MULTIMEDIA COMMUNICATIONS TECHNICAL COMMITTEE
IEEE COMMUNICATIONS SOCIETY
http://www.comsoc.org/~mmc

E-LETTER

Vol. 9, No. 3, May 2014

IEEE COMMUNICATIONS SOCIETY

CONTENTS

Message from MMTC Chair

Dear Fellow MMTC Members,

I am very happy to announce the winners of the MMTC Paper and Service Award of 2014.

MMTC best paper awards are open for nominations all year around. Any candidate paper needs to go through a rigorous three-stage process: nomination, review, and award selection. The nomination is open for any MMTC members, and self-nomination is allowed. The MMTC Review Board takes the lead of the review process, and selects award quality papers published in the bi-monthly MMTC Review Letter. The MMTC Award Board takes lead of the award selection process, by voting on the eligible papers published in the MMTC Review letter since the last award selection. Detailed procedure can be found at http://committees.comsoc.org/mmc/awards.asp.

This year, the Award Board voted on the Best Paper candidates from the MMTC Review Letter (since the April 2013 issue). The **2014 IEEE ComSoc MMTC Best Journal Paper Award** goes to the following paper:

> C.-C. Wu, K.-T. Chen, Y.-C. Chang, and C.-L. Lei, "Crowdsourcing multimedia QoE evaluation: A trusted framework," IEEE Transactions on Multimedia, vol. 15, no. 5, pp. 1121–1137, Aug. 2013.

The Award Board has decided not to give MMTC Best Conference Paper Award this year, mainly because of the very limited number of conference paper candidates in the pool this year.

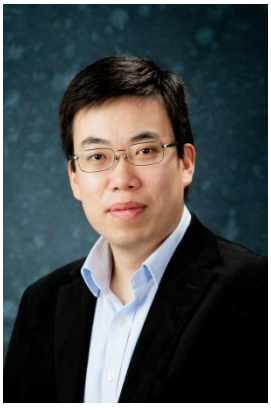
The details for the nomination and selection of the MMTC Service Awards can also be found at http://committees.comsoc.org/mmc/awards.asp. We are very happy to announce the awardees of **2014 IEEE ComSoc MMTC Outstanding Leadership Award:**

- Dr. Irene Cheng: MMTC R-letter Board Director
- Dr. Joel Rodrigues: MMTC Service and Publicity Board Director

There is no nomination for MMTC Distinguished Service Award this year.

This will be my last Chair's message, before the election of new MMTC officers in ICC 2014. I look forward to seeing many of you in Sydney, and the opportunities of continuing my contributions to MMTC in the many years ahead.

Regards,



Jianwei Huang
Chair, IEEE ComSoc Multimedia Communication Technical Committee
http://jianwei.ie.cuhk.edu.hk/

SPECIAL ISSUE ON 3D VIDEO REPRESENTATIONS & CODING: TRENDS AND CHALLENGES

Guest Editors: Petros Daras, Centre for Research and Technology Hellas/Information Technologies Institute, Greece

daras@iti.gr

During the last decade, computer graphics technology has been emerged rapidly resulting into productions of extremely high quality. Lately, due to the evolution of 3D display technologies, computer graphics are visualized in even more fascinating ways. In addition to graphics visualization, technology has shown a significant advance in natural scene capturing with High Definition (HD) cameras, HD stereo cameras and depth sensors. These technological breakthroughs introduced new types of visual media, like 3D Video (3DV) and Free Viewpoint Video (FVV) that expand user experience beyond traditional 2D video.

The capabilities, as well as the design, of a 3DV/FVV system are determined by the choice of the appropriate 3D representation of the visual content. In general, there are three representation options: i) image-based, ii) image plus depth-based and iii) geometry-based (i.e., 3D reconstructions). The choice of the 3D representation specifies the capturing requirements (e.g., number and position of cameras necessary) and on the other hand it determines the coding schemes, the transmission & rendering algorithms and capabilities such as the navigation range, robustness to occlusions, interactivity, degree of immersion, etc.

This special issue of E-Letter focuses on the recent progresses of 3D Video Representations and Coding focusing on rather new forms of 3D presentations like 3D meshes, 3D holoscopic content and Image Domain Warping in contrast to stereo/multi-view representations. It is the great honor of the editorial team to have four leading research groups, from both academia and industry laboratories, to report their solutions for meeting specific 3D Video representations and coding challenges and share their latest results.

The first article, *"A Basic Geometry Driven Mesh Coding Scheme with Surface Simplification for 3DTI"*, authored by Rufael Mekuria and Pablo Cesar, from Centrum Wiskunde Informatica, Netherlands, focuses on 3D mesh compression and transmission. The paper introduces a full real-time, geometry driven mesh codec focusing on geometry guided connectivity, which is suitable for 3D tele-immersion applications.

The second article, *"A Novel Planar Layered Representation for 3D Content and Its Applications"*, contributed by Aljoša Smolić *et al.* from Disney Research Zurich and ETH Zurich, Switzerland, presents an alternative to depth-based, 3D video processing depicted as image domain warping (IDW). IDW does not rely on dense depth or disparity maps but only requires disparity on distinct feature positions which can be estimated automatically with high accuracy and reliability. The paper also presents a novel approach for IDW coding and transmission, which was adopted by ITU/MPEG to be included in the upcoming version of the 3DHEVC standard.

Caroline Conti, Paulo Nunes and Luís Ducla Soares ISCTE – University Institute of Lisbon / Instituto de Telecomunicações, Portugal, present a multi-layer display scalable architecture for 3D holoscopic (integral) video coding, in the third article entitled *"Display Scalable 3D Holoscopic Video Coding"*. The paper also proposes a new prediction method to improve the coding efficiency when compared to independent compression of the three different display layers (simulcast case).

The last article is contributed by Marius Preda, Institut MINES TELECOM, Telecom SudParis, ARTEMIS, France, with the title *"3D content processing challenges for Mixed Reality"*. In this paper several issues, possible solutions and future trends in addressing 3D aspects in Augmented Reality (AR) and Augmented Virtuality (AV) are presented with respect to potential 3D representations. Moreover, for these representations the paper presents the MPEG endeavors concerning static mesh coding approaches (addressed by the MPEG-4 Animation Framewok eXtention framework) and mesh deformation coding. The latter can be addressed in two ways: one can either code the new vertex positions themselves as a function of time, or use some deformation controller to influence the mesh geometry, and code the controller parameters as a function of time.

While this special issue is far from delivering a complete coverage on this exciting research area, we hope that the four invited letters give the audiences a

taste of the main activities in this area, and provide them an opportunity to explore and collaborate in the related fields. Finally, we would like to thank all the authors for their great contribution and the E-Letter Board for making this special issue possible.

Petros Daras is an electrical and computer engineer (Diploma '99, MSc '02, PhD '05) graduated from the Aristotle University of Thessaloniki, Greece. He is a Researcher Grade B (Assoc. Prof.), in the Information Technologies Institute of the Centre for Research and Technology Hellas and head of the Visual Computing Lab (http://vcl.iti.gr). His main research interests include processing, retrieval, and recognition of 3D objects, 3D object reconstruction and coding, 3D image analysis, 3D medical image processing, and bioinformatics. He acts as a reviewer for many major IEEE journals and has served as a TPC member in more than 50 international conferences. He has published more than 100 papers in international conferences and journals. Dr. Daras is a Senior Member of IEEE, Chair (2012-2014) of the IEEE Interest Group on Image, Video and Mesh coding and key member (2010-2014) of the IEEE Interest Group on 3D rendering, Processing and Communications.

A Basic Geometry Driven Mesh Coding Scheme with Surface Simplification for 3DTI

Rufael Mekuria, Pablo Cesar
Centrum Wiskunde Informatica, Netherlands
Rufael.mekuria@cwi.nl p.s.cesar@cwi.nl

1. Introduction

With the rise of consumer grade depth camera's and available computational power, the computer vision and image processing communities are developing systems that can reconstruct full 3D Mesh geometry consisting of vertex and connectivity data representing a human or animal user on-the-fly. To enable interactive communication (i.e. 3D Tele-immersion: 3DTI) with this representation, efficient transmission and compression schemes are needed. While much research on the transport/compression of 3D representations focuses on multi-view video (MVV) based on multiple image views, the 3D Mesh representation has benefits for 3D immersive communications. Firstly, 3D Mesh **rendering** is directly supported in modern OpenGL compliant graphics cards. Secondly, **composite rendering (composition)** with synthetic 3D content in games and virtual worlds is much easier compared to image based MVV representations. Also, advanced **3D audio and lighting rendering** schemes can take advantage of the explicit surface position of the mesh. Thirdly, the mesh representation is common in **video games** and **virtual worlds** where controlled agents can use the geometric information to control their **AI**. Lastly, **compression efficiency** could be an advantage, as especially time-consistent 3D dynamic meshes can be compressed very efficiently, possibly reducing the bandwidth requirements over image based MVV.

Nevertheless, existing mesh transmission and compression techniques have not been designed for real-time streaming and transmission of live reconstructed content. Instead, they focus on traditional synthetic graphics and animation content, which has different requirements as all content is authored offline. Based on the large industrial potential of live 3D Tele-Immersive mesh coding, the MPEG AhG on 3D Graphics is currently exploring this new direction, gathering industry requirements and benchmarking existing technologies. In this paper we contribute by introducing a full real-time geometry driven mesh codec focusing on geometry guided connectivity coding. In Section 2 we first present related work. In section 3 we introduce the specific challenges and requirements for 3DTI. We discuss the implementation based on octree composition, surface simplification and connectivity coding in section 4. In Section 5 we present an evaluation and rate-distortion comparison of the codec to the MPEG-4 standard.

2. Related Work

Dynamic time-consistent geometry, i.e. sequences of meshes where only the geometric coordinates change (like animations), can generally be compressed well as the fixed connectivity over time allows direct coding of frame by frame vertex differences (for example the MPEG-4 FAMC [1]). In the case of live 3D capturing systems that reconstruct 3D meshes from multiple depth images, connectivity can also change over time, introducing the need for time-varying mesh compression (TVM). A standardization contribution to MPEG-3DG in [2] proposes a system for time-varying mesh compression based on the inter-frame prediction between MPEG-4 [3] static mesh coded frames. The method utilizes separately acquired skeleton motion and skinning for the prediction. Their results show some improvements compared to using the MPEG-4 codec at lower QP's. A limitation of the approach in [2], is that temporal redundancy of the geometry information over time cannot be exploited directly, as the geometry coding is guided by the connectivity. Also, based on the single rate MPEG-4 codec it is harder to implement spatial scalability in the bit-stream. To overcome these limitations, we propose a purely geometry driven approach where connectivity coding is guided by the geometric data. The geometry coding is based on an octree representation comparable to [4] which has been considered favorable for the densely sampled surfaces that TVM's represent. We further exploit the octree structure to enable basic scalability and mesh surface simplification. Also, we envision that the octree structure can in a next stage be used to exploit temporal redundancies better. In the current implementation we introduce the geometry-guided connectivity coding scheme that is simpler compared to the method from [4], making it more suitable for 3DTI. Together with an existing point cloud codec it is integrated into a complete geometry driven 3D Mesh Codec.

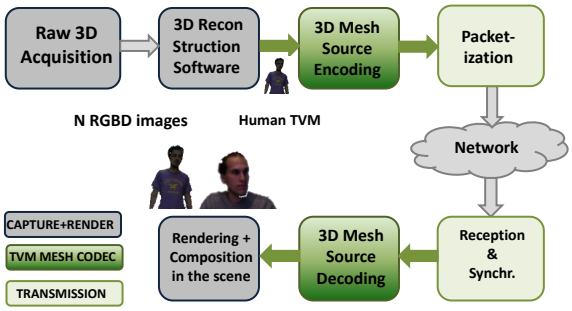


Figure 1 Mesh Based 3DTI Pipeline Example

3. 3DTI Based on Reconstructed Geometry Requirements

In Figure 1 we show a typical 3D tele-immersive pipeline based on live reconstructed geometry. The key difference compared to image based MVV is that instead of sending the raw captured 3D depth/image data the 3D reconstruction software reconstructs the full 3D surface represented as a 3D mesh $M(V,C)$ with vertex data V and connectivity data C. So instead, $M(V,C)$ is subsequently compressed and transmitted. Specifically challenging in this case compared to traditional 3D Geometry compression/streaming systems for 3D Graphics is that low encoding/decoding complexity is needed and that imperfect input data should be handled gracefully. To establish a common understanding of the requirements in this use case that is not just about compression efficiency, the MPEG AhG on 3DG has developed a document describing the use case and coding system requirements for 3DTI [5] .

4. Geometry Driven Mesh Codec

Figure 2 illustrates the geometry driven mesh coding scheme. First we encode the geometry (vertex information, possibly with colors) with an existing octree based point cloud codec that encodes only vertex/color data and is available in PCL [6] and described in [7]. This codec partitions all vertex data in voxels in a 3D bounding box that can be indexed with discrete positional indices $v_{octree}\{x,y,z\}$, we call this voxel space V_{octree}. The octree composition is also used to simplify the dense input mesh by down sampling on a per voxel basis. We only recover the voxel leaf center at the receiver instead of all enclosed vertices. This simplification allows a large data reduction with fine-grained quality control by tuning the voxel size depending on quality/bandwidth requirements.

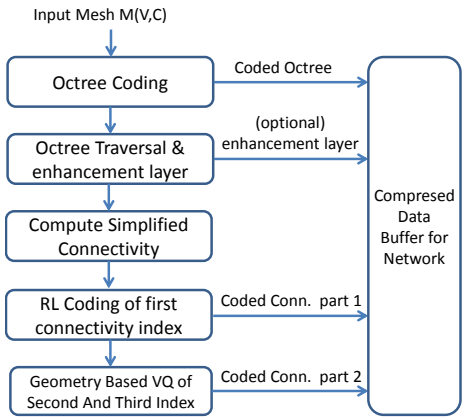


Figure 2 Basic Geometry Driven Mesh Codec

In the second stage, we traverse the octree, storing the indices I_{simp} that index a voxel leaf with a single integer. I_{simp} contains all vertex indices for the simplified connectivity. During this traversal we optionally compute an enhancement layer that refines the position of the voxel center based on the original vertex points based on a weighted average. This layer allows an increase in received mesh quality when transmitted and received properly (e.g. when enough bandwidth is available). In this traversal step we also obtain the mapping m from the original vertex indices I_{or} to the simplified leaf voxel indices I_{simp} and the mapping l between I_{simp} and the 3D voxel indices $v_{octree}\{x,y,z\}$ in the grid V_{octree} .

$$m: I_{or} \rightarrow I_{simplified}\{i_{or} \in I_{or}, m(i_{or}) \in I_{simplified}\}$$

$$l: I_{simplified} \rightarrow V_{octree}\{i_{simp} \in I_{simplified}, l(i_{simp}) \in V_{octree}\}$$

Based on the original connectivity $C_{original}$ provided by the reconstruction software where the triangles $T_{\text{unordered}}$ are provided without ordering we compute $C_{simplified}$. $C_{simplified}$ is composed by triangles $T_{ordered}$ where the vertex indices are ordered.

$$T_{unordered} = \{i_1, i_2, i_3\} : \{i_1, i_2, i_3 \in I_{original}, i_1 \neq i_2 \neq i_3\}$$

$$T_{ordered} = \{i_1, i_2, i_3\} : \{i_1, i_2, i_3 \in I_{simplified}, i_1 < i_2 < i_3\}$$

Ordering of the vertex indices i_{simp} in $C_{simplified}$ helps to preserve more of the spatial correlation that was introduced by the structured octree traversal. $C_{simplified}$ is computed by traversing the original connectivity $C_{original}$, and when distinct $m(i_{or1})$, $m(i_{or2})$, $m(i_{or3})$ are found in an original triangle $t_{\text{unordered}}$ they are stored as an ordered triangle in the new $C_{simplified}$ (note that in $T_{ordered}$ $i_1 \neq i_2 \neq i_3$ as $i_1 < i_2 < i_3$). So, $C_{simplified}$ can be described as a set of $T_{ordered}$ conditioned as:

$$C_{simplified} = \{t_{ordered}(i_1, i_2, i_3) \mid \exists t_{unord.}(i_{or1}, i_{or2}, i_{or3}) :$$

$$t_{unord} \in C_{original}, m(i_{or1}) = i_1, m(i_{or2}) = i_2, m(i_{or3}) = i_3\}$$

Next, the set $C_{simplified}$ is ordered in ascending order of i_1 and i_1 is coded via a run length coding scheme that codes the number of repetitions and increments in i_1. This is part 1 of the coded connectivity in Figure 2. Next, we construct geometry based representations T_{VQ} of each triangle $t_{ordered}$ in $C_{simplified}$ to code the second and third indices i_2, i_3. Every t_{VQ} represents $t_{ordered}$ where i_2 , i_3 are represented by a 3D shift in the octree voxel grid V_{octree} away from connected voxels $v_1 = l(i_1)$ and $v_2 = l(i_2)$. Hence, T_{VQ} is represented as:

$$T_{VQ} = \{i_1, z_2, z_3\} : \{i_1 \in I_{simplified}, z_2 \in Z^3, z_3 \in Z^3\}$$

Where z_2 and z_3 are signed discrete 3D vectors representing a shift in the octree voxel grid V_{octree}. The T_{vq} representations are obtained from T_{ordered} by F_{vq}:

$$F_{vq} : T_{ordered} \rightarrow T_{VQ} = \{i_1, l(i_2) - l(i_1), l(i_3) - l(i_2)\}$$

In each t_{vq} in T_{vq}, i_1 was already coded in part 1 and only z_2 and z_3 need to be encoded. By structuring the connectivity information in $T_{ordered}$ and F_{vq} we achieved a structure where z_2 and z_3 can be coded very efficiently via vector quantization. The prototype vectors *[-1,0,0]*, *[0,-1,0]* and *[0,0,-1]* occur over 75% of the cases z_2 and z_3 while around 15 % of the other cases are covered by the vectors *[-1,1,0]*, *[0,-1,1]*, *[1,-1,0]*. The remaining

signed binary vectors represent the last 7% of the cases. We developed an efficient VLC scheme to code these vectors with 2, 4 and 8 bit code words. In the exceptional other cases we store all components of z_2 and/or z_3. The decoder executes inverse operations, i.e the octree voxel structure is decoded with [6][7] and instead of only l we also compute the inverse l^{-1} that relates the 3D octree voxel index v_{octree} to $I_{simplified}$.i.e:

$$l^{-1}: V_{octree} \rightarrow I_{simplified}\{v_{octree} \in V_{octree}, l^{-1}(v_{octree}) \in I_{simplified}\}$$

This mapping is used to decode all $t_{ordered}$ recovering $C_{simplified}$ from all t_{vq} based on the relation F^{-1}_{vq}:

$$F^{-1}_{vq}: T_{VQ} \rightarrow T_{ordered} = \{i_1, l^{-1}(l(i_1)+z_2), l^{-1}(l(i_2)+z_3)\}$$

Where i_1 was already recovered from part 1 of the compressed connectivity data. Therefore, by recovering i_2 followed by i_3 for each t_{vq} $C_{simplified}$ is recovered.

