FLEXIBLE DEPTH MAP SPATIAL RESOLUTION IN DEPTH-ENHANCED MULTIVIEW VIDEO CODING

Payman Aflaki[a], *Miska M. Hannuksela*[b], *Moncef Gabbouj*[a]

[a]Department of Signal Processing, Tampere University of Technology, Tampere, Finland;
[b]Nokia Research Center, Tampere, Finland;

ABSTRACT

Multiview video plus depth (MVD) has proved to be a promising format enabling various 3D applications. One approach to achieve better MVD compression is to adjust the spatial resolution of depth map based on the content and the application. In this research two schemes are considered: first, multiview video coding accompanied by depth maps to improve the texture coding performance. Second, multiview video plus depth (MVD) coding targeting highest quality for synthesized views. Two algorithms to select the best spatial resolution for each scheme are proposed and the results show 10.8% and 16.5% bitrate reduction compared to the anchor case where depth map resolution is fixed.

Index Terms— 3DV, MVD, depth map, spatial resolution, synthesized views

1. INTRODUCTION

In multi-view and free-view systems, which enable perception of an arbitrary view angle of the scenes, are encumbered by storage and transmission problems resulting from huge amount of multi view data. Other than complexity of capturing 3D scenes from several different points of view, the limitations in the distribution technologies makes it not feasible to deliver a sufficiently large number of (e.g. 20-50) views to the user's side with existing compression standards. The multiview video plus depth (MVD) format [1] has been proposed as an approach to solve the described problems. In MVD format, each texture pixel is accompanied with a respective depth value and therefore, it can be served as a source to a depth image-based rendering (DIBR) [2] algorithm which produces the required number of views in the decoder side. MVD and DIBR may be used for varying baseline to adjust the depth perception on a stereoscopic display and for generating views for multiview auto-stereoscopic displays (ASDs).

The Advanced Video Coding (H.264/AVC) standard [3] has been extended with the Multiview Video Coding (MVC) extension [4], which enables stereoscopic and multiview texture video coding. More recently, H.264/AVC has been amended with two extensions enabling MVD coding, called the Multiview and Depth Video Coding (MVC+D) [5] and the Multiview and Depth Video with Enhanced Non-base View Coding (3D-AVC) 0, both developed by the Joint Collaborative Team on 3D Video Coding Extension Development (JCT-3V). MVC+D specifies the encapsulation of coded MVD data into a single bitstream. While preserving the forward compatibility with MVC specification, MVC+D enables decoding of texture views of MVC+D bitstreams with a conventional MVC decoder. 3D-AVC introduces several coding tools exploiting the dependencies and similarities between depth and texture. Both MVC+D and 3D-AVC specifications were implemented in 3DV-ATM reference software [7].

JCT-3V found to be beneficial to use 3DV-ATM with depth maps having half spatial resolution in each direction compared to texture spatial resolution [8]. Such scheme helped in reducing the bitrate while maintaining the quality of synthesized views. This paper proposes to reduce the resolution of depth even further compared to the texture views. The proposed methods were tested with 3D-AVC and 3DV-ATM software. In the decoder side, resolution of depth maps should be re-scaled to the resolution of texture views to be used as the input of DIBR algorithms for synthesizing virtual views. There are several methods available in literature to perform down or up sampling on depth maps [9, 10]. Moreover, content-adaptive texture resolution selection based on frequency domain analysis is introduced in [11] while [12] proposed depth maps compression methods taking into account their characteristics such as spatial resolution.

In this research two following topics are covered:

First, we are targeting more efficient texture coding where depth-based texture coding tools are used to improve compression performance but depth is not used as input for DIBR. For example, depth-based motion vector prediction (DMVP) and backward view synthesis prediction (BVSP) can be exploited [13]. The proposed scheme, including the bitrate required to encode both texture and associated depth maps, is compared with the anchor, i.e. MVC-coded texture views only. It is shown that best performance is achieved under different depth resolution criteria. The selection of depth map resolution is performed automatically based on the amount of high frequency components (HFCs) available in the first frame of the texture views.

