The Study of the Video Encoder Efficiency in Decoder-Side Depth Estimation Applications

Adam Grzelka adam.grzelka
Adrian Dziembowski adrian.dziembowski
Dawid Mieloch dawid.mieloch
Marek Domański marek.domanski
@put.poznan.pl
Institute of Multimedia Telecommunications
Poznań University of Technology
Polanka 3
61-131 Poznań
Poland

ABSTRACT

The paper presents a study of a lossy compression impact on depth estimation and virtual view quality. Two scenarios were considered: the approach based on ISO/IEC 23090-12 coder-agnostic MPEG Immersive video standard, and the more general approach based on simulcast video coding. The commonly used compression techniques were tested: VVC (MPEG-I Part 3 / H.266), HEVC (MPEG H part 2 / H.265), AVC (MPEG 4 part 10 / H.264), MPEG-2 (MPEG 2 part 2 / H.262), AV1 (AOMedia Video 1), VP9 (AOMedia VP9). The quality of virtual views generated from the encoded stream was assessed by the IV-PSNR metric which is adapted to synthesized images. The results were presented as a relationship between virtual view quality and the quality of decoded real views. The main conclusion from performed experiments is that encoding quality and virtual view quality are encoder-dependent, therefore, the used video encoder should be carefully chosen to achieve the best quality in decoder-side depth estimation.

Keywords

Multiview Video, Immersive Video Encoding, Depth Estimation, Virtual View Synthesis

1 INTRODUCTION

In the immersive video, a viewer has an opportunity to change his/her position and orientation in a three-dimensional scene. It enables fully immersive virtual navigation using head-mounted displays or a more simple change of viewpoint displayed on a traditional screen. In order to provide virtual views to the final user, it is required to acquire a scene from a number of views and estimate its three-dimensional geometry. As these views and geometry (usually represented in the form of depth maps) have to be sent to the renderer which generates the requested viewpoint, they usually are compressed using dedicated immersive video codecs, or simply using versatile video codecs. Lossless encoding has limited applications because even after compressing these data, the sufficient bitrate required to send it is usually in the range between 5 and 50 Mbps [Boy21][Fis20].

One of the possible solutions for decreasing the bitrate of immersive video is the estimation of geometry (depth) in the decoder, using the decoded views. This scheme of compression was already proved to be efficient in many applications [Gar21] and was included as one of the profiles of the new MPEG Immersive video (MIV) coding standard [Boy21], called MIV Geometry Absent (GA) [Mie22]. All of the profiles are codec-agnostic, i.e., after the initial pre-processing of input data, they are utilizing the traditional video encoders to encode the MIV representation.

While the MIV standard makes it possible to use any available video encoder, during the works of ISO/IEC MPEG it was mainly tested and tuned using other newest codecs from this group. The works presented in this paper were performed to find the answer to two questions related to the codec-agnosticism of MIV. First of all, what is the performance of MIV GA with other video codecs not related to MPEG standards? Secondly, how does using these different encoders impact the efficiency of different implementations of a decoder-side depth estimation scheme?

The paper is organized as follows: Section 2 describes the overview of the decoder-side depth estimation scheme and includes a description of its individual parts. Section 3 shows the methodology of experiments proposed to evaluate the DSDE in order to answer the abovementioned questions. The results of the experiments and their discussion are presented in Section 4, while the final conclusions and summary are presented in the last Section 5.

2 DECODER-SIDE DEPTH ESTIMATION

The decoder-side depth estimation (DSDE) approach shifts some of the video processing steps from the encoder to the decoder, making the decoding process more sophisticated and time-consuming. The video processing performed in the decoder operating in the DSDE approach comprises three major steps:

1. video decoding,
2. depth estimation,
3. virtual view synthesis.

When analyzing the entire data flow (not only the video sub-bitstreams), an additional step should be considered – metadata decoding. These metadata include camera parameters and other crucial information about views, or parameters used in depth maps estimation, e.g. bit depth [Gar21].

The first step of the video processing is a simple video decoding, performed by a typical 2D video decoder, e.g. VVC or HEVC. This step is crucial as it restores source views from the bitstream, but in this paper it is not considered and treated as trivial.

In the second step, the most time-consuming process is performed, allowing to estimate the geometry of the scene based on information sent within input views [Gar21] and decoded metadata of the multiview video.

