A Distributed Video Coding System for Multi View Video Plus Depth

Giovanni Petrazzuoli[1], Thomas Maugey[2], Marco Cagnazzo[1], Béatrice Pesquet-Popescu[1]

[1] *TELECOM-ParisTech, TSI department, Paris, France*

[2] *Ecole Polytechnique Federale de Lausanne (EPFL), Laboratoire de Traitement du Signal 4 (LTS4), Lausanne, Switzerland*

{petrazzu,cagnazzo,pesquet}@telecom-paristech.fr, thomas.maugey@epfl.ch

***Abstract*—Multi-view video plus depth (MVD) is gathering huge attention, as witnessed by the recent standardization activity, since its rich information about the geometry of the scene allows high-quality synthesis of virtual viewpoints. Distributed video coding of such kind of content is a challenging problem whose solution could enable new services as interactive multi-view streaming. In this work we propose to exploit the geometrical information of the MVD format in order to estimate inter-view occlusions without communication among cameras. Experimental results show a bit rate reduction up to 77% for low bit rate w.r.t. state-of-the-art architectures.**

***Index Terms*—multiview video plus depth, distributed video coding, DIBR**

I. Introduction

In the last years, interest for multiview video systems is growing up. Indeed, the users want to have the impression to being present at a real event. Then, the third dimension (the depth) is very important for the users in order to achieve immersive communication. In human sight, two eyes perceive a 3D scene, but indeed they pick up only two 2D images ; afterwards our brain processes them and deduces a 3D model. In a similar manner, in 3D-TV two or more cameras (arranged in different spatial positions) capture the same scene. The most flexible video format for multiview systems is the so-called Multiview Video plus Depth (MVD) [1]. In MVD, a number N of views is captured, along with the associated depth maps, which give, for each pixel, its distance to the camera. With N views and N depth maps, it is possible to synthesize $M > N$ new images associated to virtual/real viewpoints (i.e. viewpoints without an associated camera) by the depth-image based rendering (DIBR) algorithm [2]. The MVD format allows such services as multiview video or free viewpoints TV without the requirement of an overwhelming number of views. These considerations motivate the increasing research activity around MVD compression, which culminates in the 3D-VC standardization effort by MPEG. One of the disadvantages of these systems is that they require a huge complexity at the encoder side, and an inter-camera communication in order to extract the inter-view correlation. Distributed source coding [3] can solve these two problems. Indeed, for the Wyner-Ziv theorem [4], even if two correlated sources are separately encoded but decoded jointly, we can attain the same performance of joint coding. In the context of multiview video this two correlated sources are two adjacent views. In particular, if distributed source coding is applied to the MVD format, the information about the correlation between the views is intrinsic in depth map. In this paper, we propose a new depth-based distributed multiview video systems: one view with its depth is sufficient to reconstruct the other ones, except for the occluded areas. Then, our goal is to sent only one view in INTRA mode and to send the occluded areas of the other ones, without communication between the cameras. This system is very similar to LDV architecture proposed for MVD [5], with the advantage that our system does not need communication between the cameras. We obtain significant gain w.r.t. state-of-art multiview distributed video coding (MDVC) architectures [6]. Indeed, sometimes MDVC fails when the baseline is very large and when the cameras are not perfectly aligned. This paper is structured as follows: in the Section 2, a state of the art of related works about distributed source coding for MVD is provided. In Section 3, our new architecture is described and experimental results are illustrated in Section 4. Finally, in Section 5 we draw conclusions and future works.

II. Related works

In monoview distributed video coding the temporal stream is split into Key Frames (KFs) and Wyner-Ziv Frames (WZFs). The KFs are sent to the decoder. The WZFs are fed into a channel coder and only the parity bits are sent. An estimation of the WZF (called SI) is produced at the decoder side and, then, iteratively corrected by parity bits. This estimation is usually produced by temporal interpolation. The different tools are specified into the European project DISCOVER [6].

