VIRTUAL REALITY CONTENT STREAMING: VIEWPORT-DEPENDENT PROJECTION AND TILE-BASED TECHNIQUES

Alireza Zare, Alireza Aminlou, Miska M. Hannuksela

Nokia Technologies, Tampere, Finland
{alireza.zare, alireza.aminlou, miska.hannuksela}@nokia.com

ABSTRACT

Virtual reality (VR) head-mounted display (HMD) requires spherical panoramic contents with high-spatial and temporal fidelity to immerse the viewers into the captured scene. Hereby, VR contents are extremely bandwidth intensive and impose technical challenges for the design of a VR streaming system. A bandwidth-efficient VR streaming system can be achieved using the viewport-aware adaptation techniques, in which part of the sphere within the viewer's field of view is presented at higher quality. In this paper, two recently emerged viewport-adaptive streaming methods so-called tile-based method and truncated square pyramid (TSP) projection, a well-studied viewport-dependent projection, are compared using a proposed quality assessment methodology. The comparison is made in terms of storage and streaming bitrate performances. The simulation results indicate that the tile-based approach has slightly lower streaming performance, while offering a significant storage and encoding time saving at the server side, compared to TSP-based streaming.

Index Terms**—** Virtual reality, viewport-adaptive streaming, HEVC tiles.

1. INTRODUCTION

In the emerging virtual reality (VR), encoding and streaming of high-resolution, high-frame rate, and high-quality spherical video is becoming a big challenge. The use of such a content along with strict latency requirement is crucial to guarantee the promised full-immersive experience. In the view of this need, the Moving Picture Experts Group (MPEG) is working towards developing a rate-distortion efficient architecture for coding and streaming VR content.

The conventional way of streaming the spherical video (i.e., encoding and transmitting the whole 360° content at the same quality) requires an unnecessarily high bandwidth. To reduce the streaming bitrate, considering the limited field of view (FOV) of head-mounted displays (HMD) [1], viewport-adaptive streaming schemes have recently been studied in [2] and [3]. These techniques aim to deliver the currently viewed content (i.e., viewport) by the viewer at high quality, and the rest of the 360° content (i.e., non-viewport) at lower quality.

Viewport-adaptive streaming has been practically realized using two methods, namely, tile-based and viewport-dependent projection streaming. In the viewport-dependent projection, after the spherical video is projected onto a planar layout, multiple representations of the projected video are generated. Each representation has higher resolution/quality in a pre-defined viewport, while the non-viewport part of the content is presented at lower resolution/quality. The generated representations are then encoded, and the appropriate one is transmitted based on the viewer's current viewing orientation. Encoding and storing several representations (e.g., 30 as proposed in [4]) of the same content may be considered as a bottleneck in resource-constraint applications of VR [5]. Alternatively, in the tile-based streaming method [6], a content is divided into several regular tiles which are independently coded using motion-constraint tile set (MCTS) technique in different quality levels. Then, based on the viewer's viewing direction, an appropriate combination of tiles in different qualities is selected to be transmitted to the viewer in a single standard-compliant bitstream.

One of the challenges in viewport-adaptive streaming is rate-distortion (R-D) performance evaluation of the experienced quality, which is under study. Most of the time, the displayed content to the viewer is rendered using the viewport content. But, due to the latency in VR content delivery systems, the displayed content may be partially or fully rendered from the non-viewport content. Hence, a solid framework is required to analyze the quality of viewport-adaptive methods in different cases, including viewport and non-viewport parts. Recently, a test methodology has been agreed in [7] to analyze R-D performance of viewport-independent projections. This method targets a long term research to find a suitable projection for representing and storing spherical contents, thus, it is not suitable for analyzing viewport-adaptive streaming.

Two test methodologies are being studied for evaluation of viewport-adaptive streaming of 360° contents. The Windowed S-PSNR method in [8] was proposed and used to compare the performance of two viewport-dependent projections including truncated square pyramid (TSP) and

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

downsampled cubemap to equirectangular projection (ERP). This metric assesses the quality of viewport-dependent projection in the viewport with different rendering FOV, varying from 90° to 180°, while ignoring the quality of the content in the rest of the sphere. Alternatively, a quality assessment method is proposed in [9], which is based on rendering the viewport for several discrete pre-defined viewing orientations, uniformly distributed over the sphere. Using this method, the performance of different viewport-dependent projections were compared to ERP in [3].