There are numerous depth estimation methods described in the literature, including recent high-quality methods, e.g. graph-optimization-based methods described in [Rog19] and [Nam21], or methods based on using neural networks, e.g. GANet [Zha19] or GWCNet [Guo19]. However, as was presented in [Mie22], the most suitable method for the DSDE and overall immersive video applications is IVDE (Immersive Video Depth Estimation) [Mie20], developed by the ISO/IEC MPEG Video Coding group with its tools allowing proper depth estimation even for highly compressed input views [Mie21].

The last step of the decoding in any immersive video system, including the DSDE approach, is the rendering of viewports requested by the viewer. Such a rendering requires input views, corresponding depth maps, and camera parameters as input data, and outputs any view, created by reprojection of pixels [Dzi19a], [Fac18] followed by operations increasing the quality of rendered views such as filtering or inpainting [Jia21].

3 OVERVIEW OF THE EXPERIMENT

In order to properly assess the efficiency of different video encoders in the decoder-side depth estimation applications, two scenarios were tested. In both scenarios, the virtual views are generated from a lossy compressed multiview sequence.

The first scenario is based on using the newest ISO/IEC standard for immersive video compression: MPEG Immersive video (MIV). In the second one, a more general approach is considered, in which all the source views are separately encoded and used as the input for the standalone depth estimator and view synthesizer at the decoder side.

MPEG Immersive Video

The block diagram of the multiview video processing in the first experiment is presented in Fig. 1. As the MIV standard is codec-agnostic, any video encoder and decoder (blue blocks in Fig. 1) can be used to encode and decode the "atlases" produced by the MIV encoder (using the MIV Geometry Absent profile [Mie22]).

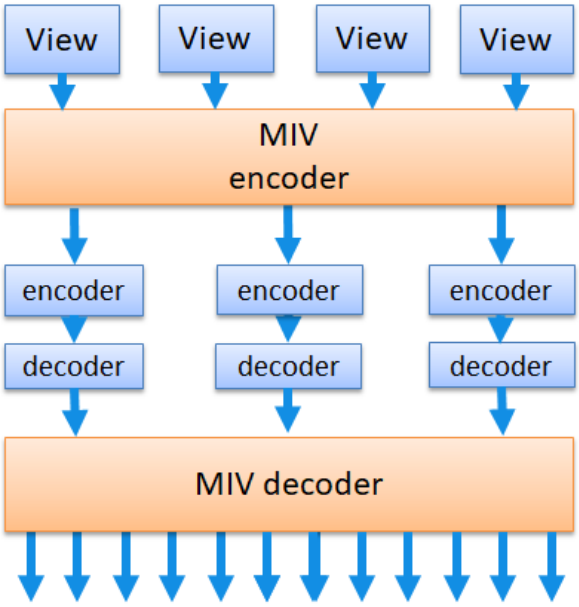

Figure 1: Block diagram of the MIV experiment.

This experiment was performed under the MIV Common Test Conditions (MIV CTC) [MPEG21b] developed by the ISO/IEC MPEG Video Coding group, which defines the entire pipeline for immersive video encoding, including detailed rules for encoding,

processing, and quality assessment of the immersive video, as well as a set of 15 miscellaneous sequences (Table 1), including both omnidirectional and perspective sequences, computer-generated content and natural sequences captured by real multicamera systems. According to the MIV CTC, for each test sequence, 17 frames were encoded.

Sequence	Resolution	Frames	Views
Carpark	1920 × 1088	250	9
Chess	2048 × 2048	300	10
ChessPieces	2048 × 2048	300	10
ClassroomVideo	4096 × 2048	120	15
Fan	1920 × 1080	97	15
Fencing	1920 × 1080	250	10
Frog	1920 × 1080	300	13
Group	1920 × 1080	99	21
Hall	1920 × 1088	500	9
Hijack	4096 × 2048	300	10
Kitchen	1920 × 1080	97	25
Mirror	1920 × 1080	100	15
Museum	2048 × 2048	300	24
Painter	2048 × 1088	300	16
Street	1920 × 1088	250	9

Table 1: Parameters of MIV CTC squenceces.