Second, MVD data is used for DIBR and depth map resolution is selected to obtain the highest quality for the synthesized views. In this scenario, different spatial resolutions are considered for different sequences based on the amount of information introduced in both depth maps and texture views. This test scheme is evaluated against depth maps with vertically and horizontally half resolution using the 3DV-ATM [7] reference software with defined specifications in relevant test configuration with respect to the JCT-3V Common Test Conditions (CTC) [8]. An automatic method based on the statistics of the first synthesized frame is used to decide which depth resolution should be selected.The performance of both schemes is

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

objectively compared to respective anchors and it is confirmed that a considerable gain is achieved in both cases.

The rest of the paper is organized as follows. Section 2 presents the proposed schemes and the algorithms to select the spatial resolution for depth maps while test material and simulation results are reported in Section 3. Finally, Section 4 concludes the paper.

2. PROPOSED CODING SCHEMES

To describe the proposed scheme, we assume two- or three-view coding scenario (C2 or C3, respectively), since these scenarios have been considered suitable for the currently available stereoscopic and multiview autostereoscopic displays [8]. As introduced in the previous section, the research reported in this paper concentrated on two different aims, namely: best texture coding performance (multiview texture coding) and best quality for synthesized views (MVD coding). Each scheme benefits from 3D-AVC coding tools designed to increase the performance of texture coding by inheriting some information from available depth maps. Furthermore, as the resolution of associated depth maps has a key influence on the efficiently of these tools, an automatic algorithm is proposed to select the superior depth map spatial resolutions. Each scheme is discussed in detail in the following sub-sections.

2.1 Multiview texture coding

In this scheme, the depth information is utilized as complementary data to improve the performance of texture coding. 3D-AVC uses depth values for depth-based motion vector prediction (DMVP) as well as backward view synthesis prediction (BVSP). In DMVP, a disparity vector is derived from the depth map associated with a texture picture and the derived disparity vector may be used under certain conditions as a motion vector predictor for inter-view prediction. In BVSP, a disparity vector is derived for each sub-block (e.g. 8x8 blocks) of a prediction block (e.g. 16x16) block and a prediction block is derived from an inter-view reference picture using the disparity vectors.

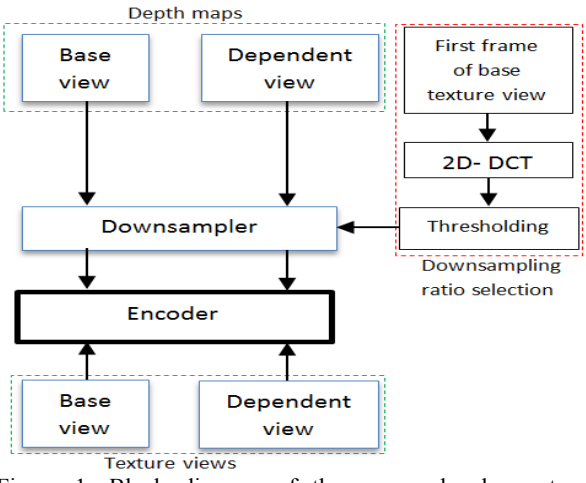

Figure 1. Block diagram of the proposed scheme targeting multiview texture coding

To select the best spatial resolution for depth maps, the statistics of texture views is considered. In particular, the amount of HFCs in the first frame of the base texture view is considered as an indicator showing the amount of the details presented in the texture views. It is assumed that the higher the amount of details there is, the more accurate depth representation is needed to obtain accurate prediction blocks with the depth-based texture prediction tools described above. Hence, the amount of HFCs is used to determine how much supplementary information should be transmitted as the depth map format. In other words, the more the amount of the HFCs, the bigger the spatial resolution of the depth maps should be. The flowchart of the proposed scheme is presented in Figure 1. In this scheme, the red dashed block first decides which spatial resolution for depth maps should be considered and then, the depth map with such resolution is used in the encoding process of texture views. The algorithm to decide the proper spatial resolution for depth maps is described in the following paragraphs.

