PERCEPTUAL QUALITY ASSESSMENT OF HEVC MAIN PROFILE DEPTH MAP COMPRESSION FOR SIX DEGREES OF FREEDOM VIRTUAL REALITY VIDEO

Sebastian Schwarz, Miska M. Hannuksela

Nokia Technologies, Tampere, Finland

ABSTRACT

The last years have shown significant advances in immersive media. Virtual reality video, in the form of spherical panoramic video, is already widely available. Such technologies will continually evolve towards the ultimate goal of a truly virtual reality experience. The first step in this direction is the support of limited translational head movement for restricted six degrees of freedom virtual reality video. This experience can be achieved by rendering virtual viewports from supplementary depth information. In this context, this paper investigates the effect of depth map compression on the perceptual quality of immersive media. The presented study is focused on near-future applications with readily available hardware. Objective quality assessments indicate possible bit rate savings of 17% through high-quality depth maps. However, these findings could not be confirmed subjectively. On the contrary, subjective evaluation shows a robustness to low-quality depth maps when viewed in a virtual reality scenario.

Index Terms— virtual reality video, six degrees of freedom, 6DOF, depth map compression, perceptual quality

1. INTRODUCTION

Virtual reality (VR) applications have gained a lot of attention in recent years. With the rise of affordable head-mounted displays (HMDs), media experiences have become more and more immersive. And with the advances in camera technology, many of these applications move away from computer generated imagery (CGI) towards real world video. While omnidirectional, or 360-degree, video has just entered the market, people are already looking at the next step towards a truly immersive media experience: six degrees of freedom (6DOF) video [1].

Conventional omnidirectional video is typically implemented as spherical panoramic video, providing the viewer with three degrees of freedom: yaw, pitch, and roll at a central viewing position, i.e. head rotation around y-, x-, and z-axis respectively [2]. For added immersion, this concept can be extended by translational movements along the y-, x-, and z-axes, delivering 6DOF video. Apparently, spherical panoramic video on its own does not carry the necessary information to support such translational movements, and additional data is required. Various 6DOF video formats are currently under consideration: depth-enhanced (multiview) video [3], volumetric representations, such as voxels [4] or point clouds [5], or light fields [6, 7]. For this paper, the emphasis is set on near-term 6DOF technology with limited translational movement. For such applications, depth-enhanced video provides the necessary information to synthesize virtual viewports as the viewer will typically experience in a seated position use case. Within this scope, various depth map encoding schemes are evaluated with respect to perceptual quality in a 6DOF viewing scenario.

The compression of depth-enhanced video has been widely studied for 3D video, and a specific video codec is available (3D-HEVC [8]). However, for this paper, the scope is further limited to standard HEVC Main Profile compatible encoding to allow easy implementation on readily available hardware. Several depth map compression approaches are evaluated, such as an increase or decrease of depth encoding quality, a decrease of depth map spatial resolution, and combinations of the above. While objective perceptual quality assessment showed up to 17% bit rate reduction, these results could not be confirmed subjectively. Nonetheless, the research presented in this paper provides valuable insights on depth map compression for 6DOF applications and the difficulties of reliably assessing perceptual quality for immersive video. To the best of our knowledge, this is the first study on the impact of depth map compression on the perceptual quality of 6DOF video.

The remainder of this paper is structured as follows. Sec. 2 addresses the limited 6DOF use case in more detail. The experiment testbed and methodology is described in Sec. 3. Objective and subjective results are discussed in Sec. 4. Results of the objective and subjective quality assessment are presented in Sec. 4 before Sec. 5 concludes this paper.

