Impact of Video Streaming Delay on User Experience with Head-Mounted Displays

Adam Grzelka
Chair of Multimedia Telecommunications and Microelectronics
Poznań University of Technology
Poznań, Poland
adam.grzelka@put.poznan.pl

Adrian Dziembowski
Chair of Multimedia Telecommunications and Microelectronics
Poznań University of Technology
Poznań, Poland
adrian.dziembowski@put.poznan.pl

Dawid Mieloch
Chair of Multimedia Telecommunications and Microelectronics
Poznań University of Technology
Poznań, Poland
dawid.mieloch@put.poznan.pl

Olgierd Stankiewicz
Chair of Multimedia Telecommunications and Microelectronics
Poznań University of Technology
Poznań, Poland
olgierd.stankiewicz@put.poznan.pl

Jakub Stankowski
Chair of Multimedia Telecommunications and Microelectronics
Poznań University of Technology
Poznań, Poland
jakub.stankowski@put.poznan.pl

Marek Domański
Chair of Multimedia Telecommunications and Microelectronics
Poznań University of Technology
Poznań, Poland
marek.domanski@put.poznan.pl

***Abstract*—Delays in the delivery of immersive video to Head Mounted Devices (HMDs) are considered in the paper. The goal of this paper is to analyze the impact of video streaming latency on user experience with HMDs. The paper reports the results of the subjective quality assessment as a function of delay. The possible practical solutions are identified and, on this base, an experimental model of the considered systems is proposed. In order to properly test the influence of the streaming delay on the quality, the extensive subjective tests have been performed. Interesting conclusions have been drawn, showing that human acceptance of delay of translation of virtual point of viewing is much stronger than for delay of rotation of virtual viewing direction in immersive video. Finally, the influence of such observations on the system architectures is concluded.**

***Keywords*—Head-Mounted Display, Virtual Reality, Quality of Experience, video streaming**

I. Introduction

The Virtual Reality (VR) market is rapidly growing and an indication of that is the forecast that even the mobile communication traffic related to virtual and augmented reality is growing by about 60% annually [17]. This enormous expansion stimulates extensive research efforts related to immersive video technologies. Such technologies include virtual navigation, free-viewpoint television [21], omnidirectional video delivery, and others. These technologies are used for services that exploit tablets, smartphones, personal computers, television sets, head-mounted displays (HMDs) or other devices as user terminals. Among them, the head-mounted devices seem to be of particular interest. Currently, such devices are mostly equipped with very limited processing power, and they rely on processing power of a computer to a large extent.

In general, one may consider two extreme models of cooperation between a server and a user device, say an HMD (Fig. 1). In the first model (Fig. 1a), the whole 3D representation of a scene is transmitted to an HMD, and the rendering is executed there with the local usage of sensor data for estimation of a virtual position of a viewer as well as its viewing direction. This approach exhibits some obvious drawbacks as whole data related to 3D representation of a scene must be transmitted to the HMD that has to store and process it. Therefore, for that solution, the HMD would need to be equipped with substantial memory and processing power that could be provided by hardware with substantial weight and strong power supply.

On the other hand, the second model (Fig. 1b) assumes that an HMD is equipped with minimum processing power, and video data is rendered by a computer. For this purpose, the server renders current viewing window using HMD sensor data transmitted from the HMD. These sensor data provide information about HMD position and viewing direction. The output of the server is a sequence of viewing windows called also as viewports estimated at each time instant (temporal sample). As compared to the model depicted in Fig. 1a, this scenario is much more realistic for the devices already present on the market. In this model, an HMD is equivalent to a stereoscopic display with an additional ability to measure its position and motion and to send these measurements to the server. Therefore, our considerations are limited to the second model (Fig. 1b) and its modifications only.

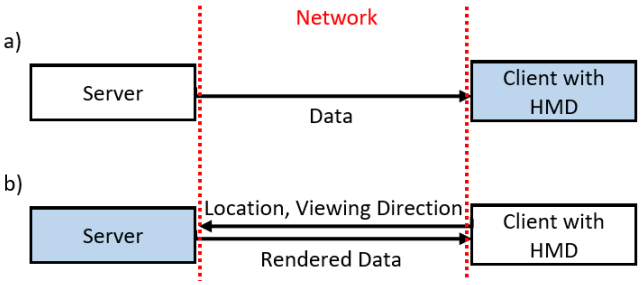

Fig. 1. Basic server and user data exchange models (video processing is performed in blue blocks).

