VIEWPORT-AWARE ADAPTIVE 360° VIDEO STREAMING USING TILES FOR VIRTUAL REALITY

Cagri Ozcinar, Ana De Abreu, and Aljosa Smolic

Trinity College Dublin (TCD), Dublin 2, Ireland.

ABSTRACT

360° video is attracting an increasing amount of attention in the context of Virtual Reality (VR). Owing to its very high-resolution requirements, existing professional streaming services for 360° video suffer from severe drawbacks. This paper introduces a novel end-to-end streaming system from encoding to displaying, to transmit 8K resolution 360° video and to provide an enhanced VR experience using Head Mounted Displays (HMDs). The main contributions of the proposed system are about tiling, integration of the MPEG-Dynamic Adaptive Streaming over HTTP (DASH) standard, and viewport-aware bitrate level selection. Tiling and adaptive streaming enable the proposed system to deliver very high-resolution 360° video at good visual quality. Further, the proposed viewport-aware bitrate assignment selects an optimum DASH representation for each tile in a viewport-aware manner. The quality performance of the proposed system is verified in simulations with varying network bandwidth using realistic view trajectories recorded from user experiments. Our results show that the proposed streaming system compares favorably compared to existing methods in terms of PSNR and SSIM inside the viewport.

Index Terms— 360° video, virtual reality, tiling, DASH, viewport-aware

1. INTRODUCTION

During the last years, significant achievements have been made regarding the rendering capacity and quality of Head Mounted Display (HMD) systems [1]. Modern HMDs can render 360° video at a sufficiently high frame-rate and resolution, allowing the viewer to be immersed in the VR environment.

Although numerous professional VR video services have emerged for streaming the 360° video, *e.g.,* YouTube 360 [2], they suffer from severe drawbacks. More clearly, each video frame, containing a 360° Field of View (FOV), is encoded and transmitted regardless of the FOV of the HMD. In fact, the existing HMDs have a viewable FOV that ranges from 96° to 110° [3], meaning they use only around one fifth of the transmitted data [4]. The area of the 360° video frame displayed by the HMD at a given time is known as the *viewport*. Hence, the perceptual quality of such video mainly depends on the viewport quality. In existing professional services, regions *outside* the viewport waste a considerable proportion of the bandwidth. This unnecessarily utilized bandwidth results in overall low-quality video streaming to comply with the present Internet and decoding limitations.

Considering its data-intensive representation and the best-effort nature of the Internet, 360° video streaming requires a bitrate adaptive solution to offer an enhanced VR experience. To this end, MPEG-Dynamic Adaptive Streaming over HTTP (DASH) [5], which is an international standard for adaptive streaming, is a key enabler in achieving a smooth VR video playback. The main aim of DASH is to provide a high-quality streaming experience based on the client bandwidth. In DASH, video streams are requested using a manifest file, Media Presentation Description (MPD), which contains a set of bitrate representations.

The significant requirements regarding resolution to ensure high-quality VR experience [3, 6] can be managed with tiling and viewport-aware solutions using DASH. For example, tiling can generate self-decodable regions, *i.e.,* tiles, by *spatially* dividing the frames. In addition, tiling can consider the importance of regions [7] and can enable parallel downloading [8] and decoding [9, 10] features. As the HMDs use only the viewport out of the captured 360° video, the visual quality of the viewport can be enhanced in a viewport-aware manner.

This paper focuses on adaptive 360° video streaming and provides improved viewport quality compared to existing professional services. A VR video streaming system, from encoding to displaying, is designed to transmit 8K 360° video and to offer an enhanced VR experience. The main contribution of this work is an end-to-end streaming system implementation that contains tiling, a novel extension of the MPD, and DASH bitrate level selection in a viewport-aware manner. Experimental results showed that the proposed system demonstrates significant quality enhancements compared with the streaming approach that is currently used by professional VR services [2, 11] in this area. Our work concentrates on adaptive distribution of bitrate and quality over the 360° video in contrast to existing professional services. The system uses the tiling concept to divide each video frame into tiles to encode, transmit, and decode them effectively. To facilitate tiled 360° video streaming, we extended the concept

