Enabling Immersive Visual Communications through Distributed Video Coding

Giovanni Petrazzuoli, Marco Cagnazzo, Béatrice Pesquet-Popescu, Fréderic Dufaux
Institut Mines-Télécom/Télécom ParisTech – CNRS/LTCI
{petrazzu, cagnazzo, pesquet, dufaux}@telecom-paristech.fr

1. Introduction

The realism of video communication tools has been steadily increasing in the last years, thanks to the advances in several fields: efficient video compression techniques allow conveying high-resolution video over network connections; high-dynamic-range imaging promises to further improve the perceived quality of videos; higher resolution and frame-rate systems are under way. In spite of all these improvements, a further enhancement of the visual communication experience is expected in the next years. One of the most anticipated new features is the addition of depth representation to visual conferencing systems. The huge expectation related to 3D video [1] is testified by the standardization effort produced by the research community: the MPEG group is finalizing the normalization of the multi-view video plus depth (MVD) representation [2]. This format consists in sending a few views, each with an associated "depth map", i.e. a single-component image representing the distance of each pixel to the camera. This can be obtained by using a certain number of so-called range cameras. The MVD format enables the generation of high quality virtual views (i.e. views that did not exist in the original video) by Depth Image Based Rendering (DIBR) [3], and thus would make it possible to have new and exciting interactive services [4] such as Free Viewpoint Television (FVT) [5] where users have the impression of being present at a real event.

Obviously, MVD has a huge redundancy: not only in time, as ordinary mono-view video, but also among views (*inter-view* correlation) and between views and depth maps (*inter-component* correlation). All these kinds of redundancies have to be exploited in order to reduce the storage space on the server and the bandwidth used for transmission [6]. These requirements are equivalent in the context of non-interactive scenarios (like TV broadcasting): the entire video stored on the server will be sent to the user. An alternative, interesting paradigm of multiple-views video is the Interactive Multiview Video Streaming (IMVS) one [7][8][9]. IMVS enables the client to select interactively the views that he/she wants to display. Given this constraint, the server will send only the data needed to display the views according to the switch pattern decided by the user. However, the video is first encoded and stored in a server and afterwards it is sent to the clients. The user will choose a pattern of viewpoints which is not known at the encoding time, and that will possibly change from a user to another. Nevertheless, we would like to minimize both the storage space demanded by the compressed video and the bandwidth needed to interactively send the requested view to the user. These requirements are conflicting in the case of IMVS, which makes the problem challenging. The fact that the user trajectory of viewpoints is unknown at the encoding time makes it difficult to employ differential coding to reduce the transmission rate.

Distributed video coding (DVC) could be an effective solution for the IMVS problem, since inter-image correlation can be exploited even if the encoder ignores the actual reference image. However DVC of MVD video involves some new challenges. In the following we introduce our recent work in using DVC in the context of IMVS and of compression of depth maps for MVD.

2. Interactive Multiview Video Streaming

In order to optimize the bandwidth use for IMVS, the server should store a huge number of versions of the original video, given by all the possible switches among views (redundant P-frames). On the other hand, in order to reduce the storage space, one could just encode all the video pictures as *Intra* frames (I-frames). However this would result in an inefficient use of the channel bandwidth. A possible solution [7], proposed in the case of multiview video without depth, is to use Distributed Video Coding (DVC). In DVC [10][11], the correlation among images is not exploited by temporal prediction at the encoder. The available frames are used to produce an estimation of the current one (for example, by motion-compensated temporal interpolation). This estimation, also called *Side Information* (SI) is considered as a noisy version of the original image. Therefore, it is improved by sending some parity bits to the decoder. Of course, the parity bits do not depend on the SI: whatever the estimation of the current image, the correction bits are always the same. This means that the server just needs to store and send a set of parity bits for a given picture. In [7], the authors find an optimal combination of I-frames, Redundant P-frames and M-frames for the view switching, but they are interested in the case of multi-