Multiview distributed video coding can be considered as a natural extension of the monoview distributed video coding. Encoding separately the different views avoids any sort of inter-camera communication and decoding them jointly ensures that the inter-view correlation is anyway exploited. However, the SI produced at the decoder side by inter-view interpolation as in DISCOVER codec is very poor, in particular for large baselines or for not aligned cameras. The

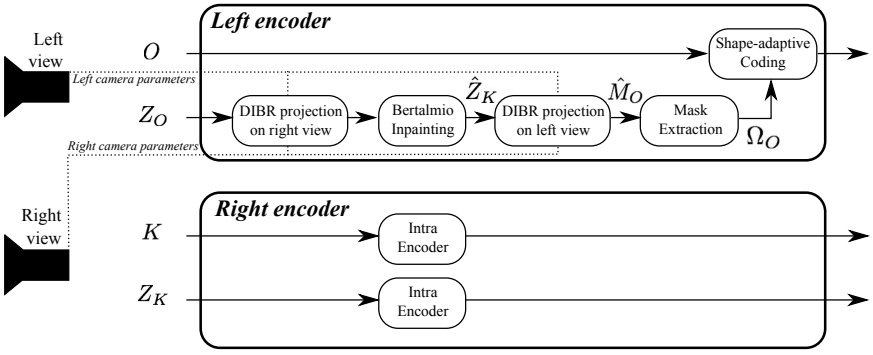

Figure 1: Structure of the encoder: the right camera is INTRA encoded. For the left camera the occlusion mask is extracted by a double DIBR run on the depth map. Only the regions selected by this mask are encoded and sent to the decoder.

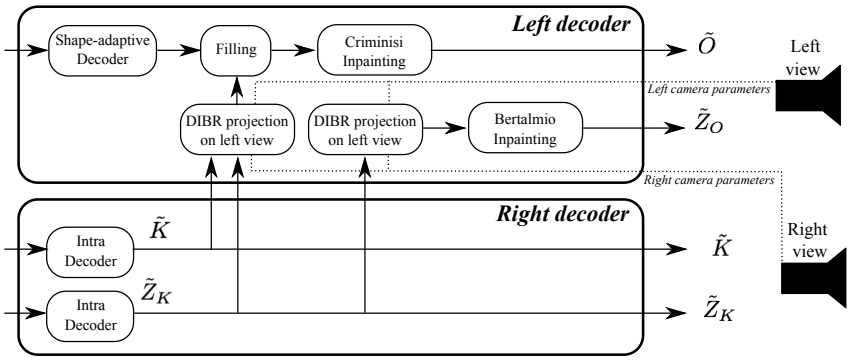

Figure 2: Structure of the decoder: the right texture and depth can reconstruct the un-occluded regions of the left view. On the other hand, the occluded regions are filled by the objects sent by the encoder.

consequence is that a great amount of parity bits are required to the turbo encoder for correcting the SI. Several solutions have been proposed in literature in order to improve the quality of inter-view interpolated image, by dense disparity algorithm [7] or by fusion technique [8]. Rarely, DVC was applied to the format multiview video plus depth. Depth map information can be used for generating any view point, by DIBR algorithm [2]. At the best of our knowledge, this information has been rarely used for the side information generation in DVC applied to multiview video plus depth. Some works [9–11] about DVC applied to multiview video plus depth have only highlighted the advantage of sharing the motion information between texture and depth. In a previous work [12], DVC is applied to MVD in order to assure continuity of the playback in the context of the interactive multiview video streaming paradigm. Depth map can be used for generating the target view when a view switching occurs, in order to have an estimation of the WZF. Indeed, when a commutation of view occurs, temporal interpolation cannot be performed, because the two reference frames are not both available at the decoder side. Unfortunately DISCOVER approach suffers that the statistical properties of the inter-view extrapolation by DIBR are different from the ones of the temporal interpolation. Indeed, one of the hypothesis for the implementation of the channel coder of DISCOVER is that the probability density function of the difference between the generated SI and the real WZF is laplacian. Moreover the statistics are estimated on each frame, by supposing that this error is ergodic. If these two hypothesis are not verified, the rate allocation of the channel coder is sub-optimal. When the estimation of the WZF is produced by DIBR from a reference view, the error is not laplacian and, moreover, no stationary. Then, the DISCOVER channel coder is not suitable for SI generated by DIBR.

