

# Joint Texture and Depth Map Video Coding Based on the Scalable Extension of H.264/AVC

Siping Tao<sup>1</sup>, Ying Chen<sup>2</sup>, Miska M. Hannuksela<sup>3</sup>, Ye-Kui Wang<sup>3</sup>, Moncef Gabbouj<sup>2</sup>, and Houqiang Li<sup>1</sup>

<sup>1</sup>University of Science and Technology  
of China  
Hefei, China  
anhuitasp@mail.ustc.edu.cn  
lihq@ustc.edu.cn

<sup>2</sup>Department of Signal Processing,  
Tampere University of Technology  
Tampere, Finland  
ying.chen@tut.fi  
moncef.gabbouj@tut.fi

<sup>3</sup>Nokia Research Center  
Tampere, Finland  
miska.hannuksela@nokia.com  
yekuiwang@huawei.com\*

**Abstract**—Depth-Image-Based Rendering (DIBR) is widely used for view synthesis in 3D video applications. Compared with traditional 2D video applications, both the texture video and its associated depth map are required for transmission in a communication system that supports DIBR. To efficiently utilize limited bandwidth, coding algorithms, e.g. the Advanced Video Coding (H.264/AVC) standard, can be adopted to compress the depth map using the 4:0:0 chroma sampling format. However, when the correlation between texture video and depth map is exploited, the compression efficiency may be improved compared with encoding them independently using H.264/AVC. A new encoder algorithm which employs Scalable Video Coding (SVC), the scalable extension of H.264/AVC, to compress the texture video and its associated depth map is proposed in this paper. Experimental results show that the proposed algorithm can provide up to 0.97 dB gain for the coded depth maps, compared with the simulcast scheme, wherein texture video and depth map are coded independently by H.264/AVC.

## I. INTRODUCTION

Due to recent advances in acquisition and display technologies, 3D video has become a reality in consumer domain with different application opportunities, such as 3D TV [1]. When transmitting 3D content based on multiple representations of 2D videos, the bandwidth constraint is an important issue, thus a compressor is required to code 3D content even with only a reasonably small number of views. However, the rendering equipment may require simultaneous providing of many views, which e.g., is the case for auto-stereoscopic displays. In order to provide an immersive user experience, it is desirable to enable the decoder to render as many and continuous views as possible. View synthesis can satisfy these requirements, by transmitting a reasonable number of views while interpolating other views at the renderer. The Depth-Image-Based Rendering (DIBR) technique [2] is a typical view synthesis algorithm in a communication system. As the high computational complexity required for depth estimation [3] is typically not acceptable at the decoder, a depth map should be transmitted [4]. In so-called video plus depth applications [5], only one view (also

named as texture video) with its associated depth map is required for view synthesis.

To utilize limited transmission bandwidth efficiently, texture video and depth map should be compressed before transmission. One standard-compliant way for compression of the texture video and the depth map is to compress them separately, by utilizing H.264/AVC. However, it is worth exploiting the correlation between the texture video and its associated depth map, to get higher coding efficiency. In [6], Grewatsh et al. proposed one MPEG-2 based method for compressing the depth map by reusing the motion vectors (MVs) of its corresponding texture video. Instead of reusing the texture video motion information without any modifications, a so-called Candidate Mode Generation process is used in [7], which can generate more accurate motion information for the depth map. Both of these cases do not need to transmit MVs for depth map. However, the previous methods have two drawbacks. First, although they are originated from MPEG-2 or H.264/AVC, they are not compliant with any existing standard. Second, they do not necessarily result in optimal compression efficiency. For example, sharing the texture video MVs with the depth map may increase the residual data when coding the depth map because the texture video MV may not be sufficiently accurate for its associated depth map. In this paper, to solve the above problems, we propose to utilize the inter-layer motion prediction tool in SVC, the scalable extension of H.264/AVC [8], to compress the texture video and its associated depth map jointly. Experimental results demonstrate that, compared with the simulcast scheme, the proposed method achieves on average 0.56 dB and up to 0.97 dB gain for the coded depth maps, which is equivalent to 22% of bit-rate reduction.

The rest of the paper is organized as follows. Section II introduces the inter-layer prediction tools in SVC. Section III analyzes the correlation between the texture video and its associated depth map. The details of the proposed method are presented in section IV. Section V describes the test scenarios and presents the simulation results, and Section VI concludes the paper.