view video coding without the depth information.
However, we remark that to the best of our knowledge, only a few papers deal with IMVS in Multiview Video plus Depth (MVD) with the constraint of ensuring that the video playback is not interrupted during switching.
In our work [8] we propose several methods for allowing arbitrary switches between two views in an MVD stream. We use the available depth-maps to produce DIBR estimation of a given image in the target view; then this estimation may or may not be corrected by parity bits. According to whether the DIBR is performed immediately (i.e. in the time instant of the view switch) or on the previous image (in advance), we have four switch strategies, in addition to the case of no-DIBR and the one where the first image of the GOP is estimated with DIBR. The best strategy depends on the position of the switching instant with respect to the GOP structure of the target view. For example, if the switch instant corresponds to an Intra image, there is no need for DIBR, while if it is in the middle of the GOP the advance DIBR + parity bits will give the best performances.
In summary, we have found the best strategy according to the temporal position of the switch. According to our study, the best single strategy is advance DIBR + parity bits: it allows an average bit-rate saving of 3.4% (Bjontegaard delta rate, [12]) with respect to the simple no-DIBR technique. Further rate savings can be obtained if the strategy is selected adaptively with respect to the temporal instant [8].

3. Wyner-Ziv Depth Map Coding

When compressing multiple video plus depth streams, one should always try to exploit not only spatial, temporal, and inter-view correlation, but also the correlation existing between textures and depth maps.
This observation is still valid when one considers *distributed* coding of MVD. In our work [9] we consider the distributed compression of depth maps. We develop several techniques for compressing depths and exploiting inter-component correlation. More precisely, we proposed new techniques for producing the side information of the current depth image. The reference scheme is the Discover SI generation method [11], used directly on the depth images. Discover performs a temporal interpolation assuming straight trajectories for object motion. A first improvement to this scheme is to consider high-order interpolation motion estimation (HOMI, [13]) for object trajectories. HOMI needs an initialization for object trajectory, and usually the linear motion of Discover is a good initialization. However, the depth images are not well suited for the motion estimation process, since they are very flat. Therefore better results can be obtained by using the motion trajectory computed on the corresponding texture images: we exploit the inter-component correlation as the movement of an object is pretty much the same in the two domains. However, since sometimes depth motion can differ from texture motion, we propose a further improvement: the trajectories are initialized with Discover on the texture, but then they are refined using HOMI on the depth data. This scheme is shown in Figure 1.

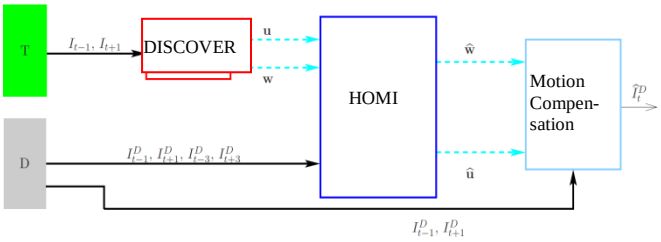

Figure 1. The proposed scheme depth SI generation

The proposed technique allows gaining up to more the 11% (Bjontegaard delta rate) with respect to the simple Discover SI generation of depth maps.

4. Conclusion

Immersive visual communications are one of most awaited technical innovations of the next few years. Several challenging problems must be solved before efficient solutions can be proposed. This study shows how DVC-inspired techniques can help in the design of effective systems for 3D video services.

References

[1] F. Dufaux, B. Pesquet-Popescu, M Cagnazzo (eds.): Emerging Technologies for 3D Video. Wiley, 2013
[2] Call for Proposals on 3D Video Coding Technology, ISO/IEC JTC1/SC29/WG11 Doc. N12036, Geneva, CH, March 2011.
[3] C. Fehn, "A 3D-TV Approach Using Depth-Image-Based Rendering (DIBR)," in *Proc. of 3rd IASTED Conference on Visualization, Imaging, and Image Processing*, Benalmadena, Spain, Sept. 2003, pp. 482–487
[4] C. Fehn, P. Kauff, O. Schreer, and R. Schäfer, "Interactive virtual view video for immersive TV applications," in *Proc. of Intern. Broadcast Conf*, 2001, pp. 14–18.
[5] M. Tanimoto, M.P. Tehrani, T. Fujii, and T. Yendo, "Free-viewpoint TV," *Signal Processing Magazine, IEEE*, vol. 28, no. 1, pp. 67 –76, 2011.
[6] M. Flierl and B. Girod, "Multiview video compression," Signal Processing Magazine, IEEE, vol. 24, no. 6, pp. 66–76, Nov. 2007.
[7] G. Cheung, A. Ortega, and N. M. Cheung, "Interactive streaming of stored multiview video using redundant frame structures," Image Proc., IEEE Trans., 2010.
[8] G. Petrazzuoli, M. Cagnazzo, F. Dufaux, B. Pesquet-Popescu. "Using distributed source coding and depth image based rendering to improve interactive multiview video access". In *IEEE International Conference on Image Processing*, vol. 1, pp. 605-608, September 2011. Bruxelles, Belgium.