5. Codec Evaluation and Comparison

The proposed scheme enables coding 3D connectivity and an enhancement layer alongside with the simplified vertex information implementing a full 3D Mesh codec. We compared the codec without enhancement to the mesh codec available in MPEG-4 presented in [3] configured for 8 bit coordinate and 4 bits color quantization (we also use 4 bits color coding in the proposed scheme) and differential prediction. We compared the rate-distortion trade-off in our scheme with the MPEG-4 codec applied on the completely enhanced simplified mesh. We use the datasets from [8] which contain dense photorealistic meshes reconstructed with Kinect based reconstruction software which are currently also used for evaluation in the MPEG-3DG group. We use the metro tool developed in [9] to compute the symmetric mean squared error metric for distortion evaluation. We tune the voxel size of the octree for surface simplification to control the rate. The results show that our method achieves only slightly higher rate-distortion compared to the MPEG-4 codec applied to the simplified mesh i.e., at very low bit-rates the mpeg-codec outperforms the scheme due to the fact that the voxel down-sampling becomes very coarse compared to the enhanced simplified mesh). Nevertheless, we believe that by future enhancements based on inter-frame prediction between the voxel data and better color coding we can achieve much better results very soon.

References

[1] K. Mamou, T. Zaharia, F. Preteux, "FAMC: The MPEG-4 standard for Animated Mesh Compression," in Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on, San Diego, 2008, pp. 2676 - 2679.

[2] A. Doumanoglou, D. Alexiadis, D. Zarpalas, P. Daras, "Towards Real-Time and Efficient Compression of Human Time-Varying-Meshes," input document m30537, ISO iec sc29wg11 (MPEG), Vienna July 2013.

[3] K. Mamou T. Zaharia, and F. Preteux, "TFAN a low complexity 3D Mesh Compression algorithm," in Computer Animations and Virtual Worlds (CASA'09) pp. 343-354, 2009.

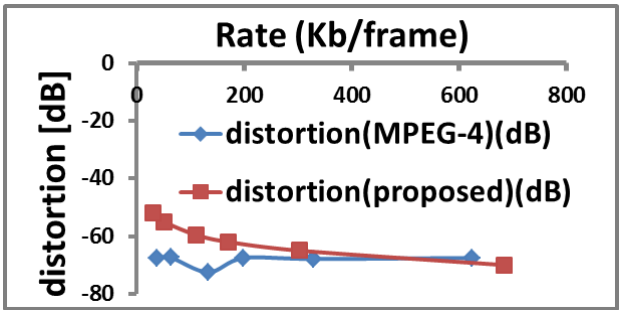


Figure 3 rate-distortion of proposed scheme

[4] J. Peng and C.-C. Jay Kuo, "Geometry-guided progressive lossless 3D mesh coding with octree (OT) decomposition," ACM Trans. Graph, vol. 24, no. 3, pp. 609-616, July 2005.

[5] R. Mekuria, P. Cesar. P, L. Bivolarsky, M. Preda, ISO MPEG N14197 Tele-immersion use case and associated draft requirements," San Jose, California, 2014.

[6] R.B. Rusu, "3D is here: Point Cloud Library (PCL)," in IEEE International Conference on Robotics and Automation, Shanghai, 2011, pp. 1-4.

[7] J. Kammerl, N. Blodow, Rusu, R.B., Gedikli, S., Beetz, M., Steinbach E., "Real-time compression of point cloud streams," IEEE International Conference in Robotics and Automation (ICRA), 2012, Saint Paul, MN, 2012, pp. 778 - 785.

[8] D. Alexiadis, D. Zarpalas, P. Daras, "Real-time, Realistic Full-body 3D Reconstruction and Texture Mapping from Multiple Kinects", 11th IEEE IVMSP Workshop: 3D Image/Video Technologies and Applications, Yonsei University, Seoul, Korea, 10-12 June 2013

[9] P. Cignoni, C. Rocchini and R. Scopigno (1998) Metro: measuring error on simplified surfaces *Computer Graphics Forum*, Blackwell Publishers, vol. 17(2), June 1998, pp 167-174

Rufael Mekuria received a Bachelor's and Master's Degree in Electrical Engineering from Delft University of Technology. He is currently with Centrum Wiskunde & Informatica in Amsterdam, the Netherlands. Since April 2013 he has been contributing to MPEG 3D Graphics group (3DG), where he is currently a co-chair of the AhG on 3DG .

Pablo Cesar leads the Distributed and Interactive Systems group at Centrum Wiskunde & Informatica: CWI. He received his PhD from the Helsinki University of Technology in 2006.

Image Domain Warping for Advanced 3D Video Applications

Aljoša Smolić[1], *Oliver Wang*[1], *Manuel Lang*[12], *Nikolce Stefanoski*[1], *Miquel Farre*[1], *Pierre Greisen*[12], *Simon Heinzle*[1], *Michael Schaffner*[12], *Alexandre Chapiro*[12], *Alexander Sorkine-Hornung*[1], *Markus Gross*[12]
[1]*Disney Research Zurich, Switzerland* [2]*ETH Zurich, Switzerland*
smolic@disneyresearch.com

1. Introduction

Stereo 3D video has gained significant attention over the past decade in both academia and industry. However, not all expectations about commercial success from 5 years ago turned into reality. S3D is well established in cinema and special applications like theme parks or the professional sector; on the other hand 3DTV and mobile usage did not take off as predicted by many. The reasons for success on one side are that technology and content creation are mature enough to create a convincing quality of experience for the user. This is not yet the case for the other application areas, which means that there is still a lot of room for research.

In this context, view synthesis has been identified as a core functionality for advanced 3D video processing [10]. This includes for instance disparity mapping [11] (display adaptation, comfort enhancement, depth corrections, artistic manipulation), 2D to 3D conversion [12], stereo to multiview conversion [13], or free viewpoint navigation [14]. Such view synthesis requires some kind of 3D reconstruction of the captured scenery in order to render novel viewpoints or to manipulate depth perception. For that purpose, depth or disparity based representations have gained a lot of interest over the last years [15], [16]. MPEG is currently working on a related standard to represent and transmit such depth based 3D video.

However, a fundamental drawback of depth based 3D video is the inherent difficulty to create the depth data with sufficient quality and robustness for high quality view synthesis. Automatic depth or disparity estimation at quality levels necessary for professional broadcast or cinema production is still an unresolved problem. Associated problems like dis-occlusions additionally require hole filling in depth image based rendering (DIBR). Therefore so far only interactive or semi-automatic applications of depth based 3D video processing are in professional use.

In this paper, we present an alternative to depth based 3D video processing depicted as image domain warping (IDW). IDW does not rely on dense depth or disparity maps but only requires disparity on distinct feature positions which can be estimated automatically with high accuracy and reliability. In the remainder of this paper we introduce basics of IDW (Section 2) and show how it is used for a variety of advanced 3D video applications (Section 3).

2. Image domain warping

Feature matching

The main design paradigm of IDW was not to rely on dense depth or disparity estimation as sufficient quality cannot be guaranteed except for interactive applications. However, *some* notion of the underlying 3D scene structure *is* necessary for 3D video processing. For that purpose, IDW relies on sparse disparity estimates that, given an initial stereo pair or multiview video, can be estimated with high accuracy and reliability. Matching of distinctive features over related images has been studied extensively, and powerful solutions are widely used in various real world applications. In principle IDW works with any feature matching algorithm. In practice we often used SBK [17], which is specifically tuned for stereo video processing.

Given a stereo pair of images or video, we estimate a sparse disparity field by feature matching, which is robust and accurate. Typically a few hundred features are enough. We constrain selection of features to get a good coverage over the whole image, while having dense populations at most distinctive regions. This sparse 3D reconstruction is part of the automatic content analysis as illustrated in Figure 1.

Non-linear locally adaptive disparity mapping

Any of the manipulations of 3D video considered here can be regarded as a mapping operation on the input disparity, which we denote as $\varphi = f(d)$, where d is the input disparity and f is a generally non-linear function. Such non-linear disparity mapping may also be locally adaptive, i.e. we have position dependent function defined as $\varphi = f(d, x, y)$. φ defines a locally adaptive non-linear goal function of what we want to achieve with our 3D video manipulation. For instance dividing all input disparities globally (i.e. no dependence on x, y) by 2, corresponds to reducing the interaxial distance by the same factor, i.e. putting the cameras closer together. Formally φ allows definition of any non-linear and locally adaptive mapping function on the input disparity to achieve a desired effect on the 3D video as we will demonstrate below.

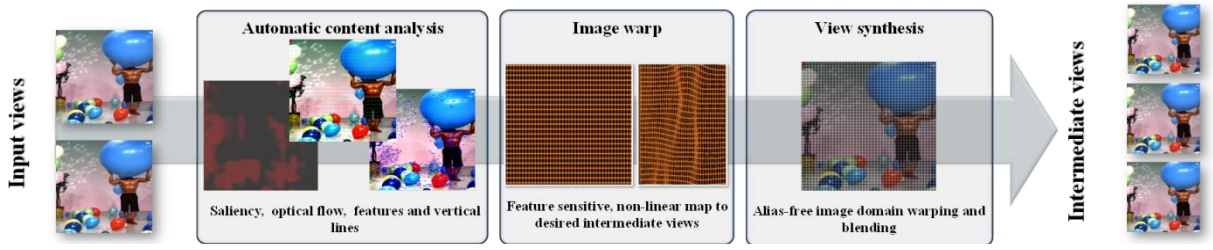


Figure 1. IDW pipeline for stereo to multiview conversion.

Saliency and edges

As we will see below, IDW means warping of images controlled by mapping functions φ. Any such content adaptive warping may result in distortions and artifacts from stretching and squeezing images, especially in horizontal direction. In order to make such unavoidable distortions least noticeable we hide them in less important image regions as far as possible. This is achieved by estimating saliency of image regions (e.g. [18]). Our constrained energy minimization (see below) will then make sure to distribute warping errors in less salient regions. We pay specific attention to vertical edges, as those may result in most visible artifacts.

Energy minimization

IDW describes non-linear deformation of input images to output images as illustrated in Figure 1. This can be formulated as quad-mesh deformation, where each vertex of the mesh corresponds to a pixel of the input image. In practice also coarser meshes (e.g. 10x10 pixels) can be used, e.g. for complexity reasons. The desired deformation function of the mesh can be formulated as solution of a constrained energy minimization problem, with the goal function φ as data term, and saliency, edges and other constraints (e.g. temporal stability) as smoothness terms. The result is a warping function as best compromise between the input constraints that will map each pixel of the input images to some position in the output images.

View synthesis

Given a set of input images and a corresponding set of warping functions, the actual view synthesis is the last step in our pipeline. This is illustrated in Figure 1 for the use case of stereo to multiview conversion (see below). Such non-linear image warping can be efficiently realized in high quality e.g. using EWA rendering, for which we developed also a dedicated hardware implementation [19].

Warp coding and transmission

In a communication scenario, warping functions may also be calculated at an encoder, then transmitted in compressed form to a decoder and display. For that purpose ITU/MPEG adopted our proposal for a corresponding warp coding and IDW framework for the upcoming 3DHEVC standard [13]. Compared to depth coding, such warp coding reaches high efficiency, while providing the same functionalities. We have shown that the overhead for warp coding on top of the color signals can be below 4% on average [13].

3. Application scenarios

Principles and pipeline for IDW introduced in the last section can be applied to a variety of advanced 3D video processing applications as described in the following. The goal function φ is in every case adjusted appropriately and also details of content analysis, energy minimization and view synthesis may differ.

Stereo 3D manipulation

This use case describes the modification of a given stereo pair or video after capture/creation, which can be real-time or offline. This is illustrated in Figure 2. Left is a stereo image as captured by the camera. On the right we see the result of a disparity mapping operation. The head of the cow is pushed backwards in z-space, such that it does not come out of the screen anymore as in the original. The background of the scene remains unchanged. As such the manipulation is locally adaptive (only cow affected) and non-linear in z-space (foreground mapped, background unchanged), which is adjusted via an appropriate goal function φ. A linear and global function φ could be used for instance to simulate changing the camera baseline as mentioned before.

We have shown that such stereo 3D manipulations work reliably and accurately also in automatic processing within a certain disparity mapping range [11]. This enables a variety of interesting application scenarios like 3D display adaptation (e.g. depth remote control), automatic and interactive stereo 3D correction, or changing depth impression in post-production.

Stereo to multiview conversion

IDW can be used to synthesize intermediate views in between available stereo views. This can be achieved

by defining appropriate goal functions φ that correspond to certain intermediate views [13]. The pipeline is illustrated in Figure 1. As IDW is content adaptive image stretching and squeezing, it does not create dis-occlusions when moving the virtual viewpoints. Therefore it can also be used in contrast to DIBR to extrapolate views beyond the initial stereo pair. This is necessary to generate compelling content with decent depth impression for autostereoscopic screens, where the overall disparity range over the multiple views is typically bigger compared to the range of the input stereo views [20].

Figure 2. Non-linear locally adaptive disparity mapping. Left: input stereo, right: manipulated stereo.

2D to 3D conversion

Conversion of 2D images and video to 3D is an important application scenario not only for legacy content but also for new productions. For professional quality generally interactive systems are used today. Typically this requires sophisticated, expensive and highly manual operation including segmentation, depth/disparity reconstruction and inpainting. We developed and interactive approach based on IDW that integrates all these steps to an instant feedback system, as illustrated in Figure 3. Sparse features are replaced by sparse disparity scribbles as user input. The artist paints those scribbles over an input 2D image and the system calculates a warp to create stereo on the fly. Through instant feedback the artist continues until he is satisfied with the result. We have shown that compelling conversion results can be achieved within minutes in a very intuitive way.

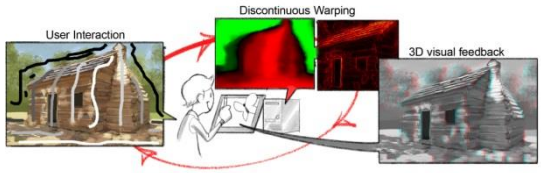


Figure 3. Concept of StereoBrush interactive 2D to 3D conversion.

4. Conclusion

We have presented IWD as a powerful core technology for a variety of advanced 3D video applications. It avoids drawbacks of alternative approaches based on depth estimation and DIBR and can achieve better automatic performance. Implementations of IDW are readily in use in professional production and may become part of consumer products in the future. The adoption of our warp coding framework for 3DHEVC may pave the way for that. Limitations remain in the possible distance of virtual views or the possible amount of disparity mapping. For such applications such as free viewpoint video [14] still a more complete 3D reconstruction in form of dense depth maps or 3D models will be necessary.

References

[10] A. Smolic, P. Kauff, S. Knorr, A. Hornung, M. Kunter, M. Mueller, and M. Lang, "Three-Dimensional Video Postproduction and Processing," Proceedings of the IEEE, vol. 99, no. 4, pp. 607-625, 2011.

[11] M. Lang. A. Hornung, O. Wang. S. Poulakos, A. Smolic, and M. Gross, "Non-linear Disparity Mapping for Stereoscopic 3D", ACM Transactions on Graphics (SIGGRAPH 2010), July 2010.

[12] O. Wang, M. Lang, M. Frei, A. Hornung, A. Smolic, M. Gross, "Stereobrush: Interactive 2d to 3d conversion using discontinuous warps", International Symposium on Sketch-Based Interfaces and Modeling (SBIM 2011), 2011

[13] N. Stefanoski, O. Wang, M. Lang, P. Greisen, S. Heinzle, and A. Smolic, "Automatic View Synthesis by Image-Domain-Warping", IEEE Trans. on Image Processing, 22 (9), Sept. 2013.

[14] A. Smolic, "3d video and free viewpoint video - From capture to display," Pattern Recognition, vol. 44, no. 9, pp. 1958-1969, 2011.

[15] K. Mueller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "View Synthesis for Advanced 3D Video Systems", EURASIP Journal on Image and Video Processing, Volume 2008, doi:10.1155/2008/438148.

