction is, that the candidate must be located in the same CTU line [3].

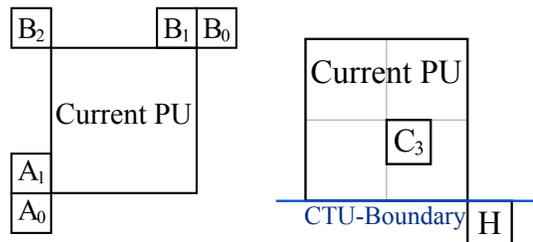


Figure 9. The spatial (left) and temporal (right) candidates that are checked for motion vector prediction as well as merge mode.

- 2) When the merge mode is not used for a PU the reference indices and motion information needs to be explicitly encoded into the bitstream. However, also in this case the motion information is not directly encoded but it is predicted and only the motion vector difference is encoded. In HEVC, Advanced Motion Vector Prediction (AMVP) is used which works quite similar to the merge mode. A list of possible prediction candidates is created using neighboring candidates as well as temporal candidates. Afterwards the chosen index as well as the motion vector difference is encoded into the bitstream [3].

III. STREAM REASSEMBLING

In this Section a high level reassembling operation is proposed. The usage of Tiles and the reassembling operation are similar to the proposed method in [6]. However, in [6] the definition of Slices and Tiles is changed in order

to enable the reassembling operation. This results in an output videostream that is not HEVC compliant and requires changes on the decoder side. The reassembling operation that is presented in this Section however, only utilizes changes on the encoder side and thus is always HEVC conforming. This allows any HEVC compliant decoder to decode the resulting bitstream.

As described in Section II, Tiles can be used to split the video stream into rectangular areas with each Tile containing one person. However, if only multiple Tiles are used, the Tiles cannot be rearranged as freely as the application requires. The problem is the *end_of_slice_flag*, which is only set for the last CTU in the last Tile. If we were to insert the last Tile with the *end_of_slice_flag* set at a different position than that of the last Tile, the set *end_of_slice_flag* would be received before all CTUs of the Slice are decoded. This results in a bitstream that is not conforming to the HEVC standard and might not be decodable. In addition the *end_of_slice_flag* is entropy coded in the bitstream and cannot be changed without entropy de-/encoding the Slice payload.

In order to circumvent this limitation we use Tiles that contain exactly one Slice as described in Section II-A. Since the concepts of Slices and Tiles coincide in this situation, we will use them synonymously hereafter. This way, each person is contained in exactly one NAL unit that contains one Tile/Slice. The people in the video stream can now be reordered by simply inserting the correct NAL units into the new bitstream while modifying the Slice headers and some parameter sets. In the Slice header the *slice_address* has to be modified to match the position of the Tile in the new video stream. Also the *first_slice_in_pic_flag* in the Slice header has to be set or reset. Furthermore, several parameters in the sequence and/or picture parameter set have to be adjusted for the new arrangement of Tiles. Concretely these are the *pic_width_in_luma_samples* and *pic_height_in_luma_samples* values as well as the *num_tile_columns_minus1*, *num_tile_rows_minus1*, *uniform_spacing_flag*, *col_width* and *col_height* syntax elements [3].

Overall, this reassembling operation is very low complex. Only a few values in the slice headers have to be changed and the entropy coded slice payload is simply copied to the output stream while in the conventional approach a full encoder as well as a full decoder is needed to create a similar result.

IV. REQUIRED ENCODER RESTRICTIONS

The definition of Tiles and Slices in HEVC allows for independent decoding of each person. The entropy coder is reset after each Tile and prediction is generally not allowed across Tile boundaries. However, this is only true for dependencies within one frame. Some dependencies on other Tiles in past frames that are in the reference buffer can



Figure 10. Decoding error after switching two Tiles due to dependencies between the Tiles (right) and the original sequence (left).

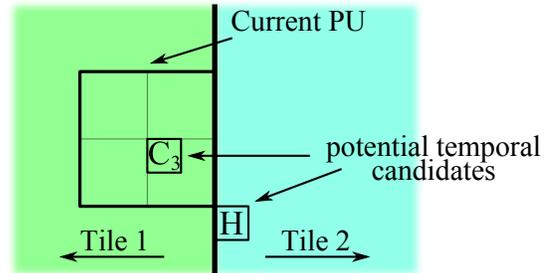


Figure 11. Temporal candidates of PU located at the edge of a Tile.

still exist. If these remaining dependencies are not removed, decoding errors as shown in Figure 10 can occur when information is referenced that changed in the reassembling operation.

