CODING OF MIXED-RESOLUTION MULTIVIEW VIDEO IN 3D VIDEO APPLICATION

Payman Aflaki[a], *Wenyi Su*[b], *Michal Joachimiak*[a], *Dmytro Rusanovskyy*[c], *Miska M. Hannuksela*[c], *Houqiang Li*[b], *Moncef Gabbouj*[a]

[a]Department of Signal Processing, Tampere University of Technology, Tampere, Finland;
[b]University of Science and Technology of China, Hefei, China
[c]Nokia Research Center, Tampere, Finland;

ABSTRACT

The emerging MVC+D standard specifies the coding of Multiview Video plus Depth (MVD) data for enabling advanced 3D video applications. MVC+D specifications define the coding of all views of MVD at equal spatial resolution and apply a conventional MVC technique for coding the multiview texture and the depth independently. This paper presents a modified MVC+D coding scheme, where only the base view is coded at the original resolution whereas dependent views are coded at reduced resolution. To enable inter-view prediction, the base view is downsampled within the MVC coding loop to provide a relevant reference for dependent views. At the decoder side, the proposed scheme consists of a post-processing scheme which upsamples of the decoded views to their original resolution. The proposed scheme is compared against the original MVC+D scheme and an average of 4% delta bitrate reduction (dBR) in the coded views and 14.5% of dBR in the synthesized views are reported.

Index Terms— 3DV, MVC, asymmetric coding, spatial resolution, synthesized views

1. INTRODUCTION

The Moving Picture Experts Group (MPEG) has recently started 3D Video (3DV) standardization to enable support of advanced 3DV applications. The concept of advanced 3DV applications assumes that users can perceive a selected stereo-pair from numerous available views at the decoder side. Examples of such applications includes varying baseline to adjust the depth perception and multiview auto-stereoscopic displays (ASDs). Considering the complexity of capturing 3D scenes and the limitations in the distribution technologies, it is not possible to deliver a sufficiently large number of (20-50) views to the user's side with existing compression standards. To solve this problem, a 3D scene can be represented in multiview video plus depth (MVD) format [1] with a limited number of views, e.g. 2-3. The MVD data is coded and served as a source to a depth image-based rendering (DIBR) [2] algorithm which produces the required number of views at the decoder side.

In March 2011, MPEG issued a Call for Proposals for 3D video coding (hereafter referred to as the 3DV CfP) [3] for a new 3DV standard enabling the rendering of a selectable number of views with respect to the available bitrate. As a result of the CfP evaluation [4], MPEG and, since July 2012, the Joint Collaborative Team on 3D Video Coding (JCT-3V) [5] have initiated development of a depth enhanced extension for MVC [6], abbreviated as MVC+D, to specify the encapsulation of coded MVD data into a single bitstream [7]. The MVC+D standard specifies MVD components (texture and depth) to have equal spatial resolution between different views and utilizes MVC technology [4] for the independent coding of texture and depth. As a result, a forward compatibility with MVC specification is preserved, and texture views of MVC+D bitstreams can be decoded with a conventional MVC decoder. The MVC+D specification was implemented in 3DV-ATM reference software [8] and was used in this study.

A possible solution to further reduce the bitrate and/or complexity of 3DV applications is to reduce the spatial resolution of a number of video views compared to the original resolution while preserving the original resolution for the remaining views. At the decoder side, views coded at the reduced resolution are upsampled to the original one using either conventional linear upsampling [9], or advanced super resolution techniques [10] that would benefit from multiview representation and the presence of depth. Being applied to texture component of MVD, this would result in a mixed-resolution texture representation and a significant bitrate reduction is hence expected.

It is obvious, that a scheme with a mixed-resolution texture representation would result in decoded views (e.g. stereoscopic image-pair) with different quality, which may affect stereoscopic perception. However, this argument can be addressed with the binocular rivalry theory [11] claiming that stereoscopic vision in the human visual system (HVS) fuses the images of an asymmetric quality stereoscopic image-pair so that the perceived quality is closer to that of the higher quality view. Several subjective quality evaluation studies have been conducted to investigate the use of the binocular rivalry theory in stereoscopic video coding [12-15]. Another work presented in [16] showed the applicability of asymmetric coding for MVC-like coding by encoding dependent views with a coarser quantization step compared to the base view. Subjective assessments confirmed that such coding scheme achieved a 20% bitrate reduction for stereoscopic image-pairs created from rendered views with no degradation in the perceived subjective quality.