This paper, for the first time to our knowledge, aims to make a comparison between tile-based streaming and TSP-based streaming [8], one of the well-studied viewport-dependent projections. The other goal of this study is to take into account the viewer's head motion and analyze the quality of streaming performance in viewport and non-viewport parts. These refer to when displayed view is rendered using the high-quality and low-quality content, respectively. The simulation results indicate that the tile-based streaming, however, has slightly lower streaming performance compared to TSP-based on average, it provides higher streaming performance in high bitrates. Tile-based streaming also offers much better quality in the case of non-viewport part, when the viewer's viewing orientation does not match with viewport of the transmitted representation.

The remaining of the paper is organized as follow. In Section 2, VR streaming technologies are reviewed. Our streaming framework and quality assessment methodology are explained in Sections 3 and 4, respectively. Simulation results are reported in Section 5, followed by conclusion.

2. VR STREAMING TECHNOLOGY

2.1. Viewport-adaptive streaming

Viewport-adaptive streaming is an interactive technique which selectively streams the viewer's current viewport based on a sensory feedback related to the viewer's viewing orientation. The idea is to reduce the transmission bitrate by delivering the minimal part of the content. The ideal case would be sending just the viewport content which is viewed by the viewer at that time. However, because of the limitations in the VR system, it is more practical to transmit the viewport at high quality and the remaining non-viewport part of the content at a lower quality. The motivation of sending the non-viewport part is that when the viewer turns his/her head to a non-viewport part of the 360° scene, due to the latency of the encoding and transmission system, it may take a while (e.g., 1 second) that the VR system can deliver the appropriate viewport to the viewer. Hence, during this short period of time, the viewer will see the lower quality content rather than the expected viewport. There are two practical methods for viewport-adaptive streaming which are reviewed below.

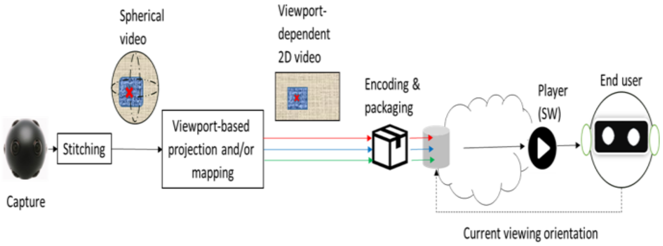
Spherical video
Viewport-dependent 2D video
Encoding & packaging
Player (SW)
End user
Capture
Stitching
Viewport-based projection and/or mapping
Current viewing orientation

Figure 1. VR viewport-dependent streaming system.

2.2. Viewport-dependent projection

Figure 1 shows an overview of a VR streaming system using viewport-dependent projection, where each representation is illustrated with a different color. In this approach, the spherical video contents are projected/mapped onto several 2D viewport-based representations of the same content, each corresponds to a viewing orientation. Each representation presents the corresponding viewport at higher quality and the non-viewport part of the sphere at lower quality. These representations are encoded and stored at the server side. Then based on the viewing orientation of the viewer, the corresponding bitstream is transmitted to the viewer.

2.3. Tile-based streaming

In this method, the video content is divided to several tiles which are coded independently using motion-constrained tile set (MCTS) technique. Tiling can be performed in several arrangements with different number of tile columns and rows. Each tile is coded into multiple versions at different quality levels and stored at the server side. Based on the viewer's viewing orientation, a number of tiles covering the viewport are transmitted at higher quality, while the other tiles corresponding to the non-viewport part are streamed at lower quality. A tile-based extractor constructs a standard compliant bitstream corresponding to the desired combination of tiles such that a standard HEVC decoder can cope with that. The tile set extraction is performed at the server end based on the viewing orientation of the viewer.