2. SIX DEGREES OF FREEDOM VR VIDEO

Fully-fledged 6DOF video applications require an abundance of data and are still a couple of years in the future. More near-term applications restrict the translational

(a)

(b) (c)

Fig 1. Monoscopic spherical panoramic input texture and depth input (a) and example of virtual viewport, with DIBR artefacts (b), and with simple artefact handling (c).

movement to a couple of centimeters around a central viewing position. The typical use case is a seated, stationary VR experience with limited head movement. Within a small viewing volume, virtual viewports can be synthesized by depth-image-based rendering (DIBR [9]). DIBR is an established concept for planar, conventional 3D-TV and can be quickly adapted for immersive video by adding supplementary depth information. Fig. 1 (a) illustrates the required input in the form of spherical panoramic texture and depth maps. Fig. 1 (b) shows a resulting virtual viewport and illustrates the principal drawback of DIBR nicely: Only areas represented in the texture and depth inputs can be synthesized. Any previously hidden parts of the scene result in holes, so-called disocclusion artifacts. Such artifacts become more prominent with larger viewport translation. Thus, the DIBR approach with only monoscopic or stereoscopic inputs is less feasible for unrestricted 6DOF video. However, in the restricted 6DOF use case, simple hole filling algorithms can achieve acceptable results, as shown in Fig. 1 (c).

3. EXPERIMENT

Regarding compression of DIBR input data, there has been a lot of work conducted on texture and depth video coding during the development of 3D-HEVC [8]. However, there are currently no readily available hardware 3D-HEVC decoders. Also, current activities on the compression of spherical panoramic texture video focus mostly on different projection formats. Therefore, the study presented in this paper addresses HEVC Main Profile compatible depth map compression alone, using the established equirectangular panorama (ERP) projection. Thus, any advances in projection formats can be translated directly to the presented results, and existing hardware solutions can be used. The testbed and the methodology are described as follows.

Tab. 1: Specifications of test sequences

sequence	source	resolution	frames	fps	zMin	zMax
BearAttack	camera	3840×1920	300	30	0.5	200
PoleVaultSide	camera	3840×1920	300	30	0.6	200
MusicVideo	camera	3840×1920	300	30	0.5	78
GT Sheriff	CGI	7680×2160	600	150	1.5	150
PartyScene	CGI	8192×4096	300	30	0.5	200

3.1. Dataset

Due to the necessary depth data, there is limited test data for DIBR-based 6DOF available. Fortunately, the Nokia OZO software provides stereoscopic 360-degree video with supplementary depth maps. A total of five test sequences was selected. Three real-world 360-degree videos, and two CGI sequences. Sequence specifications, including minimum and maximum depth range in meters, are presented in Tab. I. All sequences are stereoscopic and have a bit depth of 8 bits for texture and depth. Depth maps are quantized following the concept described in [9].

3.2 Depth map encoding test points

Texture and depth map sequences were encoded separately with HM16.12 [10], using random access encoding structure with a group of pictures (GOP) size of eight and Instantaneous Decoding Refresh (IDR). All other encoding parameters followed the common test conditions (CTC [11]). Four texture encoding test points were set by quantization parameter (QP) values of 22, 27, 32, and 37. For each texture quality test point the respective depth map sequence was encoded as stated by one of the following approaches:

- ***Reference:*** *Depth at same QP and resolution as texture.*
- ***LowQuality:*** *QP +3, +6, +9, +12, at same resolution.*
- ***HighQuality:*** *QP -1 to -6, at same resolution.*
- ***LowRes:*** *Depth downsampling by factor 2 and 4 in both direction, by factor 2 in y-direction only, and by factor 2 in x- and 4-in y-direction. Same QP value as texture.*
- ***Combination:*** *Encoding low-resolution depth (down sampled by factor 2 in both directions, and factor 2 only in y-direction) at a higher quality (QP offsets -3,-5, -7)*

It should be noted that depth down sampling was performed by nearest neighbor interpolation to avoid artificial depth values. Respective depth up sampling was performed by bicubic interpolation. No particular texture guided filtering was performed.

A total of 84 test points were evaluated, 21 depth encoding test points at four texture test points. Bit rates were calculated over the all frames, summing up stereoscopic texture and depth into a single value per test point. The calculation of the respective distortion values is described in the next subsection.