This paper presents a modified MVC+D coding scheme, where only the base view is coded at the original resolution whereas dependent views are coded at a reduced resolution. To enable inter-view prediction, the base view is downsampled within the MVC coding loop to provide a relevant reference for the inter-view predicted dependent views. At the decoder side, a post-processing scheme that performs upsampling of the decoded views back to their original resolution is proposed.

The rest of the paper is organized as follows. Section 2 presents the asymmetric texture coding schemes, while test material and simulation results are reported in Section 3. Finally, section 4 concludes the paper.

2. MVC CODING FOR MIXED-RESOLUTION TEXTURE REPRESENTATION

Let us assume that the 3DV system is coding MVD data representing a 3D scene with three viewing positions. In our description, we assume three-view (C3) coding scenario, since this is the most relevant test configuration with respect to the MPEG/JCT-3V Common Test Condition [17].

The flowchart of the proposed 3DV system with a mixed-resolution texture representation is shown in Figure 1. An arbitrary view (e.g. the center view) of the input MVD data is coded with H.264/AVC at the original resolution. According to H.264/MVC specification, this view is considered as a base view and provides reference pictures for the inter-view prediction and the coding of dependent views. In the proposed scheme, dependent views of MVD data are coded at a reduced resolution, thus the proposed scheme downsamples the data at the pre-processing stage and upsamples it back to original resolution at the post-processing stage, as shown in Figure 1.

In this study, the base view was coded at the original full resolution (FR) whereas dependent views were coded at half of the original resolution along each direction, which resulted in quarter resolution (QR) downsampled view. However, the downsampling ratio can be adjusted based on the target application.

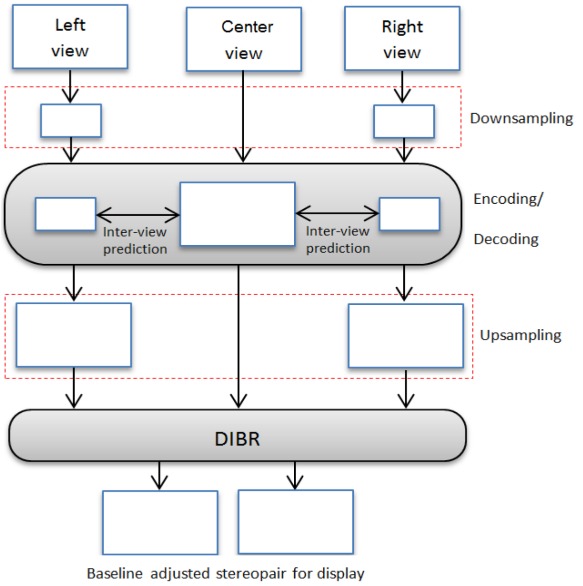

Figure 1. Block diagram of the proposed encoding scheme

Figure 2 shows a simplified flowchart of H.264/MVC scheme with the proposed modification in the in-loop operations for enabling mixed-resolution coding. Base view coding is performed with the conventional H.264/AVC technique and the decoded pictures are stored in a frame buffer. Since the resolution of the base view is different from that of the dependent view, the decoded picture of the original view cannot be used as a reference picture for coding dependent views. To enable inter-view prediction, the resolution of the base view picture should match the resolution of dependent views. There are various approaches to do this, and in this paper we tested two methods: decimation of the reference picture (marked with green line in Figure 2) and downsampling of the decoded picture (marked with blue line). The following sections present the motivation and describe the proposed schemes in details.

2.1 Low complexity Coding (Scheme 1)

The specification of H.264/MVC defines Motion Compensating Prediction (MCP) with quarter-pixel (Q-pel) resolution of motion vectors. To achieve this, in each decoded image view, which is marked to be used as a reference, undergoes in-loop interpolation by a factor of 4 in the horizontal and vertical directions. The interpolated picture is stored in a frame buffer of the corresponding view and used as a reference picture for inter-prediction (temporal MCP). In addition, the reference picture of the base view can be used as a reference for inter-view prediction when a dependent view is coded. However, in the case of mixed-resolution coding, the reference picture produced in the base view is 2x larger than the reference pictures produced in dependent views and hence cannot be used in the same MCP. To solve this problem, the inter-view reference picture (Q-pel resolution) of the base view is decimated by a factor of 2 along each direction and the subsampled version is placed in the reference frame buffer of the dependent view, shown by the green line module in Figure 2.

The algorithm proposed in this section (scheme 1) has a negligible complexity increase and introduces minimal changes to the H.264/MVC architecture. It is believed that such changes can be performed by software only update to the already deployed decoding infrastructure.