A. Modified Candidate Lists in AMVP and Merge

While creating the candidate lists for either AMVP or Merge, the possible temporal candidate *H* is located outside of the current PU and may be located outside of the current Tile (See Figure 11). If information from candidate *H* is used and *H* is located outside of the current Tile and that Tile was removed or replaced by the reassembling process, the information from candidate *H* cannot be determined by the decoder. In this case, candidate *H* must not be used for prediction. Also, it is possible that no Tile is present at position *H* while decoding. This makes all candidates after *H* unusable as well since the encoder cannot know if *H* will be available at the decoder or not and thus cannot predict how the candidate list after *H* is constructed at the decoder. If no Tile is present at the position of candidate *H* at the encoder, all candidates following the potential position of candidate *H* are unusable as well, since a Tile at the position of candidate *H* could be added during the reassembling process.

A special case occurs when the PU is located near a Tile boundary and merge mode is used to merge the motion information from a neighboring PU. In this case the encoder must not choose a PU for merging that has a motion vector which would cause the current PU to be predicted from a different Tile in the reference Frame.

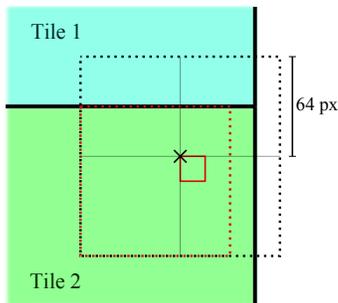


Figure 12. Motion vectors allowed by the standard (black) and the limited motion range taking into account the PU size (red).

B. Restricting Inter Prediction to Tiles

In HEVC, motion vectors are allowed to cross Tile or Slice boundaries. However, the information that is referenced by a motion vector that crosses these boundaries may have changed after reordering the Tiles. In order to only utilize information that is also available at the decoder side we limit the encoder search range near Tile boundaries. At the boundary the dimension of the PU has to be taken into account so that no part of the PU crosses the Tile boundary (See Figure 12).

V. EXPERIMENTAL RESULTS

Using Tiles in a video as well as our restrictions on motion vectors and prediction candidates results in a loss in compression efficiency. In this Section we will evaluate the loss resulting from these modifications.

A. Sequences

Corresponding to the scenario described in Section I, the test sequences are composed of smaller sub-sequences that each contain one person that has been cropped and scaled from the original sequence. An example can be seen in Figure 13. Each test sequence contains three or four sub-sequences with a spatial resolution of 256×256 pixels which corresponds to 4×4 CTUs.

B. Experiments

The proposed encoder modifications were implemented into the HEVC reference software HM version 6.0 [7]. The reassembling operation of the bitstream file was implemented in Python and the rearranged bitstream was decoded using the reference decoder to test HEVC conformance. For the test set, three different configurations were tested:

- 1) The whole sequence without Tiles
- 2) The sequence using one Tile for each sub-sequence
- 3) The sequence using one Tile for each sub-sequence and using the encoder modifications as described in Section IV

The simulations were performed using the low delay main configuration from the common test conditions [8], which



Figure 13. The test sequence Vidyo2 after cropping and scaling.

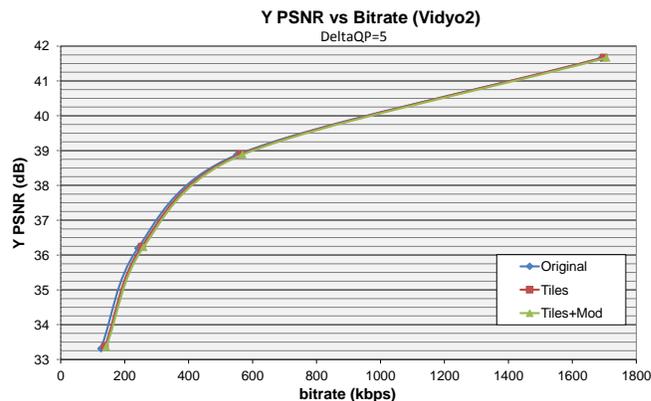


Figure 14. RD-plot of the results for the simulations of sequence “Vidyo2”.

defines a set of coding parameters that are suited for low latency video conferencing applications. The Quantization Parameter (QP) range of 22, 27, 32 and 37 was used, which spans the range of bitrates used in video conferencing applications. In Figure 14, the Rate-Distortion (RD) plot of the result for the simulations of the sequence from Figure 13 is displayed, where the other test sequences exhibit very similar results. In Table I, the Bjøntegaard Delta-rate (BD-rate) [9] overhead for the different scenarios, sequences and color components are displayed and Table II shows the average BD-rate overhead.

C. Evaluation

In general, Tiles as well as our modifications have a higher impact on the performance at lower rate points (high QP). This can be explained by the distribution of the bitrate in the coded stream. While at high QP values a high percentage of the available bitrate is used for the prediction information, this distribution is shifted for low QP values where the

Table I. BD-rate overhead when using Tiles or Tiles and our proposed encoder modifications for each test sequence.