In the scenario described above, the most common content is 360-degree video. Using the data from HMD sensors, the current location of the viewer wearing the HMD as well as its head position, i.e. the viewing direction are estimated. Using these data, the two current views are rendered, i.e. a stereoscopic viewport is generated for each time instant. Both the bi-directional transmission and rendering result in some latency in video update caused by the motion of a viewer. This latency is even increased if an HMD is more distant from its server and the respective communication is implemented through a network.

The goal of this paper is to analyze the impact of the abovementioned video streaming latency on user experience

when HMDs are used in the scenario presented in Fig. 1b. It is known that humans wearing HMDs are quite intolerant to latency of view update after a head movement. In contrary, the experiments made by the authors demonstrate that viewers using tablets are much more tolerant to delays of content updates after a motion of a finger on a touch screen that defines virtual motion of the virtual viewer or the related viewport. In the latter situation [21], a viewer accepts latency even exceeding 200 milliseconds. It is not the case with HMDs, which is researched here.

Unfortunately, in the references, there are quite few results that refer to the quality of HMD user experience as a function of video streaming latency, despite the problem of quality of experience in VR was noticed even 20 years ago [7]. The available research focuses mainly on the video or picture quality – for dense multiview video content [4] or for tiles transmission with various resolutions [1]. However many such research works show that a human requires perfect smoothness for angular resolution while space resolution can be less precise [11], [12].

The authors believe that the knowledge about the impact of video streaming latency on full user experience in virtual reality system is necessary for good understanding of the requirements and limitations that are vital for further development VR technology, in particular, the applications of HMDs.

II. Statement of the Problem

In order to comfortably use HMD, a displayed view should change instantaneously in accordance to the actual position of the head of a viewer. In an ideal case, in order to provide the best quality of immersive experience, the latency between the user movement and displaying the viewport has to be eliminated. Unfortunately, all required operations including rendering in the server (T_s), streaming (uplink and downlink – T_{Nup}, T_{Ndown}) and displaying the viewport (client processing time – Tc), take a considerable amount of time.

$$T = T_{Nup} + T_s + T_{Ndown} + T_C \qquad (1)$$

The goal of this research is to study and estimate how a delay (latency T) in various transmission scenarios impact human perception and quality of immersive experience.

III. System Architecture

Several solutions for the virtual reality system architecture have been developed recently. They can be divided into two main categories: viewport-dependent (Fig. 2) and viewport-independent (Fig. 3), described below.

In the viewport-dependent architecture (Fig. 2), the server requires whole tracking information (position and rotation) and sends to clients only the data needed for their current viewport (e.g. stereo views). In this approach, view track prediction algorithms can be used to decrease latency filling [1], [15], [16].

The second architecture is viewport-independent transmission (Fig. 3), in which server requests information on client position in a scene only (without rotation) and sends an omnidirectional video (e.g. as stereo omnidirectional view or stereo cube-map projection). Such a video has to be processed at the client side to show the current viewport.

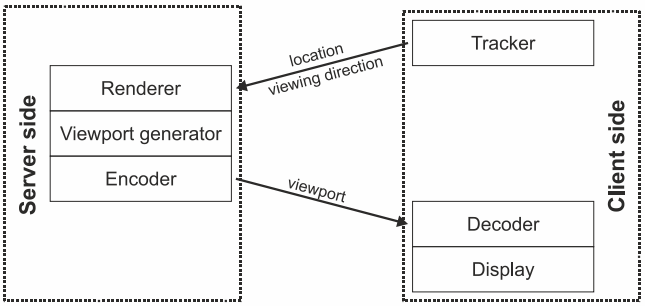

Fig. 2. Viewport-dependent scenario.

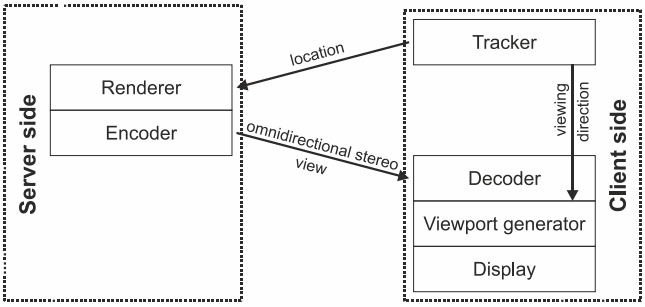

Fig. 3. Viewport-independent scenario.