In the experiment, the effectiveness of four different video encoders was assessed, including two encoders developed by ISO/IEC MPEG: VVC [Bro21] in the optimized implementation VVenC [Wie21] and fast implementation of HEVC [Sul12]: x265 [x265]; as well as two royalty-free encoders: AV1 and VP9, both implemented in FFmpeg 4.4.1 [ffmpeg].

At the decoder side, the synthesized input views (virtual views synthesized at the position of input ones) were generated using the MIV decoder, which includes the decoder-side depth estimation implemented in IVDE software [Mie20] [MPEG21c] and the renderer implemented in the TMIV 9 software (Test Model 9 for MPEG Immersive video) [MPEG21a].

The objective quality was measured as IV-PSNR [MPEG20] measured between input views and synthesized input views. The IV-PSNR was calculated for all input views and is presented as a mean value, averaged over all views and all 17 frames.

General approach

This scenario is an extension of the experiment performed by the authors of this paper and presented in [Dzi16], presenting an influence of the newest coding techniques on top of previously tested encoders.

The multiview video processing pipeline used for the second experiment is presented in Fig. 2. In this experiment, all the input views are separately encoded using four simulcast encoders, including two encoders tested

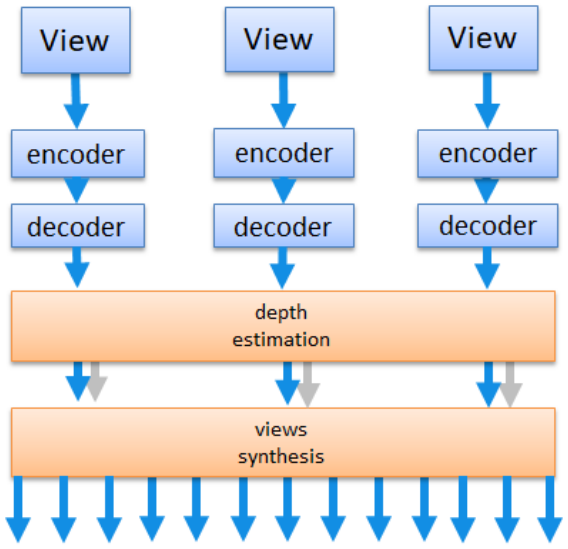

Figure 2: Block diagram of the general approach experiment.

in the first scenario: VVC and HEVC, and two older techniques: AVC in the x264 implementation [x264] and MPEG-2 implemented within the FFmpeg 4.4.1 [ffmpeg]. All used encoders are optimized and publicly available, increasing the reproducibility of presented experimental results.

On the decoder side, two multiview video processing algorithms were used. For depth estimation, the same IVDE algorithm [Mie20] was used to ensure, that the results of both performed experiments are not influenced by introducing different depth artifacts. For virtual view synthesis, the Advanced View Synthesizer described in [Dzi19a] was used. The advantage of this synthesizer is the possibility of easy optimization and fast implementation what was presented in [Sta20].

In this experiment, 8 multiview test sequences (Table 2) were used, including sequences captured by linear and circular multicamera systems [MPEG08], [MPEG15]. For all the sequences, more than 30 input views were used. For each sequence, four views were used as input ones for the entire processing (Fig. 2), while the rest was used for the quality assessment purposes, allowing proper and accurate objective quality assessment.

Sequence	Resolution	Frames	Views
BBB Butterfly Arc	1280 × 768	120	91
BBB Butterfly Lin.	1280 × 768	120	91
BBB Flowers Arc	1280 × 768	120	91
BBB Flowers Lin.	1280 × 768	120	91
BBB Rabbit Arc	1280 × 768	120	91
BBB Rabbit Lin.	1280 × 768	120	91
Dog	1280 × 960	300	80
Pantomime	1280 × 960	500	80

Table 2: Parameters of multiview sequences.

To be compliant with the first experiment, 17 consecutive frames were processed for each test sequence.