III. The Proposed Distributed Video Coding System

In this paper, we propose a new distributed architecture for MVD. The peculiarity of this architecture w.r.t. other DSC-based architecture is that no channel coding is implemented, as in [13]. Moreover, we do not need a feedback channel. This modification allows us to remove any hypothesis about the statistical property of the error between synthesized and real frame.

If we consider a stereo system (plus depth), the right [resp. left] view along with the corresponding depth is sufficient for reconstructing the left [resp. right] view except for the occluded regions. Layered Depth Video architectures [14] have already proposed to send for the other views only occluded regions. Our goal is to generate the different layers, that consists of occluded regions, without communication between

the cameras.

The proposed architecture is depicted in Fig. 1 and 2. The video stream is split into KFs and O-Frames (OFs). The latter are synthesized from KFs via the depth map by DIBR algorithm and only occluded areas are sent. Without loss of generality, we suppose for each instant in the right view we have KFs, and that for the left view we have only OFs.

Let K and Z_K be the texture and the corresponding depth map for the K view. Likewise, O and Z_O are the texture and the corresponding depth map for the O view. At the decoder side the decoded texture and depth for the K view, K and Z_K, synthesize the corresponding OF except for the occluded areas. Then, our goal is to find the occlusion areas of OF in order to send exclusively these regions, without communications between the cameras. We propose to apply a double DIBR to the depth map Z_O in order to find the occlusion mask M_O for the OF (*i.e.* the points that are visible in the OF but not in the KF). DIBR is applied to Z_O for estimating D_K. In this estimation, some holes occur due to the occlusion. They can be filled by Bertalmio inpainting, since we deal with depth maps and anisotropic diffusion is particularly suitable for the smooth characteristics of depth images [15]. DIBR is then run again and we obtain the estimate of Z_O along with the estimation of the occluded regions for the OF, *i.e.*the regions that cannot be synthesized from the K view. Then, we select the occluded regions of the O view that has to be sent. Since they are sparse data, we encode them by a shape adaptive coding technique [16, 17]. The bit rate per pixel for the occluded regions has been chosen as the same of the KF.

At the decoder side, depicted in Fig. 2, the OF is synthesized from the KF and the occluded areas are filled by the object sent by the encoder. Since the estimated occluded mask at the encoder side can be different from the one estimated at the decoder side, some holes can still occur. They can be filled by Criminisi inpainting [15].

IV. Experimental results

We have tested our algorithm with four MVD sequences at different spatial resolutions. Two of these sequences (*mobile* and *GTFly*) are synthetic and the depth data are perfect, while other ones (*balloons* and *poznan street*) are natural and depth maps have been obtained by dense disparity algorithm [18]. Then, the quality of the synthesized view from these depth maps suffers of annoying artefacts. These errors sometimes affects the performance of the whole system.

We have run our algorithm on multiview systems with three cameras. The central camera consists of KFs and the two lateral ones of OFs (as in Fig. 3). The results are compared with *All INTRA* and DISCOVER in terms of Rate-PSNR performance and in terms of Rate-SSIM performance. The *All INTRA* configuration means that all the three cameras are coded INTRA. The DISCOVER camera configuration is depicted in Fig. 3(b): the two lateral cameras are Key and the central one is Wyner-Ziv. This means that an estimation of the central camera is produced by interpolation of the two