Both of these architectures can be extended, optimized, and mixed. For example, in the second architecture, the server can receive all tracking information (position and rotation) and use it to send data to a client in a more efficient way, e.g. by improving the video quality in the virtual viewing direction and deteriorate it in other directions. This can be attained by means of adaptive choice of resolution of the transmitted video or adaptive quantization, or even adaptive choice of frame rate (e.g. using [10], [13], [14], MPEG-DASH [9] or MPEG-OMAF [11] solutions). Such optimization has one significant drawback, because when the latency is high, quality of an image presented may decrease when viewers rapidly rotate their heads. However, the rotation perception will be natural because of no latency, therefore the overall quality of experience will be much higher than for viewport-dependent scenario.

Apart from optimizations, the most important feature of the viewport-independent architecture is that a viewer is able to process received data and prepare any angle of the viewing window. Therefore the latency of video update in response to head movement depends on delays on the client side only.

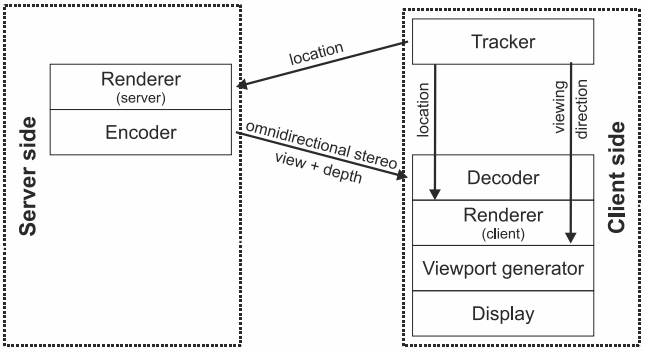

Fig. 4. Viewport-independent scenario with rendering on the client side.

Another extension of viewport-independent architecture is shown in Fig 4. The server sends omnidirectional video (related to current location of the HMD) with corresponding depth maps, and the client reprojects streamed view to a new position. In such a scenario, visible delays of the motion of virtual view depend only on client-side processing and can be similar to delays in offline games and applications. Despite the latency reduction offered by such solutions, they still show some significant disadvantage: as rendering is performed on a client side, the quality of a view may be decreased. A view provided by a server is reprojected into the position of a client which causes rendering artifacts, e.g. cracks, resolution changes and – primarily – disocclusions. There are many sophisticated algorithms, which can fill disoccluded areas (e.g. [18], [19], [20]), however in this approach, all the computations have to be performed on a client side (instead of a powerful server), obviously in the real time. The disadvantage makes the scenario presented in Fig. 4 quite impractical, because, except for slight viewer's movements, the quality of a view presented to a user may be decreased, additionally all computations should be performed in real time by a user.

IV. EXPERIMENTS

The experiments were aimed at estimation of the quality of user experience as function of the delays related to communication between server and HMD. The subjects were doing their assessments by watching static 360-degree images. As such 360-degree images were rendered according to the current position of HMD they are called sequences.

In the experiments, two scenarios (Fig. 2 and 3) were used with some minor changes. First, in order to assess only the effects related to latency and not video quality itself, we used the system without compression – encoder and decoder blocks were skipped. Secondly, the latency existing in the system in a real-world application, e.g. coming from delays introduced during compression, the position tracker, the rendering server etc. (Fig. 5) was emulated by the means of parametrized delay block. This delay block can add a delay to viewing window parameters for rotation and for position of a user independently.

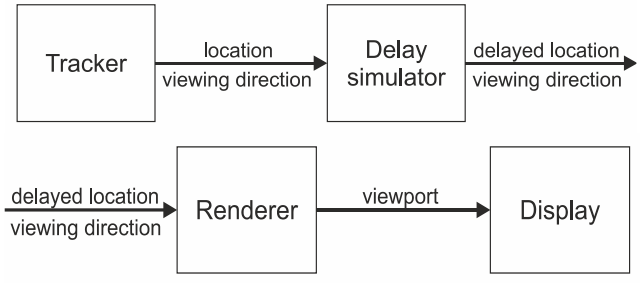

Fig. 5. Data flow in the experiment.

Switching between the viewport-dependent scenario and viewport-independent scenario is made virtually by setting delays in the head movement and the viewport position. Owing to these simplifications we were able to test two scenarios and many different latency setups during the subjective tests. The main goal of this experiment was to estimate how big latency is acceptable for both of the presented scenarios.