4 EXPERIMENTAL RESULTS

MPEG Immersive Video

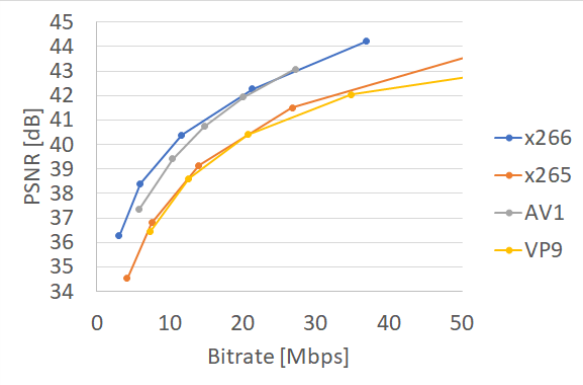

Figure 3: PSNR rate-distortion curves for decoded atlases.

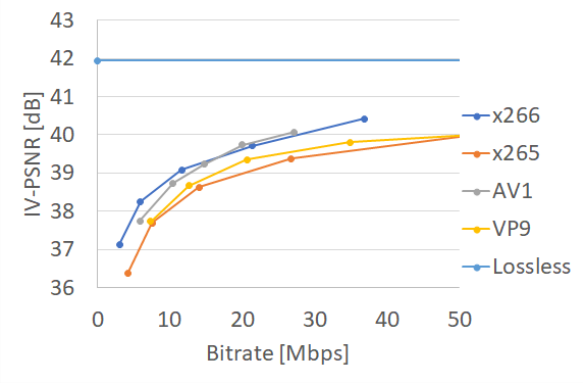

Figure 4: IV-PSNR rate-distortion curves for synthesized virtual views.

Figure 5: Atlases generated by the MIV encoder, sequence Group.

Figures 3 and 4 show the efficiency of four tested video encoders. In Fig. 3, the efficiency is presented in terms of the PSNR rate-distortion curves for decoded video (i.e., atlases, see Fig. 5). Fig. 4 presents the dependency between the total bitrate required for transmission of the immersive video encoded with different encoders and the mean quality of synthesized views (IV-PSNR averaged over 15 sequences, 17 frames, and all synthesized input views).

It should be noted that the bitrates presented in Figs. 3 and 4 are exactly the same, as they correspond to the same immersive video bitstreams.

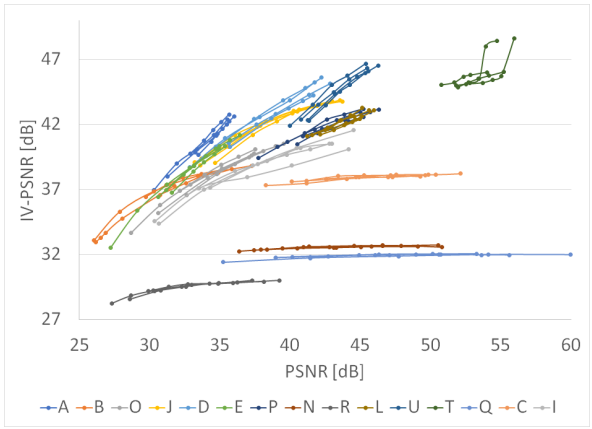

Figure 6: Dependency between decoding quality and synthesis quality for all MIV CTC sequences.

When comparing the results obtained using different encoders, some general observations can be stated. Firstly, MIV is indeed a codec-agnostic standard, and the RD-curves for all the encoders look similarly. Secondly, VVC and AV1 encoders in used fast implementations perform similarly both in terms of quality of decoded atlases and synthesized virtual views. On the other hand, the results for HEVC and VP9 seem to be more interesting. In terms of the quality of the decoded atlases, both encoders provide similar results. However, when comparing the IV-PSNR of the synthesized virtual views, a slight but noticeable advantage of VP9 can be found.

The possibility of quality assessment for two points of the decoder (before view synthesis – Fig. 3 and after view synthesis – Fig. 4) allows drawing a dependency between these two qualities, which is presented in Fig. 6. The results shown in Fig. 6 are drawn separately for each test sequence and each tested video encoder, presenting a dependency between the average PSNR of the decoded atlases and the average IV-PSNR of synthesized views. Curves for each sequence are colored differently.

As shown in Fig. 6, the majority of the curves are grouped. The only outliers can be found for sequences Q (ChessPieces), N (Chess), C (Hijack), and R (Group), for which the curves are almost horizontal. It means, that for these sequences the quality of the synthesized views does not depend on the quality of decoded atlas, thus increasing the total bitrate does not improve the user's experience.