Sequence		Y	U	V
SideBySide	Tiles	5.26%	4.13%	6.96%
	Tiles+Mod	8.69%	8.39%	7.28%
Vidyo1	Tiles	4.73%	2.13%	4.81%
	Tiles+Mod	8.47%	6.03%	8.31%
Vidyo2	Tiles	2.92%	0.06%	0.10%
	Tiles+Mod	5.02%	3.07%	1.66%

Table II. Average BD-rate overhead when using Tiles or Tiles and our proposed encoder modifications.

	Y	U	V
Tiles	4.3%	2.11%	3.95%
Tiles+Mod	7.39%	5.83%	5.75%

main part of the available bitrate is used for coding of the transform coefficients. Table II shows the average rate losses for using Tiles and for using Tiles in combination with our encoder restrictions.

VI. CONCLUSION

In this paper, a method for reordering of Tiles in an HEVC coded bitstream is proposed, that works on a very high syntax level and does not require any entropy de-/encoding of the bitstream. The resulting bitstream again conforms to the HEVC standard and can be decoded by any conforming decoder. In order to achieve this flexibility, Tiles were used in combination with Slices and some modifications to the encoder were applied to remove all remaining dependencies between neighboring Tiles.

Using Tiles and the proposed encoder modifications yields a small compression loss as shown in Section V. However, it adds the ability to arbitrarily reorder Tiles in a low complexity manner to create new bitstreams with different layouts that conform to the HEVC standard. In addition, it allows for parallel processing at the encoder as well as the decoder.

The limitations of this method result from the definition of Slices and Tiles. Since Tiles always contain a defined number of CTUs, the resolution of the Tiles has to be a multiple of the CTU size (usually 64×64 or 32×32). In addition, Tiles are defined using a grid layout, so with the proposed reordering operation, all Tiles must have the same dimensions.

Although the implementation and experiments were done using HEVC draft 6 [3], the key concepts used in the scope of our approach underwent only small changes up to the latest draft 9 [10] in a way that implementing the described approach using draft 9 is still possible. This will most likely also be true for the finished standard since only small changes are to be expected from working draft 9.

REFERENCES

- [1] M. Willebeek-LeMair and Z.-Y. Shae, "Videoconferencing over packet-based networks," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 6, pp. 1101–1114, August 1997.
- [2] A. Vetro, C. Christopoulos, and H. Sun, "Video transcoding architectures and techniques: an overview," *Signal Processing Magazine, IEEE*, vol. 20, no. 2, pp. 18–29, March 2003.
- [3] B. Bross, W.-J. Han, J.-R. Ohm, G. J. Sullivan, and T. Wiegand, "High efficiency video coding (HEVC) text specification draft 6," document JCTVC-H1003, JCT-VC, San Jose CA, USA, Tech. Rep., February 2012.
- [4] A. Fuldseth, M. Horowitz, S. X. A. Segall, and M. Zhou, "Tiles," document JCTVC-F335, JCT-VC, Torino, Italy, Tech. Rep., July 2011.
- [5] K. Misra and A. Segall, "New results for parallel decoding for Tiles," document JCTVC-F594, JCT-VC, Torino, Italy, Tech. Rep., July 2011.
- [6] P. Amon, M. Sapre, and A. Hutter, "Compressed domain stitching of HEVC streams for video conferencing applications," in *Packet Video Workshop (PV), 2012 19th International*, pp. 36–40, May 2012.
- [7] I.-K. Kim, K. Sugimoto, K. McCann, B. Bross, W.-J. Han, J.-R. Ohm, and G. Sullivan, "High efficiency video coding (HEVC) test model 6 (HM 6) encoder description," document JCTVC-H1002, JCT-VC, San Jose CA, USA, Tech. Rep., February 2012.
- [8] F. Bossen, "Common test conditions," document JCTVC-H1100, JCT-VC, San Jose CA, USA, Tech. Rep., February 2012.
- [9] G. Bjøntegaard, "Calculation of average PSNR differences between RD curves," document VCEG-M33, ITU-T Q6/16, Austin TX, USA, Tech. Rep., April 2001.
- [10] B. Bross, W.-J. Han, J.-R. Ohm, G. J. Sullivan, and T. Wiegand, "High efficiency video coding (HEVC) text specification draft 9," document JCTVC-K1003, JCT-VC, Tech. Rep., October 2012.