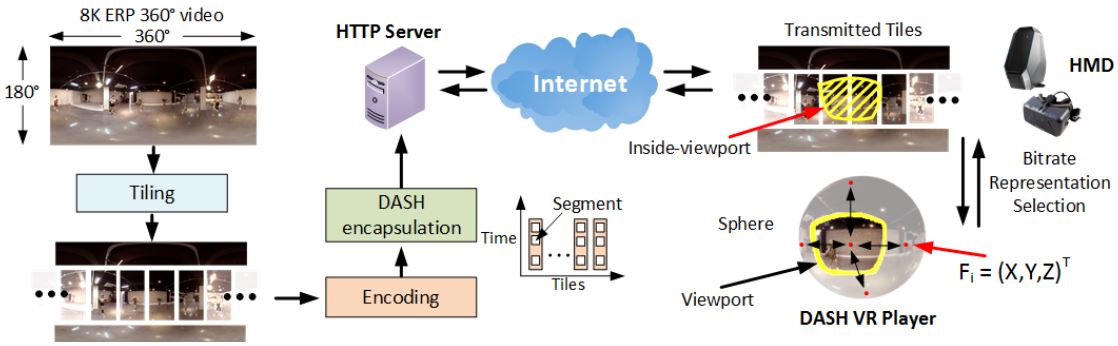

Fig. 1: Schematic diagram of our adaptive 360° video streaming system for VR.

of DASH Spatial Relationship Description (SRD) [12], and introduced a new MPD for DASH streaming. The proposed DASH player also efficiently distributes the available bandwidth to the tiles and requests the best bitrate representation for each tile in a viewport-aware manner.

The rest of this paper is organized as follows: the related work is presented in Sec. 2. In Sec. 3, we introduce the proposed solution. In Sec. 4, we present the experimental results. Finally, Sec. 5 concludes the paper.

2. RELATED WORK

Several early studies [13–16] used tiling for the aim of viewport-aware 360° video transmission, where an HMD technology was utilized without using a bitrate adaptive streaming. Although these solutions are far from achieving the expected high-quality performance because of the limitations of the early technologies, they are pioneering works in this area.

Recent advances in adaptive streaming and modern HMDs have made it feasible to deliver and render high-quality 360° video. There are two types of adaptive streaming solutions which are commonly used in this area, namely, non-tiled and tiled. In the context of non-tiled, a recent work in [17] focused on the quality impact of 360° video projections. In addition, in the context of tiling research, a short paper of Skupin *et. al* in [18] described the bandwidth problem of 360° video, and suggested to use tile-based streaming. Also, in the work in [19], high-resolution video content is transmitted in tiled fashion using fixed rectangular tiles, such as 5×5.

To design a practically applicable streaming system for 360° video and to address studies needed in this area [20], we developed a system for VR that supports tiled and viewport-aware adaptive streaming. We used unequal tile sizes and viewport-aware bitrate *distribution* using a novel *distance* criterion. To verify our method, we recorded *real* viewport trajectories from subjects in viewing sessions. With using our recorded data, we calculated the viewport quality scores and compared our proposed method with the reference solution, which is based on the existing professional adaptive streaming systems [11].

3. PROPOSED STREAMING SYSTEM

Typically 360° video is shot using a collection of cameras that cover a 360° FOV. The captured frames are then stitched together and projected onto a 2D plane using the Equi-Rectangular Projection (ERP) [21]. The overall video has to have a very high resolution such as 8K in order to provide high quality for a given viewport, which is only a portion of the overall view.

To increase the streaming performance of such content, we propose an adaptive streaming system that enhances the visual quality of the video displayed in the viewport. The proposed system divides each ERP video frame into self-decodable tiles, encodes them at various bitrates, and then encapsulates and stores them in the HTTP server. Each bitstream contains multiple self-decodable time segments for the purpose of adaptive streaming.

The proposed DASH VR player requests the most appropriate bitrate representations for each tile given the available network bandwidth and the viewport location. To avoid undesirable quality degradation during a possible sudden viewport movement, our current system streams the whole tiled ERP frame by gradually reducing the bitrate (or quality) of the outside-viewport tiles and increases the quality of the inside-viewport tiles. To this end, in order to use our approach with the DASH standard, we extended the DASH-SRD representation, and contributed a novel MPD for DASH streaming. Finally, the decoded tiles are rendered in the HMD after projecting them back into a sphere. A schematic diagram summarizing the proposed streaming system is presented in Fig. 1.