In general, all the curves can be approximated by the linear equation:

$$\text{IV-PSNR}(view) \approx a \cdot \text{PSNR}(atlas) + b \qquad (1)$$

Sequence	ID	a	b
ChessPieces	Q	0.02	31.06
Chess	N	0.03	31.43
Hijack	C	0.06	35.35
Group	R	0.13	25.15
Carpark	P	0.43	23.25
Mirror	I	0.44	21.67
Fencing	L	0.45	22.32
Kitchen	J	0.50	22.73
Hall	T	0.55	16.47
Museum	B	0.55	19.27
Fan	O	0.61	16.80
Painter	D	0.67	16.97
Street	U	0.83	8.70
Frog	E	0.90	8.88
ClassroomVideo	A	0.99	7.08

Table 3: Linear approximation results for MIV CTC sequences.

Values of parameters a and b estimated for all test sequences can be found in Table 3. An example of curves for two sequences is presented in Figs. 7 and 8. Values highlighted in red correspond to the outliers in Fig. 6. For all these sequences the correlation between the quality of decoded atlases and synthesized virtual views is extremely low. It is caused by the appearance of strong synthesis artifacts in the virtual views (Fig. 9).

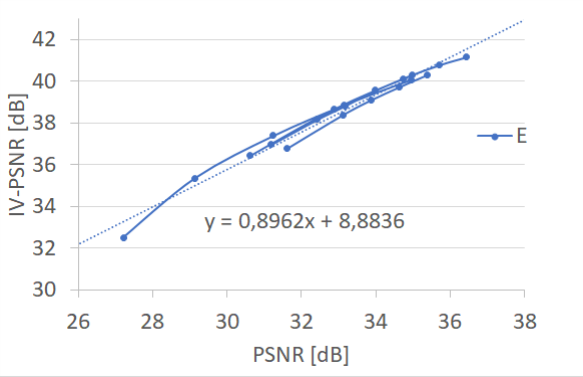

Figure 7: Linear approximation for Frog sequence.

General approach

Figs. 10 and 11 gather the results of the second experiment, presented in the same way, as for the experiment using the MPEG Immersive Video coding standard presented in the previous subsection. Fig. 10 contains the dependency between the total bitrate needed for transmission of all (four) input views and the quality of decoded views. Fig. 11 presents the results of the virtual

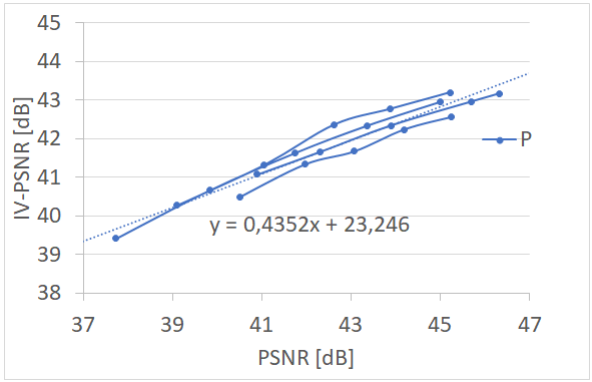

Figure 8: Linear approximation for Carpark sequence.

view synthesis, which was performed using these decoded views.

Similarly to the previous subsection, also the dependency between both qualities was measured and reported. Calculated values of parameters a and b of the linear equation bonding the IV-PSNR of the virtual view with PSNR of the decoded input views are

Figure 9: Synthesized views with strong artifacts, sequences ChessPieces and Group.

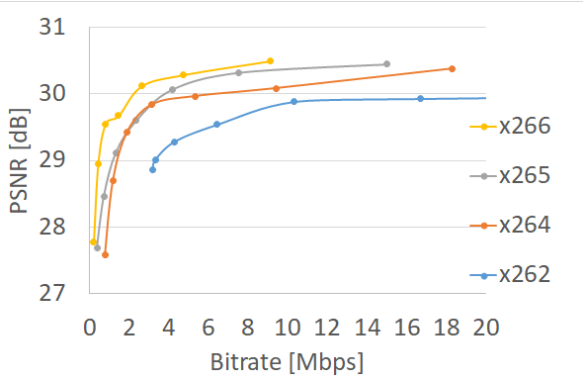

Figure 10: PSNR rate-distortion curves for decoded source views.