[16] K. Mueller, P. Merkle, and T. Wiegand, "3-D Video Representation Using Depth Maps," Proceedings of the IEEE, vol. 99, no. 4, pp. 643 - 656, 2011.

[17] F. Zilly, C. Riechert, P. Eisert, and P. Kauff, "Semantic Kernels Binarized – A Feature Descriptor for Fast and Robust Matching," Visual Media Production (CVMP), 2011 Conference for, 2011.

[18] C. Guo, Q. Ma, and L. Zhang, "Spatiotemporal saliency detection using phase spectrum of quaternion Fourier transform," Computer Vision and Pattern Recognition (CVPR), IEEE Conference on, 2008.

[19] P. Greisen, M. Schaffner, S. Heinzle, M. Runo, A. Smolic, A. Burg, H. Kaeslin, M. Gross, "Analysis and VLSI Implementation of EWA Rendering for Real-time HD Video Applications", IEEE Trans. on TCSVT, Vol. 22, No. 11, pp. 1577-1589, November 2012.

[20] A. Chapiro, S. Heinzle, T.O. Aydın, S. Poulakos, M. Zwicker, A. Smolic, and M. Gross, "Optimizing Stereo-to-Multiview Conversion for Autostereoscopic Displays, " Eurographics 2014, Strasbourg, France, April 2014.

Display Scalable 3D Holoscopic Video Coding

Caroline Conti, Paulo Nunes and Lu ís Ducla Soares
ISCTE – University Institute of Lisbon / Instituto de Telecomunica çõ es, Portugal
{*caroline.conti, paulo.nunes, lds*}*@lx.it.pt*

1. Introduction

Motivated by the notable success of three-dimensional (3D) movies in the last decade, 3D display technologies have been extensively researched and there has been a significant effort for promoting the adoption of three-dimensional television (3DTV). In this context, 3D holoscopic imaging, also known as integral imaging, is an autostereocopic 3D capture and display approach which has become recently an appealing solution for providing a more natural 3D viewing experience than the current stereoscopic or multiview solutions [*1*].

However, in order to gradually introduce this technology into the consumer market and to efficiently deliver 3D holoscopic content to end-users, backward compatibility with legacy displays is essential. Consequently, to enable 3D holoscopic content to be delivered and presented on legacy displays, a display scalable 3D holoscopic coding approach is required. This would mean that a legacy two dimensional (2D) device (or a legacy 3D stereo device) that does not explicitly support 3D holoscopic content should be able to play a 2D (or 3D stereo) version of the 3D holoscopic content, while a more advanced device should play the 3D holoscopic content in its entirety. Therefore, this letter presents a multi-layer display scalable architecture for 3D holoscopic video coding, where each layer represents a different level of display scalability. Furthermore, a proposed prediction method to improve the coding efficiency when compared to independent compression of the three different display layers (simulcast case) is also presented.

Experiments were conducted by incorporating this novel prediction method into a display scalable version of the ISO/IEC & ITU-T High Efficiency Video Coding (HEVC) standard [*2*]. From the obtained results, it is shown that this display scalable coding scheme can significantly outperform the simulcast solution based on the original HEVC prediction schemes.

2. Scalable 3D Holoscopic Coding Architecture

To support display scalability, a multi-layer architecture for 3D holoscopic coding was designed, whose layers represent different levels of display scalability:

- *Base Layer*: The base layer represents a single 2D view, which can be used to deliver a 2D version of the 3D holoscopic content to 2D displays;
- *First Enhancement Layer*: This layer represents the necessary information to obtain an additional view (representing a stereo pair) or various views (representing multiview content). It intends to allow stereoscopic displays or multiview displays to support 3D holoscopic content;
- *Second Enhancement Layer*: This layer represents the additional data needed to support full 3D holoscopic video content.

Hence, to be able to display 3D holoscopic content on 2D and 3D multiview displays, it is necessary to produce adequate versions of the content for the *Base Layer* and the *First Enhancement Layer*. This means to generate various 2D views from 3D holoscopic content. In [*3*], algorithms to generate 2D images from a 3D holoscopic image are proposed in the context of richer 2D image capturing systems. These algorithms take into account the principles of 3D holoscopic imaging and the structure of the 3D holoscopic image itself.

Basically, the acquisition system for 3D holoscopic imaging comprises an array of small spherical micro-lenses in the optical structure, known as a "fly's eye" lens array. Each micro-lens can be seen as an individual small low resolution camera, recording a different perspective of the scene at slightly different angles. As a result, the 3D holoscopic image consists of a 2D array of micro-images, comprising planar light intensity and direction information. An example of a 3D holoscopic image is shown in **Error! Reference ource not found.**.

(a) (b)

Figure 1 Holoscopic image captured with a 250 µm pitch micro-lens array: (a) full image with resolution of 1920×1088; (b) enlargement of 196×140 pixels showing the array of micro-images.

In this letter, two of the algorithms proposed in [*3*], referred to here as *Basic Rendering* and *Weighted Blending*, are considered to generate the data for each layer in the proposed scalable coding architecture. Briefly, the both *Basic Rendering* and *Weighted Blending* algorithms are based on the fact that since each micro-image can be seen as a low resolution view of the scene, it is possible to choose suitable portions

(patches) from each micro-image and to properly stitch them to compose a 2D image view. It is also shown in [*3*] that by choosing proper patches sizes, it is possible to control the plane of focus in the generated 2D image view (i.e., which objects will appear in sharp focus). Moreover, by varying the relative position of the **extracted** patch in the corresponding micro-image, it is possible to generate various 2D views with different viewing angles (i.e., different scene perspectives).

3. Combined Prediction Scheme for Scalable 3D Holoscopic Content Coding

Based on the presented scalable coding architecture, a combined prediction method is proposed to take full advantage of the redundancies existing in the presented scalable coding architecture for efficiently handling the 3D holoscopic content. This is accomplished by combining an inter-layer prediction method with a self-similarity compensated prediction.

Therefore, this proposed combined prediction was integrated into a display scalable 3D holoscopic video codec based on the emerging HEVC standard [*2*]. For this, an inter-layer (IL) and a self-similarity (SS) prediction picture are introduced as new reference frames to be used by the existing inter-frame prediction modes of the HEVC. Consequently, no changes in the lower levels of the syntax and decoding process of HEVC are needed.

Inter-Layer Prediction Method

The inter-layer (IL) prediction method (previously proposed in [*4*]) aims to exploit the existing redundancy between the multiview and the 3D holoscopic content. To accomplish this, a prediction picture, referred to as IL reference, is built by using: the set of reconstructed 2D views from the previous coding layers; information from the 3D holoscopic capturing process (such as the resolution of micro-images and of the micro-lens array); and also information from the 2D view generation process.

Thus, it is possible to distinguish two steps when generating an IL reference:

- *Patch Remapping*: This first step is an inverse process of the 2D view generation algorithms (*Basic Rendering* and *Weighted Blending*). In other words, it corresponds to an inverse mapping (referred to as remapping) of the patches from all coded and reconstructed 2D views to their original positions inside the 3D holoscopic image. A template of the 3D holoscopic image assembles all patches and the output of this process is a sparse 3D holoscopic image, as shown in Figure 2.

- *Micro-Image Refilling*: Due to the small angular disparity between adjacent micro-lenses in the 3D holoscopic content acquisition process, a significant cross-correlation exists between neighboring micro-images. This inherent cross-correlation is emulated in this process to fill the holes in the sparse 3D holoscopic image (built in the *Patch Remapping* step) as much as possible. An illustrative example of this process can be seen in Figure 3 for only three neighboring micro-images. The output of this process is the IL reference as shown in Figure 3.

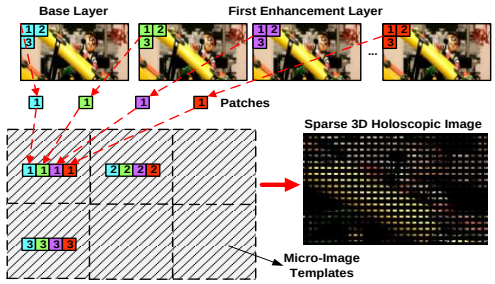


Figure 2 The *Patch Remapping* step of IL prediction

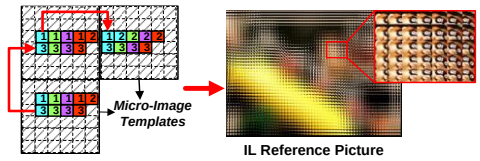


Figure 3 The *Micro-Image Refilling* step of IL prediction

Self-Similarity Compensated Prediction Method

A method for a self-similarity (SS) compensated prediction was proposed by the authors in [*5*] to efficiently handle 3D holoscopic content coding using the HEVC. This method is also integrated into the proposed combined prediction scheme so as to explore the redundancy inherent to the 3D holoscopic content.

Similar to motion compensation, the SS compensation uses as reference the previous coded and reconstructed area of the current frame itself. This reference frame is referred to as the SS reference. Then, instead of compensating the temporal redundancy between objects in neighboring frames, it tries to compensate the aforementioned significant cross-correlation existing between neighboring micro-images.

4. Performance evaluation

This section assesses the Rate Distortion (RD) performance of the scalable 3D holoscopic prediction scheme proposed in this letter. For this purpose, the test conditions are firstly introduced and, then, the obtained results are presented.

Test Conditions

Three different 3D holoscopic test images with different spatial resolutions were used so as to achieve representative RD results: *Plane and Toy* (1920×1088), *Scene1* (5426×3564), and *Laura* (7240×5432).

To generate the content for the first two scalability layers, the three test images were processed with both

algorithms cited in Section 2 (*Basic Rendering* and *Weighted Blending*) where the patch size was chosen to let the main object appear in sharp focus. In this process, nine different 2D views were generated – one for the *Base Layer* and eight for the *First Enhancement Layer*. The nine views of each test image were coded independently with the HEVC using the "*Intra, main*" configuration.

Afterwards, each set of nine coded and reconstructed 2D views is processed to generate a corresponding IL reference, which is ready to be included in the reference picture buffer of the HEVC together with the SS reference.

Finally, the proposed combined prediction scheme for scalable 3D holoscopic coding (*Scalable+SS*) is compared against the simulcast case (*HEVC*), where each 3D holoscopic test image is coded with HEVC with "*Intra, main*" configuration.

Experimental Results

The performance was evaluated in terms of Bjontegaard Delta (BD) measurement method [*6*] of PSNR and bitrate (BR) for each test sequence, as shown in **Error! Reference source not found.**. The esults for each view generating algorithm (*Basic Rendering* and *Weighted Blending*) are properly presented in different columns as it refers to different content in the first two hierarchical layers of the scalable coding architecture.

From the results in **Error! Reference source not ound.**, it can be seen that the proposed combined scalable coding scheme, *Scalable+SS*, always outperforms the simulcast case (*HEVC*) with considerable gains and bitrate savings (up to 1.58dB and -21.16%), independently of the used view generating algorithm.

Table 1 BD-PSNR and BD-BR test results

	Basic Rendering		***Weighted Blending***	
	BD-PSNR	**BD-BR**	**BD-PSNR**	**BD-BR**
Plane and Toy	1.48 dB	-20.13 %	1.58 dB	-21.16 %
Scene 1	0.36 dB	-9.88 %	0.30 dB	-8.42 %
Laura	2.26 dB	-30.05 %	2.24 dB	-29.87 %

5. Conclusion

This letter presented a three-layer coding solution for 3D holoscopic content to enable display scalability with legacy displays. Based on this architecture, a combined prediction scheme is presented where an inter-layer prediction is combined with a self-similarity compensated prediction. It was shown that the proposed display scalable coding scheme always outperforms the simulcast solution based on the original Intra HEVC prediction scheme.

References

[1] Aggoun, E. Tsekleves, D. Zarpalas, P. Daras, A. Dimou, L. Soares, P. Nunes., "Immersive 3D Holoscopic Video System," *IEEE MultiMedia*, vol. 20, no. 1, pp. 28-37, Jan.-March 2013.

[2] G. J. Sullivan, J.-R. Ohm, W.-J. Han, T. Wiegand, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649-1668, December 2012.

[3] T. Georgiev and A. Lumsdaine, "Focused plenoptic camera and rendering," *Journal of Electronic Imaging*, vol. 19, no. 2, pp. 021106–021106, April 2010.

[4] Conti, P. Nunes, and L.D. Soares, "Inter-Layer Prediction Scheme for Scalable 3-D Holoscopic Video Coding," *IEEE Signal Processing Letter*, vol. 20, no. 8, pp. 819,822, Aug. 2013.

[5] Conti, P. Nunes, and L. D. Soares, "New HEVC Prediction Modes for 3D Holoscopic Video Coding," in *ICIP 2012*, Orlando, USA, October, 2012.

[6] G. Bjontegaard, "Calculation of average PSNR differences between RD curves," ITU-T VCEG, Austin, TX, USA, Document VCEG-M33, April, 2001.

Caroline Conti (S'11) received her B.Eng in Electrical Engineering from University of São Paulo (USP), Brazil, in 2010. She is currently a Ph.D. candidate in Information Science and Technology at University Institute of Lisbon (ISCTE–IUL), Portugal. Her research interests include video coding, 3D video representation and 3D holoscopic video coding.

Paulo Nunes (M'07) received the B.S., M.Sc. and Ph.D. degrees in electrical and computer engineering from the Instituto Superior Técnico (IST), Universidade Técnica de Lisboa, Portugal, in 1992, 1996, and 2007, respectively. He is currently an Assistant Professor at the Information Science and Technology Department, University Institute of Lisbon (ISCTE-IUL), Portugal, and a member of the Research Staff of Instituto de Telecomunicações, Portugal. His current research interests include 3D video coding and multimedia communications.

Luís Ducla Soares (S'98-M'01) received the B.Sc. and Ph.D. degrees in electrical and computer engineering from the Instituto Superior Técnico (IST), Universidade Técnica de Lisboa, Portugal, in 1996 and 2004,

respectively. He is currently an Assistant Professor at the Information Science and Technology Department, University Institute of Lisbon (ISCTE-IUL), Portugal, and a member of the Research Staff of Instituto de Telecomunicações, Portugal. His current research interests include 3D video coding and processing.

3D content processing challenges for Mixed Reality

Marius Preda
Institut MINES TELECOM, Telecom SudParis, ARTEMIS, France
marius.preda@institut-telecom.fr

1. Introduction

Mixed Reality is a term that refers to real-time combinations of reality - captured by various sensors and represented in digital form, with computer generated assets. Milgram [1] stipulates that there is a continuum of such combinations, from all-real to all-virtual. In this continuum, two points are relatively easy to understand (Figure 1): the Augmented Reality (AR), defined as mixing graphical objects with the images captured by a camera, and Augmented Virtuality (AV), defined as adding real images and videos in a synthetic scene.

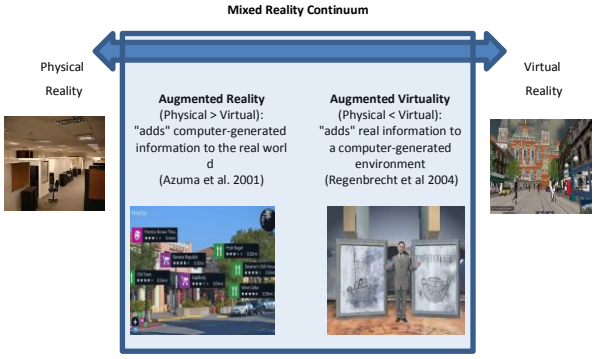


Figure 4 — The Mixed Reality Continuum, extracted from Mixed and Augmented Reference Model [2].

In both cases, creating fully interactive and dynamic experience requires appropriate 3D representation and processing. In AR, for realistic augmentation, the real 3D space should be captured and correctly interpreted. In AV, the real 3D object should be captured and represented such as generating various views from different perspectives should be possible during the rendering phase.

In this paper we are presenting several issues, possible solutions and future trends in addressing 3D aspects in AR (section 2) and AV (section 3).

2. Capturing and mixing 3D for Augmented Reality experiences

Creating interactive multimedia applications is a complex task: not only is content composed of different media assets (images, 2D/3D videos, 2D/3D graphics), but one must also provide a programming layer to implement the application behaviour, and tools to capture and interpret different kinds of input. Creating immersive applications is even less developer-friendly due to the high expectations users have on content quality (in the case of virtual reality) or to the "disbelief" caused by an imperfect match between natural and digital content (in the case of augmented reality). Indeed, like in all interactive applications, the level of user involvement depends strictly on the quality of the scenario. 3D processing is needed in two kinds of operations: on one side for capturing the real space, to be further on used in real time as an augmentation support and, on another side, for creating compelling graphics assets.

For the 3D reconstruction of captured real environments, several cameras are typically used. Classically, this is based on monoscopic color cameras calibrated by computer vision methods [3] for estimating extrinsic and intrinsic camera parameters. Stereo and multi-view matching among the cameras can be performed by finding feature points of identical scene content in each image [4] and disparities between feature points can be estimated to generate depth data [5]. With multiple feature points, projection matrices can be calculated. Additionally, depending on the nature of this initial knowledge, different auto-calibration algorithms can be used [6]. From the calibrated cameras, a 3D approximation of the visible scene content can be extracted [7] and be transformed into a wireframe surface model, where the surface geometry of the scene is modelled by a polygonal mesh and the recorded images mapped onto it [8]. More recently, additional depth sensors such as time-of-flight cameras [9], are used for direct acquisition of depth and range information. A combination of colour camera and depth sensor is also used in Microsoft Kinect, which was conceived primarily for motion sensing, but can also obtain depth information for up to 3,5 m. While various approaches were proposed in the literature, reconstructing (even partially) useful 3D representation of spaces, especially outdoor and unconstrained environments, remains challenging.