3.3. Objective quality assessment

Various distortion metrics are currently under consideration to evaluate the effect of compression on spherical panoramic video content [12, 13]. However, none of these do consider 6DOF application. To assess the effect of depth compression on DIBR, one has to look at the resulting virtual viewports. To complicate things further, low-quality depth data and hole filling algorithms may lead to synthesis artifacts which obscure objective quality metrics. The following procedure was applied to regulate objective quality assessment for 6DOF content and limit the effect of synthesis artifacts on objective scores:

For each test point, 625 virtual viewports were synthesized, representing 25 different head rotations at 25 different head positions. The head positions cover a limited viewing volume of 12 cm to the left and right and 6 cm in the remaining four directions. Each viewport had a 90° field of view (FOV) and a resolution of 1024×1024 pixels. A reference viewport was synthesized from uncompressed texture and depth data. The virtual viewport for each test point was then compared to this reference using luma peak signal-to-noise ratio (PSNR). To reduce the effect of synthesis artifacts, no hole filling was applied, and any unknown values in either the reference or test point were excluded from the mean squared error (MSE) calculations. To limit the number of syntheses to a reasonable amount, only the first nine frames of each sequence were evaluated. PSNR values were averaged to derive a distortion value per test point. The impact of the depth compression approaches listed in Section 3.2 on the synthesized viewport quality was then assessed using the Bjontegaard delta bit rate (BD-BR [14]).

3.4. Subjective quality assessment

Since any objective metric may have difficulties in identifying common DIBR artifacts, a subjective quality assessment was performed. Because subjective testing is time-consuming, it was limited to three test points and two test sequences: reference depth encoding plus the two best-performing depth encoding test points found in the objective quality assessment, for the sequences *GT Sheriff* and *MusicVideo*.

Tab. 2: Objective depth coding performance

approach	test point	avg. Y-PSNR BD-BR
LowQuality	QP+3	18.4%
LowQuality	QP+6	38.9%
LowQuality	QP+9	72.2%
LowQuality	QP+12	123.8%
HighQuality	QP-1	-3.1%
HighQuality	QP-2	-7.2%
HighQuality	QP-3	-11.2%
HighQuality	QP-4	-13.5%
HighQuality	**QP-5**	**-17.2%**
HighQuality	QP-6	-14.0%
LowRes	0.5X / 0.5Y	35.7%
LowRes	0.25X / 0.25Y	107.0%
LowRes	1X / 0.5Y	20.3%
LowRes	0.5X / 0.25Y	73.1%
Combination	0.5X / 0.5Y, QP-3	32.2%
Combination	0.5X / 0.5Y, QP-5	11.7%
Combination	0.5X / 0.5Y, QP-7	7.3%
Combination	1X / 0.5Y, QP-3	21.5%
Combination	1X / 0.5Y, QP-5	3.6%
Combination	**1X / 0.5Y, QP-7**	**-2.9%**

Tab. 3: Subjective depth coding performance

	MOS BD-BR (mean)		
test point	*GT Sheriff*	*MusicVideo*	average
QP-5	20.2%	-5.4%	7.4%
1X / 0.5Y, QP-7	**-2.6%**	**-27.7%**	**-15.1%**

The subjective quality assessment was performed on an HMD (Oculus Rift CV1, 110° FOV, 1080×1200 pixel per eye). Virtual stereoscopic viewports were rendered in real-time according to the subjects head movement. Depth-weighted interpolation was applied to fill disocclusion and mapping artifacts, as illustrated in Fig. 1 (c). Because the test platform does not support uncompressed stereoscopic inputs, only a static frame was evaluated (frame 6 for all sequences). The subjective quality assessment was performed by ten subjects (9 male and one female) with an average age of 37.5 years and an average interpupillary distance (IPD) of 64.5mm. All subjects were screened for adequate visual acuity, as well as color and stereoscopic vision. Perceptual quality was assessed following the double stimulus impairment scale (DSIS [15]). Each subject was asked to rate the visual impairment between virtual viewports rendered from uncompressed texture and depth inputs and from the evaluated encoding test point inputs, on a scale from 5 (imperceptible) to 0 (very annoying). The order of test points was randomized for each subject to avoid