Only the first frame of the base texture view is considered in the process of HFC calculation. In this process, the two dimension-discrete cosine transform (2D-DCT) of the respective frame is calculated first. This is reported in Eq. (1):

$$I_{DCT}(m,n) = DCT_{2d}\,[I_{Pixel}(m,n)] \quad (1)$$

where:

$I_{Pixel}(m,n)$ is the input frame pixel values;

m,n are the width and the height of the frame, respectively;

$I_{DCT}(m,n)$ includes the output transform coefficients;

$DCT_{2d}[]$ is the function to calculate the 2D-DCT

Following this, the absolute value of all HFCs in $I_{DCT}(m,n)$ are zigzag scanned [14] into the vector HFC_{zigzag}. Coefficients located at the end of this vector, i.e. from a certain point till the end, are selected to be an indicator of the amount of HFCs in the I_{Pixel}, as shown in Eq. (2):

$$\begin{aligned} HFC_{zigzag} &= \text{Zigzag}(\text{abs}(I_{DCT}(m,n))) \\ HFC_{Selected} &= HFC_{zigzag}(m\times n\times w : m\times n) \end{aligned} \quad (2)$$

where:

abs(x) is the function to calculate absolute value of x

HFC_{zigzag} returns the zigzag scanned coefficients into a vector

w is a positive threshold smaller than 1 indicating the range of HFCs in HFC_{zigzag}

Finally, a threshold is applied on the average value of selected coefficients in $HFC_{Selected}$ targeting proper selection of spatial resolution for depth maps, as shown in Eq. (3).

$$\text{Ratio} = \begin{cases} \frac{1}{4} & , \quad Mean(\text{HFC}_{Selected}) \geq th_1 \\ \frac{1}{8} & , th_1 > Mean(\text{HFC}_{Selected}) \geq th_2 \\ \frac{1}{16} & , th_2 > Mean(\text{HFC}_{Selected}) \end{cases} \quad (3)$$

where:

Ratio is the downsampling ratio applied to spatial resolution of depth maps in both directions
Mean(X) returns average value of X.

2.2 MVD coding

In this scheme, the depth maps are exploited by DIBR algorithms and in the view synthesis process. Therefore, the targeted spatial resolution decrease in depth maps should not be less than a particular value to guarantee an acceptable quality for rendered views. Based on several characteristics of texture views and depth maps, the spatial resolution of depth maps can vary aiming the best performing RD curves for synthesized views.

To make the decision about the spatial resolution of depth maps, an analysis is performed on the first frame of each sequence. In this algorithm, the original first texture frame is exploited along with respective depth maps down and up sampled with different ratios. The Bjontegaard delta bitrate and delta Peak Signal-to-Noise Ratio (PSNR) metrics [15] of synthesized views according to the C3 scenario specified in JCT-3V CTC [8] are used to evaluate the performance of each downsampling ratio. The general flowchart of this scheme is depicted in Figure 2. The performance of each spatial resolution of depth maps is compared to anchor having the depth maps with quarter spatial resolution compared to texture views (half resolution in each direction). Therefore, if the dBR% achieved for synthesized views is less than a specific threshold, then that resolution is selected, otherwise, the anchor depth spatial resolution is selected.

3. TEST MATERIAL AND SIMULATION RESULTS

Both schemes proposed in this paper (as presented in sub-sections 2.1 and 2.2) were tested with 3DV-ATM v8.1