An essential element of the AR experience is the mobile application (or AR browser), a software component enabling the augmentation of the physical environments and their fusion with digital artefacts. The AR browser should be multi-platform, targeting heterogeneous mobile terminals (smartphones and tablets). High complexity of the current AR systems quickly exceeds the performance of mobile

technologies. In fact, the parallel execution of very complex operations, such as the continuous terminal pose estimation, the object tracking and the rendering of computationally demanding graphical effects, e.g., the blended presentation of 2D/3D synthetic and 2D/3D natural video objects, represent major issues that tend to saturate the available processing capabilities of the terminal; this, in turn, has a substantial impact on the user experience. To convey an immersive and satisfactory user experience, improvements to the state of the art need to be achieved. Several researchers investigated hybrid, server-client approaches in order to reduce the computation load of the mobile terminal. Therefore, it is possible to deport object detection [10], computation of camera parameters [11], 3D simulation [12] or even the whole process, including the rendering of the mixed content. These approaches require in general a very good network connection. Therefore, research on aspects deemed of primary importance for reducing the player computational load should be conducted, namely: the design of efficient user interaction paradigms, the usage of sensor information available on the mobile terminals, the development of optimised algorithms for object searching and tracking, and the implementation of lightweight data streaming. Within this context one more element in this software architecture deserves a special mention: the fully immersive experience. Three dimensional video and audio rendering on mobile terminals should be achieved including efficient algorithms for the real-time adaptation of 3D media content to the terminal capabilities. Mechanisms for integration of 2D and 3D, synthetic and natural, video and audio contents, into a harmonised and graceful multi-media scene, are key elements for a successful AR experience.

Within this context, the MPEG consortium published beginning of 2014 the first version [13], and is currently working on the second version, of a standard that has as an objective to provide a full formalism for expressing AR experiences. This formalism is a pivot element, able to create the bridge between the creation of AR experiences by powerful authoring tools, and their consumption within AR browsers, while allowing to the two ends to take advantages of specific platform capabilities or access to distributed resources. This vision is illustrated in Figure 2.

2. Capturing and representing 3D for Augmented Virtuality experiences

Since 1963, when Sutherland created the first computer program for drawing, the field of synthetic graphics has followed, as many others involving computers, more or less the same exponential growth foreseen by Moore in 1965 for the semiconductor industry. In the early years,

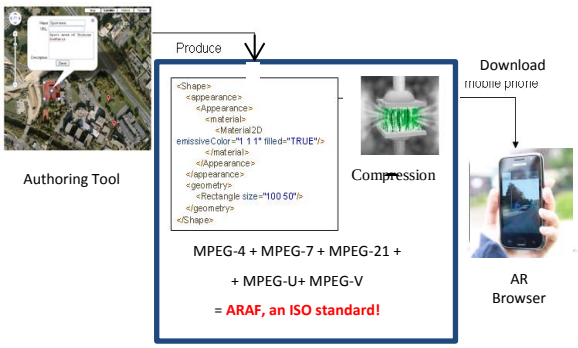


Figure 5 — ARAF, the bridge between authoring and consumption.

the progress was pushed up by scientific interest, the real boom starting when the need for special effects and 3DG content came from the film industry, and later from the videogames one. Of those two industries, the former leads the development of technologies for the production of high quality images, and the latter, which has already outgrown the former in terms of market share, that of efficient techniques for real-time processing and rendering.

Even from the beginning, the strategy for developing 3DG solutions was based on three requirements: faster, cheaper, better (image quality). Traditionally, the 3DG chain was focused on the two ends: production and consumption, each side having its own strong requirements. Additionally, in many applications such as games, dealing with complex 3D environments, there is an iterative loop involving content creators and programmers or beta-testers, making sure that the content fits a specific end-user terminal configuration. The complexity of creating 3DG content leads to the coexistence of several authoring tools, each specialized in one or more particular tasks. More recently, by using 3D reconstruction tools, it is possible in some extend to obtain, in an automatic manner and without the need of specific knowledge of the 3D authoring tools, representations of Digital Artefacts (DAs) present in physical spaces. DAs can be integrated in the repository, indexed, connected to descriptions available on specialised websites, introduced into the AR scenes, and later presented to end users. End users themselves should be able to augment DAs yet again, à la “AR 2.0”, either with respect to their representation (by taking new photos and videos who will be further processed by the reconstruction tools) or to their semantics (by providing comments, recommendations, etc.). A key aspect of dealing with DAs is their representation, since not all of them are user or processing friendly. Since only the object surface shape and appearance are taken into account by most

rendering engines, "truly 3D" models describing solids are much less frequent than "merely 2,5D" ones describing surfaces (immersed in 3D space, but with just two degrees of freedom). In traditional surface representation techniques, the shape is described first, and then an appearance is linked to it. For a triangle mesh (the most used method for representing the surface), this is done by assigning attributes to its elements (triangles, vertices, etc.). Examples of such attributes are color, local normal vector to the surface, or, most typically, a pair of texture coordinates. Some of the mesh features may be continuously updated, therefore transforming the static object into a dynamic one. Most commonly, two kinds of features are animated over time, and both concern the geometry data: either the position and/or orientation of the whole set of vertices is changed globally, which means that the object moves and/or rotates inside the scene, or the locations of some vertices are altered relative to that of others, which yields a shape deformation. When the objects are build by authoring tools, mesh connectivity remains always consistent over time, and so do textures and attributes. Unhappily, this is not the case when the object is captured with several cameras, in real-time as usually required in the case of Augmented Virtuality. While some approaches are using a template and modify it with respect to the captured information, therefore keeping the same connectivity, a large set of methods consider the non-consistent connectivity. This simplifies significantly the process, by deploying multiple depth-cameras from different angle being possible to reconstruct the mesh for each frame by using general-purpose GPU and parallel computing [14, 15]. Each frame can be therefore encoded as a static mesh. Various static mesh coding approaches are available in the literature, including standardized by ISO as part of MPEG-4 Animation Framewok eXtention and we refer to [16] for a detailed description of these methods. As for image or video coding, the concepts of resolution scalability and of single-vs. multi-rate coding are important for 3D mesh coding. In some scenarios, a single-rate coding of a 3D mesh may be enough, but its progressive compression may be desirable, especially if it is a complex mesh to be transmitted over a network with a restricted bandwidth, or to terminals with limited processing power. Traditionally, mesh deformation coding is addressed in two ways: one can either code the new vertex positions themselves as a function of time, or use some deformation controller to influence the mesh geometry, and code the controller parameters as a function of time. In both cases, classic signal processing techniques such as space-frequency transforms or predictive and entropy coding can be used to reduce the data size. Usually, the animation data is highly redundant, so compressing it may easily reduce the size by two orders of magnitude.

The challenge remains for compressing more effective non-consistent connectivity meshes. It is obvious that the temporal redundancy can be exploited since from one frame to another, only few parts of the object changes (including connectivity) and recent exploration activities carried out by the 3DG sub-group of MPEG demonstrated significant potential.

3. Conclusions

In the last decade, several efforts have been made to develop representation and processing methods to deal with 3D objects and scenes. Various requirements should be fulfilled, starting from editable representations to fast rendering and there is no unique ways to address all of them. Built on top of VRML97, MPEG-4 contained, already in its first two versions published by ISO in 1999 and 2000, tools for the compression and streaming of 3DG assets, enabling to describe compactly the geometry and appearance of generic, but static objects, and also the animation of human-like characters. Since then, MPEG has kept working on improving its 3DG compression toolset and published already four Editions of MPEG-4 Part 16, "Animation Framework eXtension (AFX)" [5], which addresses the requirements above within a unified and generic framework. More recently MPEG published the first version of its Augmented Reality Application Format, integrating in a consistent manner video, audio and graphics elements with data coming from various set of environmental sensors. MPEG is currently working on the second version of ARAF and on non-consistent connectivity meshes compact representation in order to provide a full solution for 3D representation to be used in Milgram MAR continuum.

References

[21] Paul Milgram; H. Takemura, A. Utsumi, F. Kishino (1994). "Augmented Reality: A class of displays on the reality-virtuality continuum". Proceedings of Telemanipulator and Telepresence Technologies. pp. 2351–34.

[22] Preda, Marius; Kim Gerry, Candidate WD 4.0 Mixed and Augmented Reference Model, w14250, 107th MPEG Meeting, San Jose, January 2014.

[23] R. Hartley and A. Zisserman, "Multiple View Geometry in Computer Vision", Cambrigde Universitity Press, 2003.

[24] C. Cigla, X. Zabulis, and A. A. Alatan, "Region-based dense depth extraction from multi-view video", Proc. IEEE Intl. Conf. on Image Processing (ICIP'07), vol. 5, pp. 213-216, Sep. 2007.

[25] S.-B. Lee and Y.-S. Ho, "View Consistent Multiview Depth Estimation For Three-dimensional Video

Generation", Proc. IEEE 3DTV Conf., Jun. 2010.
[26] P. Carballeira, J.I. Ronda and A. Valdés, "3D reconstruction with uncalibrated cameras using the six-line conic variety", Proc. IEEE Intl. Conf. on Image Processing (ICIP'08), pp. 205-208, Oct. 2008.
[27] W. Waizenegger, I. Feldmann, and O. Schreer, "Real-time Patch Sweeping for High-Quality Depth Estimation in 3D Videoconferencing Applications", Proc. SPIE Conf. on Real-Time Image and Video Processing, vol. 7871, pp. 78710E-10, Jan. 2011.
[28] K. Müller, A. Smolic, M. Droese, P. Voigt, and T. Wiegand, "3D Reconstruction of a Dynamic Environment with a fully Calibrated Background for Traffic Scenes", IEEE Trans. on Circuits and Systems for Video Technology, vol. 15, no. 4, pp. 538-549, Mar. 2005.
[29] S. Gokturk, H. Yalcin, and C. Bamji, "A time-of-flight depth sensor system description, issues and solutions", Proc. IEEE Computer Vision and Pattern Recognition (CVPR'04) Workshop, vol. 4, pp. 35-43, Jun. 2004.
[30] S. Gammeter, A. Gassmann, L. Bossard, T. Quack, and L. Van Gool. Server-side object recognition and client-side object tracking for mobile augmented reality. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPRW '10, pages 1–8, 2010.
[31] T. Verbelen, T. Stevens, P. Simoens, F. De Turck, and B. Dhoedt, Dynamic deployment and quality adaptation for mobile augmented reality applications. Published in the Journal of Systems and Software (JSS), 84(11):1871–1882, 2011.
[32] B. Kevelham, N. Nijdam, G. Papagiannakis, M. Lim, N. Magnenat-Thalmann, "Remote Augmented Reality for Virtual Fashion Models", Workshop on Hyper-media 3D Internet,Geneva, October 14 2008
[33] ISO/IEC 23000-13 Information technology - Multimedia application format (MPEG-A) -- Part 13: Augmented reality application format, ISO, 2014.
[34] Alexiadis D., Kordelas D.S., G. Apostolakis, K.C. ; Agapito, J.D. ; Vegas, J.M. ; Izquierdo, E. ; Daras, Reconstruction for 3D immersive virtual environments (2012). 3th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS) pp.1-4
[35] [7] Vasudevan R., Kurillo K., Lobaton E., Bernardin T.,Kreylos O., Bajcsy R., Nahrstedt, K. High Quality Visualization for Geographically Distributed 3-D Tele-immersive Applications(2011). IEEE Transactions on Multimedia, Vol. 13, NO 3, June 2011 pp. 573-584
[36] Francisco Morán Burgos and Marius Preda, MPEG 3D Graphics Representation, Chapter 11, in The MPEG Representation of Digital Media, Leonardo Chiariglione (Editor), Springer; 2012 edition (October 28, 2011).

Marius Preda received a B.E. degree from "Politehnica" University, Bucharest and Ph.D. degree from "Paris 5" University, France, in 1998 and 2002 respectively. He is Associate Professor at Institut MINES-Telecom and Chairman of the 3D Graphics group of ISO's MPEG (Moving Picture Expert Group). He contributes to various ISO standards with technologies in the fields of 3D graphics, virtual worlds and augmented reality and has received several ISO Certifications of Appreciation. He leads a research team with a focus on Augmented Reality, Cloud Computing, Games and Interactive media.

SPECIAL ISSUE ON BIG DATA ANALYTICS FOR MULTIMEDIA SYSTEMS

Big Data Analytics for Multimedia Systems

Guest Editor: Zhu Liu, AT&T Labs Research, USA

zliu@research.att.com

There is no doubt that we are at the dawn of a big data era. With the ease of creating, capturing, transmitting, processing, and storing all kinds of data anytime and anywhere, the big bang of the data universe just started. Big data is defined by not only lots of data, but also high velocity and variety, which reflect the fast movement and the diverse structure aspects in the big data ecosystem. A spectrum of new platforms and technologies (Hadoop, NoSQL database, toolkits for natural language understanding, machine learning, and visualization, just to name a few) paved the way for analyzing the heterogeneous data source and exploring the valuable business intelligence. While the big data tidal wave brings us unprecedented opportunities for data analytics, new challenges also emerge. Factors including reliability, scalability, security, privacy, cost, etc., may impede the progress of big data, and further fundamental research is required to address all of them to unleash the full potential and power of big data.

Video and multimedia data, which accounts for more than 60% of the Internet traffic, is obviously a major constituent of the data universe. With more than 100 hours of videos uploaded to YouTube per minute, about 9 billion photos sent to Facebook per month, hundreds of HD TV channels being broadcast, and millions of public/private surveillance cameras monitoring the world, multimedia data is growing blindingly fast. The key to make insight from such a big volume of multimedia data is a set of effective analytics tools that can understand the embedded semantics correctly and swiftly.

This special issue of E-Letter focuses on the recent progresses of big data analytics for multimedia systems. It is the great honor of the editorial team to have six leading research groups, from both academia and industry laboratories, to report their solutions for meeting these challenges and share their latest results.

In the first article titled, "*Detecting Complex Events from Big Video Data",* Jingen Liu, Omar Javed, Hui Cheng, and harpreet Sawhney from SRI International presented their large scale complex event detection system. The system achieved promising results in the TRECVID 2012 and 2013 multimedia event detection (MED) and multimedia event recounting (MER) evaluation tasks hosted by NIST. The reported approach employed both the state of the art low-level visual features and the semantic concept features and explored various fusion strategies to improve the detection performance. To take advantage of distributed computing resources and provide high scalability for processing big video data, the system was built on HTCondor, a high-throughput computing framework.

Zhigang Ma, Shoou-I Yu, and Alexander Hauptmann from CMU authored the second article, "*How to Efficiently Handle Large-Scale Multimedia Event Detection*". The participants of the multimedia event detection (MED) evaluation task of TRECVID 2014 will process 8000 hours of video within a short period of time. Such a big video corpus poses a real challenge for systems that pursues high detection accuracy while paying less attention to the computational cost. The authors reviewed existing techniques that can potentially reduce the computational complexity of a event detection system. Specifically, efficient methods in four categories were surveyed, including feature extraction, kernel computation, learning classification model, and using fewer training samples. Combination of these tactics will further boost the system efficiency, which is essential for handling the large-scale video data used in the MED task.

The third article is contributed by David Gibbon and Lee Begeja from AT&T Labs Research, and the title is *"Distributed Processing for Big Data Video Analytics"*. The authors gave an overview of video analytics with an emphasized consideration on the distributed architecture that can take advantage of a camera-to-cloud processing path. Constituents, including advanced IP-connected cameras, distributed video processing architectures, and standards for video metadata representation, are vital for deploying practical video analytics systems. To support additional novel information services, the video analytics platform also needs to incorporate traditional structured or unstructured big data sources including location, text based media, etc.

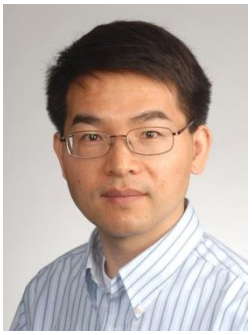
Ye Xing, Vivek Gupta, and Tao Wu presented the fourth article, *"A Survey on Personal Digital Photo Management"*. This article provided a high-level review on two main techniques adopted in automatic personal photo organizers: content-based image

analysis and the use of photo metadata. The purpose of image analysis is to detect objects or people in the photo, and the rich metadata, including time, location, temperature, etc. is able to provide useful contextual information. The main motivation of applying these techniques in photo management system is to reliably detect the three key features: event, person, and time, which are most frequently used by consumers when sorting their collections of photos.

The fifth article is *"Dynamic Structure Preserving Map (DSPM) for Human Action Primitive Modeling"*, from Qiao Cai and Hong Man at Stevens Institute of Technology. The authors introduced their recent work on human action detection using a method called dynamic structure preserving map (DSPM), which is an extension to self organizing map (SOM). The three unique properties of DSPM, including learning low-level features automatically, aggregating the spatial-temporal clustering, and reducing the dimensionality of raw feature set, make it suitable for the challenging human action recognition task. Experimental results on multiple complex datasets, especially the UCF sport dataset, demonstrate that the DSPM approach is effective and robust.

