Evaluation of Depth-Based Super Resolution on Compressed Mixed Resolution 3D Video

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

[a] Tampere University of Technology, Tampere, Finland
[b] Nokia Research Center, Tampere, Finland

Abstract. The MVC+D standard specifies coding of Multiview Video plus Depth (MVD) data for enabling advanced 3D video applications. MVC+D defines that all views are coded with H.264/MVC encoder at equal spatial resolution. To improve compression efficiency it is possible to use mixed resolution coding in which part of texture views are coded at reduced spatial resolution. In this paper we evaluate the performance of Depth-Based Super Resolution (DBSR) on compressed mixed resolution MVD data. Experimental results show that for sequences with accurate depth data the objective coding performance metric increases. Even though some sequences, with poor depth quality, show slight decrease in coding performance with respect to objective metric, subjective evaluation shows that perceived quality of DBSR method is equal to symmetric resolution case. We also show that depth re-projection consistency check step of the DBSR can be changed to simpler consistency check method. In this way the DBSR computational complexity is reduced by 26% with 0.2%dBR average bitrate reduction for coded views and 0.1% average bitrate increase for synthesized views. We show that proposed scheme outperforms the anchor MVC+D coding scheme by 7.2% of dBR on average for total coded bitrate and by 10.9% of dBR on average for synthesized views.

1 Introduction

3D video consumer devices, including video cameras and displays, start to emerge on the market. To store 3D video data efficiently new compression methods are required. As a response to the growing need for 3D video compression the Moving Picture Experts Group (MPEG) initiated 3D video standardization process [5], that has been continued by the Joint Collaborative Team on 3D Video Coding (JCT-3V) since July 2012. 3D video consists of a set of 2D video sequences, registered by cameras synchronized in time. Video acquisition using many cameras is challenging and some displays like autostereoscopic displays (ASD) require many views on input. However, encoding and transmission of many views would require a great amount of processing power and bandwidth. With the help of DIBR [16] techniques it is possible to register and encode a lower amount of views and synthesize missing views from decoded texture and corresponding depth data. In addition, depth-enhanced 3D video allows for baseline adjustment on the receiver side. Variable baseline might be required to adjust to properties of some displays, viewing conditions or user preferences [20]. The Advanced Video Coding (H.264/AVC) standard [3] was earlier amended by a Multiview Video Coding (MVC) extension [3, 13], and has been recently amended by a multiview-and-depth coding extension (MVC+D) [3, 12]. MVC+D uses the MVD format and specifies that texture and depth data are coded independently but put in the same bitstream. In addition, it specifies that texture views have an equal resolution, while the depth views also share the same resolution, which may differ from the texture resolution.

In mixed resolution, stereoscopic video coding, first introduced in [21], one of the two views is coded at lower spatial resolution than the other view. According to the binocular suppression theory [9] the Human Visual System (HVS) is able to fuse stereoscopic images in way that the perceived quality is close to that of the higher quality view. The theory has been verified by the systematic subjective viewing experiments, such as [7] and [24]. Mixed resolution stereoscopic video coding allows a reduction of computational and storage resources due to one of the views having a lower spatial resolution. Thus, mixed resolution stereoscopic video makes it possible to keep the video quality close to that of full-resolution, symmetric, stereoscopic video, while the computational complexity is reduced.

In the case when the MVD format is used with texture views having mixed resolution, it is possible to improve the perceived quality by upsampling low-resolution views using a Depth-Based Super Resolution (DBSR). Since all views represent the same scene observed from different viewpoints, the video content, present in each view, is highly correlated. Thus, in case of mixed resolution 3D video, a low-resolution view can be enhanced using a neighboring high-resolution view. The first approach to improve low-resolution images in mixed resolution, stereoscopic image sequences was proposed in [22]. The method was based on a weighted averaging of the pixel intensity values coming from corresponding pixel positions in a high-resolution view. Another approach [17, 19] assumes that low-frequency components can be restored with high fidelity during the upsampling process of the low-resolution view. After that, missing high-frequency components can be extracted from the high-resolution neighboring view. Both methods project a high-resolution view to the position of the low-resolution view and use the projected view data to improve the upsampling of the low-resolution view. The resulting stereoscopic pair has one view in full resolution without any modifications and one low-resolution view upsampled with the use of projected pixel data from the full-resolution view.