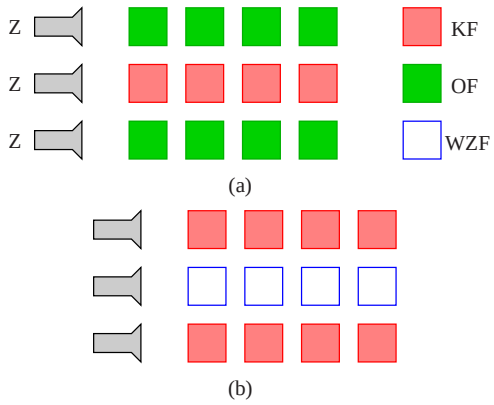

Figure 3: Two DVC camera configurations: OKO (a) and DISCOVER (b)

adjacent Key Frames. This estimation is the SI, that will be corrected by the channel coder.

We have introduced also SSIM metric because it seems more consistent with human sight: a small misplacement of the pixels (due to DIBR) will decrease significantly the PSNR, while the SSIM is nearly constant. As an example, in Fig. 4 we show a couple of images for the sequence *mobile*: at right we have the KF and at left one of the two OFs. The latter is synthesized from the K one and we can remark that between the two images we have a PSNR loss of 8.26 dB, that translate in a loss of 0.004 in terms of SSIM. Also by a subjective analysis of the two images, we are not able to see the a loss of 7.36 dB in PSNR from the right image to the left one.

In Fig. 5 and 6, we show the Rate Distortion curves for two sequences: *mobile* and *balloons*. The first one is a synthetic sequence and the depth data are perfect. The sequence *balloons* is natural and depth maps are computed by dense disparity algorithm and they can suffer from errors. As a consequence, the synthesized frames from that data suffers of annoying artefacts. Our technique is always better than All INTRA for the sequence mobile both in terms of PSNR and SSIM. On the other hand, for the *balloons* sequence, we can remark that even it seems that no gain in PSNR is obtained with our technique, a significant gain in SSIM is obtained in particular for low bit rate. For both sequences we can remark that the we have a PSNR and SSIM improvement w.r.t. DISCOVER. Since for several sequences we have a different behavior for low and high bit rate, we have computed the Bjontegaard metric [19] by separating these two scenarios. The results are in Tab. I and II. For *balloons* sequence, we remark that no gain w.r.t. DISCOVER is achieved for high bit rate (14.32% of bit increase), while for low bit rate we have a bit reduction of 30.01%. In general for synthetic sequences, we obtain always a gain w.r.t. DISCOVER and All INTRA both for low bit rate and for high bit rate. On the other hand, for natural sequences we obtain a gain in PSNR only for low bit rate. We have also computed the gain in PSNR and the corresponding gain in SSIM for a fixed bit rate per pixel w.r.t. DISCOVER. In Tab. III, we show those gains. We can remark