3.1. Tiling and encoding

Tiling divides video frames into several self-decodable streams (tiles) to deliver and decode high-resolution 360° video effectively. Consequently, selection of the tile size plays an important role regarding streaming performance. For instance, using large tiles may increase the compression efficiency, but it may contain many pixels that are not a part of the current viewport. In contrast, using small tiles may reduce the compression efficiency because of exploiting less spatial redundancies. Consequently, in the context of this work, we pay attention to the typically lower importance and low-motion characteristics of the poles and the dominant

viewing adjacency of the equator [22]. First, each ERP video frame is vertically divided into three parts: the equator and two poles. The equator represents the middle segment, and the two poles stand for the top and the bottom sections of the frame. As the poles occupy the largest regions of redundant pixels [23], in those areas, larger tile resolution size can be used to compress them using a lower bitrate. Additionally, as the equator is associated with the most dominant viewing adjacency [22], it is further divided horizontally into several tiles to efficiently transmit them using the proposed viewport-aware representation selection, as shown in Fig. 1.

Each tiled video frame may then be encoded e.g. with either the H.264/Advanced Video Coding (AVC) [24] or the High Efficient Video Coding (HEVC) standard (H.265/MPEG-HEVC). As the proposed system is a codec agnostic, the player can decode each tile stream with either of the coding standards, along with several possible individual decoders for AVC or a single decoder for HEVC encoded tiles. Importantly, tiling offers additionally parallel downloading and decoding opportunities to transmit and decode the high-resolution 360° video effectively.

3.2. MPD extension for 360° video

In order to transmit the tiled video frames to the VR client, the DASH standard is used for adaptive streaming. To this end, Each encoded tile is divided into self-playable time segments, and an MPD is delivered before starting a streaming session. The MPD describes the structure of bitrate representations for each tile.

In order to use our streaming approach with the DASH standard, a novel MPD structure is introduced for 360° video content. For that, the standard MPD structure for SRD is extended to support our viewport-aware approach. The proposed MPD includes the tiles' Identification Numbers (ID), their resolutions, their positions and their center locations in terms of the spherical Cartesian coordinates. This way, each tile can be requested with its ID and then mapped to a sphere with the help of its resolution and its position on the sphere. To facilitate viewport-aware streaming, a center location is added in the MPD for each tile. This way, each i^{th} tile contains a center point, which is represented as $\mathbf{F}_i = (X, Y, Z)^T$, as illustrated in Fig. 1.

3.3. Viewport-aware representation selection

Given the current viewport, an optimal bitrate representation is selected for each tile. This work is interested in providing more importance to the tiles in the viewport, meaning higher bitrate. Consequently, the total bits to be assigned to the tiles that are *outside* of the viewport are reduced. The perceived visual quality is enhanced by increasing the bitrate of the viewport. The bitrate of the tiles outside of the viewport is gradually reduced based on the *distance* between the spherical center location of the viewport and each outside-viewport tile.

The proposed DASH VR player requests the optimum bitrate representation for each tile using the designed MPD. To this end, we define $\boldsymbol{V}$, $\boldsymbol{S^{in}}$ and $\boldsymbol{S^{out}}$ for the viewport, a set of the tiles inside the viewport, and a set of tiles outside of the viewport, respectively. The bitrate assigned to the i^{th} tile in the viewport is the following:

$$R_{V_i} = (\gamma R_{cur})\omega_i \quad i \in \boldsymbol{V}, \quad i \notin \boldsymbol{S^{out}}, \tag{1}$$

where the tile represents as i, $i \in \mathbb{Z}$ and $i \in [1, N]$. N denotes the total number of tiles. γ is a client-defined constant term, which is $\gamma \in [0, 1]$. In this work, γ is selected as 0.8 empirically. R_{cur} is the current available bandwidth, and ω_i is the weight of the i^{th} tile in $\boldsymbol{S^{in}}$. This weight is calculated for the i^{th} tile as:

$$\omega_i = \frac{\#\text{ of pixels in } (\boldsymbol{V} \cap S_i^{in})}{\rho_{tot}}, \tag{2}$$

where ρ_{tot} is the total number of pixels in $\boldsymbol{V}$.