Disocclusion Handling Using Depth-Based Inpainting

Suryanarayana M. Muddala, Roger Olsson and Mårten Sjöström

Dept. of Information Technology and Media

Mid Sweden University

Sundsvall, Sweden

suryanarayana.muddala@miun.se, roger.olsson@miun.se and marten.sjostrom@miun.se

Abstract—Depth image based rendering (DIBR) plays an important role in producing virtual views using 3D-video formats such as video plus depth (V+D) and multi view-video-plus-depth (MVD). Pixel regions with non-defined values (due to disoccluded areas) are exposed when DIBR is used. In this paper, we propose a depth-based inpainting method aimed to handle disocclusions in DIBR from V+D and MVD. Our proposed method adopts the curvature driven diffusion (CDD) model as a data term, to which we add a depth constraint. In addition, we add depth to further guide a directional priority term in the exemplar based texture synthesis. Finally, we add depth in the patch-matching step to prioritize background texture when inpainting. The proposed method is evaluated by comparing inpainted virtual views with corresponding views produced by three state-of-the-art inpainting methods as references. The evaluation shows the proposed method yielding an increased objective quality compared to the reference methods, and visual inspection further indicate an improved visual quality.

Keywords—3D; video plus depth; warping; depth-image-based rendering; inpainting;

I. INTRODUCTION

In recent years, Three Dimensional Television (3DTV) and Free Viewpoint Television (FTV) have become hot topics in the 3D research area. A common way to transmit the 3D content required for these applications is to use video-plus-depth (V+D) and multi view-plus-depth (MVD) formats, as these ensure a display agnostic rendering of virtual views for both stereoscopic and autostereoscopic multiview displays. A required tool for V+D and MVD formats is view synthesis, which creates content suitable for each specific display type. A fundamental view synthesis method is depth-image-based rendering (DIBR), which produces virtual views using pixel dense texture and depth information. Unfortunately DIBR brings inherent artifacts, mainly caused by disocclusions [1]. Disocclusions are areas that are occluded in an original view that is stored in the format, which become visible in rendered virtual views. Although MVD permits virtual views to be rendered using information from not one but two or more V+D data sets, there still exists a disocclusion problem that needs to be addressed. Mainly due to content with a baseline that significantly differs from that required by a specific display.

Inpainting methods aim to solve the disocclusion problem by filling the unknown regions using neighborhood informa-

tion. Disoccluded areas can be considered as missing texture information alone, as is being done by texture synthesis methods [2]. Criminisi et al. proposed an efficient image inpainting technique that combines the structural and textural propagation into the missing regions [3]. However, this method was not aimed at V+D or MVD formats and thereby could not recognize the differences between foreground (objects closer to the camera) and background parts (objects away from the camera) in a virtual view. As a result it propagates foreground information into the disoccluded areas, which should only contain background information. Daribo et al. extended the exemplar based inpainting to address this limitation by introducing the depth constraint. However, this method only reduces the problem to a degree as it still partly propagates the foreground information into disoccluded regions [4]. Gautier et al. extended the Criminisi method by considering the 3D structure tensor as a data term that identifies the strongest structure in the neighborhood, and added the depth information to calculate the required inpainting priorities [5]. Worth noting with these previous work is that both Daribo et al. and Gautier et. al relies on having true depth map available at the rendered virtual view position. This assumption is in general not feasible or realistic since the depth map of the virtual view also must be estimated.

This paper proposes a novel method to inpainting for V+D and MVD based DIBR. The proposed method relies on the fundamental method introduced in [3] but enhanced using the available depth information. In contrast to [4], [5], we have not relied on having access to a true depth map but instead considered a more general case with having a warped depth map available in our inpainting process.

The outline of the paper is as follows: The related work is briefly reviewed in Section II and the proposed inpainting method is presented in Section III. The test arrangement and evaluation criteria are described in Section IV. The results and analysis are given in Section V and finally we conclude the work in Section VI.

II. RELATED WORK

Criminisi et al. introduced the exemplar based texture synthesis, which effectively replicates both structure and

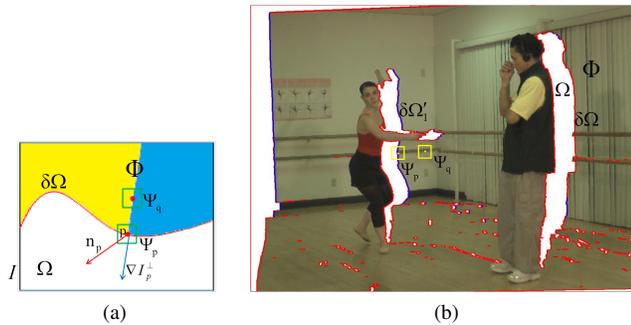


Figure 1. Schematic illustration: (a) notation diagram; (b) warped view with notations.

texture by using the advantages of partial differential equations (PDE) based inpainting method and non-parametric texture synthesis. The quality of the inpainted image is highly dependent on the order in which filling is performed.