[9] G. Petrazzuoli, M. Cagnazzo, F. Dufaux, B. Pesquet-Popescu. "Wyner-Ziv coding for depth maps in multiview video-plus-depth". In *IEEE International Conference on Image Processing*, vol. 1, pp. 1857-1860, September 2011. Bruxelles, Belgium.

[10] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," Proc. IEEE, vol. 93, no. 1, pp. 71–83, Jan. 2005.

[11] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M. Ouaret, "The DISCOVER codec: Architecture, techniques and evaluation, picture coding symposium," in Coding of Audio-Visual Objects, Part 10: Advanced Video Coding, 1st Edition, 2007

[12] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," in VCEG Meeting, Austin, USA, Apr. 2001.

[13] G. Petrazzuoli, M. Cagnazzo, and B. Pesquet-Popescu, "High order motion interpolation for side information improvement in DVC," in *IEEE Int. Conf. on Acoustics Speech and Signal Processing*, 2010, pp. 2342 –2345.

Giovanni Petrazzuoli obtained the PhD degree from Telecom-ParisTech in January 2013, with a thesis on distributed video coding for multi-view and multi-view plus depth video. His research interests also cover interactive streaming and shape adaptive coding.

Marco Cagnazzo obtained the Ph.D. degree from Federico II University and the University of Nice-Sophia Antipolis, Nice, France in 2005. Since 2008 he has been Associate Professor at Telecom ParisTech within the Multimedia team. His research interests are scalable, robust, and distributed video coding, 3D and multi-view video coding, multiple description coding, and video delivery over MANETs. He is the author of more than 70 journal articles, conference papers and book chapters, and a co-editor of the book "Emerging Technologies for 3D Video: Creation, Coding, Transmission, and Rendering", Wiley Eds., (to appear in 2013).
Dr. Cagnazzo is an Area Editor for Elsevier Signal Processing: Image Communication and Elsevier Signal Processing. He is an IEEE Senior Member, a Signal Processing Society member and an EURASIP member.

Beatrice Pesquet-Popescu received the Ph.D. thesis from the École Normale Supérieure de Cachan in 1998. Since Oct. 2000 she is with Télécom ParisTech, first as an Associate Professor, and since 2007 as a Full Professor, Head of the Multimedia Group. She is also the Scientific Director of the UBIMEDIA common research laboratory between Alcatel-Lucent Bell Labs and Institut Mines Télécom. She serves as an Editorial Team member for IEEE Signal Processing Magazine, and as an Associate Editor for IEEE Trans.CSVT, IEEE Trans Multimedia, IEEE Trans. Image Processing and for Elsevier Signal Processing: Image Communication. She holds 23 patents in video coding and has authored more than 270 book chapters, journal and conference papers in the field. She is a co-editor of the book to appear "Emerging Technologies for 3D Video: Creation, Coding, Transmission, and Rendering", Wiley Eds., 2013. Beatrice Pesquet-Popescu is an IEEE Fellow.

Frédéric Dufaux is a CNRS Research Director at Telecom ParisTech. He is also Editor-in-Chief of Signal Processing: Image Communication. He received his M.Sc. in physics and Ph.D. in electrical engineering from EPFL in 1990 and 1994 respectively. Frédéric has over 20 years of experience in research, previously holding positions at EPFL, Emitall Surveillance, Genimedia, Compaq, Digital Equipment, MIT, and Bell Labs. He has been involved in the standardization of digital video and imaging technologies, participating both in the MPEG and JPEG committees. He is currently co-chairman of JPEG 2000 over wireless (JPWL) and co-chairman of JPSearch. He is the recipient of two ISO awards for these contributions. Frédéric is an elected member of the IEEE Image, Video, and Multidimensional Signal Processing (IVMSP) and Multimedia Signal Processing (MMSP) Technical Committees.
His research interests include image and video coding, distributed video coding, 3D video, high dynamic range imaging, visual quality assessment, video surveillance, privacy protection, image and video analysis, multimedia content search and retrieval, and video transmission over wireless network. He is the author or co-author of more than 100 research publications and holds 17 patents issued or pending. He is a co-editor of the book to appear "Emerging Technologies for 3D Video: Creation, Coding, Transmission, and Rendering", Wiley Eds., 2013.