Even though the first DBSR approach was introduced in [17], its performance was not evaluated on compressed mixed resolution 3D video. In this paper we evaluate the performance of the DBSR upsamling [17] on mixed resolution 3D video, encoded with a modified MVC+D codec [8], based on the recent 3D video coding standard [12]. We show that this coding and upsampling concept provides objective gains by means of the Bjontegaard delta bitrate reduction (dBR) [10] when compared to both the MVC+D coding with symmetric resolution between

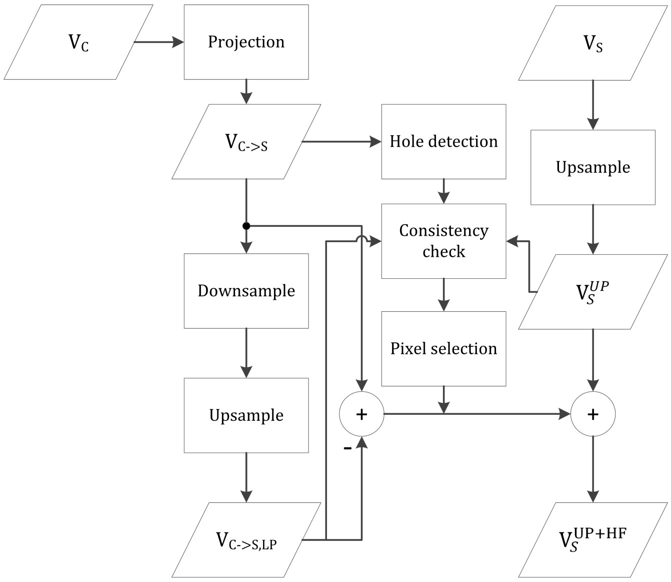

Fig. 1. Flow chart of the modified depth based super resolution algorithm.

views and mixed resolution coding with conventional upsampling. Moreover, we propose a simplified consistency check and show that depth re-projection consistency check is not only slower but, in some cases, can impact the quality of upsampled view. Finally, we show that even if the objective quality in terms of Peak Signal-to-Noise Ratio (PSNR) does not show improvement for sequences with low quality depth maps, the subjective evaluation proves that DBSR produces mixed resolution 3D video with quality equal to the symmetric resolution 3D video.

The rest of the paper is organized as follows. Section 2 presents the DBSR upsampling with the proposed consistency check. Section 3 describes coding simulation and subjective evaluation conditions. Section 4 summarizes simulation and evaluation results. Section 5 concludes the paper.

2 Depth-Based Super Resolution

The rendering of 3D video on a 3D display requires views with the same resolution. In order to render mixed resolution 3D video at high resolution, the decoded low-resolution, dependent views have to be upsampled to the base-view resolution. Since depth data and camera parameters corresponding to each view are encoded in a 3D video bitstream it is possible to use the DBSR to improve the quality of upsampling.

The flowchart for the DBSR with modified consistency check is presented in Fig. 1. The dependent, low-resolution, side view V_S is first upsampled using a conventional upsampling method, either with usage of the H.264/AVC motion interpolation filter [4] or scalable HEVC (SHVC) upsampling filter [11]. The resulting upsampled image is missing high frequency components due to low-pass filtering performed during downsampling. Since the base view V_C is transmitted at full resolution it is not low-pass filtered and high-frequency components are preserved. The depth and camera data for the base view V_C, as well as the base view V_C itself, are used to synthesize a novel view at the camera position corresponding to V_S. The synthesized view $V_{C->S}$ is downsampled and upsampled to obtain a low-pass filtered version. To extract a high-frequency component for pixel in $V_{C->S}$ image it is enough to subtract from that image its low-passed version $V_{C->S,LP}$. The consecutive downsample and upsample steps realize low-pass filtering using filters with the same coefficients as the ones used for side views during pre-processing. Since the high-frequency components of the synthesized view $V_{C->S}$ and the upsampled side view V_S^{UP} represent scene at the same camera viewpoint they can be added up to create enhanced, side view V_S^{UP+HF}.

Depth maps contain errors appearing for example due to imprecise depth estimation process [20], occlusions between camera views, transform-based coding and quantization. Due to these errors a subset of pixels from synthesized view might be projected to wrong locations. To mitigate the impact of these errors on the view synthesis quality and, consequently, on the super resolution enhancement quality, the consistency check based on minimizing the depth map reprojection error was proposed in [17].