For an input image I with an empty region Ω , also known as hole, the source region Φ (the remaining part of the image except the empty region) is defined as $\Phi = I - \Omega$. The boundary between Φ and Ω is denoted as $\delta\Omega$ (see Fig. 1). The basic steps of Criminisi's algorithm are (i) Computing the priorities on the boundary region and (ii) Finding the best match using patch matching. Suppose a patch Ψ_p centered at a pixel p for some $p \in \delta\Omega$ and the priority is computed as the product of two terms:

$$P(p) = C(p) \cdot D(p), \quad (1)$$

where $C(p)$ is the confidence term indicating the amount of non-missing pixels in a patch and the data term $D(p)$ gives importance to the isophote direction.

Once all priorities on boundary $\delta\Omega$ are computed, the highest priority patch $\Psi_{\hat{p}}$ centered at \hat{p} is selected to be filled first. A block matching algorithm is used to find the best similar patch $\Psi_{\hat{q}}$ from which to fill-in the missing pixels:

$$\Psi_{\hat{q}} = \arg \min_{\Psi_q \in \Phi} \{d(\Psi_{\hat{p}}, \Psi_q)\}, \quad (2)$$

where d is the distance between two patches defined as sum of squared difference (SSD). After the most similar patch $\Psi_{\hat{q}}$ is found, the values of the hole pixels in the target patch $\hat{p} | \hat{p} \in \Psi_{\hat{p}} \cap \Omega$ are copied from their corresponding pixels inside $\Psi_{\hat{q}}$. Once the patch $\Psi_{\hat{p}}$ is filled, the confidence term $C(p)$ is updated as follows:

$$C(q) = C(\hat{p}), \forall q \in \Psi_{\hat{p}} \cap \Omega. \quad (3)$$

Daribo et al. extended the Criminisi method first by introducing a depth regularity term in the priority term calculation in (1). The depth regularity term is defined as the

inverse variance of the depth patch centered at p . Their depth regularity term is described as controlling the inpainting process such that the filling order favors the background. Furthermore, the patch matching step is modified by searching for a best patch in both the texture and the depth domain.

Gautier et al. followed the Darios method in considering depth map to help the inpainting process, but introduced a 3D tensor as a data term in the priority calculation of (1) and a one-sided priority to restrict the filling direction. In the patch matching step they also used a weighted combination of the best patches as the final selected patch.

III. PROPOSED INPAINTING METHOD

The novelty of our proposed depth-based inpainting method can be described in three steps:

- A. Depth guided directional priority
- B. Depth included curvature data term
- C. Depth-based source region selection

Fig. 2 shows how these steps relate to the general inpainting process. Step A, consists of defining a depth guided directional priority that selects background patches to be filled first. In Step B, we adopt the Curvature Driven Diffusion (CDD) model similarly to [6] as data term $D(p)$, and extend the CDD model by incorporating depth information. Finally, Step C excludes foreground information from the source region, using depth constraints derived from the warped depth. In the patch matching, a weighted combination of $-N$ best patches is used to define the target patch.

A. Depth guided direction priority

In this step, the boundary extraction block of Fig. 2 is improved by using depth information to guide the filling such that it starts from the background. This because disocclusions result from depth discontinuities between foreground and background, which makes filling the disocclusion from the horizontal background side reasonable. The background side of the disocclusion is obtained as follows. First, a one sided boundary $\delta\Omega_1$ of the disocclusion area is obtained by applying the convolution operation on a disocclusion map (DM) as given in (4). Second, the directional priority selection is further improved by using a depth constraint on $\delta\Omega_1$, such that pixels whose depth values are less than M percent of the maximum depth value in the warped depth map are selected (see the blue colored border in Fig. 1(b)):

$$\delta\Omega_1 = DM * H \quad (4)$$

$$\delta\Omega'_1 = \delta\Omega_1(q) |_{q \in \delta\Omega_1 \cap (Z(q) < M \cdot \max(Z))}, \quad (5)$$

where $\delta\Omega'_1$ is the depth guided boundary, Z is the depth map and $Z(q)$ is the depth value at pixel location q . The convolution kernel H is defined as follows, depending of from which direction the warp is performed:

$$H = \begin{cases} \begin{bmatrix} 0 & 0 & 0 \\ 1 & -8 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} & \text{if left warped view;} \\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & -8 & 1 \\ 0 & 0 & 0 \end{bmatrix} & \text{if right warped view.} \end{cases} \quad (6)$$

Once the hole boundary is obtained, using (4) and (5), priorities are calculated according to (1) utilizing the proposed data term (10). Then the holes in the background regions are filled using the depth guided direction priority. The filling process continues with the one sided boundary priority and finally the holes which are not filled using the one sided boundary priority are processed with total boundary extraction.