The last article of this special issue is from Trista Chen at Cognitive Networks, and the title is *"Multimedia Big Data: From Large-Scale Multimedia Information Retrieval To Active Media"*. This article provides an insightful overview of recent research trends in multimedia big data. The author first discussed the efforts on reducing the size of the visual features and increasing their indexing effectiveness, and then explained the advantageous impact on the adopted machine learning methods by the massive amount of multimedia data. Finally, a novel class of applications emerging in the era of multimedia big data, named active media, was introduced. Active media includes interactive marketing, voting and polling, program promotions, etc., which do not require users to submit specific queries for multimedia content of interest.

While this special issue is far from delivering a complete coverage on this exciting research area, we hope that the six invited letters give the audiences a taste of the main activities in this area, and provide them an opportunity to explore and collaborate in the related fields. Finally, we would like to thank all the authors for their great contribution and the E-Letter Board for making this special issue possible.



Zhu Liu received the B.S. and M.S. degrees in Electronic Engineering from Tsinghua University, Beijing, China, in 1994 and 1996, respectively, and the Ph.D. degree in Electrical Engineering from Polytechnic Institute of New York University, Brooklyn, NY, in 2001. He joined AT&T Labs Research, Middletown, NJ, in 2000, and is currently a Principle Member of Technical Staff in the Video and Multimedia Technologies Research Department. He was an adjunct professor of the Electrical Engineering Department of Columbia University from 2010 to 2013. His research interests include multimedia content analysis, multimedia databases, video search, pattern recognition, machine learning, and natural language understanding. He holds 45 U.S. patents and has published more than 70 conference, workshop, and journal papers. Dr. Liu is on the editorial board of the IEEE Signal Processing Letter and the Peer-to-peer Networking and Applications Journal. He was also on the organizing committee and technical committee for a number of IEEE Conferences. Dr. Liu is a senior member of IEEE and a member of ACM.

Detecting Complex Events from Big Video Data

Jingen Liu, Omar Javed, Hui Cheng, Harpreet Sawhney
SRI International, Princeton, NJ, USA
{first-name.last-name}@sri.com

1. Introduction

In this fast paced digital age, a vast number of videos including movies, TV programs, and personal consumer videos, are produced every day. For instance, as March of 2014, about 100 hours videos were uploaded to YouTube every minute [6]. Detecting complex events from this never-ending growing big video data is very challenging due to the fact of large-scale, as well as the characteristics of events and videos "in the wild". For example, a very specific event, such as "wedding ceremony", "birthday party", "parkour", "getting a vehicle unstuck", and so on, usually covers a great diversity of contents involving various objects, atomic human actions, scenes, and audio information. On the other hand, open source videos are unconstrained which are typically recorded under uncontrolled conditions with large variations in camera motion, illumination, object appearance and scale, as well as viewpoint [7]. Therefore, to capture various aspects of an event, we develop various low-level features (LLFeat) and semantic concept features (i.e., high-level feature, HLFeat) in our system, and explore different fusion strategies to inquire discriminative information from all aspects of an event.

Our system is developed on HTCondor [16], which supports high throughput computing on distributed computing resources, and is also able to configure the job dependency. It has been worked with over 150K videos (about 4000 hours), and this number is increasing. The system obtains promising performance on multiple challenging tasks during the TRECVID MED evaluation 2012 [14] and 2013 [15]. The entire data processing flow of system is demonstrated in Fig. 1. In the following sections, we discuss them in detail.

2. LLFeat Based Event Representation

Low-level features are designed to acquire the first-hand characteristics of an event, such as the object appearance and color information, and scene structure. Our system incorporates three types of low-level features: static features, dynamic motion features, and audio features. The static features include Sparse SIFT [11], Dense SIFT, ColorSIFT [12], and Transformed Color Histogram [13]. The dynamic motion features include STIP [9], Dense Trajectory Feature (DTF) [10], and MoSIFT. The detailed information about these features is discussed in our previous work [4].

The above set of features is computed either on the key-

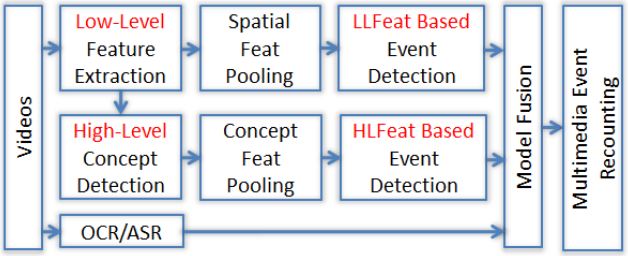


Figure 1: System data processing flow. Low-Level feature based event recognition and high-level concept based event recognition are two major components of the system.

frames or on spatial-temporal windows of frames throughout a given video. The event happening in the video is represented as an aggregate feature, which is the histogram of "words" corresponding to each feature type computed over the entire video. This is popularly known as a "Bag-of-Words"(BoW) representation. The feature specific vocabularies are first learned using K-means clustering of raw features. Once the features in a video are quantized using the respective vocabularies, a BoW vector is computed per feature for a video. As an alternative, Fisher Vector is also employed to aggregate features in a video by GMM based soft quantization. Event models are learned using SVM with intersection kernel. We thoroughly evaluated the performance of each feature in [4].

The BoW is treated as an average feature pooling over the holistic video. However, a specific event typically bears some Region of Interest that captures the most discriminative information of the event. Hence, we propose a new strategy for spatial pooling of the LLFeat. The basic idea is similar to Spatial Pyramid Match [2], which constructs the pyramid structure for an image. The matches at the fine level contribute more to the final match score.

3. Concept-Based Event Representation

In general, the LLFeat Based Event Representation needs a large number of training samples to obtain better model generalization. This is due to the fact that the visual contents of an event are usually very diverse. For instance, a " wedding ceremony" consists of various concepts including actions such as "hugging" and "kissing", scenes such as "church" and "garden",

and objects such as "cake" and "ring". In this scenario, it is difficult to solve the intra-class variability with only LLFeat, especially when the number of training examples is small. In addition, the numeric LLFeat has no semantic meaning, which makes it infeasible to perform high-level video understanding, such as event recounting. In our system, we propose to represent events with their concept attributes. Not only does it overcome the aforementioned limitations of LLFeat model, but it enables MED without training examples.

We embed the event videos in a semantic space consisting of concepts related to actions, scenes and objects. Action concepts are typically atomic and localized motion and appearance patterns, which are strongly associated with some specific event. Both scene and object concepts captured through key-frames provide an environment of an event. Within this semantic space, we are able to depict an event by various concept attributes. Our system builds a semantic space with about 1,700 concepts [1]. We collect two types of concepts: data-relevant concepts and data-irrelevant concepts. The former is annotated from the event videos. To contrast, the latter is acquired from a third party, such as ImageNet scenes/objects, TRECVID SIN concepts, and UCF-101. They may contain both event related and unrelated concepts.

However, the human manual annotation is very laborious, which is not scalable to detecting events from big video data. To release annotators from this time-consuming job, we also explore a novel strategy to achieve semi-automated concept annotation (SAA) via PageRank technique. A regular process of concept annotation over consumer videos starts with downloading relevant videos of one concept using search queries and then annotators start to localize the start and end times of a concept. We notice that, given a specific well-defined concept, the majority of video shots in the collected videos are relevant. By applying PageRank to the video-shot similarity network, we are able to select the most relevant video shots as positive video clips.

We employ well-established techniques for action, scene and object detection for building our concept detectors. In particular, static features (i.e., SIFT), dynamic features (i.e., STIP and DTF), and the BoW models defined over vocabularies of these features are used to represent action, scene and object concepts. Binary SVM classifiers with Histogram Intersection kernel/Linear Kernel are used for concept training.

The trained concept detectors are applied to an event video, and then various concepts features, including Max Concept Detection Scores (MAX), Statistics of ConceptScores (SCS), Bag of Concepts (BoC), Co-occurrence Matrix (CoMat), and Max Outer Product (MOP), are derived from the detection to represent an vent video. In [3, 1], we systematically demonstrate the advantages of HLFeat in applications of video event detection and understanding. Specifically, it has better generalization capability, which enables to much better MED with a few training examples than that of LLFeat. In addition, it enables to recognize a novel event without training examples. Furthermore, it provides a straightforward way for event recounting/ understanding.

4. Multimedia Event Recounting

A video event is a complex activity occurring at a specific time. Such a video may contain a lot of irrelevant information. Thus, for each recognized event occurrence in a video, the goal of recounting is to describe the details of the occurrence. The recounting includes key observations regarding the scene, people, objects, and actions pertaining to the event occurrence. Such recounting provides user a semantic description that is useful to perform further analysis. As concept features that we use by definition contain semantic information, concept features are more appropriate for recounting purpose than low-level features. In [5, 3], we have detailed discussions on this topic.

5. Conclusions

To successfully detect complex event from the big video data, we develop a novel system which employs the state-of-the-arts low-level features and semantic concept features. The proposed novel representation of events defined in terms of a semantic space of action, scene and object concepts makes our system scalable to big video data. The concept based event representation (CBER) requires less number of training examples versus low-level features for similar event classification performance. It also enables to detect a new event without any training examples. Our system has been evaluated on the challenging TRECVID MED, which is completely unconstrained, and large scale dataset from open sources. It obtains promising results, especially for the EX10 task (having 10 training examples) and EX0 task (without training examples). The CBER also enables our system to perform multimedia event recounting.

Acknowledgment

This work has been supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11-PC20066. The U.S. Government is authorized to reproduce and distribute reprints for

Governmental purposes not with-standing any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.

References

[1] H. Cheng, J. Liu, S. Ali, O. Javed, and et al. Sri-sarnoff aurora system at trecvid 2012: Multimedia event detection and recounting. In TRECVID, 2012.

[2] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In CVPR, 2006.

[3] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, and H. Sawhney. Video event recognition using concept attributes. In WACV, 2013.

[4] A. Tamrakar and et al. Evaluation of low-level features and their combinations for complex event detection in open source videos. In CVPR, 2012.

[5] Q. Yu, J. Liu, H. Cheng, A. Divakaran, and H. Sawhney. Multimedia event recounting with concept based representation. In ACM MM, 2012.

[6] http://www.youtube.com/yt/press/statistics.html

[7] J. Liu, J. Luo and M. Shah, Recognizing Realistic Actions from Videos "in the Wild". In CVPR 2009.

[8] J. Liu, H. Cheng, and et. al, SRI-Aurora Team at TRECVID 2013, Multimedia Event Detection and Recounting. In TRECVID 2013.

[9] I. Laptev and T. Lindeberg. Space-time interest points. In ICCV 2003.

[10] H. Wang, A. Klser, C. Schmid, and C. L. Liu. Action recognition by dense trajectories. In CVPR, 2011.

[11] D. Lowe, Distinctive image features from scale invariant key-points. IJCV, pp. 91-110. 2004

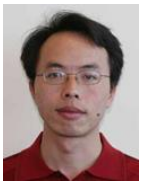
[12] K. E. Sande, T. Gevers, C. G. Snoek, Evaluating color descriptors for object and scene recognition. TPAMI, 2010.

[13] G.J. Burghouts and J.M. Geusebroek, Performance Evaluation of Local Color Invariants, CVIU, vol. 113, pp. 48-62, 2009.

[14] www-nlpir.nist.gov/projects/tv2012/tv2012.html

[15] www-nlpir.nist.gov/projects/tv2012/tv2013.html

[16] http://research.cs.wisc.edu/htcondor/

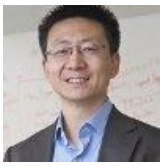


Jingen Liu is a Senior Computer Scientist at SRI International, who received PhD degree in Computer Science from University of Central Florida in 2009. He was a Research Fellow at University of Michigan (Ann Arbor) between 2010 and June 2011.

His research interests include action recognition, large-scale video event recognition, scene understanding, and moving object detection. He has authored more than 30 papers on prestigious computer vision conferences such as CVPR, ICCV and ECCV. He co-chairs the THUMOS 2013 and 2014 challenge on action recognition with a large number of classes. Currently, he is a senior member of IEEE.

Omar Javed is a principal scientist and technology leader at the Vision and Learning Lab at SRI International. His areas of interest include video event understanding, object tracking, multi-sensor surveillance and online machine learning.

Javed is the author of the book Multi-Camera Surveillance Algorithms and Practice and his article "Object Tracking: A Survey" was ranked #1 in ACM's Most Popular Magazine and Computing Survey articles in 2007. His paper on "Modeling Inter-Camera Space-Time and Appearance Relationships for Tracking across Non Overlapping Views" was listed as the top-10 most cited paper in the Computer Vision and Image Understanding Journal from 2007-2011.



Hui Cheng is the Program Director of Cognitive and Adaptive Vision Systems Group in the Center of Vision Technology of SRI International. He received his Ph.D. degree in Electrical Engineering from Purdue University. His research focuses in the area of image/video understanding, data mining, machine learning, pattern analysis, cognitive systems and informatics. He has led a number of programs in the area of full motion video, wide-area surveillance video and web video exploitation, summarization and indexing.

He is also the technical lead in automated behavior analysis and performance evaluation for military training. Dr. Cheng has more than 50 publications and 20 patents. He is a senior member of IEEE, a member of IEEE Technical Committee on Multimedia Systems and Applications and the Chair of Princeton/Central Jersey Chapter, IEEE Signal Processing Society.

Harpreet S. Sawhney, PhD, is CTO – Vision Technologies, and Technical Director, Vision & Learning, at SRI International in Princeton, NJ, and has over 20 years of experience in the field of Computer Vision. He has led the development of algorithms and systems in the areas of Security and Surveillance, Video Data Mining, Situational Awareness, and Training, and has done extensive work in Video and 3D analysis, Object & Event Recognition and Computer Vision. He was a key developer of video indexing within the context of IBM's pioneering QBIC

system for content based image retrieval. He was a key technical contributor to three Sarnoff start-ups/ventures: VideoBrush, LifeClips, and VideoFlashlights & VisionAlerts. He is a past Associate Editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence. He has over 100 publications and 51 issued patents. He received his Ph.D. in Computer Science from the University of Massachusetts, Amherst, and an MTech and BTech in Electrical Engineering from the Indian Institute of Technology, Kanpur, India. He was elected an SRI Fellow in 2011 and an IEEE Fellow in 2012 for his contributions to computer vision and video algorithms.

How to Efficiently Handle Large-Scale Multimedia Event Detection

Zhigang Ma, Shoou-I Yu and Alexander G. Hauptmann
Carnegie Mellon University, USA
{kevinma,iyu,alex}@cs.cmu.edu

1. Introduction

With ever expanding multimedia collections, large-scale multimedia content analysis is becoming a fundamental research issue for many applications such as indexing and retrieval, etc. Multimedia content analysis aims to learn the semantics of multimedia data. Several topics within this field have been studied in recent years, either at concept level or event level. A "concept" means an abstract or general idea inferred from specific instances of objects, scenes and actions such as fish, outdoor and boxing. Concepts are lower level descriptions of multimedia data which usually can be inferred with a single image or a few video frames. As shared video collections have explosively proliferated these years, video event analysis is gradually attracting more research interest. An "event" refers to an observable occurrence that interests users, e.g. celebrating the New Year. Compared with concepts, events are higher level descriptions of multimedia data. A meaningful event builds upon many concepts and the video can last up to a few hours. In 2010, the TRECVID community launched the task of "Event detection in Internet multimedia (MED)" which aims to encourage new technologies for detecting generic and complicated events, e.g., landing a fish. The video archive used in MED task keeps growing these years which reaches 8000 hours of videos for evaluation in 2014. This is so far the largest video corpora but only roughly 6 weeks are given to process these videos. The large scale and the tight schedule pose a great challenge for efficiency. Besides, how to learn reliable detection models with less computational cost is also important in the MED systems. In this paper, we therefore review different existing approaches on dealing with this large-scale issue, mainly including the techniques on feature extraction, kernel computation, learning classification model and using few training exemplars.

2. Feature Extraction

The feature extraction is the most time consuming part of MED task. Several teams have developed effective ways to make it more efficient. Video rescaling is a common approach to make the feature extraction more efficient. It can be realized by spatial rescaling or temporal rescaling. For example, feature extraction of motion features such as MoSIFT [1] is one of the main bottlenecks in the MED system from CMU. In order to expedite the feature extraction process, the researchers have proposed to lower the resolution of the videos such that the maximum width is less than 320 [2]. Experiments on the MoSIFT feature extraction have shown that lowering the resolution increases the speed by more than three times and does not suffer any performance loss. Some other methods have been developed by different research groups as well. For instance, BBN extended kernel descriptors to the spatio-temporal domain to model salient flow, gradient and texture patterns in video and then extracted features from different color channels [3]. In particular, they developed a fast algorithm for kernel descriptor computation of O(1) complexity for each pixel in each video patch, producing two orders of magnitude speedup over conventional kernel descriptors and other popular motion features. In [4], a system is developed to generate a temporal pyramid of static visual semantics. Kernel optimization and model subspace boosting are then applied to customize the pyramid for each event. Particularly, model subspace boosting helps reduce the size of the model significantly, which improves the efficiency for dealing with large-scale MED data.

3. Kernel Computation

It has been shown that using a Chi-square kernel is quite helpful for boosting the MED performance. However, computing a Chi-square kernel is expensive. To improve the efficiency, the research team from IBM worked on how to efficiently solve exponential Chi-square kernel [5]. Chi-square kernel is an additive kernel and can be approximated by mapping the feature into a high dimensional space as suggested by [7]. The Nystrom's approximation is used to construct the mapping function explicitly. To approximate the kernel, the researchers have employed explicitly kernel mapping [7]. The experiments have demonstrated that the new method significantly reduces the computational complexity.