However, the consistency check based on depth has two limitations. First - the depth re-projection is computationally costly and second - instead of measuring the inconsistencies in the image domain the depth domain is taken into account. We propose an alternative, computationally simple, solution for a consistency check realized in two consecutive steps. Firstly, hole areas detected during view synthesis are excluded since they do not contain projected pixels that can be compared to their correspondents in the dependent view. After that, a pixel-wise similarity, derived from luminance component, between the upsampled version of side view V_S^{UP} and the low-pass filtered version of center view, projected to the side view position $V_{C->S,LP}$ is calculated. The motivation behind this procedure is that these images contain corresponding pixels at the same low frequency spectrum. The similarity check is calculated within the matching window according to the following formula:

$$\sum_{x=x_0-w}^{x_0+w} \sum_{y=y_0-h}^{y_0+h} |V_{C->S,LP}(x,y) - V_S^{UP}(x,y)| < T, \quad (1)$$

where $2w$ and $2h$ are width and height of the matching window. The pixels for which the absolute difference is smaller than a threshold T are considered similar and the enhancement, using high-frequency component extracted from the synthesized view, is executed. The pixels for which similarity condition does not hold are kept intact and copied from V_S^{UP}. Our empirical study shows that threshold T is not a function of a quantization parameter and thus it is kept constant over all quantization parameters. The consistency check based on pixel-wise similarity check does not rely solely on depth re-projection and it is expected to have much lower computational complexity. Simulation results confirm these expectations.

3 Subjective Test Setup and Simulation Environment

The modified 3DV-ATM reference software [8] was used to encode mixed resolution 3D video. The base view is coded at original resolution and dependent views are coded at half resolution in vertical and horizontal direction. Hence, the base views of the created bitstreams conform to H.264/AVC, while the dependent views are compatible with the MVC extension [3] except for the re-sampling of the decoded base view pictures for inter-view prediction, as described in [8]. Texture and depth views were coded into the same bitstream using the MVC+D bitstream syntax. The coding and view synthesis simulations were executed under the Common Test Conditions (CTC) [1] on the test sequences specified by the same document. The three-view (C3) coding scenario was selected in which three texture views accompanied by corresponding depth data are encoded with center view coded as the base one. A set of three additional views between each pair of adjacent coded views was synthesized using the view synthesis reference software (VSRS) described in [23], intended to serve for multiview autostereoscopic displays. The synthesized views PSNR is calculated against the views at same camera positions, rendered from uncompressed texture and depth data.

The anchor bitstreams were created with the 3DV-ATM software [6, 18], using the MVC+D configuration with full resolution for the base and dependent views, in C3 scenario. The major settings for 3DV-ATM are summarized in Tab. 1. To obtain low resolution versions of side views they were downsampled with use of the 12-tap low-pass filter [15] with a cut-off frequency of 0.9π and the following coefficients:

$$[2 \quad -3 \quad -9 \quad 6 \quad 39 \quad 58 \quad 39 \quad 6 \quad -9 \quad -3 \quad 2 \quad 0]/128 \quad . \tag{2}$$

After that, the coding process of mixed resolution 3D video was conducted using the modified 3DV-ATM software [8] with the same settings as specified in

Table 1. Major 3DV-ATM configuration settings.

Coding Parameter	Setting
Texture : Depth width ratio	1 : 0.5
Texture : Depth height ratio	1 : 0.5
Inter-view prediction structure	PIP
Inter prediction structure	HierarchicalB, GOP8
QP settings for texture and depth	26, 31, 36, 41
Encoder optimization settings	RDO ON, VSO ON
View Synthesis in post-processing	Fast 1D VSRS [19]

Tab.1. The only difference was to set width and height of the dependent, texture views to half comparing to the anchor bitstream settings specified in CTC [1].

Two types of upsampling algorithms were executed on the decoded dependent views of the mixed resolution 3D video. First type was 8-tap upsampling filter used in the scalable HEVC [11] with the following filter coefficients:

$$[-1 \quad 4 \quad -11 \quad 40 \quad 40 \quad -11 \quad 4 \quad -1]/64 \quad . \tag{3}$$

In the second upsampling scheme, the modified DBSR algorithm presented in Section 2. was used. In the first upsampling step that creates V_S^{UP}, the same 8-tap scalable HEVC filter [11] was used. This approach guarantees fair comparison between the two tested upsampling schemes. A threshold $T = 15$ for the proposed consistency check serving for the DBSR scheme was empirically found the best from the set $\{5, 10, 15, 20\}$.

The execution times of the tested algorithms were measured in seconds on the same computer with 1 process per CPU unit. The simulation framework was set up on the Linux operating system with minimial set of services and graphical user interface shut down.