3. PROPOSED STREAMING FRAMEWORK

In the viewport-dependent projection, considering 90°x90° FOV for viewport of each representation, at least 6 viewport representations are required to cover the whole sphere, four representations along the equator in addition to the two ones covering the north and south poles. Typically, a higher number of pre-defined uniformly distributed viewport representations are generated over the sphere to achieve a finer representation and better quality of experience. In [8], the whole sphere is presented with 30 viewport representations using TSP projection. In order to keep the number of calculation reasonable, similar to [9], this work considers only 12 viewport representations located along the equator, each 30 degrees apart. Note that most of the

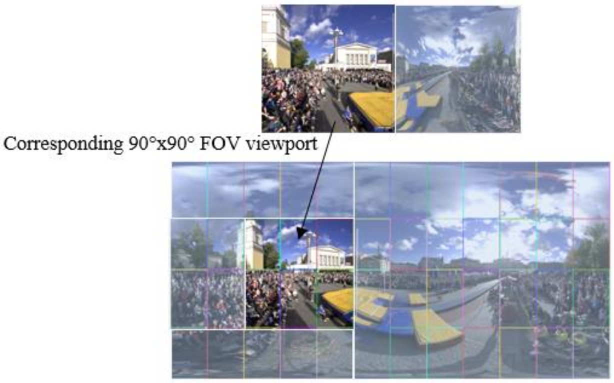

Figure 2. A tile-based viewport representation and its corresponding TSP-based viewport representation.

panorama contents have less interesting content in top and bottom parts which are less likely to be seen. We characterize each representation by its center at high-quality viewport, which are shown in Figure 3 with blue marks.

In the tile-based method, the goal is to use a tile grid which is aligned with the generated TSP representations. This means that each tile-based representation has a high-quality 90°x90° FOV viewport which matches with that of the corresponding TSP representation, as illustrated in Figure 2. To feasibly achieve the abovementioned alignment requirement in both streaming techniques, a 12x4 tiling is used such that every 3x2 tiles cover 90°x90° FOV. Hence, similar to the TSP-based method, in tile-based streaming 12 viewport representations is defined along the equator. For each viewport representation, a 3x2 tile set is transmitted at high quality while the remaining 42 tiles are picked from the low-quality versions.

4. QUALITY MEASUREMENT METHODOLOGY

In a VR system, different directions of a 360° content have equal chance to be viewed by the viewer. In order to keep the number of calculations reasonable, it is suggested to consider a limited set of discrete viewing orientations for measuring the quality of experience [9]. Each viewing orientation is called quality assessment view (QAV) which is characterized by the center of its rendered viewport, shown with red marks in Figure 3. The center of a QAV's viewport may match the center of one of the viewport representations or may not. The closest viewport representation is used to render an appropriate representation for each QAV.

This work assumes streaming VR content to HMDs which are different in their FOV and optics. Thus, in practice, the rendering pipelines are HMD-wise. We assume that a rectilinear view with 90° FOV is close to the real-world perspective and sufficiently estimates the views presented on different HMDs. Furthermore, when there is head motion, the viewer may see the non-viewport content before switching to a new viewport representation. Hence, for a better analysis of the experienced quality, the viewport content needs to be separated from the non-viewport part. Considering the above points, for each QAV, a cubemap is rendered using the associated viewport representation. Then, the front face of the rendered cubemap is considered for viewport quality measurement, while the other five faces are used for quality evaluation of the non-viewport part. Figure 3 illustrates centers of the QAVs used in this study for the quality assessment process. Along the equator, 36 (i.e., 360°/10°) uniformly distributed QAVs are defined, where 12 of those match with the selected 12 viewport representations. Similarly, the same number of QAVs are defined in ±15° latitudes, resulting in 108 QAVs in total. Therefore, each viewport representation is associated with 9 QAVs, as shown in Figure 3.

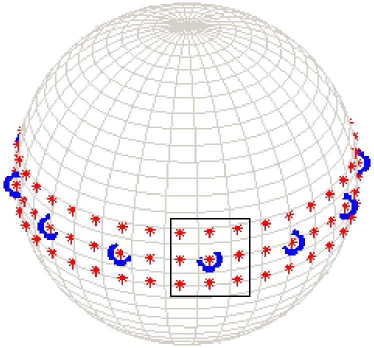

Figure 3. QAVs used in the simulation. Blue marks: center of viewport representations, Red marks: center of QAVs.