During the tests, the Oculus Rift [8] device was used. The content was rendered by a dedicated application with the use of the Nvidia GTX 970 graphics card. The frame rate of the content was 90 fps for each eye of a user, which limited the possible delay changing step to 11ms. This corresponds to the delay of viewport parameters, i.e. location and viewing direction of HMD by a single frame. A delay caused by capturing and processing data by Oculus devices and software was unknown but similar to values observed in dedicated offline games and applications.

In the tests, two variants of the delay were tested: delay of both rotation and translation (what corresponds to the viewport-dependent scenario in Fig. 2), and translation delay only (viewport-independent scenario - Fig. 3).

The range of the tested delays was set before the presented experiment and was based on similar tests conducted earlier by the authors on the smaller group of viewers over a wider range of delays.

During one session, the participants assessed the experience related to one sequence only. At the beginning of sessions, viewers were instructed with two versions of the delays: one without delay and one with 110 ms delay (the case of very uncomfortable rotation and translation). Next versions were mixed for each participant independently. Every sequence was shown in 11 delay variants: one without delays (hidden reference) and five for each tested scenario. The session content is presented in Table 1. The participants were asked to give a score in 11-point MOS (Mean Opinion Score) scale: from 0 (very bad quality of experience) to 10 (excellent quality of experience).

TABLE 1. SESSION CONTENT

Presentation order	Rotation delay [ms]	Translation delay [ms]	Scenario
1	0	0	reference
2	110	110	anchor
R[a]	0	0	hidden reference
R	11	11	dependent
R	22	22	dependent
R	33	33	dependent
R	44	44	dependent
R	55	55	dependent
R	0	44	independent
R	0	88	independent
R	0	132	independent
R	0	176	independent
R	0	220	independent

[a] Random order for every participant

All sequences were tested by at least 15 persons. The participants could walk within a 2 by 2 meter square. In order to improve the quality of the results, the outliers were removed: if a subject rated a sequence with higher delay significantly better (2 or more in 11-point scale), their results were ignored.

The first frame of three omnidirectional sequences was used during the tests. Two of these sequences (Classroom-Video and TechnicolorMuseum) are recommended by ISO/IEC MPEG for 3DoF+ research [5], and one (Poznan-People360) is from the test set proposed by Poznań University of Technology and Electronics and Telecommunications Research Institute [6]. TechnicolorMuseum sequence was slightly preprocessed before the experiment – hemispherical views were merged into one omnidirectional view. The resolution of tested sequences was downsampled to 2048×1024 to reduce processing time and to ensure temporal smoothness of presented content.

V. RESULTS

In Fig. 6 the results for the system with delayed rotation and translation (viewport-dependent scenario) are presented. The distribution of scores given by participants for different delays is represented by colors.

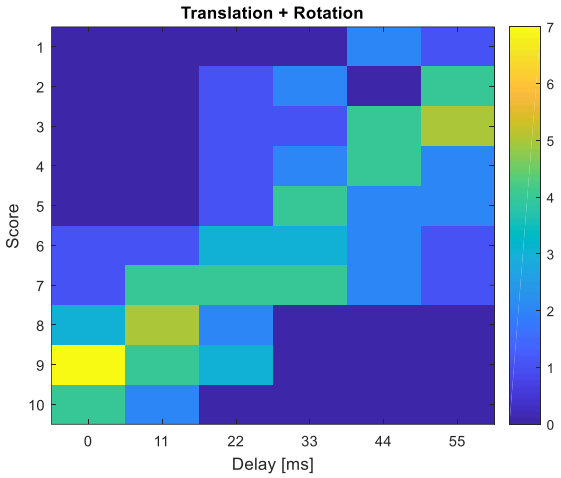

Fig. 6. The histogram of score distribution for different translation and rotation delays (ClassroomVideo test sequence).

In Fig. 7 the subjective quality averaged over all the participants is presented. The bars represent the 95% confidence interval.

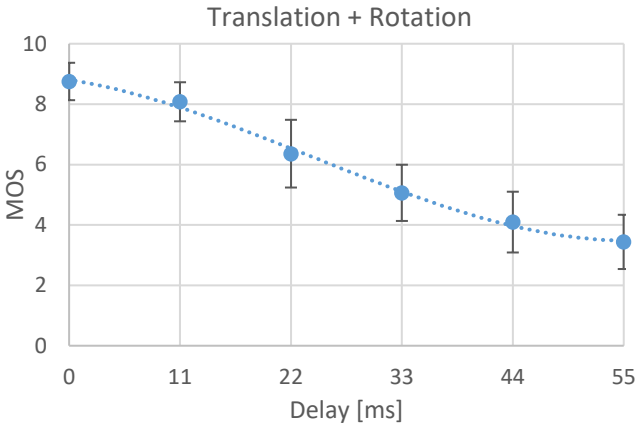

Fig. 7. Mean Opinion Score for delayed translation and rotation (ClassroomVideo test sequence).