To gradually distribute the remaining bandwidth among the outside-viewport tiles, the Euclidean distance, δ_i, is calculated between the middle point of the $\boldsymbol{V}$, and each defined $\mathbf{F}_i$ point for each outside-viewport tile, $\boldsymbol{S_i^{out}}$. The bitrate estimation for the i^{th} tile in $\boldsymbol{S^{out}}$ is calculated as follows:

$$R_{S_i^{out}} = \widehat{\kappa_i}((1-\gamma)R_{cur}) \quad i \in \boldsymbol{S^{out}}, \quad i \notin \boldsymbol{V}, \tag{3}$$

where $\widehat{\kappa_i}$ is calculated as $\widehat{\kappa_i} = \frac{\kappa_i}{\sum_i \kappa_i}$ and $\kappa_i = \frac{\max_i \delta_i}{\delta_i}$.

Finally, the client requests a bitrate representation for each tile, which can be obtained as follows:

$$J \Leftarrow \min_J |r_J - R_i| \quad i \in (\boldsymbol{S^{in}} \cup \boldsymbol{S^{out}}), \tag{4}$$

where R= $R_{S^{out}} \cup R_{S^{in}}$, J is the selected DASH representation ID, $J \in \{1, \ldots, \epsilon\}$ given ϵ is the total number of representations, and r_J is the bitrate of the J^{th} representation.

4. EXPERIMENTS

4.1. Setup

Two 360° video sequences from the MPEG video exploration experiments [25] were used: *Stiched_left_Driving360_8K* and *Stiched_left_Dancing360_8K*. Their resolutions are 8K, 8192×4096. We focused on the browser-based VR use-case that is one of the core experiments in the ongoing standardization activity for this subject [20]. Since AVC is the only implemented decoder in current browsers that handle HMDs, AVC encoded streams were tested over a real network. To this end, the x264 software (*ver.* r2643) [26] was used for encoding purposes. We used 0.9, 2, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, and 25 *Mbps* as *target bitrates* for the proposed and reference methods. The MP4Box [27] was used to wrap the encoded content within an MP4 header file. Then this tool was utilized to create 2 *sec.* time segments for each tile. Also, a server-client test-bed was implemented to analyze the performance of the proposed system in a realistic network

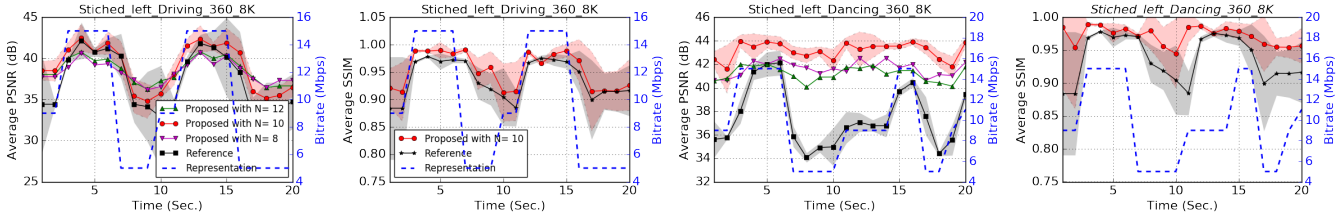

Fig. 2: Quality over time comparison between the proposed and the reference adaptive streaming solutions.

environment. We evaluated the algorithms at a perceptually acceptable quality level for VR video applications [4]. The connection between the server and the client was a local wired connection, with varying bandwidth between 4 and 22 *Mbps*.

The proposed method, denoted as *proposed*, divides each video frame into N tiles. Two tiles were utilized for the poles, and $N-2$ tiles were used for the equator. In this test, we experimented with several settings of N. Encoded bitrate for each tile is equally distributed by dividing the *target bitrate* to the N tiles. The DASH VR player was implemented using three APIs, namely, three.js [28], WebVR [29], and dash.js [30]. On the contrary, the reference system, denoted as *reference*, which works in similar principle in existing professional streaming services, neither uses viewport-aware nor tiling techniques. The *reference* transmits each 8K ERP video frame as a single tile using DASH.

PSNR and SSIM measures were employed to evaluate the objective viewport quality. For that, we recorded real view trajectories of 8 users viewing the content on Oculus Rift DK2 [1]. Each participant session started after a 10 *sec.* training video. Average inside-viewport quality scores were calculated over time using the participants' trajectories.

4.2. Performance Evaluation

To evaluate the quality performance of the *proposed* and the *reference* adaptive streaming methods over varying network throughput, average qualities were calculated using the participants' changing viewport over the time. N = 12, 10 and 8 were tested to investigate the impact of the number of tiles. Fig. 2 shows the comparison between the *proposed* and the *reference* streaming solutions in terms of *average* PSNR and SSIM with their *variations* over time. In the figure, the filled area shows the variation, and the left and right axes represent the objective measures and the bitrate representation, respectively. The dashed-line denotes the selected bitrate representation by the DASH client.