4. Learning classification model

From the learning of classification aspect, several methods have been proposed recently. When classifying the widely used and very effective bag-of-word vectors, histogram intersection and Chi-square kernel SVMs (Support Vector Machines) are mostly used [2][6] as they significantly outperform the linear SVMs though the latter is more efficient. Kernels can be viewed as the dot products in a high dimensional space, which is often of infinite dimensions. The big problem with kernel SVMs is that training and

prediction are significantly slower than linear SVMs on large-scale data. Two directions have been proposed to alleviate this issue in existing MED systems, i.e., Explicit Feature Maps [7] and Fisher Vectors [8][9][10]. Vedaldi et al. have proposed to explicitly compute a finite-dimensional estimation for a feature vector which is mapped into the infinite-dimensional Chi-square kernel space [7]. The linear dot-product in the finite-dimensional space approximates the dot-product in the Chi-square kernel space. Therefore, given the finite-dimensional estimation, a linear SVM can be directly used to perform training and prediction. Experiments in [7] show no drop in performance and at least five times speedup in training. Systems using Fisher Vectors [8] have been the leading performers in the Multimedia Event Detection task [9]. Fisher Vectors are not only more effective than bag-of-words representation, but also with the convenience that a linear SVM can be directly used for these vectors. This makes the prediction very efficient as only a dot product is needed. However, the training phase could still be slow as the dimensions of the Fisher Vectors are usually 20 times higher than traditional bag-of-words. In [10], Aly et al. have proposed to reduce the number of dimensions in the Fisher Vector by representing spatial information using Gaussian Mixtures instead of the traditional spatial pyramid method [11]. Empirically, the feature dimensions can be reduced by at least 80%, thus accelerating the training phase significantly. Some other methods have also been proposed to make the classification more efficient. For example, Tang et al. utilize a conditional model trained in a max-margin framework that is able to automatically discover discriminative and interesting segments of video to understand the temporal structure of multimedia events [12]. Their model can perform fast, exact inference using dynamic programming, which is able to process a very large number of videos quickly and efficiently. In [13], low-rank representation (LRR) is optimized to scale to the massive sizes of modern datasets. This is beneficial for the scalability of subspace segmentation, which helps obtain order-of-magnitude speed ups for MED.

5. Using few training exemplars

Another way to address the efficiency issue is to use only few training exemplars. The CMU team has proposed to leverage knowledge adaptation from other multimedia resources to facilitate MED using only few positive examples for training [14]. Sun et al. proposed to apply Fisher Kernel techniques so that the concept transitions exploiting activity concept transitions of the videos can be encoded into a compact and fixed length feature vector very efficiently [15]. Their method works well with a very small number of training samples. Liu et al. propose to use concept-based event representation (CBER) for MED, which is able to detect events with a few training examples [16]. In addition, CBER makes it possible to detect a novel event without training examples (i.e., zero-shot learning). There is some other work which focuses on using related videos to mine more information so that only few positive examples are used for training. In [17], the Weighted Margin SVM formulation is modified so that related class observations can be effectively incorporated in the training of the event detector. The CMU team has proposed a regression model to assign soft labels to related examples to use such extra informative cues in the learning of the event detector [18].

6. Discussion

Multimedia event detection in large-scale video archives is an emerging research topic. On one hand, processing large-scale videos itself is a very computationally expensive process which requires more advanced video processing techniques to make it more efficient. Current efforts are mainly focused on speeding up the feature extraction. We can do it by rescaling the videos or developing faster descriptor generating algorithms. On the other hand, the efficiency of the learning of the classification models can be also improved by a variety of approaches such as more efficient kernel calculation, Explicit Feature Maps/Fisher Vector/low-rank representation based classifier learning. Further, as the events to be detected are more complicated than traditional objects, scenes and human actions, a large number of positive exemplars are usually required for reliable performance. However, it is difficult to collect the videos with precise event descriptions in practice so it is beneficial to investigate how to attain reasonable performance by using only few positive exemplars. This direction has also been explored by several researchers and leveraging knowledge adaptation is one of the effective techniques. In the future, it would be interesting to combine the various advanced techniques reviewed in this paper to further improve the efficiency. Even though current MED systems are mainly focused on better accuracy, efficiency to scale will continually increase in importance.

Acknowledgements

This research was partially supported by Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068 and by the National Science Foundation under Grant Number IIS-12511827. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained

herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, the National Science Foundation or the U.S. Government.

References

[1] M. Chen, et al, "Mosift: Recognizing human actions in surveillance videos", CMU Technical Report 2009.

[2] W. Tong, et al, "E-LAMP: integration of innovative ideas for multimedia event detection", Machine Vision and Applications 2014.

[3] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, U. Park, R. Prasad, P. Natarajan. Multi-channel Shape-Flow Kernel Descriptors for Robust Video Event Detection and Retrieval. In ECCV, 2012.

[4] N. Codella, A. Natsev, G. Hua, M. Hill, L. Cao, L. Gong, J. R. Smith. Video Event Detection Using Temporal Pyramids of Visual Semantics with Kernel Optimization and Model Subspace Boosting. In ICME, 2012.

[5] L. Cao, S.-F. Chang, N. Codella, C. Cotton, D. Ellis, L. Gong, M. Hill, G. Hua, J. Kender, M. Merler, Y. Mu, J. R. Smith, F. X. Yu. IBM Research and Columbia University TRECVID-2012 Multimedia Event Detection (MED), Multimedia Event Recounting (MER), and Semantic Indexing (SIN) Systems. In Trecvid Workshops, 2012.

[6] S. Lazebnik, et al, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories", CVPR 2006.

[7] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps", PAMI 2012.

[8] F. Perronnin, et all, "Improving the fisher kernel for large-scale image classification", ECCV 2010.

[9] J. Krapac, et al, "Modeling spatial layout with fisher vectors for image categorization", ICCV 2011.

[10] Robin Aly, et al, "The AXES submissions at TrecVid 2013", TRECVID Video Retrieval Evaluation Workshop 2013.

[11] S. Lazebnik, et al, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories", CVPR 2006.

[12] K. Tang, L. Fei-Fei, D. Koller. Learning Latent Temporal Structure for Complex Event Detection. In CVPR, 2012.

[13] A. Talwalkara, L. Mackeyb, Y. Mu, S.-F. Chang, M. I. Jordan. Distributed Low-Rank Subspace Segmentation. In ICCV, 2013.

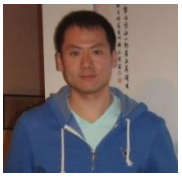
[14] Z. Ma, Y. Yang, Y.Cai, N. Sebe, A. G. Hauptmann. Knowledge Adaptation for Ad Hoc Multimedia Event Detection with Few Exemplars. In ACM MM 2012.

[15] C. Sun, R. Nevatia. ACTIVE: Activity Concept Transitions in Video Event Classification. In ICCV, 2013.

[16] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, H. Sawhney. Video Event Recognition Using Concept Attributes. In WACV, 2013.

[17] C. Tzelepis, N. Gkalelis, V. Mezaris, I. Kompatsiaris. Improving event detection using related videos and Relevance Degree Support Vector Machines. In ACM MM, 2013.

[18] Y. Yang, Z. Ma, Z. Xu, S. Yan, A. G. Hauptmann. How Related Exemplars Help Complex Event Detection in Web Videos? In ICCV, 2013.



Zhigang Ma received the Ph.D. in computer science from University of Trento, Trento, Italy, in 2013. He is now a Postdoctoral Research Fellow with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. His research interest is mainly on machine learning and its applications to multimedia analysis and computer vision.

Shoou-I Yu received the B.S. in Computer Science and Information Engineering from National Taiwan University, Taiwan in 2009. He is now a Ph.D. student in Language Technologies Institute, Carnegie Mellon University. His research interests include computer vision and multimedia retrieval.

Alexander G. Hauptmann received the B.A. and M.A. degrees in psychology from Johns Hopkins University, Baltimore, MD, the degree in computer science from the Technische Universität Berlin, Germany, in 1984, and the Ph.D. degree in computer science from Carnegie Mellon University (CMU), Pittsburgh, PA, in 1991. He is currently with the faculty of the Department of Computer Science and the Language Technologies Institute, CMU. His research interests include several different areas: man-machine communication, natural language processing, speech understanding and synthesis, video analysis, and machine learning.

Distributed Processing for Big Data Video Analytics

David Gibbon and Lee Begeja
AT&T Labs Research, USA
{dcg,lee}@research.att.com

1. Introduction

The growing number of live video streams available today, combined with advances in system integration capabilities and video processing methods, provides a rich source of quantitative information that can act as a source stream for Big Data ingest to enable a wide range of new applications. While the primary application area for video analytics (VA) has been surveillance [1], some of the methods can be applied to entertainment video sources and consumer generated video as well. This e-letter will provide an overview of VA with a focus on distributed architecture considerations for practical deployment.

While video analytics (VA) typically refers to less constrained computer vision problems than manufacturing inspection [2], the extraction of quantitative information from a manufacturing process via video monitoring followed by analysis of that data clearly uses some of the same building blocks and has similar benefit. VA is used in tasks such as retail inventory control and occupancy monitoring in large retail environments [3]. More recently, consumer applications for VA have opened up due to the widespread availability of requisite infrastructure such as broadband internet access, home WiFi and advances in IP camera technology. While USB cameras connected to PCs and cameras integrated into laptops have been prevalent for many years, it is only recently that affordable standalone IP cameras have become available. The precursors to today's smart IP cameras were standalone web cams that supported video streaming and, in some cases, pan-tilt-zoom control from web clients via HTTP. More basic implementations allowed only for JPEG frames at low frame rates. With this precedent established, today's IP cameras are targeted for more private applications such as monitoring a baby's room, a pet, or for home monitoring. Device capability has improved and costs have dropped so H.264 encoding of 1080p HD video is practical. To be competitive, camera vendors have differentiated on features such as inclusion of microphone/speakers and video quality. To improve quality, vendors have focused on correcting problems such as poor lighting conditions and unstable cameras that plague home monitoring and similar surveillance applications. Wide dynamic range (WDR) or High dynamic range (HDR) algorithms [4], digital image stabilization (DIS) [5] and geometric distortion correction can vastly improve image quality in certain situations.

When used to identify segments of interest from seemingly endless video streams, VA can provide a great benefit to professionals as well as home owners. VA also holds the promise of easing the burden of monitoring an aging parent and raising an alert in the case of a potentially life-threatening fall. Aggregated VA stream storage can provide valuable information on historical patterns and can enable anomaly detection. Systems can incorporate estimates of basic demographic information based on attribute detection (such as age range) in a manner that does not rely on identification of individuals.

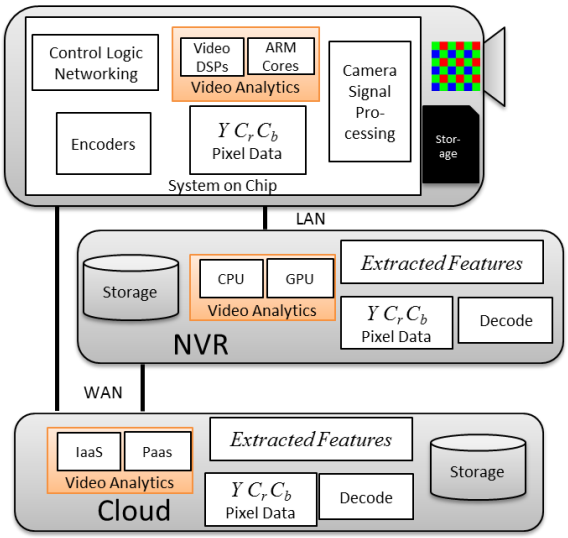


Figure 1. Distributed video analytics architecture

2. Architectures

Cost effective implementation of these capabilities requires a high degree of integration and system-on-chip (SoC) architectures [6] are used which typically employ fully customized blocks for tasks such as H.264 encoding along with general purpose computing blocks such as ARM cores. The SoC can implement traditional camera signal processing such as color correction, geometric distortion correction, and noise reduction in addition to the more advanced camera processing mentioned above (HDR, DIS). Given the SoC approach, some level of video analytics can be included with very little additional cost to the overall product. Available for a long time in electronic photography applications, face detection based on the

Voila-Jones [7] algorithm is now becoming available in consumer IP cameras for home monitoring applications. Another form of VA suitable for SoC implementation is video motion detection where a fixed camera is assumed and a background removal algorithm is used to detect moving objects. False alarms are reduced by specifying regions-of-interest or 'trip lines' which objects must cross before an alarm is triggered. More advanced approaches for motion video analytics [8] provide more accurate results and may be implemented as camera processing capabilities improve.

In enterprise applications, several cameras are monitored centrally by a dedicated PC with recording capability. Low-cost IP Network Video Recorders (NVR) for the consumer market are now available. Here, traditional CPU-based VA implementations or those that leverage GPU can play a role, in cases were cameras fall short in processing capabilities. However, in this scenario, these CPU/GPU based implementations will have to handle a number of streams simultaneously. Cloud approaches offer some promise of reduced cost, but considering that numerous streams of HD video would require high upload bandwidth, this may not be a good fit. Further, if this approach were taken, a significant component of the computational load would be fixed and would not take full advantage of the elasticity of the cloud. Camera SoC implementations also have a clear advantage in that they have access to the pixel data prior to compression (see Fig. 1). A hybrid approach where data reduction occurs in several phases seems to be the best choice. The figure suggests the components required for pixel-based processing at different access nodes along the video acquisition to cloud storage path. Algorithms that operate in the compressed domain do not require decoding, and systems that perform feature extraction on the camera advantageously do not require the pixel data (either compressed or decoded) at the later stages.

In order to optimize the processing capability for doing real time VA it is critical that the individual components of the system are designed in an integrated manner. One solution would be for a single entity to design the entire structure from end to end. An alternative is to design the end to end system using standard methods, data structures and APIs through the entire process. The difficulty in that solution is getting a standard established in a timely fashion.

Despite these difficulties, the hybrid approach is the best bet to succeed because it will combine ubiquitous distributed computing using local devices (laptops, tablets, phones) with cloud distributed computing. The need for the cloud is the need to incorporate the benefits of big data into the mass of local computing using ever improving networking.

3. Event Detection

Event detection is the goal of many VA applications, but detailed metadata streams intrinsic to this process can be tapped for additional value. For surveillance tasks, accurately determining when a person enters a zone or leaves an object such as a backpack are typical design goals and have clear utility if they can be achieved with low false alarm rates. However, platforms designed for this can also be used to generate streams of metadata describing the visual content of the scene quantitatively. Analyzed in aggregate, this data can describe trends such as pedestrian or vehicular traffic patterns. Video camera data can be augmented with other sensor data and for some application areas, low-cost depth sensors can provide very robust object segmentation without relying on background removal algorithms.

One branch of current research is focused on reliably detecting events under more challenging conditions such as in public spaces with camera views that encompass a large number of people. The TRECVID Surveillance Event Detection (SED) [9] evaluates algorithms' performance for a range of events such as a person running, placing an object, or putting a cellphone to their ear. Top performing algorithms extract several types of motion features that are often computationally expensive and apply these in a learning framework for event detection [8]. The primary focus is detection accuracy; further work will be required to improve efficiency so that these novel approaches can be brought to widespread application.

Big data approaches to event detection will not only have to be top performing, they will have to be efficient, and of course respect privacy concerns. The definition of efficiency can be contextual. If processing is distributed locally then there can be various layers of event detection, from local in the camera, to in an aggregating device (like an NVR) and finally to the cloud. Each layer would process and pass up analytics results and meta-data. The higher layer could focus on the most interesting events and use computationally expensive but highly accurate algorithms.

4. Metadata Representation

Open standards for exchange of media and metadata from IP camera systems to Big Data analytics services and storage lakes are critical for the success of this ecosystem. For media, adoption of H.264 is well-established both for camera encoding and mobile device decoding but transcoding may be required to

match display device capabilities. Newer cameras support multiple simultaneous encoding streams to address this situation. H.265 HEVC and approaches targeted at surveillance video [10] will offer improved coding efficiencies and suggest that analytics preprocessing (background segmentation) can be performed jointly with coding. At the transport layer, RTSP is used for low latency applications while adaptive HTTP streaming such as HLS is typically used for off-line retrieval of recorded clips. MPEG DASH provides an open standard alternative, but has not yet seen wide adoption in mobile devices.

MPEG-7 provides a flexible framework for content description [11] and can be used to represent extracted metadata. While it can also encode intermediate-level image and video features, true interoperability at the feature level has not been embraced by vendors and systems implementers. Even at the higher level, application-specific profiles or alternative schema may provide developers with a lightweight option to metadata exchange. The industry forum, ONVIF™, has drafted a specification for VA services [12] that includes its own syntax for scene and object description, with spatial and temporal relations. While adoption of open standards in this area is at the early stages and specifications are still under development, the progress to date shows promise toward truly interoperable VA systems.

5. Conclusions

Given the intrinsic bitrates involved with representing high resolution sensor data, video promises to be the biggest of the big data in terms of its computing and networking demands. In order to take advantage of video big data, several elements must come together including advanced IP-connected cameras, distributed video processing architectures, and standards for representation of extracted meta-data of video events. Further work is needed in developing systems and algorithms that can be easily distributed and take advantage of a camera-to-cloud processing path. Combining aggregated video analytics with traditional structured and unstructured big data sources such as location and text based media, facilitated by exchange in agreed-upon metadata representations, with appropriate privacy protections will yield a platform capable of supporting a wide range of novel information services.

References

[1] C.S. Regazzoni, A. Cavallaro, Y. Wu, J. Konrad, A. Hampapur, "Video Analytics for Surveillance: Theory and Practice [From the Guest Editors]," *Signal Processing Magazine*, IEEE , vol.27, no.5, pp.16,17, Sept. 2010.