\*The work of Houqiang Li and Siping Tao is partially supported by NSFC General Program under contract No. 60572067, NSFC General Program under contract No. 60672161, and NSFC Key Program under contract No. 60736043. The work of Ying Chen is partly supported by the Nokia Foundation Award granted by Nokia Research Center (NRC). The work of Ye-Kui Wang was carried in NRC and he is currently with Huawei Technologies, Bridgewater, NJ, USA.

## II. INTER-LAYER PREDICTION IN SVC

The Scalable Video Coding (SVC) extension of the H.264/AVC standard has been developed by the Joint Video Team (JVT). The layered coding approach is employed in SVC, wherein more than one dependency layer can be presented and each layer is identified by a dependency identifier. The layer with the lowest dependency identifier value is called the base layer and the other layers are called enhancement layers in this paper. To remove the redundancy between different spatial or quality layers, three inter-layer prediction tools are used, namely motion prediction, intra texture prediction, and residual prediction. The inter-layer motion prediction tool is described in the next paragraph. For an introduction to the other inter-layer prediction tools, which are not employed in the proposed method, please refer to [8].

Co-located macroblock (MB) in the base layer can be used to derive predictors for the MVs in an MB of an enhancement layer. It is called inter-layer motion predictor. For dyadic spatial scalability case, the base layer MVs need to be simply scaled. While for Coarse Granularity Scalability (CGS), an MV in the base layer can be directly used as the predictor. Note that an MV difference can be signaled for each MB or MB partition to further refine the inter-layer motion predictor to a better one in terms of e.g., rate-distortion performance.

### III. MOTION INFORMATION CORRELATION BETWEEN TEXTURE VIDEO AND DEPTH MAP

A texture video consists of three components, namely one luma component Y, and two chroma components U and V, whereas the depth map only has one component representing the distance between the object pixel and the camera. Generally, a texture video is represented in YUV 4:2:0 format and a depth map is regarded as luma-only video in YUV 4:0:0 format. Fig. 1 is an example of a texture video frame and its associated depth map. Obviously, the color information of a pixel and its distance from the camera are less relevant. However, from Fig. 1, it can be observed that both the texture video and its associated depth map have similar object silhouette, so they should have similar object boundary and movement. To confirm this observation, the following experiment was performed.



Figure 1. A texture video frame and its associated depth map.

The texture video and its associated depth map were coded with H.264/AVC, and their motion fields in the unit of the 4x4 block were extracted. Let  $\vec{t}$  and  $\vec{d}$  be the motion field of a specific frame of the texture video and its associated depth map, respectively. The correlation coefficient between the two motion fields is calculated as

$$\rho_i = \frac{Cov(\vec{t}, \vec{d})}{\sqrt{Var(\vec{t})Var(\vec{d})}}, \quad (1)$$

wherein

$$Cov(\vec{t}, \vec{d}) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} [(\vec{t}_{m,n} - \vec{t}_{avg}) \cdot (\vec{d}_{m,n} - \vec{d}_{avg})] \quad (2)$$

$$Var(\vec{t}) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \|\vec{t}_{m,n} - \vec{t}_{avg}\|^2 \quad (3)$$

$\vec{t}_{avg}$  is the average MV of  $\vec{t}$ ,  $\vec{d}_{avg}$  is that of  $\vec{d}$ , M/N is the picture height/width divided by 4, “ $\cdot$ ” denotes the inner product, and  $\|\cdot\|$  denotes the norm operator. The correlation coefficients were calculated for 100 frames in the *Ballet* test sequence, and the curve of correlation coefficient per frame is plotted in Fig. 2. It can be observed that, the texture video motion field and its associated depth map motion field are well correlated. Therefore, coding efficiency would be improved if one can efficiently use this correlation. In the next section, a new method of compressing texture video and depth map jointly using SVC is proposed.

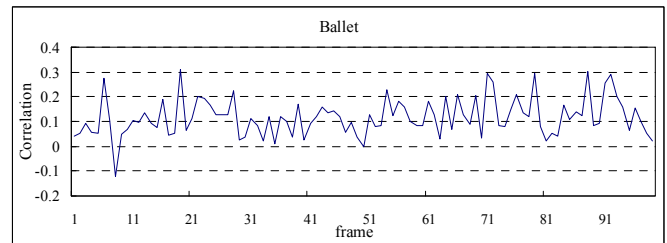


Figure 2. Correlation coefficient of motion fields per frame.