The results show that the *proposed* with N=10 considerably increases the streaming quality compared with the *reference* at all times. Also, the results show a quality drop after each 1 *sec.* segment playback. The reason is the viewport movement before delivering new segments. Because we used a distance-based bitrate distribution, visual quality was preserved effectively, compared with the *reference*, for each content. The bitrate distribution of the *Stiched_left_Driving360_8K* sequence is illustrated in Fig 3, where the inside-viewport tiles use the highest bitrates. In addition, as it can be seen in Fig 2, the *proposed* with N=10 achieves 1.66/0.02 and 5.72/0.04 average PSNR(dB)/SSIM gains with respect to the *reference* method for the *Stiched_left_Driving360_8K* and *Stiched_left_Dancing 360_8K* sequences, respectively.

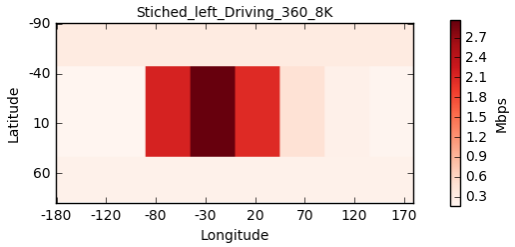

Fig. 3: Bitrate distribution of the 4^{th} participant at 4 *sec.* with N = 10. Selected representation and calculated viewport quality are 9 *Mbps* and 40.156 *dB*, respectively.

In this work, we demonstrated that the *proposed* streaming performance depends on the viewport movement activity and content type. For example, for the *Stiched_left_Driving360_8K* sequence, the *proposed* with N = 12 and 8 methods preserves complex details at low bitrate, and provides higher visual quality. Also, in the *proposed* with $N = \{12, 10, 8\}$ methods, a considerable quality gain was achieved with respective to the *proposed* for the *Stiched_left_Dancing360_8K* sequence. In this sequence, the *proposed* with N = 10, which uses the optimum N size in our system, achieved the highest quality performance relative to using $N = \{12, 8\}$.

5. CONCLUSIONS

This paper introduced a novel end-to-end streaming system for VR, which resulted in enhanced viewport quality under varying bandwidth and different viewport trajectories. The proposed system includes tiling, a novel MPD for DASH, and viewport-aware bitrate level selection methods. The quality performance of the proposed system was verified in simulations with varying network bandwidth using realistic view trajectories recorded from user experiments. Experimental results showed that significant quality enhancement was achieved by the proposed method compared with the reference solution that is currently used by professional VR streaming services. Future research will be devoted to investigating tile size optimization and tile discarding, as they are expected to further increase the visual quality.

6. REFERENCES

[1] Oculus VR, "Oculus rift," https://www3.oculus.com/en-us/rift/, Accessed: 2017-1-5.

[2] YouTube, "Virtual reality," https://www.youtube.com/channel/UCzuqhhs6NWbgTzMuM09WKDQ, Feb 2017.

[3] "Requirements for high quality for VR," Tech. Rep. MPEG 116/M39532, JTC1/SC29/WG, ISO/IEC, Chengdu, CN, Oct. 2016.

[4] Huawei Technologies co., LTD., *Whitepaper on the VR-Oriented Bearer Network Requirement*, 15 Sept. 2016.

[5] ISO/IEC 23009-1, "Information technology — dynamic adaptive streaming over HTTP (DASH) — part 1: Media presentation description and segment formats," Tech. Rep., ISO/IEC JTC1/SC29/WG11, 2014.

[6] J. D. Moss and E. R. Muth, "Characteristics of head-mounted displays and their effects on simulator sickness," *Human Factors*, vol. 53, no. 3, pp. 308–319, April 2011.

[7] M. Meddeb, M. Cagnazzo, and B. Pesquet-Popescu, "ROI-based rate control using tiles for an HEVC encoded video stream over a lossy network," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 1389–1393.

[8] D. Stenberg, "HTTP2 explained," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 3, pp. 120–128, 2014.

[9] K. Misra, A. Segall, M. Horowitz, S. Xu, A. Fuldseth, and M. Zhou, "An Overview of Tiles in HEVC," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 6, pp. 969–977, Dec 2013.