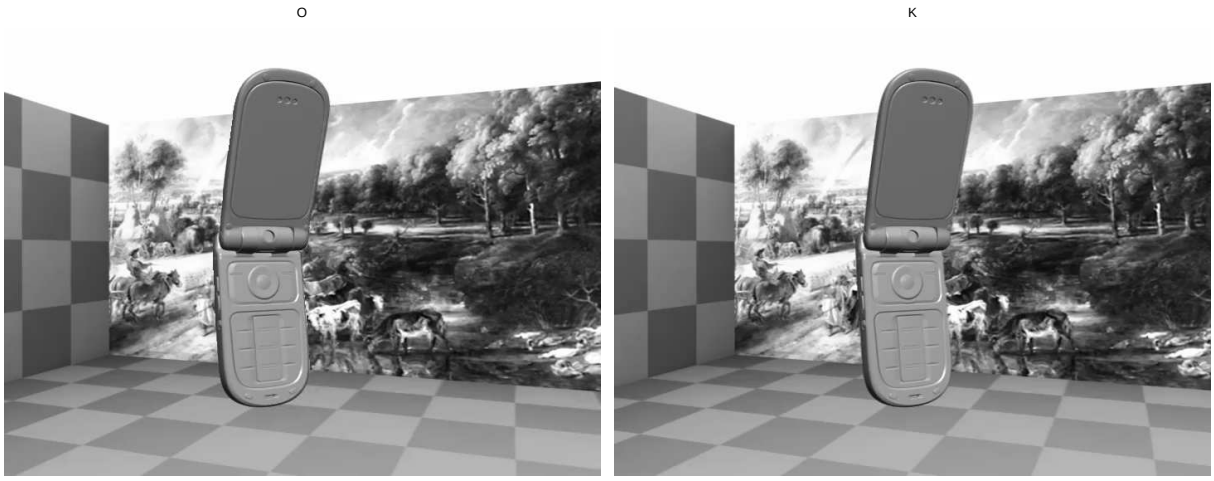

Figure 4: Synthesized left view (OF) - PSNR = 36.70 dB - SSIM = 0.985 and INTRA coded right view (KF) - PSNR = 44.06 dB - SSIM = 0.989

	w.r.t. DISCOVER		w.r.t. All INTRA	
	Δ_R [%]	Δ_{PSNR} [dB]	Δ_R [%]	Δ_{PSNR} [dB]
mobile	-51.96	3.18	-55.32	3.29
GTFly	-40.11	1.62	-26.32	0.95
balloons	14.32	-0.98	43.69	-2.45
poznan street	-1.02	0.22	20.41	-0.78

Table I: Bjontegaard metric for comparison of our technique w.r.t. DISCOVER and All INTRA for high bit rate. Distortion is measured in terms of PSNR

	w.r.t. DISCOVER		w.r.t. All INTRA	
	Δ_R [%]	Δ_{PSNR} [dB]	Δ_R [%]	Δ_{PSNR} [dB]
mobile	-66.08	3.77	-70.81	3.82
GTFly	-77.42	3.35	-55.19	1.89
balloons	-30.01	1.40	3.33	-0.34
poznan street	-32.34	1.37	-14.53	0.72

Table II: Bjontegaard metric for comparison of of our technique w.r.t. DISCOVER and All INTRA for low bit rate. Distortion is measured in terms of PSNR

that even if for some sequence we have a PSNR loss, we obtain a SSIM gain, particularly for high bit rate. For example, for the *poznan street* sequence for a bit rate of 0.2 bpp, we have a PSNR loss of 0.60 dB, but a SSIM gain of 0.02. Moreover, at the decoder side the complexity of our algorithm is much more low than the one of DISCOVER. The complexity of DIBR projection is negligible w.r.t. DISCOVER interpolation, that needs a full search algorithm of block matching. In our architecture, we also suppress the iterative channel coding as well as the feedback channel: then, the decoded does not need to send parity bits requests to the encoder. As a consequence, the complexity is further reduced

V. Conclusion and future works

In this paper, we have proposed a new distributed architecture for the MVD format. One view along with its depth is sufficient for reconstructing any viewpoints, except

	R = 0.1 bpp		R = 0.2 bpp	
	Δ_{PSNR} [dB]	Δ_{SSIM} .	Δ_{PSNR} [dB]	Δ_{SSIM}
mobile	3.80	0.10	3.43	0.06
GTFly	1.96	0.02	0.76	0.02
balloons	1.97	0.04	-1.52	0.01
poznan street	1.45	0.05	-0.60	0.02

Table III: Gain in PSNR Δ_{PSNR} and in SSIM Δ_{SSIM} for two fixed bit rate per pixel w.r.t. DISCOVER

for the occluded areas. Then, we have proposed to identify the occluded areas for each camera via a double DIBR and to send only those ones. These regions are coded by shape adaptive techniques. We have obtained significant gain w.r.t DISCOVER multiview codec, in particular for low bit rate and for synthetic sequences, where the depth data are perfect. With our method, we are able to have a bit rate reduction up to 77.42% for low bit rate w.r.t. DISCOVER. We have also remarked that when images are synthesized by DIBR, SSIM seems to be more consistent with human sight. As future works, we can perform a more efficient rate allocation between OFs and KFs, by taking into account that the quality of KFs significantly affects the whole performance.