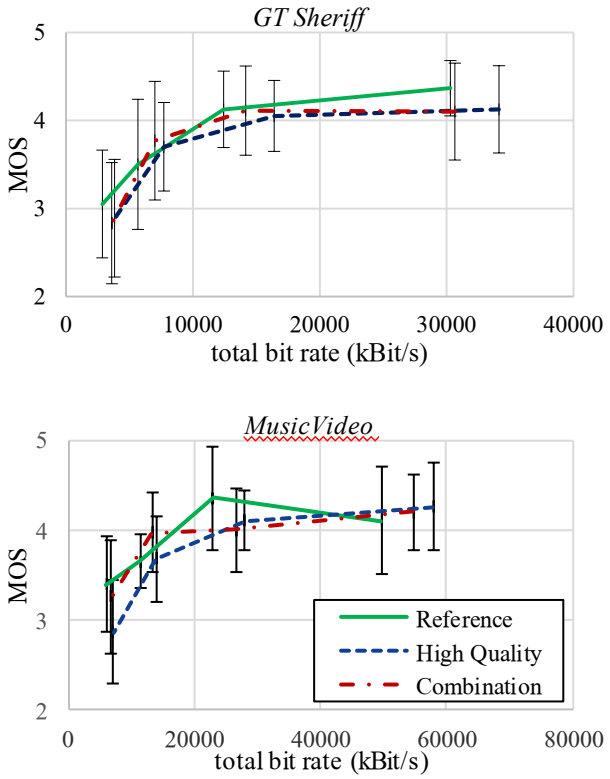

Fig 2. Rate-quality curves with 95% confidence intervals for the performed subjective perceptual quality assessment of 6DOF depth map compression.

any contextual effects on the subjective scores. Four training sequences were displayed beforehand to accustom the subject to the 6DOF viewing scenario and stabilize the quality scores. In addition, hidden references were added for each test sequence to help detect possible outliers. Mean opinion scores (MOS) and 95% confidence intervals were calculated for each test point. Then statistical outlier detection was performed, following ITU recommendations [15]. No outliers had to be removed.

4. RESULTS

Tab. 2 summarizes the objective evaluation results regarding BD-BR. Encoding depth at a lower quality or lower resolution as the corresponding texture yields an increase in required bit rate. The impact on the objective quality of the rendered viewports is so severe that it outweighs the benefits of the reduced bit rate for the depth maps. Depending on the sequence, depth maps require only around 10-30% of the texture bit rate. Thus respective overall savings are rather small. Consequently, higher quality depth map encoding has less negative impact on the overall bit rate and the resulting better synthesis quality results in an actual increase in required bit rate. As much, that distortions introduced by slight depth down sampling can be compensated by encoding depth maps at a higher quality. Based on these results, the two test points marked in bold were selected for subjective quality assessment.

BD-BR results calculated from the average MOS value for each test point with respect to the average MOS for the reference are presented in Tab. 3. The respective rate-quality curves including 95% confidence intervals are shown in Fig. 2. It is evident that the confidence intervals overlap and the objective evaluation results cannot be confirmed. Furthermore, looking at the average MOS values ignoring statistical confidence, the results in Tab. 3 are in clear contradiction to the objective test results in Tab. 2. The effect of low-quality depth maps on the virtual view synthesis seems to become less relevant when evaluated subjectively, essentially reversing the findings from the objective quality assessment. Then again, the confidence intervals are so large that no certain conclusion can be drawn. In the case of the *MusicVideo* test sequence, there is hardly any significant statistical difference between the lowest and highest quality test points in any of the evaluated approaches. This issue could be related to the actual quality of the used depth maps, as this is a real-world sequence with estimated depth maps. Whereas *GT Sheriff*, a CGI sequence with perfect depth maps, shows smaller confidence intervals and less overlap. Considering this effect together with the results in Tab. 3, it seems that real-world estimated depth maps are more robust to compression if evaluated subjectively in a 6DOF scenario.