[2] N. Gagvani, Challenges in Video Analytics, in *Embedded Computer Vision* B. Kisacanin, et al, ed., pp 237-256, Springer, 2009.

[3] A. Leykin, *Visual human tracking and group activity analysis: A video mining system for retail marketing*. ProQuest, 2007.

[4] W.C. Kao, "High Dynamic Range Imaging by Fusing Multiple Raw Images and Tone Reproduction," *IEEE Transactions on Consumer Electronics*, vol.54, no.1, pp.10-15, February 2008.

[5] L. Marcenaro, G. Vernazza, C.S. Regazzoni, "Image stabilization algorithms for video-surveillance applications," in *Proceedings of Image Processing*, pp.349,352, 2001.

[6] Texas Instruments, TMS320DM6467T Digital Media System-on-Chip (Rev. C), Jul. 2012.

[7] P. Viola, M. Jones, "Robust real-time object detection," International Journal of Computer Vision 4 (2001): 34-47.

[8] X. Yang et al., "AT&T Research at TRECVID 2013: Surveillance Event Detection," in *Proceedings of TRECVID 2013*.

[9] P. Over et al., "TRECVID 2013 -- An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics," in *Proceedings of TRECVID 2013*.

[10] X. Zhang, T. Huang, Y. Tian, W. Gao, "Background-Modeling-Based Adaptive Prediction for Surveillance Video Coding," IEEE Transactions on Image Processing, vol.23, no.2, pp.769,784, Feb. 2014.

[11] ISO/IEC 15938-5:2003, Information technology – Multimedia content description interface – Part 5: Multimedia description schemes, 2003.

[12] ONVIFTM Video Analytics Service Specification, Version 2.1.1, Jan. 2012.

David Gibbon is Lead Member of Technical Staff in the Video and Multimedia Technologies and Services Research Department at AT&T Labs Research. His current research focus includes multimedia processing for automated metadata extraction with applications in media and entertainment services including video retrieval and content adaptation. In 2007, David received the AT&T Science and Technology Medal for outstanding technical leadership and innovation in the field of Video and Multimedia Processing and Digital Content Management and in 2001, the AT&T Sparks Award for Video Indexing Technology Commercialization. David has contributed to standards efforts through the Metadata Committee of the ATIS IPTV Interoperability Forum. He serves on the Editorial Board for the Journal of Multimedia Tools and Applications and is a senior member of the ACM

and the IEEE. He joined AT&T Bell Labs in 1985 and has and holds over 55 U.S. patents in areas such as multimedia indexing, streaming, and video analysis. He has written a book on video search, several book chapters and encyclopedia articles as well as numerous technical papers. David is an adjunct professor at Columbia University where he teaches a graduate level digital image processing course.

Lee Begeja is Principle Member of Technical Staff in the Video and Multimedia Technologies and Services Research Department at AT&T Labs Research. His current research focus includes distributed processing for visual analysis with applications in home security and surveillance. His most recent awards are from 2012 when Lee received the AT&T IP Bronze Award for generating shareholder value from the license of intellectual property and 2011 when he received the AT&T Social Innovation Award. Lee holds 69 patents in various areas including network optimization, audio enhancement and video analysis. He has written multiple book chapters and numerous technical papers.

A Survey on Personal Digital Photo Management

Ye Xing, Vivek Gupta, Tao Wu
Nokia, Boston, USA
{ye.xing, vivek.10.gupta, tao.a.wu}@nokia.com

1. Introduction

Traditional photo organizers provide software platforms allowing users to browse photos, make albums and annotate their photos. Users can retrieve or sort photos based on time, annotation or albums. Traditional photo organizers cannot fill the need of consumers when the number of personal digital photos increases dramatically with the ubiquitous use of digital cameras. Therefore, automatic photo organizers that help users manage, browse and share their photo collections with little or no input have been developed.

Based on a user study which analyzed users' usage pattern in sorting photos, the frequent features that consumers used to sort photos are found to be 'event', 'person' and 'time'[1]. To identify these three key features for sorting photos, content-based image analysis and use of metadata are the two main techniques to build automatic photo organizers.

2. Content-based image analysis

The aim of content-based image analysis is to extract objects or people in the photos, and then use the detected features to classify and categorize the photos.

The Photo Tourism system [2] has 3D views of major world sites, object-based photo browsing and transfers the annotations from similar images automatically. The basis of this system is to recognize objects across large photo collections by using key-point detection and matching algorithms. Photos containing the same objects can be grouped and the annotations can be transferred from the labeled one to the rest, thus users do not need to annotate every photo taken with the same background. The Photo Tourism system is built on worldwide major place of interests, so it does not identify users' common locations in their personal photo collections.

The main limitation of content-based image analysis is that the gap between semantic expression and the low-level features (color, texture, shape) is still wide [3]. Some efforts have been put to enhance content-based image analysis for automatic annotation. ESO Game [4], LabelMe [5] and autotag.me [6] encourage users to label images on the web, so the collected databases can facilitate the progress of content-based image analysis for annotation.

Although the content-based image analysis cannot be generally applied to personal photo collections, the use of content-based image analysis in face detection and face recognition can aid in tag recommendations, which is a very important part in identifying people in photo management. When face recognition algorithms are used in photo organizers photos from the user and their friends are scanned to recognize faces and then tags of people's name are suggested for the faces by matching them with users' profile photos or other previously tagged photos.

Recently, Facebook published DeepFace, which represents a big step forward in face recognition [7]. Most traditional face recognition algorithms have difficulties when the photo is not clear or does not have frontal views of the face. DeepFace uses deep learning to create 3D models of faces and can recognize faces from the side. The accuracy of recognizing faces, are comparable to that of human brain (97.25% vs. 97.53%). The model was built on four million face images belonging to more than 4000 identities, more than 120 million parameters was involved in the deep learning process. The computational time during the training is 3 days using GPU-based engine, but with the trained model, it only takes 0.33 seconds to classify an image on a standard CPU-based computer.

3. Use of Metadata from photos

Metadata is collected automatically by most smart phones and digital cameras when a photo is taken and can contain information about creation time, location, flash use, etc., which provides rich contextual information about the photos. This metadata has been shown to be effective and helpful in organizing photos automatically.

The use of timestamp has been long favored by researchers. The assumption of this type of work is photos belonging to one 'event' share similar temporal features. Cooper developed a general, unsupervised photo organizer by clustering the temporal features of the photos [8]. PhotoTOC clusters the photos based on time and color histogram, detecting the event boundaries using an adaptive local threshold method [9]. It can also automatically choose one representative image per cluster to better represent the content of each cluster.

In addition to time-based photo management tools algorithms using both timestamp and location data

have been widely developed. The organizer in [10] uses both location and timestamp to provide hierarchical clustering on personal photos. Hierarchical clustering provides a better organization scheme for collecting together many events that occur during a single long trip. PhotoCamps [11] not only uses location to cluster photos, it also uses location and timestamp to annotate photos with a place name and time to supply the full automatically-derived context, e.g. "Boston Common, Boston, July 4th, 2013". Nokia's Storyteller application is a Windows Phone app that automatically sorts photos and videos into 'stories' based on time and geo-location [12]. The organization of the photos can be viewed either from timeline or on a map. It can also download and sort photos from a variety of online services, which can help group photos taken from multiple cameras.

Going beyond timestamp and location, some researchers collect metadata from external resources to aid in photo organization. PhotoCampas was extended by including more metadata [13]: time of day and light status can provide annotation suggestions like day, dusk, night or dawn; weather status and temperature can provide additional annotation suggestions such as freezing, rainy and the access to personal calendar can detect the essence of an 'event' such as birthday/ wedding, which are unlikely to be detected by other algorithms.

The main limitation of using metadata for photo management comes from the fact that the automatic sorting process is usually an unsupervised clustering algorithm, which can mis-cluster some photos, especially for photos taken on the boundaries of an event. Providing a user –friendly UI to let users manually change the mis-clustered photos is the usual way to ease this problem. Also, validation of the usefulness of tags suggested by external sources (weather, calendar, time of day) needs to be further evaluated.

4. Combination of metadata and content-based image analysis

Some works combine the metadata and content-based image analysis into one photo organizer system, so that 'event', 'time' and 'location' can be identified by metadata, while 'person' can be identified by content-based analysis.

In [14], the photo organizer uses timestamp and color to hierarchically cluster the photos into different events, also it applies the torso detection algorithm [15] to group photos with the similar torsos together, so that users can apply bulk annotation.

VidCat provides a cloud service system that stores the photos based on time, detected face and face similarity, and detailed photo features [16]. VidCat's main features are the fast retrieval of related photos ability to detect near-duplicate images that are grouped to minimize redundant sets of photos.

5. Conclusions

The main trend in automatic photo management has been illustrated in this paper. The main goal for automatic photo management tools is to use algorithms to reliably detect 'event', 'person' and 'time' which are the key features in sorting photos. The algorithms based on content image analysis and metadata have their advantages in detecting person, time and event respectively. But further improvement on the accuracy of the detection should be one of the areas for future work. Some of the content-based image analysis algorithms need strong computational backend support. How to simplify and apply these computationally expensive algorithms to personal computers or mobile phones would be another challenging problem.

References

[1] K. Rodden and K. Wood, "How do people manage their digital photographs" ACM conference on human factors in computing systems, 2003

[2] N. Snavely, S. Seitz, R. Szeliski, "Photo Tourism: Exploring Photo Collections in 3D", ACM Transactions on Graphics, 25(3), 2006

[3] R.C. Veltkamp and M. Tanase, "Content-based image retrieval systems: A survey". Technical Report TR UU-CS-2000-34, Department of Computing Science, Utrecht University, October 2002.

[4] L. von Ahn and L. Dabbish, "Labeling Images with a Computer Game". ACM 2004

[5] B. Russell, A. Torralba, K. Murphy, W. Freeman, "LabelMe: A Database and Web-Based Tool for Image Annotation" Int J Comput Vis (2008) 77: 157–173

[6] autotag.me

[7] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, Lior Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification", CVPR, 2014

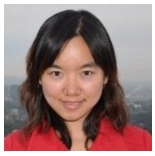
[8] M. Cooper, J. Foote, A. Girgensohn and L. Wilcox 'Temporal Event Clustering for Digital Photo Collections' ACM, 2003

[9] J. Platt, M. Czerwinski, B. Field, "PhotoTOC: Automatic Clustering for browsing personal photographs." Proc. Fourth IEEE Pacific Rim Conference on Multimedia, 2003

[10] A. Pigeau and M. Gelgon, "Building and tracking hierarchical geographical & temporal partitions for image collection management on mobile devices" MM'05

[11] M. Naaman, Y.J.Song, A.Paepcke, and H. Garcia-Molina, "Automatic organization for digital photographs with geographic coordinates" In Proceedings of the Fourth ACM/ IEEE-CS Joint Conference on Digital Libraries, 2004

[12] http://www.windowsphone.com/en-us/store/app/nokia-storyteller-beta/b0940143-e67e-4f74-8f68-16b7ad872dd2
[13] M. Naaman, S. Harada, Q. Wang, H. Garcia-Molina, A. Paepcke, "Context Data in Geo-refenced digital photo collections" ACM, 2004
[14] B. Suh, B. Bederson, "Semi-Automatic Image Annotation Using Event and Torso Identification".
[15] Lienhart, R and Maydt, J, "An extended Set of Haar-like Features for Rapid Object Detection." IEEE ICIP 2002, Vo;/ 1. Pp 900-903, Sep 2002
[16] L. Begeja, E. Zavesky, Z. Liu, D. Gibbon, R. Gopalan, B. Shahraray, "VidCat: An image and video analysis service for personal media management." SPIE 2013



Ye Xing received her B.S degree in Biomedical Engineering from Tsinghua University (Beijing, China) in 2003, M.S degree in Electrical Engineering in 2006 and PhD degree in Biomedical Engineering from University of Pennsylvania in 2010 respectively.
Currently, she is a Data Scientist at Nokia, Burlington, MA, where her work focuses on machine learning algorithms design and development for big data analytics. One of her first authored paper won the best paper runner up in MICCAI 2008.

Vivek Gupta received his B.S. Degree in Math/Computer Science from Carnegie Mellon University (Pittsburgh, PA, USA) in 1992, M.S. Degree in Computer Science from Indiana University (Bloomington, IN, USA) in 1994 and is working on his PhD in Computer Science at the University of Massachusetts (Lowell, MA, USA).
Currently, he is a Principal Engineer at Nokia, Burlington, MA, where his work focuses on data analytics and visualization.

Tao Wu is Chief Data Scientist at Nokia's Data & Analytics organization. Prior to his current position, he was responsible for Nokia's analytics platform architecture, and was a research affiliate at MIT Computer Science and Artificial Intelligence Laboratory. His current interests are large scale data analysis and data-driven production creation. Tao has more than 20 publications and more than 10 inventions in distributed systems and speech recognition. He received B.S. degree from Tsinghua University, M.S. degree from Rice University and Ph.D. degree from Boston University, all in electrical and computer engineering.

Dynamic Structure Preserving Map (DSPM) for Human Action Primitive Modeling

Qiao Cai and Hong Man
Stevens Institute of Technology, Hoboken NJ, USA
{qcai, hman}@stevens.edu

1. Introduction

Human action recognition has attracted much attention in the fields of computer vision and machine learning in recent years [1]. Many previous works have focused on augmenting the feature descriptions, such as proposing stronger feature sets and combining different features [2], or improving action recognition models, such as clustering and classification for scene analysis or abnormal events detection [3]. The analysis of human actions in a video sequence is challenging, because the recognition system is required to extract implicit properties including spatio-temporal coherence, behavior dynamics, and shape deformation. The action feature extraction from a video sequence is different from static image analysis, since spatio-temporal variation might result in meaningful behavior patterns. For example, the changes in human motion orientation or gesture during a specified time interval may indicate what actions might have occurred. In real world applications, irregular behaviors or environments should also be taken into account, which requires the dynamic model to adapt to unexpected factors.

2. Dynamic Structure Preserving Map

In this work we introduced a dynamic structure preserving map (DSPM) for action clustering and recognition, with emphasis on unsupervised clustering. DSPM is an extension to self organizing map (SOM) [4] in capturing spatial-temporal dependency in video sequences. DSPM has several unique properties.

1) DSPM is able to learn low-level features and produce a generative model to represent the dynamic topological structure. Instead of extracting carefully selected features, our method can automatically learn intrinsic characteristics from raw optical flow field for action recognition. Extending to the conventional SOM models, DSPM accumulates dynamic behavior of best-matching units (BMUs) to adjust their synaptic neuron weights, which can effectively capture the temporal information.

2) DSPM can aggregate the spatial-temporal clustering while simultaneously preserve underlying topological structure. Characterized by the parameters of latent neural distribution and neighborhood kernel function, the highly relevant spatial-temporal correlations for each action feature set are adaptively preserved in a 2-D lattice of neurons.

3) DSPM provides an effective way to reduce the dimensionality of input raw feature set, such as dense optical flow, to represent human motions in videos. Through the non-linear mapping procedure, DSPM can reduce the computational cost and data redundancy in action recognition.

The DSPM method contains a series of operations. Fields of optical flow are first calculated from consecutive video frames. Each field vector is mapped to one neuron in the DSPM according to competitive and adaptive learning rules. An adaptive neuron merging scheme is applied for cluster optimization. Based on the clusters on the spatial-temporal feature map defined in DSPM, the parameters of the latent space Markov model are estimated. The ensemble learning based on EM further enhances the dynamic model to yield better performance. The classifier with highest likelihood will be selected to predict class label. Normally there are significant amount of redundancy in action video sequences, especially at the beginning and the end of the sequence. A frame down sampling and simple motion based frame selection is applied in the pre-processing stage. The basic learning procedure of DSPM is illustrated in Figure 1.

3. DSPM for Action Recognition

It is a complex process to analyze the correlation and variation across space and time. There are limitations on the estimation of traditional state-space models, since the high dimensional parameters may lead to complex dependency structures. Based on the clusters on the spatio-temporal feature map defined in DSPM, the parameters of the latent space model are estimated. The ensemble learning based on EM further enhances the dynamic model to yield better performance. The classifier with highest likelihood will be selected to predict class label.

In SOM, the neighborhood function can be only used to preserve the spatial topology. Several extensions to SOM, including Temporal Kohonen map (TKM) and recurrent self-organizing map (RSOM) [5], have been proposed to adaptively model a data distribution over time on non-stationary input sequences. DSPM intends to capture the whole dynamic patterns within the data sequence. We improve the learning rule of DSPM. The input sequential samples, after some simple cleaning

operation, have the same importance and contribute evenly to the model from the beginning to the end. The resulting DSPM with the complete spatio-temporal information is then used in classification. In particular, the neuron transition probabilities in DSPM can describe the temporal dynamics from the training video sequences. DSPM models sequential dynamics by introducing Markov process to capture neuron transition probabilities between every two time samples. It is similar to Markov random walk [6] on graph, where at each step the walk jumps from one place to another based on specified probability distribution. The parameters of Markov process are used in neuron update and model classification.