In Figs. 8 and 9 the results for the system with translation delay (viewport-independent scenario) are presented. Fig. 8 contains score distribution of subjective quality experienced by individual viewers, Fig. 9 – the same quality values averaged over all participants.

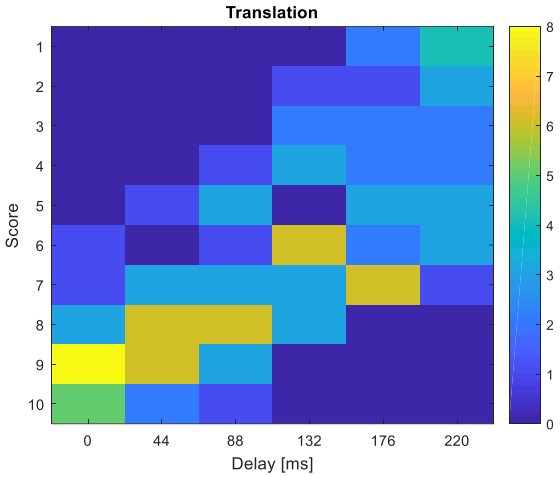

Fig. 8. The histogram of score distribution for different translation delays (ClassroomVideo test sequence).

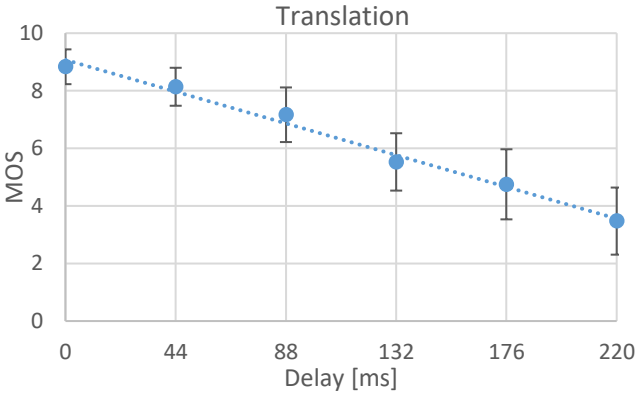

Fig. 9. Mean Opinion Score for delayed translation (ClassroomVideo test sequence).

In Figs. 10 and 11 the Mean Opinion Scores for all tested omnidirectional sequences are presented. Fig. 10 contains results obtained for the system with delayed translation and rotation (viewport-dependent scenario), Fig. 11 – system with delayed translation (viewport-independent scenario). The confidence interval bars are skipped in order to preserve the clarity of presented data. However, it can be observed that for all used sequences the confidence interval bars are similar thus they are independent on the content.

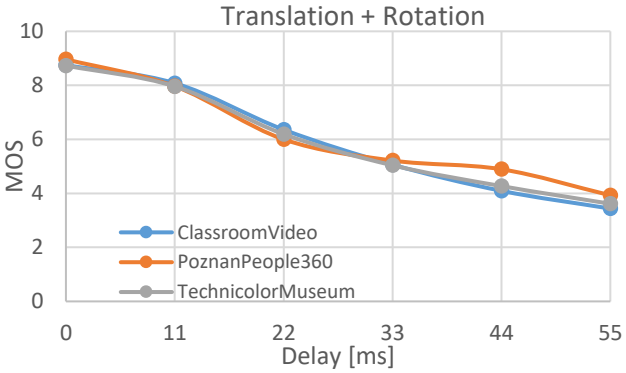

Fig. 10. Mean Opinion Score for delayed translation and rotation.

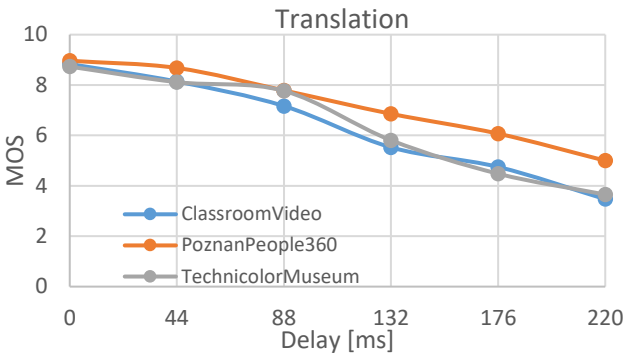

Fig. 11. Mean Opinion Score for delayed translation.