The subjective evaluation experiment was carried out using four sequences: Poznan Hall2, Poznan Street, Kendo and Balloons [1]. Four quantization parameters (QP) were used according to CTC. For the purpose of subjective test the sequences encoded with QPs 26 and 36, namely the lowest and second highest, were selected. The naive subjects compared 3D stereoscopic videos generated from views specified according to [1]. The polarized Sony Bravia 55" 3D display was used for subjective viewing. The viewing distance was equal to 4 times the displayed image height (2.72m). Subjective quality assessment was conducted according to the Double Stimulus Impairment Scale (DSIS) method [2] with a discrete scale from 0 to 10 for quality assessment. Prior to each test, subjects were familiarized with the test task, the test sequences, and the expected variation in quality. The subjects were instructed that 0 stands for the lowest quality and 10 for the highest. Moreover, duration of each session was limited to half an hour to prevent the subjects from experiencing fatigue or eye strain. The test sequences were played in a random order and each video clip was played twice to increase the accuracy of the evaluation. Subjective viewing was conducted with 20 subjects, (14 males and 6 females), aged between 23 and 32 years with mean 27 years. All subjects passed the stereovision test prior to the 3D viewing.

4 Results and Discussion

To compare the objective quality of the test sequences the Bjontegaard delta bitrate (dBR) [10] and Peak Signal-To-Noise (PSNR) ratio were used. The dBR results, for each scheme tested, are shown in two columns and were calculated as following. For the 'coded views' column an aggregated bitrate of coded texture and depth data versus average PSNR of the coded texture was used. For the 'synthesized views' column an aggregated bitrate of coded texture and depth data versus average PSNR of three synthesized views between each adjacent

Table 2. Performance of mixed resolution 3D video coding with use of the SHEVC upsampling method.

Sequence	Index	Coded views		Synthesized views	
		dBR [%]	dPSNR [dB]	dBR [%]	dPSNR [dB]
PoznanHall2	S01	-18.36	0.65	-20.33	0.79
PoznanStreet	S02	0.61	-0.10	-5.77	0.16
UndoDancer	S03	21.26	-0.91	3.49	-0.24
GhostTownFly	S04	1.75	-0.40	-5.46	0.05
Kendo	S05	-12.61	0.60	-15.14	0.70
Balloons	S06	-13.54	0.69	-16.08	0.80
Newspaper	S08	-2.74	0.07	-6.58	0.22
Shark	S10	-3.13	0.07	-9.37	0.37
Average		**-3.35**	**0.08**	**-9.41**	**0.35**

pair of coded views for a given sequence was used. The objective results of mixed resolution coding are presented in Tab. 2 and Tab. 3. The anchor was generated using the 3DV-ATM software [6, 18] that uses symmetric-resolution for texture coding. The results in Tab. 2 correspond to the scalable HEVC upsampling. In Tab. 3 results pertain to the DBSR method [17] and its variation presented in Section 2. It can be seen that both DBSR upsampling methods improve coding efficiency significantly both for coded and synthesized views. Our simulations show that the proposed DBSR with simplified consistency check executes 26% faster for all sequences in the test set. Furthermore, gain for sequences with high accuracy depth, such as Undo Dancer and Ghost Town Fly, is higher in case of the proposed consistency check, which signals better performance of the proposed consistency check.

Table 3. Performance of mixed resolution 3D video coding with use of the two versions of DBSR.

Seq.	DBSR [17]				proposed DBSR			
	Coded views		Synthesized views		Coded views		Synthesized views	
	dBR[%]	dPSNR[dB]	dBR[%]	dPSNR[dB]	dBR[%]	dPSNR[dB]	dBR[%]	dPSNR[dB]
S01	-17.21	0.59	-19.19	0.74	-17.14	0.59	-19.02	0.73
S02	-0.46	-0.05	-4.46	0.11	-0.72	-0.03	-4.40	0.11
S03	0.75	-0.09	-7.67	0.23	-0.53	-0.04	-8.04	0.25
S04	-7.26	0.19	-11.71	0.42	-9.22	0.30	-12.54	0.46
S05	-11.64	0.54	-13.86	0.63	-11.02	0.50	-13.32	0.60
S06	-12.83	0.64	-14.96	0.74	-12.17	0.60	-14.34	0.70
S08	-1.68	0.02	-5.41	0.17	-0.99	-0.01	-4.75	0.15
S10	-5.86	0.26	-10.74	0.46	-5.76	0.25	-10.52	0.45
Average	**-7.02**	**0.26**	**-11.00**	**0.44**	**-7.19**	**0.27**	**-10.87**	**0.43**