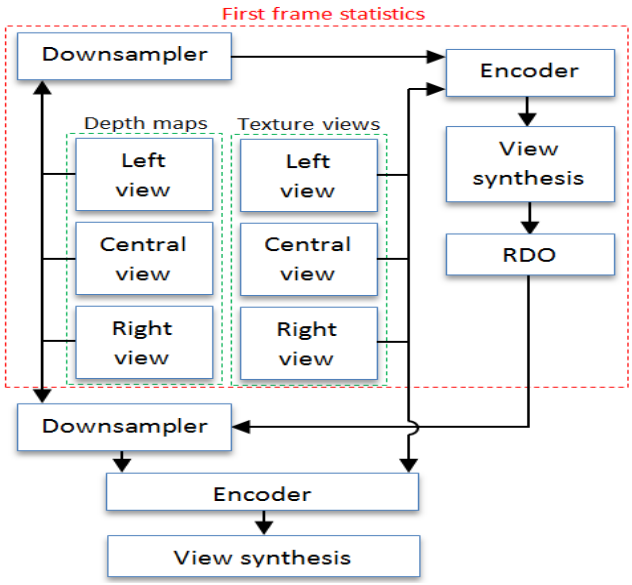

Figure 2. Block diagram of the proposed scheme targeting MVD coding

TABLE I. CONFIGURATION OF 3DV-ATM CONFIGURED THE ANCHOR (MVC+D) AND PROPOSED SCHEME

Coding Parameters	Settings
Compatibility Mode	0 (MVC+D)
Multi-view scenario	Three/two views (C3/C2)
Inter-view prediction structure	PIP
Inter prediction structure	HierarchicalB, GOP8
QP settings for texture & depth	26, 31, 36, 41
Encoder settings	RDO ON, VSO OFF
View Synthesis in post-processing	Fast_1D VSRS [16]
Test sequences and coded, synthesized views	As specified in [8]

software and compared against the anchor scheme (MVC+D). Simulations were conducted under the specifications of C2 and C3 scenario in JCT-3V CTC [8] for multiview texture coding and MVD coding, respectively, and JCT-3V MVD test sequences were utilized. In these scenarios two and three depth-enhanced texture views are encoded and then several possible in-between views are synthesized to be exploited in stereoscopic image-pair creation.

The full resolution MVC+D coding, as implemented in 3DV-ATM [8], and 3DV VSRS [16] were utilized to produce a full resolution anchor results. Table I summarizes the major parameters used for the anchor 3DV-ATM configuration, whereas complete configuration files for MVC+D are available in [8].

The simulation framework for the proposed schemes is specified as shown in Table I and input depth maps were selected based on the algorithms introduced in sub-sections 2.1 and 2.2. Moreover, the 3DV-ATM was changed to support flexible texture to depth resolution ratio and the same software was used in proposed schemes and anchor.

The compression efficiency of the proposed schemes was evaluated according to the CTC [8] specification and against MVC+D scheme as anchor. The Bjontegaard delta bitrate and delta PSNR metrics [15] were utilized for these purposes. For multiview texture coding scheme targeting best texture coding performance, following initiations were applied: $w = 0.7$, $th_1 = 1$, $th_2 = 0.75$. To confirm the performance of the proposed algorithm in sub-section 2.1 to select the best depth spatial resolution, we run the full length simulations using depth maps with different resolutions too. The spatial resolution selections based on both methods are reported in Table II. The results confirmed that the proposed algorithm succeeds to select the depth map spatial resolution correctly for all sequences except for Newspaper where the proposed algorithm selected a higher resolution than the best performing one (see Table II). The reason for Newspaper to be an outlier might be due to two reasons. First, Newspaper has a relatively larger disparity compared to all other sequences and therefore, even a higher spatial resolution depth map (as suggested by the proposed algorithm) might be incapable of providing good support for texture coding. Second, the relatively lower depth map quality of

TABLE II. DEPTH TO TEXTURE SPATIAL RESOLUTION RATIO SELECTION WITH PROPOSED ALGORITHM AND SIMULATIONS WITH FULL LENGTH SEQUENCES – MULTIVIEW TEXTURE CODING SCHEME

	Depth to texture spatial resolution ratio	
Sequence	**Proposed method**	**Full simulations**
Poznan Hall2	1/8	1/8
Poznan Street	1/4	1/4
Undo Dancer	1/4	1/4
Ghost Town Fly	1/4	1/4
Kendo	1/16	1/16
Balloons	1/16	1/16
Newspaper	***1/4***	1/16

TABLE III. PERFORMANCE OF MULTIVIEW TEXTURE CODING SCHEME COMPARED TO THE ANCHOR

	Coded views	
	dBR, %	dPSNR,dB
Poznan Hall2	-18.38	0.65
Poznan Street	-3.56	0.11
Undo Dancer	-9.65	0.35
Ghost Town Fly	-11.48	0.45
Kendo	-10.65	0.57
Balloons	-15.07	0.87
Newspaper	-6.71	0.31
Average	**-10.79**	**0.47**

Newspaper sequence confirms that the change in its spatial resolution might have less effect on the efficiency of the exploited coding tools [13].