5. CONCLUSION

This paper presented the objective and subjective evaluation of HEVC main profile compatible depth map compression on the perceptual quality of 6DOF immersive video. In this context, limited 6DOF was achieved by DIBR. Objective results pointed at a benefit for encoding depth at a higher quality as the corresponding texture. However, the subjective evaluation could not confirm this. On the contrary, depth map quality seems to have a lower impact on the 6DOF experience when evaluated subjectively. Thus, bit rate savings of up to 30% could be achieved without a major effect on perceptual quality.

In conclusion, the study presented in this paper provides insights into the challenges faced with immersive media quality assessments. Future work in this area could concentrate more on the effect of low-quality depth maps in 6DOF applications and the development of more reliable objective quality metrics for such use cases.

REFERENCES

[1] G. Teniou, “Video formats for VR: A new opportunity to increase content value… But what is missing today?” MPEG Workshop on Global Media Technology Standards for an Immersive Age, Geneva, 2017.

[2] B. Choi, Y.-K. Wang, M. M. Hannuksela, Y. Lim, “Text of ISO/IEC 23000-20 CD Omnidirectional Media Application Format,” ISO/IEC JTC1/SC29/WG11 N16636, Geneva, 2017.
[3] P. Merkle, A. Smolic, K. Muller, T. Wiegand, “Multi-View Video Plus Depth Representation and Coding,” IEEE International Conference on Image Processing, 2007.
[4] C. Loop, C. Zhang, Z. Zhang, “Real-Time High-Resolution Sparse Voxelization with Application to Image-Based Modelling,” High-Performance Graphics Conference, 2013.
[5] R. Mekuria, K. Blom, P. Cesar, “Design, Implementation and Evaluation of a Point Cloud Codec for Tele-Immersive Video,” IEEE TCSVT, Vol. PP, Issue 99, 2016.
[6] I. Ihrke, J. Restrepo L. Mignard-Debise, “Principles of Light Field Imaging: Briefly Revisiting 25 Years of Research,” IEEE Signal Processing Magazine, Vol. 33-5, 2016.
[7] G. Alves, F. Pereira, E. A. B. da Silva, “Light Field Imaging Coding: Performance Assessment Methodology and Standards Benchmarking,” 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2016.
[8] G. Tech, Y. Chen, K. Müller, J.-R. Ohm, A. Vetro, and Y.-K. Wang, “Overview of the Multiview and 3D Extensions of High Efficiency Video Coding”, IEEE TCSVT, Vol. 26-1, 2015.
[9] C. Fehn, “3D-TV Using Depth-Image-Based Rendering (DIBR),” International Picture Coding Symposium (PCS), 2004.
[10] HEVC Test Model [online] https://hevc.hhi.fraunhofer.de
[11] K. Sharman, K. Sühring, “Common Test Conditions for HM,” JCTVC-X1100, Geneva, 2016.
[12] J. Boyce, E. Alshina, A. Abbas, Y. Ye, “Common Test Conditions and Evaluation Procedures for 360° Video Coding,” ISO/IEC JTC1/SC29/WG11 N16515, Chengdu, 2016.
[13] E. Upenik, Martin Rerabek, T. Ebrahimi, “A Testbed for Subjective Evaluation of Omnidirectional Visual Content,” International Picture Coding Symposium (PCS), 2016.
[14] G. Bjøntegaard, “Calculation of average PSNR differences between RD-curves,” VCEG-M33, Austin, TX, USA, 2001
[15] ITU-R BT.500-13, “Methodology for the Subjective Assessment of the Quality of Television Pictures,” ITU 2012.