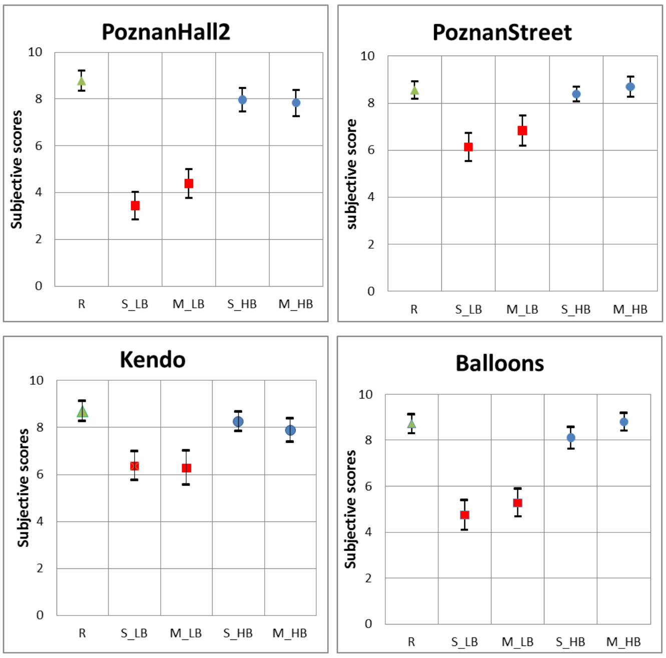

Fig. 2. Viewing experience MOS with 95% confidence interval. Abbreviations correspond to R: Reference, S: Symmetric resolution, M: Mixed resolution, LB: Low bitrate, HB: High bitrate.

The results tabularized in Tab. 2 and Tab. 3 show also that the DBSR upsampling performance, measured by an objective dBR metric, does not show improvement for sequences with less accurate depth quality such as Poznan Hall2, Kendo, Balloons and Newspaper. In order to verify the performance of DBSR upsampling for these sequences the subjective quality evaluation test was executed with results presented in Figure 2. The values in plots correspond to mean opinion scores (MOS) with 95% confidence interval. The reference sequences, abbreviated as R, were generated using original not compressed texture and depth data. The symmetric resolution sequences, abbreviated as S were generated using unmodified 3DV-ATM software [6, 18] with same resolution parameters for all texture views. The mixed resolution sequences, abbreviated as M, were generated using modified 3DV-ATM software [8] with mixed resolution texture views. The results are presented for low bitrate (LB) wih QP parameter equal to 36 and high bitrate with QP equal to 26. Results in Figure 2 show that perceived quality of DBSR upsampling is perceived equal or, in some cases slightly better then symmetric resolution coding. To test the significant difference between the scores the Wilcoxon's signed rank test [25] is used. The significance level was set to $p = 0.05$. The results of the Wilcoxon's test, tabularized in Tab. 4. show

Table 4. Wilcoxon's signed rank test. Pairwise comparison between symmetric and mixed resolution MOS scores with statistical difference level $p < 0.05$.

Sequence	p - value	
	low bitrate	high bitrate
Poznan Hall2	0.00	0.46
Poznan Street	0.02	0.20
Kendo	0.65	0.22
Balloons	0.15	0.00

that mixed resolution for 3 cases was perceived better than symmetric resolution and in all other cases there was no statistically significant difference in perceived quality between tested schemes. Based on these results it can be concluded that perceived quality between two schemes: symmetric resolution and mixed resolution with DBSR upsampling, cannot be subjectively differentiated.

5 Conclusions

Since the introduction of the DBSR algorithm [17] its performance was not analyzed on compressed mixed resolution 3D video. This paper presents the objective and subjective evaluation of the DBSR upsampling scheme executed on compressed mixed resolution 3D video. The experiments show that even though, some sequences with low accuracy depth maps, show slight decrease in coding efficiency, the perceived quality of DBSR-upsampled views is equal to the symmetric case. In addition improvement to the consistency check step is introduced that speeds up the total execution time of the DBSR algorithm by 26% and slightly improves coded views delta bitrate. We believe that complete, objective and subjective, performance analysis of the DBSR upsampling approach shows that it is a viable solution for the mixed resolution 3D video coding.

Acknowledgement. Authors would like to thank professor Marek Domanski et al. for sharing the 3D video test sequences [14] and Maryam Homayouni for the help with subjective test arrangement.