### IV. TEXTURE VIDEO AND DEPTH MAP COMPRESSION USING SVC

In the proposed method, two layers are coded: the texture video is coded as the base layer, and the depth map is coded as the CGS enhancement layer. The texture video is coded using the same mechanism as H.264/AVC single layer coding, while the depth map is coded using SVC inter-layer motion prediction in addition to single layer coding techniques. It should be noted that the other two inter-layer prediction tools are disabled in the proposed method. In the depth map motion estimation process, the conventional spatial MV predictor or the inter-layer MV predictor can be chosen for each MB in the enhancement layer. Furthermore, when the co-located MB in the base layer is inter coded, the MB in the enhancement layer can adaptively choose the base mode in addition to the conventional H.264/AVC modes in the mode decision process. After motion estimation and mode decision, transform and quantization are applied to the enhancement layer as in SVC.

The detailed enhancement layer mode decision process for the inter frame coding is illustrated in Fig. 3. When the co-located MB in the base layer is intra coded, inter and intra modes without inter-layer prediction will be checked. Otherwise, the base mode without residual prediction will be checked first, then the inter modes using inter-layer motion

predictors (without residual prediction) are checked next, as well as inter and intra modes without inter-layer prediction.

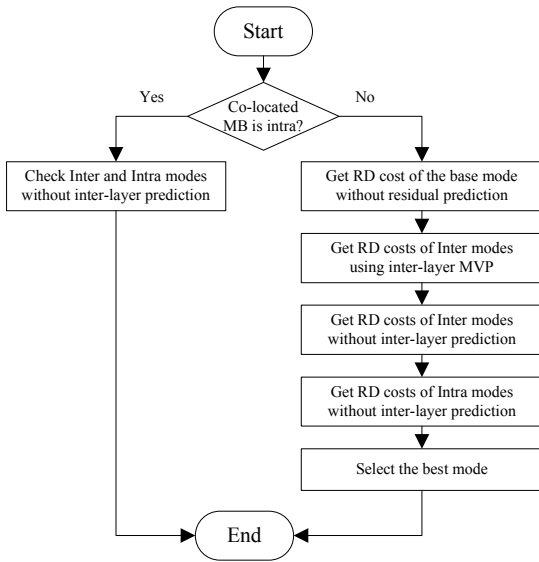


Figure 3. Enhancement layer mode decision process.

## V. SIMULATIONS

The proposed algorithm was implemented in SVC reference software JSVM\_9\_13 [9], and the simulation conditions were as follows:

- IPPP coding structure and hierarchical B picture coding structure with GOP size equal to 16 were tested separately.
- 100 frames were encoded.
- Spatial / temporal resolution: 1024x768@15Hz.
- The CABAC entropy coding method was used.
- Intra picture refresh was turned off.
- 8x8 transform was turned on.
- The base layer QP (QP0) varies in {24, 28, 32, 36}. The enhancement layer QP (QP1) was selected to result in depth map bit-rate approximately in the range of 10% to 20% compared with the bit-rate of the texture video, which has been considered to provide sufficient quality for DIBR [10].

Two sequences, namely *Ballet* and *Breakdancers* [11], were tested. These two sequences were selected due to the fact that accurate depth maps were available for them.

The following four methods were tested in the simulations:

- The proposed method as described in section IV.
- Simulcast: texture video and depth map were independently compressed.
- All inter-layer prediction (AP): the proposed method as described in section IV but additionally all the inter-layer prediction tools were used adaptively.

- Forced motion prediction (FP): inter-layer motion prediction (without motion refinement) was always used for the depth maps, whereas the other inter-layer prediction tools and spatial/temporal MV prediction were disabled for the depth maps.

The FP method is similar to the method in [6]. Because the texture video MV may not be accurate for its associated depth map, both of the methods in [6] and [7] have worse efficiency than Simulcast in medium and high bit-rates, and the overall efficiency improvement is limited. In contrast, the simulation results showed that the efficiency gain of the proposed method is consistent over all the bit-rates.