However, this algorithm does not take into consideration parameters of downsampling performed to the dependent view at the post-processing stage, e.g. the low pass filter (LPF) phase, and its performance may suffer from a possible mismatch in the pixel location grid used in the base and dependent views, and aliasing, since the decimation procedure does not apply any LPF. This may lead to sub-optimal performance of the MCP in the inter-view prediction.

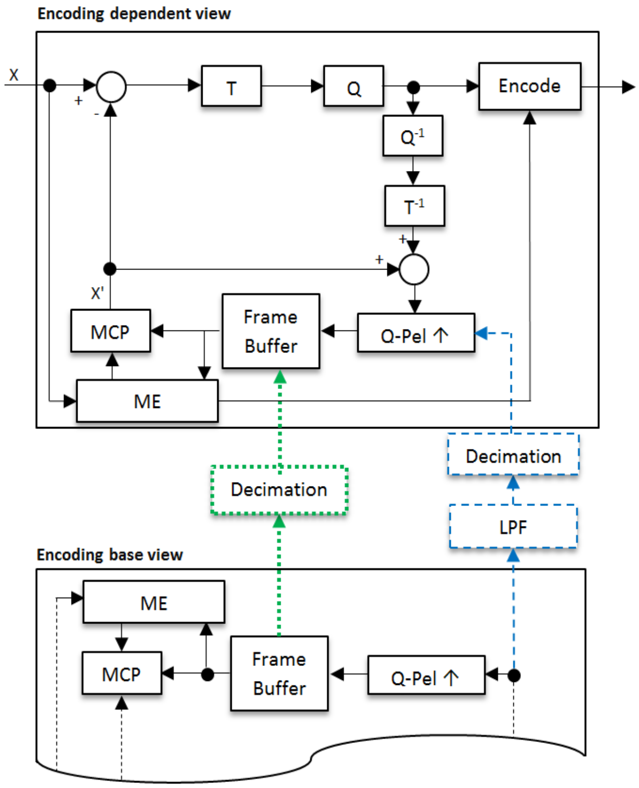

Figure 2. MVC coding for mixed-resolution video, where the proposed Scheme 1 is depicted in green box and Scheme 2 in blue

2.2 High performance Coding (Scheme 2)

To overcome the problem raised in the previous sub-section, the reference picture of the base view which is to be used for inter-view prediction should be downsampled with a proper antialiasing low pass filtering applied prior to decimation. The decoded picture of the base view is downsampled and undergoes Q-pel interpolation in the dependent view, and thus it is handled independently from the MCP chain of the base view. The proposed alternative solution is shown in Figure 2 with processing modules marked with blue dashed lines.

It is essential for in-loop downsampling applied to pictures of the base view to use an identical filter as the one used in the preprocessing of the dependent views. This will require adequate signaling in the Sequence Parameters Set: however, it will ensure an identical pixel location grid for the dependent view and the reference picture of the base view.

The algorithm proposed in this section (scheme 2) has a larger computational complexity in comparison to scheme 1, since it performs antialiasing low pass filtering and additional Q-Pel interpolation. However, the absence of aliasing artifacts along with no mismatch in pixel location grid between the coded and the reference images are expected to contribute towards efficient inter-view prediction. The simulation results provided in the next section confirm these expectations.

3. TEST MATERIAL AND SIMULATION RESULTS

Both schemes proposed in this paper (Scheme 1 and Scheme 2) were integrated to the 3DV-ATM software and compared against the anchor scheme (MVC+D). Simulations were conducted under the specifications of C3 scenario of 3DV Common Test Condition (CTC) [17] and JCT-3V/MPEG MVD test sequences were utilized. In this scenario 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 [18] were utilized to produce a full resolution anchor results. Table I summarizes the major parameters used for the 3DV-ATM configuration, whereas complete configuration files for MVC+D are available in [17].

The simulation framework for the proposed schemes (Scheme 1 and Scheme 2) is specified as shown in Table I and the following changes were introduced.

The following pre-processing and post-processing stages as shown in Figure 1 were utilized to produce simulation results for the proposed Scheme 1 and Scheme 2.