B. Depth included data term

As the data term in the general inpainting process we adopt, and add depth to, the CDD model in order to consider the depth curvature along with the texture. The CDD model uses the strength and geometry of an isophote [7], where the latter obtained using scalar curvature. The CDD model is defined as follows:

$$g(s) = s^\alpha, s > 0, \alpha \geq 1 \quad (7)$$

$$k_p = \nabla \cdot \left(\frac{\nabla I_p}{|\nabla I_p|} \right) \quad (8)$$

$$\frac{\partial I_p}{\partial t} = \nabla \cdot \left(\frac{g(|k_p|)}{|\nabla I_p|} \nabla I_p \right), \quad (9)$$

where k_p is the curvature of the isophote through some pixel p , $\nabla \cdot$ is the divergence at p , and g is the control function to adjust the curvature. The conductive coefficient of CDD model is influenced by the isophote strength and curvature. By incorporating the CDD model as a data term in the proposed method and setting $\alpha = 1$ in (7), the data term becomes:

$$D(p) = \left| \nabla \cdot \left(\frac{k_p}{|\nabla I_p|} \nabla I_p \right) \right|, \quad (10)$$

The depth information is considered as an additional channel along with R, G, and B when calculating the curvature and isophote values.

C. Depth-based source region selection

The patch-matching step in the proposed inpainting method is an improvement of the method of [4] and [5]. The improvement consists of classifying the source region using depth information, in order to select similar patches from the

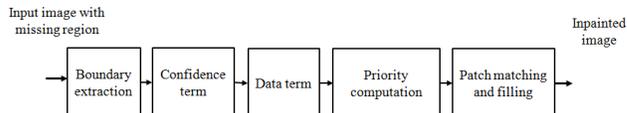


Figure 2. Block diagram of the inpainting method.

nearest depth range. The idea of separating the background region has been previously employed by [8] using patch averages. However, here we classify the source region to enhance the patch-matching step. By considering Φ to be the known source region, which contains both foreground and background regions we avoid patch selection from foreground region by sub-dividing Φ using depth threshold Z_c according to:

$$\Phi_b = \Phi - \Phi_f, \quad (11)$$

where Φ_f is the source region whose depth values are higher than the depth threshold Z_c .

The depth threshold has two different values depending on the variance of the depth patch. If the variance of the depth patch is greater than the threshold γ , the patch might contain unwanted foreground values. The average value of the depth patch is then instead chosen to deduct the foreground parts. Otherwise, the patch contains the constant or continuous area values and so the maximum value in the depth patch is used as the depth threshold to get the best patch according to the depth level. So the depth threshold Z_c is defined as follows:

$$Z_c = \begin{cases} \overline{Z_{\hat{p}}} & \text{if } var(Z_{\hat{p}}(\mathbf{q})|_{\mathbf{q} \in \Psi_{\hat{p}} \cap \Phi}) > \gamma; \\ \max(Z_{\hat{p}}) & \text{otherwise.} \end{cases} \quad (12)$$

$\Psi_{\hat{p}}$ is the highest priority patch, $Z_{\hat{p}}$ is the depth patch centered at \hat{p} ; and $\overline{Z_{\hat{p}}}$ is the average value of the depth patch. $Z_{\hat{p}}(\mathbf{q})$ is the depth value at pixel \mathbf{q} and γ is the depth variance threshold.

Once the highest priority patch $\Psi_{\hat{p}}$ and depth-based source region Φ_b defined in (11) are computed, we search for the best N number of patches within the source region.

$$\Psi_{\hat{q}} = \arg \min_{\Psi_{\mathbf{q}} \in \Phi_b} \{d(\Psi_{\hat{p}}, \Psi_{\mathbf{q}}) + \beta \cdot d(Z_{\hat{p}}, Z_{\mathbf{q}})\}, \quad (13)$$

where d is SSD, and β is a parameter to emphasize the depth. The depth map is considered in the patch matching process to find the similar patches in the depth domain and simultaneously fill the disocclusion in the depth map along with the texture.

The best N number of patches obtained from the patch matching step are not equally reliable [9]. Therefore, we adopt a weighted average of N patches when fill the missing information of the disocclusion.