5. EXPERIMENTAL RESULTS

5.1. Experimental conditions

The MCTS encoding technique was implemented in the HEVC reference software (HM) version 16.7 [10] which was used for tile-based encoding. The same HM version was used for coding of the TSP sequences. For ERP to TSP conversion, the 360Lib tool [7] developed by Joint Video Exploration Team (JVET) was used. The simulations were conducted using the Main profile random access configuration [11] with 49 frames for each sequence. In the tile-based technique, the tiles covering the viewport were picked from bitstreams coded with the quantization parameter (QP) values of the JCT-VC random access common test conditions [11]. The non-viewport tiles were coded with higher QP values of QP+7 compared to the viewport tiles. The decoding refresh type was set to instantaneous decoding refresh (IDR) picture with period of 32. Test sequences include 8 monoscopic panorama clips in ERP format at resolution 4K and 8K listed in Table I, which are from JVET test contents for 360° video coding [7]. The compression performance and streaming bitrate saving measurements are presented in terms of Bjøntegaard Delta-rate (BD-rate) criterion [12] for luma pictures, where positive/negative values indicate how much the bitrate is increased/decreased for the same peak signal-to-noise ratio (PSNR). The streaming bitrate performance of viewport and non-viewport parts are analyzed separately. The bitrate and PSNR values are averaged over the all QAVs.

5.2. Result analysis

Table I presents the comparison between the two streaming techniques in terms of streaming bitrate performance. As can be seen, the tile-based method introduces about 7% loss when considering the viewport only, on average over the test sequences. For the non-viewport streaming, however, the tile-based approach achieves an average streaming bitrate saving of 42% when compared to the TSP-base method. With further optimization, the viewport bitrate loss in tile-based streaming can be compensated by lowering the quality of the non-viewport part.

In Figure 4 and Figure 5, the R-D curves of PoleVault sequence are shown for viewport and non-viewport part, respectively. It can be observed that the tile-base method has higher streaming bitrate performance at higher bitrates compared to the TSP-based approach. This trait is appreciated in VR streaming which typically consumes high-quality content. Although, only the R-D curves related to PoleVault sequence is presented, similar behavior is observed for the other video sequences. In fact, the TSP loss in higher bitrates is because of removing of high-frequency information due to the extra ERP to TSP conversion. This is why the PSNR values in the R-D curves of the TSP-based method get saturated, especially for the non-viewport part.

In the viewport-dependent projection, all the viewport representations have to be encoded and stored at the server side. In contrast, in tile-based streaming, each tile is coded in different quality levels once, where different viewport representations are generated by combining the tiles on the fly. In order to provide some insight into the storage requirement, the TSP-based method was compared to the tile-based method in which two versions of video bitstreams coded with QP and QP+7 were employed. In this experiment, the storage requirement of the tile-based method is about 29% of that of the TSP-based method. This is for the case of using 12 TSP-based representations which were used in this work. However, when considering a complete set of viewport representations covering the whole sphere, the ratio would considerably be lower. For example, in the case of 30 viewport representations recommended in [4] the ratio is estimated to be around 12% (=29%x12/30). Thus, the tile-based method brings a significant storage space reduction when compared to TSP-based streaming. Similarly to storage benefit, the tile-based approach is also beneficial for transmission over content delivery networks and caching.

Table I. Streaming bitrate comparison between tile-based and TSP-based methods.

	Viewport		Non-viewport	
Test sequences	BD-Rate (%)	BD-PSNR (dB)	BD-Rate (%)	BD-PSNR (dB)
AerialCity	0.68	0.12	-54.90	3.31
DrivingInCity	4.78	-0.05	-66.60	3.77
DrivingInCountry	-2.07	0.16	-54.15	2.56
PoleVault	5.52	-0.08	-60.02	3.52
Harbor360	13.89	-0.44	-40.82	2.28
KiteFlite360	19.07	-0.77	-27.22	1.65
Skateboard_trick	3.59	-0.09	-16.86	0.63
Train	12.20	-0.43	-14.65	0.83
Average	7.21	-0.20	-41.90	2.32

6. CONCLUSION

In this paper, a comparison was made between the two recently emerged viewport-adaptive streaming techniques called tile-based and viewport-dependent projection in terms of storage and streaming R-D performances. The simulation results indicate that the streaming performance of the tile-based method is slightly lower than that of the viewport-dependent projection. However, the tile-based method is further capable of providing better performance by optimizing the non-viewport part. Furthermore, the tile-based method consumes much less pre-processing and encoding time to generate multiple versions of the content and requires considerably less storage space at the server side, when compared to viewport-dependent projection. This makes the tile-based approach a suitable candidate for viewport-adaptive streaming of VR content.