Pre-processing:

Texture views of MVD data marked to be coded as dependent views were downsampled at the pre-processing stage. The downsampling was performed with a lowpass filter used in [19]. The LPF is designed with a cut-off frequency of 0.9π and has 12 filter taps . The filter coefficients are as follows:

$$h1 = [2\ -3\ -9\ 6\ 39\ 58\ 39\ 6\ -9\ -3\ 2\ 0]/128 \qquad (1)$$

Post-processing:

Following the decoding and prior to the DIBR, the decoded dependent views were upsampled by a factor of 2 in the horizontal and vertical directions back to the original resolution. The upsampling was performed with the 6-tap H.264/AVC interpolation filter [9]. The coefficients of this

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 views (C3)
MVD resolution ratio (Texture : Depth)	1:0.5
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 [18]
Test sequences and coded, synthesized views	As specified in [17]

filter are as follows:

$$h2 = [1\ -5\ 20\ 20\ -5\ 1]/32 \qquad (2)$$

Proposed schemes:

Integration of Scheme 1 to the 3DV-ATM software was straightforward and its details were given in sub-section 2.1. Scheme 2 as described in Section 2.2 was integrated to 3DV-ATM and the filter given in equation (1) was used.

The compression efficiency of the proposed schemes was evaluated according to the CTC [17] specification. The Bjontegaard delta bitrate and delta Peak Signal-to-Noise Ratio (PSNR) metrics [20] were utilized for these purposes and the MVC+D scheme was used as the anchor. The delta bitrate reduction (dBR) is presented for the total coded views (the total bitrate of the texture and depth coding along with PSNR of the texture views) and the synthesized views (the total bitrate of the texture and depth coding along with the PSNR of the synthesized views). The PSNR of the synthesized views at the decoder side were computed against the reference view synthesis results, as specified in CTC [17] and achieved from the original uncompressed texture and depth information. The results comparing the proposed schemes against the MVC+D anchor are reported in Tables II and III. Moreover, rate-distortion (RD) curves achieved with Scheme 2 and for the synthesized views of Poznan Hall 2 sequence are depicted in Figure 3. These curves well match with dBR values presented in Table III, confirming higher efficiency of Scheme 2 against anchor.

As reported in Tables II and III, both proposed MVC+D schemes with mixed-resolution texture representation outperformed the full resolution MVC+D anchor. The low complexity Scheme 1 reduces the average coded bit rate by 1.35% compared to the anchor, whereas the average compression gain for all natural sequences (excluding Ghost Town Fly and Dancer) is more than 10% of dBR. For the synthesized views, Scheme 1 provides 12.63% dBR on average (synthetic sequences included). A possible explanation for this effect is the fact that DIBR operates at the low resolution depth map, see Table I, and therefore, rendering becomes less accurate and high frequency components of synthetic sequences may not bias the final PSNR of synthesized views.

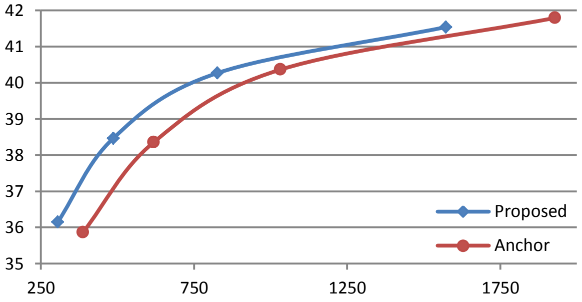

Figure 3. Rate-distortion curve for synthesized views for the sequnce Poznan Hall used in Scheme 2 against the anchor

TABLE II. PERFORMANCE OF THE PROPOSED MIXED-RESOLUTION SCHEME 1 COMPARED TO THE ANCHOR

	Coded views		Synthesized views	
	dBR, %	dPSNR, dB	dBR, %	dPSNR, dB
Poznan Hall2	-18.12	0.60	-20.22	0.75
Poznan Street	-2.16	0.00	-8.96	0.27
Undo Dancer	30.74	-1.22	-12.39	0.32
Ghost Town Fly	10.47	-0.83	-6.43	0.11
Kendo	-12.14	0.59	-14.46	0.69
Balloons	-13.35	0.68	-15.47	0.77
Newspaper	-4.90	0.17	-10.45	0.39
Average	**-1.35**	**0.00**	**-12.63**	**0.47**

TABLE III. PERFORMANCE OF THE PROPOSED MIXED-RESOLUTION SCHEME 2 COMPARED TO THE ANCHOR

	Coded views		Synthesized views	
	dBR, %	dPSNR, dB	dBR, %	dPSNR, dB
Poznan Hall2	-18.29	0.62	-20.58	0.78
Poznan Street	0.04	-0.09	-8.39	0.25
Undo Dancer	18.98	-0.88	-18.27	0.54
Ghost Town Fly	1.58	-0.41	-12.96	0.40
Kendo	-12.36	0.60	-14.88	0.71
Balloons	-13.44	0.68	-15.86	0.79
Newspaper	-4.90	0.15	-10.71	0.40
Average	**-4.06**	**0.10**	**-14.52**	**0.55**