In Table I, the comparison of the proposed method and Simulcast is presented. It can be observed that the proposed method offers up to 0.97 dB gain in terms of depth map PSNR compared with Simulcast, and 0.56 dB gain on average. Fig. 4 shows the RD curves of the above four methods. It can be found that the proposed method has similar performance as AP; this is because inter-layer texture prediction and residual prediction have little contribution to improve the coding efficiency. However, comparing with AP, the proposed method does not need to check the modes with residual prediction, as well as the modes with texture prediction. Note that those are about half of the modes tested in JSVM and requires a substantial large part of the encoding computations in the reference JSVM encoder. Therefore, the proposed method has much lower encoding complexity than AP. From Fig. 4, we can see that the FP method has the worst performance among the tested methods, about 3.6 dB loss on average in terms of depth map PSNR compared with Simulcast. The cause for the inferior performance of the FP method is illustrated in Fig. 5, which presents the percentage of MBs using inter-layer motion prediction for the proposed method. The percentage of MBs using inter-layer motion prediction is about 10%. Thus, nearly 90% of the MB modes of the depth maps are not optimal in the FP method.

TABLE I. COMPARISON OF THE PROPOSED METHOD AND SIMULCAST.

Sequence (GOP)	Proposed vs. Simulcast Gain	
	□ PSNR (dB)	□ Bitrate (%)
<i>Ballet</i> (16)	0.81	-20.81
<i>Breakdancers</i> (16)	0.36	-7.72
<i>Ballet</i> (1)	0.97	-21.87
<i>Breakdancers</i> (1)	0.08	-1.59

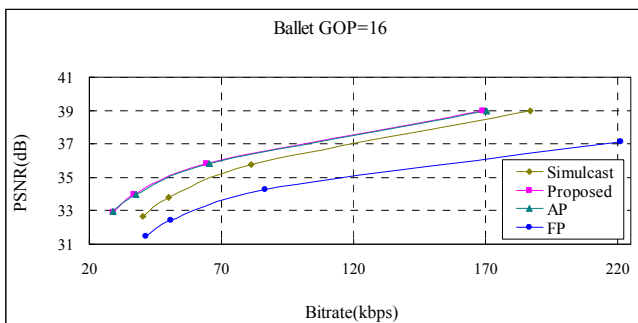
## VI. CONCLUSIONS

In this paper, a new coding algorithm was proposed for joint texture and depth map video coding. The method encodes texture video as the base layer of a scalable video bitstream, and depth map as the enhancement layer. Inter-layer motion vector prediction is adaptively used to improve the compression efficiency compared with coding texture video and depth map as two independent bitstreams. The proposed method complies with the Scalable Video Coding (SVC) standard, hence facilitating the reuse of SVC implementations for Depth-Image-Based Rendering. Moreover, the fact that the base layer remains H.264/AVC compliant enables phased introduction of depth maps into existing H.264/AVC-based services, because devices that do not have 3D functionality

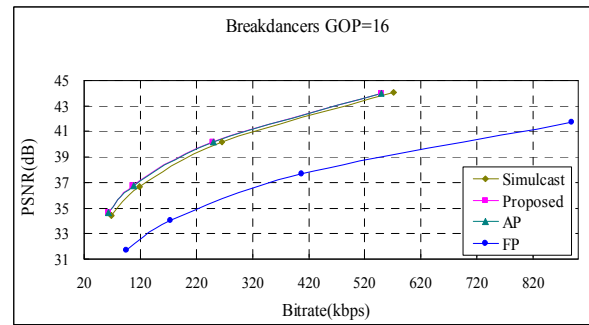
can decode the base layer to provide a conventional 2D video sequence. Simulation results showed that compared with simulcast coding of texture video and depth map, the proposed method can bring significant coding gain for the associated depth map.