References

[1] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Proceed. of IEEE Intern. Conf. Image Proc.*, San Antonio, TX, 2007.

[2] C. Fehn, "A 3D-TV Approach Using Depth-Image-Based Rendering (DIBR)," in *Proceedings of 3rd IASTED Conference on Visualization, Imaging, and Image Processing*, Benalmádena, Spain, Sep. 2003, pp. 482–487.

[3] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. 19, pp. 471–480, Jul. 1973.

[4] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the receiver," *IEEE Trans. Inform. Theory*, vol. 22, pp. 1–11, Jan. 1976.

[5] K. Muller, A. Smolic, K. Dix, P. Kauff, and T. Wiegand, "Reliability-based generation and view synthesis in layered depth video," in *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*, oct. 2008, pp. 34 –39.

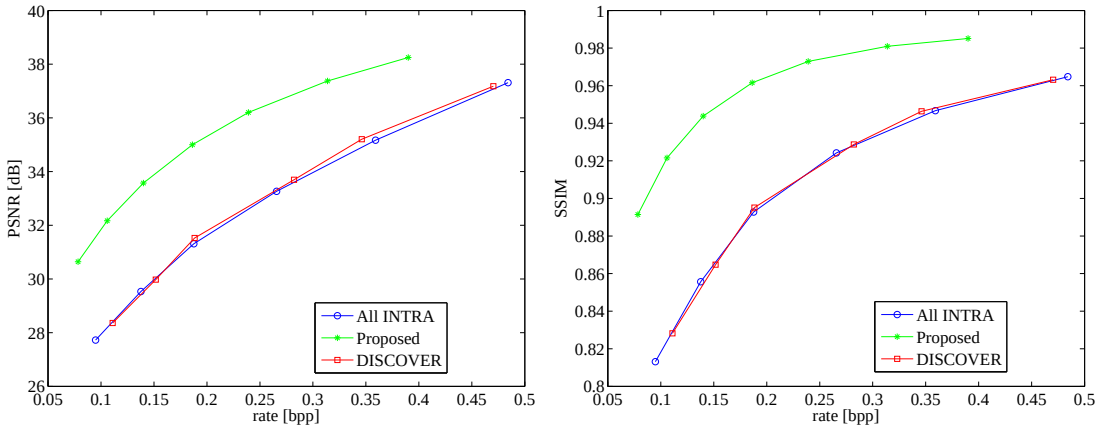

Figure 5: Rate-PSNR and Rate-SSIM curves for the sequence mobile

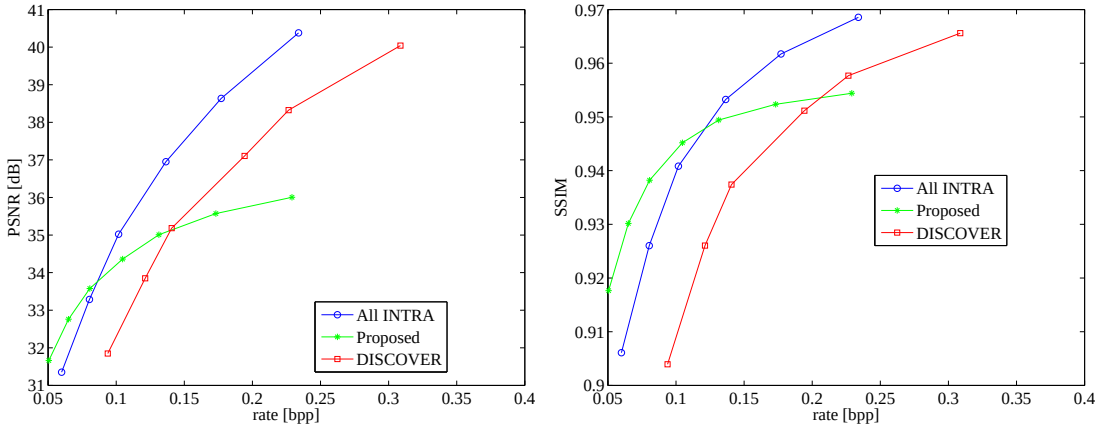

Figure 6: Rate-PSNR and Rate-SSIM curves for the sequence balloons

[6] C. Guillemot, F. Pereira, L. Torres, T. Ebrahimi, R. Leonardi, and J. Ostermann, "Distributed monoview and multiview video coding: Basics, problems and recent advances," *IEEE Signal Processing Mag.*, pp. 67–76, Sep. 2007.

[7] W. Miled, T. Maugey, M. Cagnazzo, and B. Pesquet-Popescu, "Image interpolation with dense disparity estimation in multiview distributed video coding," in *International Conference on Distributed Smart Cameras*, Como, Italy, 2009.

[8] T. Maugey, W. Miled, M. Cagnazzo, and B. Pesquet-Popescu, "Fusion schemes for multiview distributed video coding," in *Proceed. of Europ. Sign. Proc. Conf.*, Glasgow, Scotland, 2009.

[9] G. Petrazzuoli, M. Cagnazzo, F. Dufaux, and B. Pesquet-Popescu, "Wyner-ziv coding for depth maps in multiview video-plus-depth," in *Proceed. of IEEE Intern. Conf. Image Proc.* Bruxelles, Belgium: IEEE, 2011, pp. 1817–1820.