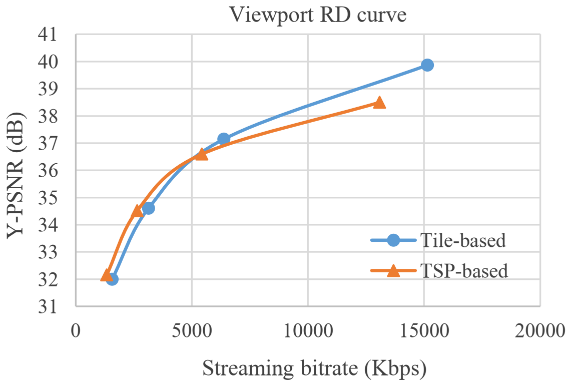

Figure 4. Viewport R-D curve, PoleVault test sequence.

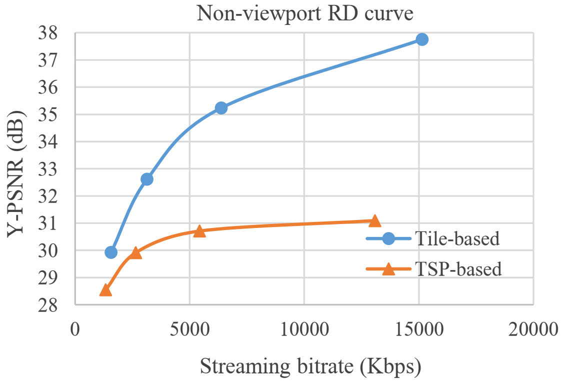

Figure 5. Non-viewport R-D curve, PoleVault test sequence.

7. REFERENCES

[1] W. Mason, 2017. VR HMD Roundup: Technical Specs. Accessed April 20 from http://uploadvr.com/vr-hmd-specs/.

[2] A. Zare, K. Kammachi Sreedhar, V. K. Malamal Vadakital, A. Aminlou, M. M. Hannuksela, and M. Gabbouj, "HEVC-compliant viewport-adaptive streaming of stereoscopic panoramic video," 2016 Picture Coding Symposium (PCS 2016), Nuremberg, Germany, Dec. 2016.

[3] K. Kammachi Sreedhar, A. Aminlou, M. M. Hannuksela, and M. Gabbouj, "Viewport-adaptive encoding and streaming of 360-degree video," The IEEE International Symposium on Multimedia (ISM), California, USA, Dec. 2016.

[4] E. Kuzyakov and D. Pio, "Next-generation video encoding techniques for 360 video and VR," Facebook, [Online]. Available:https://code.facebook.com/posts/1126354007399553. [Accessed Jan. 2017].

[5] S. Lederer, "Today's and future challenges with new forms of content like 360°, AR, and VR, " invited talk in MPEG workshop Global Media Technology Standards for an Immersive Age, Jan. 2017, http://mpeg.chiariglione.org/sites/default/files/events/06_Lederer.pdf.

[6] A. Zare, A. Aminlou, M. M. Hannuksela, and M. Gabbouj, "HEVC-compliant tile-based streaming of panoramic video for virtual reality applications," Proceedings of the 2016 ACM on Multimedia Conference, Amesterdam, Netherlands, Oct. 2016.

[7] J. Boyce, E. Alshina, A. Abbas, Y. Ye. "JVET common test conditions and evaluation procedures for 360° video," ITU-T Joint Video Exploration Team (JVET), document JVET-D1030, Oct. 2016.

[8] G. V. d. Auwera, M. Coban and H. Mart, "Truncated Square Pyramid Projection (TSP) for 360 video," ITU-T Joint Video Exploration Team (JVET), 4th meeting, document: JVET-D0071, Chengdu, 2016.

[9] A. Aminlou, K. Kammachi-Sreedhar, A. Zare and M. M. Hannuksela, "Testing methodology for viewport-dependent encoding and streaming". ITU-T Joint Video Exploration Team (JVET), document JVET-D0079, Chengdu, Oct. 2016.

[10] High Efficiency Video Coding (HEVC), Fraunhofer Institut for Telecommunications, Heinrich Hertz Institute, DOI= https://hevc.hhi.fraunhofer.de/, Accessed Apr. 2016.

[11] F. Bossen, "Common test conditions and software reference configurations," Joint Collaborative Team on Video Coding (JCT-VC), JCTVC-F900, Jul. 2011.

[12] G. Bjøntegard, "Calculation of average PSNR differences between RD-curves," document VCEG-M33, Austin, 2001.