The performance of multiview texture coding scheme compared to MVC+D anchor, as specified in C2 scenario in CTC [8], is reported in Table III. Since only the texture coding performance is considered in this scheme, only the texture coding bitrates were considered for anchor too. An average 10.8% dBR reduction shows that changing the spatial resolution of depth maps has a significant effect on the performance of the codec.

Considering MVD coding scheme targeting best quality for synthesized views, C3 scenario in CTC [8] was considered and dBR% threshold was fixed to 6%. Moreover, we performed the simulations using full length sequences with different depth map resolutions to confirm the validity of the proposed algorithm in sub-section 2.2. Table IV reports both spatial resolution ratio values of texture to depth achieved with proposed algorithm and full length sequence simulations. It can be seen that the spatial resolution of only Ghost Town Fly is not correctly selected using the proposed algorithm. The reason is that in this sequence we have strong change of content between starting frames and the rest of the sequence. Therefore, any algorithm considering only the statistics of the first frame(s) is not able to accurately estimate the application of different coding schemes. The performance of the coding MVD coding scheme compared to MVC+D anchor is presented in Table V. As the quality of synthesized views is targeted in this scheme, the anchor statistics include the bitrate required to encode depth maps too. Considering the quality of synthesized views, the 16.5% dBR reduction shows that the spatial resolution of depth maps has a significant effect.

In tables III and V the coding performance of the resolutions selected from full simulations (as mentioned in Tables II and IV) are reported. However, the performance of the resolutions selected by the proposed automatic methods was under performing these results by only 0.69% and 0.3% dBR for multiview texture coding and MVD coding, respectively. Therefore, the proposed methods can be considered as a good replacement to find the best depth map resolution for different scenarios. Moreover, we limited our experiments to dyadic ratios between texture and depth as non-dyadic ratios tended to provide very marginal gain.

TABLE IV. DEPTH TO TEXTURE SPATIAL RESOLUTION RATIO SELECTION WITH PROPOSED ALGORITHM AND SIMULATIONS WITH FULL LENGTH SEQUENCES – MVD CODING SCHEME

	Depth to texture spatial resolution ratio	
Sequence	**Proposed method**	**Full simulations**
Poznan Hall2	1/4	1/4
Poznan Street	1/4	1/4
Undo Dancer	1/2	1/2
Ghost Town Fly	***1/4***	1/8
Kendo	1/4	1/4
Balloons	1/4	1/4
Newspaper	1/2	1/2

TABLE V. PERFORMANCE OF THE PROPOSED MVD CODING SCHEME COMPARED TO THE ANCHOR

	Coded views		Synthesized views	
	dBR, %	dPSNR, dB	dBR, %	dPSNR, dB
Poznan Hall2	-32.27	1.22	-22.95	0.73
Poznan Street	-16.13	0.52	-9.66	0.29
Undo Dancer	-18.91	0.74	-11.22	0.38
Ghost Town Fly	-23.47	0.97	-21.03	0.74
Kendo	-25.37	1.38	-19.34	0.88
Balloons	-29.16	1.72	-23.49	1.20
Newspaper	-15.73	0.72	-7.80	0.28
Average	**-23.00**	**1.04**	**-16.50**	**0.64**

4. CONCLUSIONS

In this paper, we exploited few encoding tools contributing to better texture coding based on the information presented in depth maps. We considered reducing the spatial resolution of depth maps targeting better texture coding performance and higher quality of synthesized views. The simulation results confirmed that having flexible spatial resolution ratio between depth maps and texture views, enables higher performance of codec for both texture coding and quality of synthesized views. Two algorithms were introduced to predict the best spatial resolution of depth maps based on the first frame statistics of each sequence. Each method was capable of well estimating the best spatial resolution.