References

1. Common test conditions of 3DV core experiments. ISO/IEC JTC1/SC29/WG11 JCT3V-E1100 (2013)
2. ITU-R Rec. BT.500-11, Methodology for the subjective assessment of the quality of television pictures. (2002)
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) 2013
4. JSVM Software http://ip.hhi.de/imagecom_G1/savce/downloads/SVCReference-Software.htm

5. MPEG Video and Requirement Groups: Call for Proposals on 3D Video Coding Technology. MPEG output document N12036, Geneva, Switzerland, (2011)
6. Test model for AVC based 3D video coding, ISO/IEC JTC1/SC29/WG11 MPEG2012/N12558, (2012)
7. Aflaki, P., Hannuksela, M. M., Gabbouj, M.: Subjective quality assessment of asymmetric stereoscopic 3D video. Springer Journal on Signal, Image and Video Processing (2013)
8. Aflaki, P., Su, W., Joachimiak, M., Rusanovskyy, D., Hannuksela, M. M., Gabbouj, M.: Coding of mixed-resolution multiview video in 3D video application. IEEE International Conference on Image Processing (ICIP), (2013)
9. Asher, H.: Suppression theory of binocular vision. British Journal of Ophthalmology **37(1)** (1953) 37-49
10. Bjontegaard, G.: Calculation of average PSNR differences between RD-Curves. ITU-T SG16 Q.6 document VCEG-M33 (2001)
11. Chen, J., Boyce, J., Ye, Y., Hannuksela, M.: Scalable HEVC (SHVC) Test Model 4 (SHM 4) ISO/IEC JTC1/SC29/WG11 MPEG2013/N13939 (2013)
12. Chen, Y., Hannuksela, M. M., Suzuki, T., Hattori, S.: Overview of the MVC+D 3D video coding standard. Journal of Visual Communication and Image Representation (2013)
13. Chen, Y., Wang, Y. K., Ugur, K., Hannuksela, M. M., Lainema, J., Gabbouj, M.: The emerging MVC standard for 3D video services. EURASIP Journal on Applied Signal Processing **8** (2009)
14. Domanski, M., Grajek, T., Klimaszewski, K., Kurc, M., Stankiewicz, O., Stankowski, J., Wegner, K.: Poznan Multiview Video Test Sequences and Camera Parameters. ISO/IEC JTC1/SC29/WG11 MPEG 2009/M17050 (2009)
15. Dong, J., He, Y., Ye, Y.: Downsampling filters for anchor generation for scalable extensions of HEVC. ISO/IEC JTC1/SC29/WG11 MPEG2012/M23485 (2012)
16. Fehn, C.: 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, CA, U.S.A., **5291** (2004) 93104
17. Garcia, D. C., Dorea, C., Queiroz, R. L.: Super resolution for multiview images using depth information. IEEE Trans. Circuits Syst. Video Technology **22(9)** (2012) 1249-1256
18. Hannuksela, M. M., Rusanovskyy, D., Su, W., Chen, L., Li, R., Aflaki, P., Lan, D., Joachimiak, M., Li, H., Gabbouj, M.: Multiview-video-plus-depth coding based on the advanced video coding standard.. IEEE Transactions on Image Processing **22(9)**, (2013) 3449-3458
19. Joachimiak, M., Hannuksela, M. M., Gabbouj, M.: View synthesis quality mapping for depth-based super resolution on mixed resolution 3D video. In 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON) (2014)
20. Kauff, P., Atzpadin, N., Fehn, C., Muller, M. , Schreer, O., Smolic, A., Tanger, R.: Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability. EURASIP International Journal on Signal Processing **22(2)** (2007)
21. Perkins, M. G.: Data compression of stereopairs. IEEE Transactions on Communications **40(4)** (1992) 684-696
22. Sawhney, H. S., Guo, Y., Hanna, K., Kumar, R., Adkins, S., Zhou, S.: Hybrid Stereo Camera: An IBR Approach for Synthesis of Very High Resolution Stereoscopic Image Sequences. in Proc. of the 28th Conf. on Computer Graphics and Interactive Techniques, New York, USA (2001) 451-460

23. Schwarz, H., et al: Description of 3D Video Technology Proposal by Fraunhofer HHI (MVC compatible). ISO/IEC JTC1/SC29/WG11 MPEG2011/M22569 (2011)
24. Tam, W. J.: Image and depth quality of asymmetrically coded stereoscopic video for 3D-TV. Joint Video Team document JVT-W094 (2007)
25. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics **1** (1945) 80-83