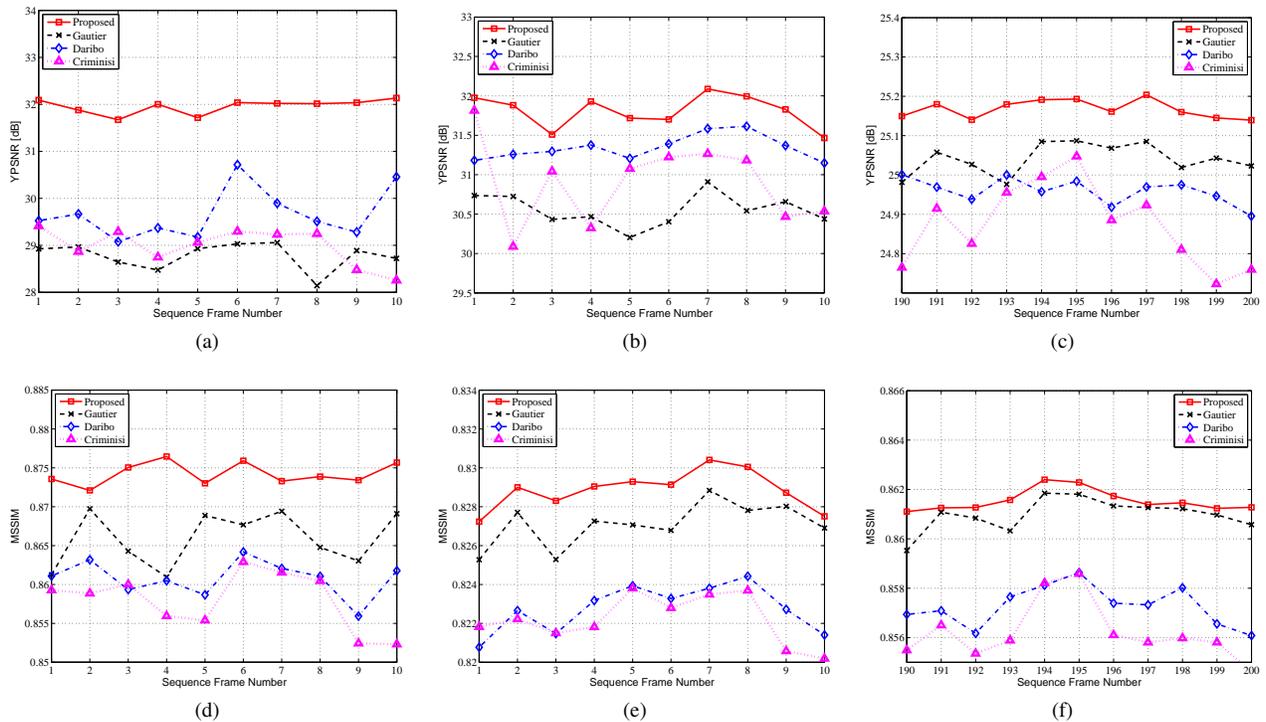


Figure 3. Objective metrics PSNR and MSSIM of the investigated sequences; PSNR for each rendered frame at view position 4 of “Ballet” (a), at view position 4 of “Break dancers” (b), at view position 4 of “Lovebird1” (c); MSSIM for each rendered frame at view position 4 of “Ballet” (d), at view position 4 of “Break dancers” (e) and at view position 4 of “Lovebird1” (f).

IV. TEST ARRANGEMENT AND EVALUATION CRITERIA

Results from the proposed method are evaluated by objective measurements as well as visual inspection. A set of 10 frames are selected from the three MVD sequences “Ballet”, “Break dancers” and “Lovebird1” for objective evaluation. All three sequences have a spatial resolution of 1024x768 pixels. The two first sequences are captured with 8 cameras and a baseline of 300 mm and 400 mm respectively [10]. The third sequence is captured with 12 cameras and a baseline of 35 mm [11]. The chosen MVD sequences have characteristics that make them suitable for testing different disocclusion filling attributes of inpainting methods. The “Ballet” sequence has large depth discontinuities at two different depth ranges, which results in big disocclusion areas at different depth levels. The “Break dancers” sequence has a large number of objects located in almost the same depth level. The “Lovebird1” sequence has complex texture and more structured background, with larger depth discontinuities.

All sequences are used in a DIBR of V+D scenario with full reference evaluation possible, i.e. access to ground truth texture and depth is available. More specifically, the first two sequences renders view 4 from view 5 and in the third “Lovebird1” sequence, view 4 is rendered from view 6. Post processing is applied on the rendered view and the depth

to remove the cracks and ghosting artifacts before starting the inpainting process. Important parameters of the proposed inpainting method is a patch matching window size of 120 pixels, $M = 0.4$, $\gamma = 80$ in (12), $\beta = 3$ in (13), and $N = 5$. For evaluation purposes, two objective evaluation metrics are considered: peak signal to noise ratio of the luminance component (Y-PSNR) and mean structural similarity index (MSSIM).