REFERENCES

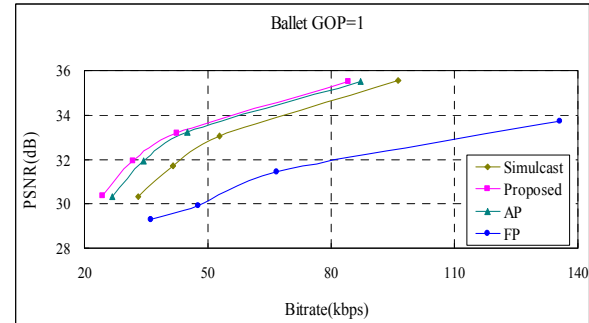
- [1] L. Onural, T. Sikora, and A. Smolic, "An overview of a new European consortium: integrated three-dimensional television—capture, transmission and display (3D TV)," Proc. European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (EWIMT), London, UK, Nov. 2004.
- [2] W. R. Mark, "Post-rendering 3-D image warping: visibility, reconstruction, and performance for depth-image warping," PhD thesis, University of North Carolina at Chapel Hill, NC, USA, 1999.
- [3] N. Fukushima, T. Yendo, T. Fujii and M. Tanimoto, "Free viewpoint image generation using multi-pass dynamic programming," Proc. SPIE Stereoscopic Displays and Virtual Reality Systems XIV, vol 6490, pp. 460-470, Feb. 2007.
- [4] "Description of exploration experiments in 3D video coding," ISO/IEC JTC1/SC29/WG11, Doc. W9991, Hannover, Germany, 2008.
- [5] C. Fehn, P. Kauff, M. Op de Beeck, F. Ernst, W. Ijsselstein, M. Pollefeys, L. Van Gool, E. Ofek, and I. Sexton, "An evolutionary and optimized approach on 3D-TV," Proc. IBC 2002, pp. 357-65, Amsterdam, Netherlands, Sept. 2002.
- [6] S. Grewatsh and E. Muller, "Sharing of motion vectors in 3D video coding," Proc. IEEE International Conference on Image Processing, vol 5, pp. 3271-3274, Oct. 2004.
- [7] H. Oh and Y. S. Ho, "H.264-based depth map sequence coding using motion information of corresponding texture video," Spinger Berlin/Heidelberg, Advances in Image and Video Technology, vol. 4319, pp. 898-907, 2006.
- [8] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," IEEE Trans on Circuits and Systems for Video Technology, vol. 17, No. 9, pp. 1103-1120, Sept. 2007.
- [9] J. Reichel, H. Schwarz, and M. Wien, Joint Scalable Video Model 11 (JSVM 11), Joint Video Team, Doc. JVT-X202, Jul. 2007.
- [10] A. Smolic, K. Mueller, N. Stefanoski, J. Ostermann, A. Gotchev, G. B. Akar, G. Triantafyllidis, and A. Koz, "Coding algorithms for 3DTV-a survey," IEEE Trans. On Circuits and Systems for Video Technology, vol. 17, no. 11, pp. 1606-1621, Nov. 2007.
- [11] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," ACM SIGGRAPH and ACM Trans. on Graphics, Los Angeles, CA, pp. 600-608, Aug. 2004.
- [12] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," VCEG-M33, Mar. 2001.



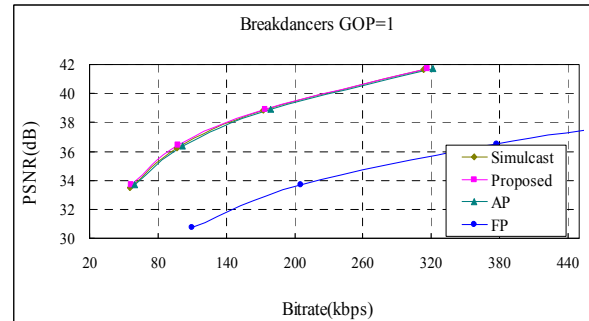
(a) Ballet with GOP=16



(b) Breakdancers with GOP=1



(c) Ballet with GOP=1



(d) Breakdancers with GOP=1

Figure 4. RD performance of the proposed method, Simulcast, AP and FP.

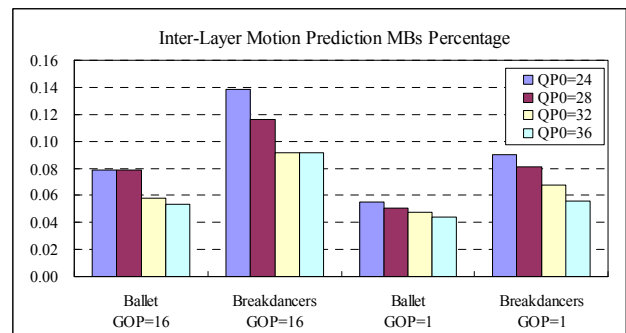


Figure 5. Percentages of MBs using inter-layer motion prediction.