5. ACKNOWLEDGEMENT

The authors would like to thank M. Domański, et al. for providing Poznan sequences and Camera Parameters [17].

6. REFERENCES

[1] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-view video plus depth representation and coding," Proc. of IEEE International Conference on Image Processing, vol. 1, pp. 201-204, Oct. 2007.

[2] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," in Proc. SPIE Conf. Stereoscopic Displays and Virtual Reality Systems XI, vol. 5291, CA, U.S.A., Jan. 2004, pp. 93–104.

[3] ITU-T and ISO/IEC JTC 1, "Advanced video coding for generic audiovisual services," ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), 2010.

[4] Y. Chen, Y.-K. Wang, K. Ugur, M. M. Hannuksela, J. Lainema, and M. Gabbouj, "The emerging MVC standard for 3D video services," EURASIP Journal on Advances in Signal Processing, vol. 2009, Article ID 786015, 2009. doi:10.1155/2009/786015.

[5] Y. Chen, M. M. Hannuksela, T. Suzuki, and S. Hattori, "Overview of the MVC+D 3D video coding standard," Journal of Visual Communication and Image Representation, Apr. 2013.
Online: http://dx.doi.org/10.1016/j.jvcir.2013.03.013

[6] M. M. Hannuksela, Y. Chen, T. Suzuki, J.-R. Ohm, and G. Sullivan (ed.), "3D-AVC draft text 8," Joint Collaborative Team on 3D Video Coding Extension Development, document JCT3V-F1002, Nov. 2013.

[7] M. M. Hannuksela, D. Rusanovskyy, W. Su, L. Chen, R. Li, P. Aflaki, D. Lan, M. Joachimiak, H. Li, and M. Gabbouj. "Multiview-Video-Plus-Depth Coding Based on the Advanced Video Coding Standard," *IEEE transactions on image processing*, vol. 22, no. 9 (2013).

[8] "Common test conditions for 3DV experimentation," ISO/IEC JTC1/SC29/WG11 MPEG2012/N12560, Feb. 2012.

[9] P. Aflaki, M. M. Hannuksela, D. Rusanovskyy, and M. Gabbouj, "Non-Linear Depth Map Resampling for Depth-Enhanced 3D Video Coding, " IEEE Signal Processing Letters, Vol. 20, issue 1, pp. 87-90, Jan. 2013

[10] JSVM software
http://p.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm

[11] P. Aflaki, D. Rusanovskyy, M. M. Hannuksela, and M. Gabbouj. "Frequency based adaptive spatial resolution selection for 3D video coding," Proceedings of the 20th European Signal Processing Conference (EUSIPCO), pp. 759-763, 2012.

[12] V. Nguyen, M. Dongbo, and M. Do. "Efficient Techniques for Depth Video Compression Using Weighted Mode Filtering, " IEEE Trans. Circuits Syst. Video Technology, vol. 23, no. 2, pp. 189-202, 2013

[13] D. Rusanovskyy, M. M. Hannuksela, and W. Su, "Depth-based coding of MVD data for 3D video extension of H. 264/AVC," Journal of 3D Research, 4:6, p.p. 1-10, June 2013.

[14] Wen-Hsiung Chen; Pratt, W., "Scene Adaptive Coder," Communications, IEEE Transactions on , vol.32, no.3, pp.225,232, Mar 1984

[15] G. Bjøntegaard, "Calculation of average PSNR differences between RD-Curves," ITU-T SG16 Q.6 document VCEG-M33, April 2001

[16] H. Schwarz, et al., "Description of 3D Video Technology Proposal by Fraunhofer HHI (MVC compatible)," ISO/IEC JTC1/SC29/WG11 MPEG2011/M22569, Nov, 2011.

[17] M. Domañski, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, and K. Wegner, "Poznan Multiview Video Test Sequences and Camera Parameters", ISO/IEC JTC1/SC29/WG11 MPEG 2009/M17050, Xian, China, October 2009.