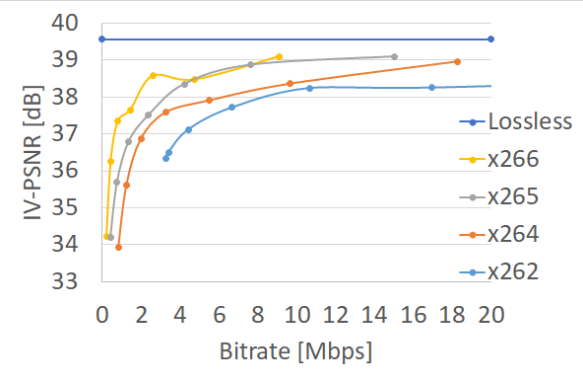

Figure 11: IV-PSNR rate-distortion curves for synthesized virtual views.

reported in Table 4. PSNR/IV-PSNR curves obtained for two test sequences are presented in Figs. 13 and 14.

Four sequences were highlighted in red in Table 4. For these sequences, the correlation between synthesis IV-PSNR and decoding PSNR is very low. For these sequences, many disturbing artifacts can be found in the synthesized virtual views, as presented in Fig. 12. Such a dependency is consistent with the observations taken for the first experiment, showing the relevance of both tested scenarios.

Sequence	a	b
BBB Flowers Lin.	0.1	23.04
BBB Flowers Arc	0.1	22.48
BBB Butterfly Lin.	0.15	29.92
Pantomime	0.16	32.29
BBB Butterfly Arc	0.22	30.89
Dog	0.25	26.31
BBB Rabbit Lin.	0.32	25.71
BBB Rabbit Arc	0.34	23.98

Table 4: Linear approximation results for 8 multiview sequences.

Obtained results follow the expectations, as newer and more advanced encoding standards perform better than the older ones, both in terms of the decoding quality and the quality of synthesized virtual views.

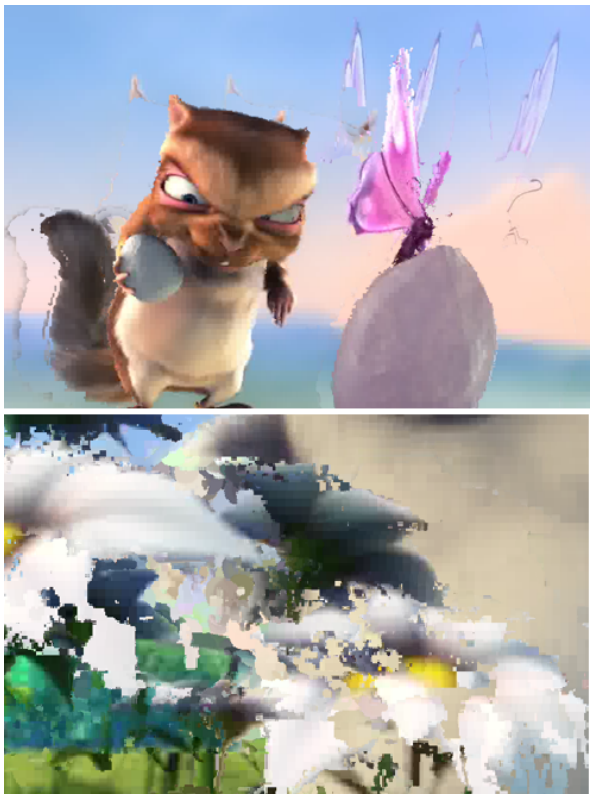

Figure 12: Synthesized views with strong artifacts, sequences BBB Flowers Lin. and BBB Butterfly Lin.

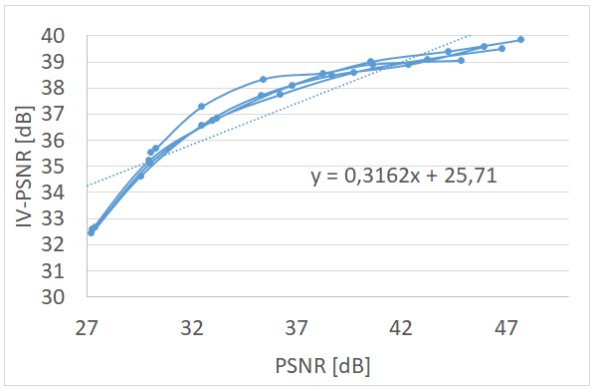

Figure 13: Linear approximation for BBB Rabbit Lin.