The high performance Scheme 2, as expected, provides a larger coding gain, outperforming the MVC+D anchor by 4.06% of dBR on average for coded bitrates and by 14.52% of dBR on average for synthesized views. It should be noted that Scheme 2 significantly outperforms Scheme 1 for synthetic sequences, where the impact of aliasing artifacts and mismatch in pixel grid seem to degrade inter-view prediction in Scheme 1. On the other hand, coding performance for natural sequences seems to be very close for both Scheme 1 and Scheme 2, giving about 10% of dBR for coded views and about 14% of dBR gain for synthesized views against the MVD anchor, respectively.

4. CONCLUSIONS

The paper proposed a novel modified MVC+D coding scheme that supports the coding MVD data with a mixed-resolution texture representation. We proposed to encode only the base view at the original resolution whereas the spatial resolution of dependent views is reduced. At the decoder side, the proposed scheme consists of a post-processing scheme that performs upsampling of decoded views back to their original resolution. To enable inter-view prediction, the base view is downsampled within the MVC coding loop to provide a relevant reference for dependent views. The proposed scheme was compared against the original MVC+D and objective coding gains of 4% of average delta bitrate reduction (dBR) and 14.5% of dBR on synthesized views were reported.

5. ACKNOWLEDGEMENT

The authors would like to thank M. Domański, et al. for providing Poznan sequences and Camera Parameters [21].

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] MPEG Video and Requirement Groups, “Call for Proposals on 3D Video Coding Technology”, MPEG output document N12036, Geneva, Switzerland, March 2011

[4] http://mpeg.chiariglione.org/working_documents/explorations/3dav/3d-test-report.zip

[5] T. Suzuki, M. M. Hannuksela, Y. Chen, S. Hattori, and G. J. Sullivan (ed.), “MVC extension for inclusion of depth maps draft text 4,” Joint Collaborative Team on 3D Video Coding Extension Development, document JCT3V-A1001, July 2012

[6] 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.

[7] M. M. Hannuksela, Y. Chen, and T. Suzuki (ed.), “3D-AVC draft text 3,” Joint Collaborative Team on 3D Video Coding Extension Development, document JCT3V-A1002, Sep. 2012

[8] “Test model for AVC based 3D video coding,” ISO/IEC JTC1/SC29/WG11 MPEG2012/N12558, Feb. 2012

[9] JSVM Software http://ip.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm

[10] P. Milanfar, Ed., “Super-Resolution Imaging”. Boca Raton, FL: CRC Press, 2010.

[11] R. Blake, “Threshold conditions for binocular rivalry,” Journal of Experimental Psychology: Human Perception and Performance, vol. 3(2), pp. 251-257, 2001.

[12] P. Aflaki, M. M. Hannuksela, J. Häkkinen, P. Lindroos, M. Gabbouj, “Subjective Study on Compressed Asymmetric Stereoscopic Video,” Proc. of IEEE Int. Conf. on Image Processing (ICIP), Sep. 2010.

[13] W. J. Tam, “Image and depth quality of asymmetrically coded stereoscopic video for 3D-TV,” Joint Video Team document JVTW094, Apr. 2007.

[14] P. Seuntiens, L. Meesters, and W. IJsselsteijn, “Perceived quality of compressed stereoscopic images: effects of symmetric and asymmetric JPEG coding and camera separation,” ACM Transactions on Applied Perception, vol. 3, no. 2, pp. 95–109, Apr. 2006.

[15] H. Brust, A. Smolic, K. Müller, G. Tech, and T. Wiegand, “Mixed-resolution coding of stereoscopic video for mobile devices” 3DTV Conference, May 2009.

[16] P. Aflaki, D. Rusanovskyy, T. Utriainen, E. Pesonen, M. M. Hannuksela, S. Jumisko-Pyykkö, and M. Gabbouj, “Study of asymmetric quality between coded views in depth-enhanced multiview video coding,” International Conference on 3D Imaging (IC3D), Dec. 2011.

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

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

[19] J. Dong, Y. He, Y. Ye, “Downsampling filters for anchor generation for scalable extensions of HEVC,” ISO/IEC JTC1/SC29/WG11 MPEG2012/M23485, Feb. 2012.

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

[21] 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.