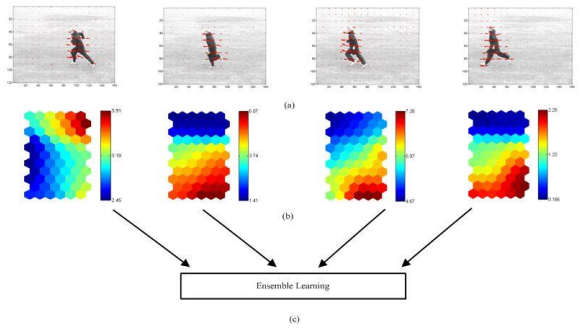


Figure 1. DSPM learning process. (a) Optical flow is extracted from each action video sequences. Given two consecutive frames, optical flow is computed at each pixel, and sampled with a 10×10 grid. For instance, the frame size of KTH data set is 160×120, after optical flow computing, the size of optical flow field for each frame is 16×12×2. The third dimension 2 indicates the magnitude and direction of optical flow. (b) Example DSPMs describing spatio-temporal patterns. The colors of grid represent the distances of various motions on DSPM. (c) The EM based ensemble learning is adopted to predict the action class.

We take the frame samples of ``bend" action from 9 persons in the Weizmann dataset in Figure 2. DSPM can extract the key feature information by spatio-temporal knowledge and statistically measure the dependency by Markov transition probability. The green color grid is the output of DSPM, which can aggregate the key features into the clustering. The x coordinate represents temporal feature in frame number and the y coordinate means the cost on distance between the input video frame and its best matching neuron in DSPM. The red marked circles represent the corresponding cluster in the DSPM. We can see that key features with sparse distribution have a high cost on distance.

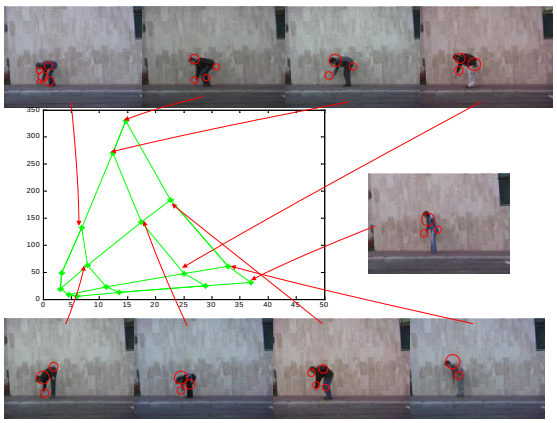


Figure 2. Spatio-temporal dependency analysis on key regions.

Figure 3 shows an example of the dynamic neuron trace in a Markov random walk.

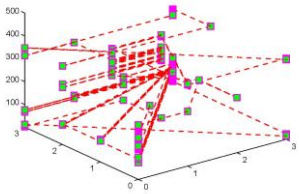


Figure 3. The neuron trace of a Markov random walk for action "run" in KTH dataset

4. Experimental Results

Table 1 shows the recognition results of many comparable approaches based on KTH, Weizmann and UCF dataset, respectively. On KTH dataset, Wu et al. [7] and Kovashka et al. [8] achieved the best performance with 94.5%. Our method can achieve 94.2% on average. On Weizmann dataset, Blank [9] and Fathi [10] achieved 100%, and our method achieved 98.7%. On the most challenging UCF dataset, Kovashka et al. [11] and Wu et al. [12] achieved 87.3% and 91.3%, respectively. Our method with 91.6% performs better than these methods. The performance of our method is comparable with these state of the art methods on these action datasets. Particularly, for more complex dataset, such as UCF sport dataset, our method can effectively improve the recognition performance. But more importantly our method can adaptively learn from low level features, such as optical flow, rather than using strong features. This improves model robustness, and requires less human intervention.

Table 1 Average accuracy on KTH, Weizmann, UCF, and YouTube datasets

Method	KTH	Weizmann	UCF	YouTube
Fathi *et al.*	90.5%	100%	-	-
Dollar *et al.*	81.2%	86.7%	-	-
Niebles *et al.*	81.5%	90.0%	-	-
Zhang *et al.*	91.3%	92.9%	-	-
Blank *et al.*	-	100%	-	-
JHuang *et al.*	91.7%	98.8%	-	-
Schuldt *et al.*	71.7%	-	-	-
Laptev *et al.*	91.8%	-	-	-
Klaser *et al.*	91.4%	84.3%	-	-
Campos *et al.*	91.5%	96.7%	80.0%	-
Wang *et al.*	89.0%	97.8%	83.3%	-
Wu *et al.*	94.5%	-	91.3%	-
Kovashka *et al.*	94.5%	-	87.3%	-
Liu *et al.*	93.8%	-	86.5%	71.2%
Le *et al.*	93.9%	-	86.5%	75.8%
Our method	94.2%	98.7%	91.6%	76.5%

4. Conclusions

In this paper, we proposed DSPM as an efficient spatio-temporal approach to recognize human actions from video sequences. Through learning on low level features, DSPM automatically extracts the implicit spatio-temporal patterns from the video sequence. DSPM improves learning rule based on dynamic information of BMU, which helps to preserve spatio-temporal dynamic topological structure. Through the non-linear mapping, DSPM can reduce computational cost and data redundancy for action recognition. The ensemble learning based on EM is adopted to estimate the latent parameters. In the future work, we will continue to improve DSPM to efficiently recognize more complex human actions from the real-world video datasets.

References

[1] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," IEEE Trans. Circuits and Systems for Video Technology, vol. 18, no. 11, pp. 1473–1488, 2008.

[2] H. Lee, R. Grosse, R. Ranganath, and A. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," ICML, 2009.

[3] T. Duong, H. Bui, D. Phung, and S. Venkatesh, "Activity recognition and abnormality detection with the switching hidden semi-markov model," CVPR, pp. 838–845, 2005.

[4] T. Kohonen, "Self-organizing maps," Springer, 2001.

[5] M. Varsta, J. Heikkonen, J. Lampinen, and J. Millan, "Temporal kohonen map and recurrent self-organizing map: analytical and experimental comparison," Neural Processing Letters, vol. 13, pp. 237–251, 2001.

[6] M. Szummer and T. Jaakkola, "Partially labeled classification with markov random walks," NIPS, 2001.

[7] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," CVPR, 2011.

[8] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," CVPR, 2010.

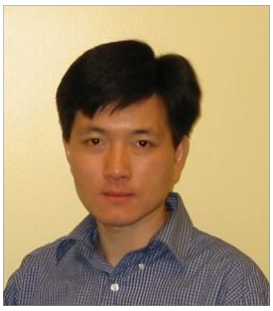
[9] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," CVPR, 2005.

[10] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," CVPR, 2008.

[11] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," CVPR, 2010.

[12] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," CVPR, 2011.

Qiao Cai received the B.S. degree in electrical engineering from Wuhan University, China, in 2005, the M.S. degree in electrical engineering from Huazhong University of Science and Technology, China, in 2008. In May 2013, he completed his PhD degree in the Department of Electrical and Computer Engineering, Stevens Institute of Technology. His current research interests include machine learning, data mining, computational intelligence, and pattern recognition. He is a student member of IEEE.



Hong Man received his Ph.D. degree in Electrical Engineering from Georgia Institute of Technology in December 1999. He joined Stevens Institute of Technology in January 2000. He is currently an associate professor in the Electrical and Computer Engineering Department. He is also the director of the Computer Engineering program, and the director of the Visual Information Environment Laboratory at Stevens. Over the past several years, he served on the Organizing Committees of IEEE WOCC, ICME, MMSP, and on the Technical Program Committees of IEEE Globecom, ICC, ICIP, ICME, VTC, WOCC and ACM SIGMAP. His research interests include image processing, pattern recognition, data mining, wireless communications, and data networking, on which he has published more than 120 technical journal and conference papers. He is a senior member of IEEE.

Multimedia Big Data: From Large-Scale Multimedia Information Retrieval To Active Media

Trista P. Chen
Cognitive Networks, U.S.A.
trista.chen@ieee.org

1. Introduction

The explosion of multimedia data (image, video, 3D, etc.) from mobile image captures, social sharing, the web, TV shows and movies, and the availability of large amount of metadata have created unprecedented opportunities and fundamental challenges to multimedia signal processing. Web users are uploading 100 video-hours to YouTube per minute. In a month, social media users are posting 9 billion photos to Facebook, 150 million photos using Instagram, and 50 million photos to Flickr [1]. Multimedia is big data, not just because there is a lot of it. Multimedia is big data because they are not just big in volume, but also unstructured, multi-modal (big in variety) and comes in fast (big in velocity).

The emergence of big data has brought about a paradigm shift to all fields of computing. Most big data systems currently in use are very restrictive in nature in that it is quite difficult to handle data types other than text or numbers. Moreover, text and numbers are handled using pre-defined fields only in the database. It is extremely difficult to store/retrieve multimedia data. This is attributed to the fact that although we have seen remarkable advances in computing power, storage capacity and network speed/reach, the underlying technology to retrieve multimedia data is not only hard, but often ill-defined.

This article provides an overview of recent trends of research in multimedia big data, including compact feature descriptors for large-scale multimedia information retrieval [2-4], powerful insights provided by simple machine learning methods with a lot of multimedia data [5-7], and active media applications, that were not possible before, enabled by large-scale automatic multimedia content recognition [8].

2. Compact feature descriptors

One of the key hinder factors of multimedia information retrieval is the semantic gap between existing low-level visual features and high-level semantic concepts. The big volume and big velocity characteristics of multimedia big data only make it more challenging. Efforts have been on reducing the size of the visual features and increasing the effectiveness of indexing [2][4]. Unlike traditional multimedia metadata that requires labor-intensive annotations, social media is readily to be harvested with massive amount of metadata to complement visual features [3].

3. Simple feature and simple model

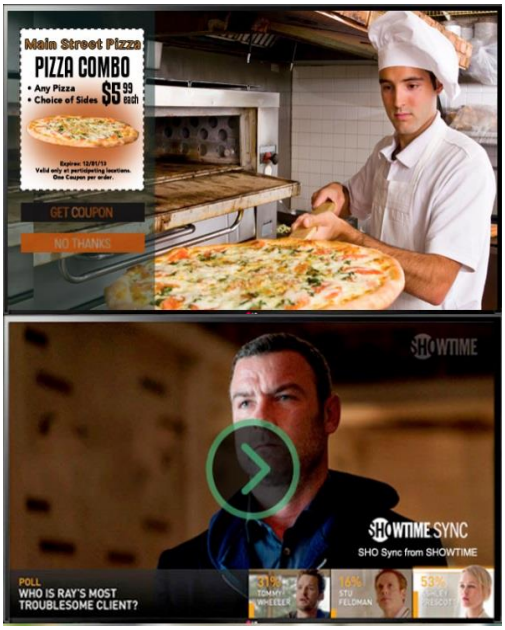
Given a small amount of data, the subspace of an object or a concept usually cannot be described by a simple model such as a linear model very well. However, when the amount of data is massive, a simple model not only wins in computing performance, but also closely approximates the optimal model [5]. For richly represented classes, such as people, the performance is comparable to leading class-specific detectors. An analogy to this phenomenon is Newton's method, where linear convergence is shown when the delta of data is small.

Similarly, the patch data, which is an array of RGB pixel values, without any sophisticated transformation, works well in a real-world multimedia big data matching system [6].

An ambitious project, NEIL – Never Ending Image Learner, exploits the visual knowledge gathered from the Internet to automatically build a large structured visual knowledge base which not only consists of labeled instances of scenes, objects, and attributes but also the relationships between them [7]. Joint discovery of relationships and labeling at a gigantic scale improve semi-supervised learning in NEIL. The end result is a computer program that runs 24 hours per day and 7 days per week to automatically extract visual knowledge from the Internet data.

4. Active media

While most of the multimedia big data research has been on information retrieval and discovery, a new class of applications has been born called active media. Active media was not possible before the era of multimedia big data. The volume and speed of multimedia has to be so huge that it is impossible for human labelers or traditional complicated models to work. With the new class of features and learning methods that are suitable for multimedia big data problems, new media services such as social TV, interactive marketing, voting and polling, program promotions are now in reality [8]. It is "active" because it does not require users to deliberately submit a query of the concerned multimedia content.



(a)

(b)

Figure (a) shows a interactive marketing app dynamically embedded to the TV content, which enables immediate product engagement. Figure (b) shows a voting and polling app that is invariant to delayed viewing. It strengthens the connection between real TV viewers and the content.

Other active media examples include automatic scene completion and crowd-sourced 3D scene reconstruction Photosynch.

5. Conclusions

Multimedia is increasingly becoming the biggest big data due to the three V's: volume, variety and velocity. Not only is it a challenging domain of research, but it enables new fields of applications. Insights we didn't have before with "small data" suddenly arise. Problems we didn't see before are becoming critical. Algorithms that were ill fitted before are now attractive. In this article, we briefly described current research directions in compact features and simple methods as well as provided examples of the new active media enabled by multimedia big data.

References

[1] Hunter Schwarz, "How Many Photos Have Been Taken Ever", BuzzFeed, September 24, 2012.

[2] David Chena, Sam Tsaia, Vijay Chandrasekhara, Gabriel Takacs, Ramakrishna Vedantham, Radek Grzeszczuk, Bernd Girod, "Residual enhanced visual vector as a compact signature for mobile visual search", Signal Processing, Volumn 93, Issue 8, August 2013, pp. 2316-2327.

[3] Damian Borth, Tao Chen, Rongrong Ji, Shih-Fu Chang, SentiBank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content", 2013 ACM international conference on Multimedia, New York, NY, U.S.A., pp. 459-460.

[4] Guangnan Ye, Dong Liu, Jun Wang, Shih-Fu Chang, Large-Scale Video Hashing via Structure Learning", International Conference on Computer Vision, Sydney, Australia, December 2013.

[5] Antonio Torralba, Rob Fergus and William T. Freeman, "80 million tiny images: a large dataset for non-parametric object and scene recognition", IEEE Transactions on Pattern Analysis and machine intelligence, November 2008, pp. 1958-1970.

[6] Zeev Neumeier, Edo Liberty, "Methods for identifying video segments and displaying contextual targeted content on a connected television", US patent 8595781.

[7] Xinlei Chen, Abhinav Shrivastava, Abhinav Gupta, "NEIL: Extracting Visual Knowledge from Web Data", International Conference on Computer Vision, Sydney, Australia, December 2013.

[8] Cognitive Networks, http://www.cognitivenetworks.com.

Trista P. Chen received the Ph.D. degree in Electrical and Computer Engineering from Carnegie Mellon University (CMU), Pittsburgh, PA, U.S.A., and M.S. and B.S. degrees in Electrical Engineering and Physics from National Tsing Hua University, Hsinchu, Taiwan, in 1997 and 1999 respectively. Currently, Trista is the Director of Core Application Engineering at Cognitive Networks, San Francisco, CA, where she focuses on real-time large-scale multimedia content recognition in the cloud and big data analytics. She is also a big data consultant in machine learning for big data. She was previously the founder of a computer vision startup that helped monetizing image and video assets, researcher at Gracenote and Intel, and architect of Nvidia's first video processor.

Her multidisciplinary research interests include multimedia big data, media content recognition, computer vision, augmented reality, processor architecture, and multimedia communications. She co-authored 20+ publications and 10 issued and pending patents, and gave keynote and invited lectures at IEEE conferences and universities.

Call For Papers
IEEE MultiMedia Magazine

http://www.computer.org/portal/web/computingnow/multimedia/

IEEE MultiMedia magazine was founded in 1994, and is the first IEEE publication in the multimedia area. IEEE MultiMedia serves the community of scholars, developers, practitioners and students who are interested in multiple media types, used harmoniously together, for creating new experiences. As such, it fills a lacuna that exists between several fields such as image processing, video processing, audio analysis, text retrieval and understanding, data mining and analysis and data fusion.

The information consists of articles, product reviews, new product descriptions, book reviews and announcements of conferences and workshops. Articles discuss research as well as advanced practice in multimedia hardware, software, systems and their applications, and span the range from theory to working systems.

Scope

We encourage our authors to write in a conversational style, presenting even technical material clearly and simply. You can use figures, tables, and sidebars to explain specific points, summarize results, define acronyms, guide readers to other sources, or highlight items. Assume an educated general audience, and you will successfully communicate your ideas to generalists and specialists alike.

IEEE MultiMedia serves both users and designers of multimedia hardware, software, and systems. Its readers work in industry, business, the arts, and universities. Some are generalists and some specialists in specific areas of multimedia.

We invite articles on multimedia systems and their applications, as well as those that present theories and/or relate practices. High quality technical papers that propose new concepts and are forward-looking are encouraged. We particularly welcome tutorials, surveys, overview papers, and special topics, such as:

- Big and broad multimedia
- Emotional and social signals in multimedia
- Interactive multimedia
- Media transport and delivery
- Multimedia and society
- Multimedia and the crowd
- Multimedia art, entertainment and culture
- Mobile multimedia
- Multimedia cloud computing
- Multimedia HCI and QoE
- Multimedia search and recommendation
- Multimedia security, privacy and forensics
- Multimedia systems and middleware
- Multimodal and cross-modal analysis and description
- Music, speech and audio processing in multimedia
- Pervasive multimedia
- Social media and collective online presence

For information about manuscript preparation and submission, please visit
http://www.computer.org/portal/web/peerreviewmagazines/multimedia

MMTC OFFICERS

CHAIR

Jianwei Huang
The Chinese University of Hong Kong
China

STEERING COMMITTEE CHAIR

Pascal Frossard
EPFL, Switzerland

VICE CHAIRS

Kai Yang
Bell Labs, Alcatel-Lucent
USA

Chonggang Wang
InterDigital Communications
USA

Yonggang Wen
Nanyang Technological University
Singapore

Luigi Atzori
University of Cagliari
Italy

SECRETARY

Liang Zhou
Nanjing University of Posts and Telecommunications
China

E-LETTER BOARD MEMBERS