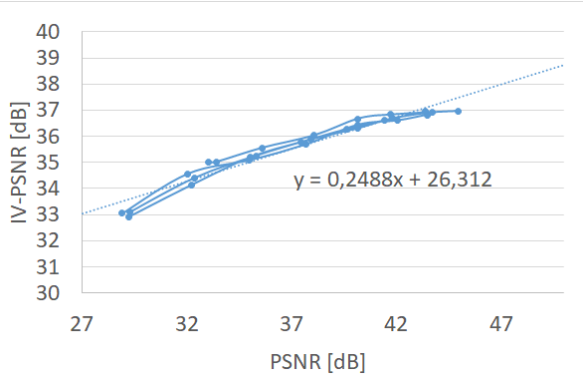

Figure 14: Linear approximation for Dog sequence.

5 CONCLUSIONS

In the paper, we have analyzed the influence of the lossy compression introduced by various video encoders on the depth map estimation process. Such research is very important and allows us to make some crucial observations.

At first, there is a strong correlation between the quality of the decoded input views and the quality of virtual views synthesized based on them.

Secondly, the newest ISO/IEC standard for immersive video compression – the results prove that MPEG Immersive Video (MIV) is indeed a "codec-agnostic" technique and any video codec can be used with it, nevertheless, the used codec significantly impacts the quality of synthesized virtual views thus the viewer's experience.

The third observation is that VP9 and the optimized implementation of the HEVC encoder (x265) provide similar quality. However, when comparing the IV-PSNR of the synthesized virtual views, a slight but noticeable increase the final quality for VP9 can be found.

At last, the dependency between the quality of the decoded input views and the quality of the synthesized views can be expressed by a linear approximation. The slope of this linear approximation can suggest if a sequence is easy to be properly synthesized. The steep trend line suggests, that the virtual view is visually consistent; if the trend line is almost horizontal, the virtual view has noticeable rendering artifacts.

All of the presented observations and conclusions suggest that efficient decoder-side depth estimation is possible.

6 ACKNOWLEDGMENTS

This work was supported by the Ministry of Education and Science of Republic of Poland.

7 REFERENCES

[Boy21] Boyce, J. et al. "MPEG Immersive Video Coding Standard," in Proceedings of the IEEE, vol. 109, no. 9, pp. 1521-1536, Sept. 2021, doi: 10.1109/JPROC.2021.3062590.

[Bro21] Bross, B. et al. "Overview of the Versatile Video Coding (VVC) standard and its applications," IEEE Tr. on Circ. and Syst. for Vid. Tech., 2021, doi: 10.1109/TCSVT.2021.3101953.

[Dzi16] Dziembowski, A. et al. "The influence of a lossy compression on the quality of estimated depth maps," 2016 International Conference on Systems, Signals and Image Processing (IWSSIP), 2016, pp. 1-4, doi: 10.1109/IWSSIP.2016.7502730.

[Dzi19a] Dziembowski, A. et al. "Virtual View Synthesis for 3DoF+ Video," 2019 Picture Coding Symposium (PCS), 2019, pp. 1-5, doi: 10.1109/PCS48520.2019.8954502.

[Dzi19b] Dziembowski, A. and Domański, M. "Objective quality metric for immersive video", ISO/IEC JTC1/SC29/WG11 MPEG2019/M48093, July 2019, Göteborg, Sweden

[Fac18] Fachada, S. et al. "Depth image based view synthesis with multiple reference views for virtual reality," 2018 - 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2018, pp. 1-4, doi: 10.1109/3DTV.2018.8478484.

[ffmpeg] FFmpeg encoder available at: www.ffmpeg.org.

[Fis20] Fischer, R., et al. "Improved Lossless Depth Image Compression", Journal of WSCG 28, 2020, 168-176, doi: 10.24132/JWSCG.2020.28.21.

[Gar21] Garus, P. et al. "Immersive Video Coding: Should Geometry Information be Transmitted as Depth Maps?," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 5, pp. 3250-3264, May 2022, doi: 10.1109/TCSVT.2021.3100006.