[10] M. Salmistraro, L. L. Rakêt, M. Zamarin, A. Ukhanova, and S. Forchhammer, "Texture side information generation for distributed coding of video-plus-depth," in *IEEE International Conference on Image Processing, ICIP*, Melbourne, Australia, 2013.

[11] M. Salmistraro, M. Zamarin, and S. Forchhammer, "Wyner-Ziv coding of depth maps exploiting color motion information," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2013.

[12] G. Petrazzuoli, M. Cagnazzo, F. Dufaux, and B. Pesquet-Popescu, "Using distributed source coding and depth image based rendering to improve interactive multiview video access," in *Proceed. of IEEE Intern. Conf. Image Proc.*, vol. 1, Bruxelles, Belgium, Sep. 2011, pp. 605–608.

[13] W. Daio, G. Cheung, N.-M. Cheung, A. Ortega, and O. C. Auo, "Rate-distortion optimized merge frame using piecewise constant functions," in *Proceed. of IEEE Intern. Conf. Image Proc.* Melbourne, Australia: IEEE, 2013.

[14] X. Cheng, L. Sun, and S. Yang, "A multi-view video coding approach using layered depth image," in *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*. Chania, Greece: IEEE, 2007, pp. 143–146.

[15] I. Daribo and B. Pesquet-Popescu, "Depth-aided image inpainting for novel view synthesis," in *Proceed. of IEEE Worksh. Multim. Sign. Proc.*, Saint Malo, France, Oct. 2010, pp. 167–170.

[16] M. Cagnazzo, S. Parrilli, G. Poggi, and L. Verdoliva, "Costs and advantages of object-based image coding with shape-adaptive wavelet transform," *EURASIP Journal on Image and Video Processing*, vol. 2007, pp. Article ID 78 323, 13 pages, 2007, doi:10.1155/2007/78323.

[17] M. Cagnazzo, G. Poggi, L. Verdoliva, and A. Zinicola, "Region-oriented compression of multispectral images by shape-adaptive wavelet transform and SPIHT," in *Proceed. of IEEE Intern. Conf. Image Proc.*, vol. 4, Singapore, Oct. 2004, pp. 2459–2462.

[18] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," in *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3. ACM, 2004, pp. 600–608.

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