[10] O. A. Niamut, E. Thomas, L. D'Acunto, C. Concolato, F. Denoual, and S. Y. Lim, "MPEG DASH SRD: Spatial relationship description," in *7th International Conference on Multimedia Systems*. 2016, MMSys '16, pp. 5:1–5:8, ACM.

[11] "360 video and virtual reality (VR) in JW player part 1 : State of the industry," https://www.jwplayer.com/blog/360-vr-part1-state-of-the-industry/, Feb 2017.

[12] "Spatial relationship description, generalized URL parameters and other extensions," Tech. Rep., ISO/IEC 23009-1:2014/Amd 2, 2015.

[13] C. Grunheit, A. Smolic, and T. Wiegand, "Efficient representation and interactive streaming of high-resolution panoramic views," in *2002 International Conference on Image Processing (ICIP)*, Sept. 2002, vol. 3, pp. III–209–III–212 vol.3.

[14] A. Smolic and P. Kauff, "Interactive 3-D video representation and coding technologies," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 98–110, Jan 2005.

[15] P. R. Alface, J. F. Macq, and N. Verzijp, "Evaluation of bandwidth performance for interactive spherical video," in *2011 IEEE International Conference on Multimedia and Expo*, July 2011, pp. 1–6.

[16] S. Heymann, A. Smolic, K. Mueller, Y. Guo, J. Rurainsky, P. Eisert, and T. Wiegand, "Representation, coding and interactive rendering of high-resolution panoramic images and video using MPEG-4," in *Panoramic Photogrammetry Workshop*, Berlin, Germany, Feb. 2005, pp. 24–25.

[17] X. Corbillon, A. Devlic, G. Simon, and J. Chakareski, "Viewport-adaptive navigable 360-degree video delivery," *arXiv:cs.MM 1609.08042v1*, vol. cs.MM, no. 1609, Sep. 2016.

[18] R. Skupin, Y. Sanchez, C. Hellge, and T. Schierl, "Tile based HEVC video for head mounted displays," in *IEEE International Symposium on Multimedia (ISM)*, San Jose, CA, USA, Dec 2016, Accessed: 2017-1-16.

[19] J. Le Feuvre and C. Concolato, "Tiled-based adaptive streaming using MPEG-DASH," in *7th International Conference on Multimedia Systems*, New York, NY, USA, 2016, MMSys '16, pp. 41:1–41:3, ACM.

[20] "Descriptions of core experiments on DASH amendment," Tech. Rep. MPEG2016/ N16224, JTC1/SC29/WG, ISO/IEC, Geneva, Switzerland, June 2016.

[21] J. P. Snyder, *Flattening the Earth: Two Thousand Years of Map Projections*, University of Chicago Press, 1993.

[22] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *2015 IEEE International Symposium on Mixed and Augmented Reality*, 2015, pp. 31–36.

[23] M. Budagavi, J. Furton, G. Jin, A. Saxena, J. Wilkinson, and A. Dickerson, "360 degrees video coding using region adaptive smoothing," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 750–754.

[24] "Information technology — coding of audio-visual objects —part 10: Advanced Video Coding," Tech. Rep., ISO/IEC 14496-10, 2010.

[25] G. Bang, G. Lafruit, and M. Tanimoto, "Description of 360 3D video application exploration experiments on divergent multiview video," Tech. Rep. MPEG2015/ M16129, ISO/IEC JTC1/SC29/WG11, Chengdu, CN, Feb. 2016.

[26] "VideoLAN," http://www.videolan.org/developers/x264.html, Feb 2017.

[27] J. Le Feuvre, C. Concolato, J.-C. Dufourd, R. Bouqueau, and J.-C. Moissinac, "Experimenting with multimedia advances using GPAC," in *Proceedings of the 19th ACM International Conference on Multimedia*, New York, NY, USA, 2011, MM '11, pp. 715–718, ACM.

[28] "JavaScript 3D library. https://threejs.org/," https://github.com/mrdoob/three.js/, Feb 2017.

[29] "WebVR: Bringing virtual reality to the web," https://webvr.info/, Feb 2017.

[30] "A reference client implementation for the playback of MPEG DASH via javascript and compliant browsers.," https://github.com/Dash-Industry-Forum/dash.js, Feb 2017.