[Guo19] Guo, X. et al. "Group-Wise Correlation Stereo Network," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3268-3277, doi: 10.1109/CVPR.2019.00339.

[Jeo21] Jeong, J.Y. et al. "DWS-BEAM: Decoder-Wise Subpicture Bitstream Extracting and Merging for MPEG Immersive Video," 2021 International Conference on Visual Communications and Image Processing (VCIP), 2021, pp. 1-5, doi: 10.1109/VCIP53242.2021.9675419.

[Jia21] Jia, B. et al. "Virtual view synthesis for the nonuniform illuminated between views in surgical

video." Multim. Tools Appl. 80 (2021): 20619-20639, doi: 10.1007/s11042-021-10732-3.

[Mie20] Mieloch, D. et al. "Depth Map Estimation for Free-Viewpoint Television and Virtual Navigation," in IEEE Access, vol. 8, pp. 5760-5776, 2020, doi: 10.1109/ACCESS.2019.2963487.

[Mie21] Mieloch, D. et al. "Point-to-Block Matching in Depth Estimation," International Conference on Computer Graphics, Visualization and Computer Vision WSCG 2021, doi: 10.24132/CSRN.2021.3002.15.

[Mie22] Mieloch, D. et al. "Overview and Efficiency of Decoder-Side Depth Estimation in MPEG Immersive Video," in IEEE Transactions on Circuits and Systems for Video Technology, doi: 10.1109/TCSVT.2022.3162916.

[MPEG08] "1D Parallel Test Sequences for MPEG-FTV," ISO/IEC JTC 1/SC 29/WG11 M15378, Apr. 2008.

[MPEG15] "[FTV AHG] Big Buck Bunny light-field test sequences," ISO/IEC JTC 1/SC 29/WG11 M35721, Feb. 2015.

[MPEG20] "Software manual of IV-PSNR for Immersive Video," ISO/IEC JTC 1/SC 29/WG04 N0013, Oct. 2020.

[MPEG21a] "Test Model 9 for MPEG Immersive Video," ISO/IEC JTC 1/SC 29/WG04 N0084, May 2021, Online.

[MPEG21b] "Common Test Conditions for MPEG Immersive Video," ISO/IEC JTC 1/SC 29/WG04 N0085, May 2021, Online.

[MPEG21c] "Manual of IVDE 3.0," ISO/IEC JTC1/SC29/ WG04 N0058, Jan. 2021.

[Nam21] Nam, D.Y. and Han, J.K. "Improved Depth Estimation Algorithm via Superpixel Segmentation and Graph-cut," 2021 IEEE International Conference on Consumer Electronics (ICCE), 2021, pp. 1-7, doi: 10.1109/ICCE50685.2021.9427631.

[Rog19] Rogge, S. et al. "MPEG-I Depth Estimation Reference Software," in 2019 International Conference on 3D Immersion (IC3D), 2019, pp. 1-6, doi: 10.1109/IC3D48390.2019.8975995.

[Sta20] Stankowski, J., Dziembowski, A., "Fast View Synthesis for Immersive Video Systems", Journal of WSCG 28, 2020, 137-144, doi: 10.24132/CSRN.2020.3001.16.

[Sul12] Sullivan, G. et al. "Overview of the High Efficiency Video Coding (HEVC) standard," IEEE Tr. on Circ. and Syst. for Vid. Tech., vol. 22, 2012, doi: 10.1109/TCSVT.2012.2221191.

[Wie21] Wieckowski, A. et al. "VVenC: An Open And Optimized VVC Encoder Implementation," Proc. IEEE International Conference on Multimedia Expo Workshops (ICMEW), 2021, doi: 10.1109/ICMEW53276.2021.9455944.

[x264] "x264 encoder" available at: https://www.videolan.org/developers/x264.html.

[x265] "x265 encoder" available at: https://x265.com/.

[Xie21] Xie, Y. et al. "Performance analysis of DIBR-based view synthesis with kinect azure," 2021 International Conference on 3D Immersion (IC3D), 2021, pp. 1-6, doi: 10.1109/IC3D53758.2021.9687195.

[Zha19] Zhang, F. et al. "GA-Net: Guided Aggregation Net for End-To-End Stereo Matching," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 185-194, doi: 10.48550/arXiv.1904.06587.