V. RESULTS AND ANALYSIS

The rendered and inpainted virtual views were generated and compared for disocclusion handling using methodology presented in the previous. Results from the objective evaluation are shown in Fig. 3. The PSNR and MSSIM graphs consistently demonstrate that the proposed depth-based inpainting method performs better than the Criminisi, Daribo and Gautier methods. Fig. 4 shows the rendered views with disoccluded areas (denoted with white color) and inpainting methods results of the “Ballet” and “Lovebird1” images for visual comparison. Note that the disocclusion regions in Fig. 4(c) and (d) are filled with foreground information since no depth is available to assist the filling process. Although the Daribo and Gautier methods are aided with true depth information, there still exists artifacts in the virtual views disocclusions. The proposed inpainting method shows visual improvements with respect to all the

reference methods, although it is operating in a more realistic setting where only warped depth information is available. The results from Fig. 4(i) and (j) show that the proposed method propagates the required neighboring information into the disocclusions region, retaining both smooth areas (at the left side of the “Ballet” image) and continuing neighborhood structure (on the curtain in the “Ballet” image and at the head of the women in the “Lovebird1” image). The proposed method still show some jaggedness effects at object boundaries, which is due to constraints on the source region selection and patch matching. In summary, the proposed method performs better than the reference methods both objectively and visually, which is a result of utilizing the depth-based direction priority, the depth included data term and search constraints incorporating depth information.

VI. CONCLUSION

We have proposed a new depth-based inpainting method to fill disocclusions in a virtual view by employing a depth guided directional term, a depth enhanced curvature driven diffusion model and depth searching constraints in the exemplar based texture synthesis. The results of the proposed method have been compared with the inpainting method of Criminisi, Daribo and Gautier using objective quality metrics and visual inspection. Both ways of evaluating consistently demonstrates that the proposed method offers an improved quality. In future work, we will focus on reducing the computational time that is inherent with processing large disocclusions, temporal consistency, and more elaborate subjective tests to further validate our results.

ACKNOWLEDGMENT

This work has been supported by grant 00156702 of the EU European Regional Development Fund, Mellersta Norrland, Sweden, and by grant 00155148 of Lnsstyrelsen Vsternorrland, Sweden. We would like to acknowledge J.Gautier et. al [5] for providing their software.

REFERENCES

- [1] C. Fehn, “Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV,” *Proc. SPIE Stereoscopic Displays and Virtual Reality Systems XI*, Jan. 2004, pp. 93–104.
- [2] Z. Tauber, Z. N. Li, and M. S. Drew, “Review and preview: Disocclusion by inpainting for image-based rendering,” *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, vol. 37, no. 4, 2007, pp. 527–540.
- [3] A. Criminisi, P. Pérez, and K. Toyama, “Region filling and object removal by exemplar-based image inpainting,” *IEEE Transactions on Image Processing*, vol. 13, 2004, pp. 1200–1212.
- [4] I. Daribo and B. Pesquet-Popescu, “Depth-aided image inpainting for novel view synthesis,” in *Multimedia Signal Processing*, 2010, pp. 167–170.
- [5] J. Gautier, O. L. Meur, and C. Guillemot, “Depth-based image completion for view synthesis,” in *3DTV conference*, 2011, pp. 1–4.
- [6] S. Li, R. Wang, J. Xie, and Y. Dong, “Exemplar image inpainting by means of curvature-driven method,” in *Computer Science and Electronics Engineering (ICCSEE)*, vol. 2, march 2012, pp. 326–329.
- [7] T. F. Chan and J. Shen, “Non-texture inpainting by curvature-driven diffusions (cdd),” *J. Visual Comm. Image Rep.*, vol. 12, 2001, pp. 436–449.
- [8] I. Ahn and C. Kim, “Depth-based disocclusion filling for virtual view synthesis,” in *ICME*, 2012, pp. 109–114.
- [9] Y. Wexler, E. Shechtman, and M. Irani, “Space-time completion of video,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, 2007, pp. 463–476.
- [10] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, “High-quality video view interpolation using a layered representation,” *ACM Trans. Graph.*, vol. 23, no. 3, Aug. 2004, pp. 600–608.
- [11] G. M. Um, G. Bang, N. Hur, J. Kim, and Y. S. Ho, “3d video test material of outdoor scene,” *ISO/IEC JTC1/SC29/WG11/M15371*, April 2008.



(a)



(b)



(c)



(d)



(e)



(f)



(g)



(h)



(i)



(j)

Figure 4. Illustration of the different inpainting method results for the investigated sequence frames “Ballet” first frame in the coulumn1 and “Lovebird1” 190th frame in column 2; (a)(b) rendered view images (disocclusions are represented with white regions); (c)(d) The results of Criminisi method; (e)(f) The results of Daribo method; (g)(h) The results of Gautiers method; (i)(j) The results of Proposed method.