VI. Conclusions

In the paper, we have presented novel results related to the impact of delays on quality of experience in immersive video transmission for head-mounted displays. The possible practical solutions have been identified and basing on them we have proposed an experimental model of the considered systems.

The performed experiments led to interesting conclusions. Firstly, human acceptance of delay of translation of virtual point of viewing is much stronger than for delay of rotation of virtual viewing direction in immersive video. For example, in order to attain good quality (MOS score 8), the latency of about 44 ms is acceptable for translation delay, but when rotation and translation are delayed, the latency must be limited to about 11 ms.

Secondly, by increasing delay, MOS is decreasing faster for translation and rotation MOS. For translation and rotation scenario, for delay of 22 ms, we get MOS equal 6. In the translation scenario, MOS is still over 7 for delay of 88 ms.

In conclusion, we state that omnidirectional view transmission can give the same quality when the delay is 44ms instead of 11 ms when the stereo-view transmission is used. However, it must be also noted, that scenario with omnidirectional view reprojection can provide high quality of experience even when the latency is higher than 44 ms.

Acknowledgement

The research was supported by the Ministry of Science and Higher Education.

References

[1] Y. S. de la Fuente, G. S. Bhullar, R. Skupin, C. Hellge and T. Schierl, "Delay impact on MPEG OMAF's tile-based viewport-dependent 360° video streaming," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 18-28, March 2019. doi: 10.1109/JETCAS.2019.2899516

[2] J. M. P. Van Waveren, "The asynchronous time warp for virtual reality on consumer hardware", in *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology, VRST '16*, pp. 37–46. ACM, New York, NY, USA, 2016. doi: 10.1145/2993369.2993375

[3] J.P. Freiwlad, N. Katzukis, P. Steincke, "Camera Time Warp: Compensating Latency in Video See-ThroughHead-Mounted-Displays for Reduced Cybersickness Effects", in *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology, VSRT '18*, Tokyo, Japan, 2018. doi: 10.1145/3281505.3281521

[4] J. Cubelos, P. Carballeira López, J. Gutiérrez and N. García, "QoE analysis of dense multiview video with head-mounted devices," *IEEE Transactions on Multimedia*, doi: 10.1109/TMM.2019.2924575, early access

[5] "Common Test Conditions on 3DoF+ and Windowed 6DoF," ISO/IEC JTC1/SC29 WG11 MPEG 124th meeting, Macau, October 2018, Doc. N18089.

[6] O. Stankiewicz, K. Wegner, A. Dziembowski, M. Lorkiewicz, G. Lee, J. Seo, M. Domański, "Proposed test materials for 3DoF+ or Omnidirectional 6DoF," ISO/IEC JTC1/SC29 WG11 MPEG 124th meeting, Macau, October 2018, Doc. M44461.

[7] J. J. LaViola, Jr., "A discussion of cybersickness in virtual environments," *SIGCHI*, vol. 32, no. 1, pp. 47–56, 2000.

[8] Oculus, [Online]: https://www.oculus.com. Accessed in 2019.

[9] M. Hosseini and V. Swaminathan, "Adaptive 360 VR Video Streaming Based on MPEG-DASH SRD," in *2016 IEEE International Symposium on Multimedia (ISM)*, San Jose, CA, 2016, pp. 407-408.doi: 10.1109/ISM.2016.0093

[10] C. Díaz et al., "Viability analysis of content preparation configurations to deliver 360VR video via MPEG-DASH technology," in *2018 IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas, NV, 2018, pp. 1-2.

[11] D. Runde, "How to realize a natural image reproduction using stereoscopic displays with motion parallax," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 3, pp. 376-386, April 2000.

[12] F. Speranza, W. Tam, T. Martin, L. Stelmach and C.-H. Ahn, "Perceived smoothness of viewpoint transition in multi-viewpoint stereoscopic displays," *Proceedings of SPIE - The International Society for Optical Engineering*. Vol. 5664, 2005. doi: 10.1117/12.587170.

[13] K. K. Sreedhar, A. Aminlou, M. M. Hannuksela and M. Gabbouj, "Viewport-adaptive encoding and streaming of 360-Degree video for virtual reality applications," in *2016 IEEE International Symposium on Multimedia (ISM)*, San Jose, CA, 2016, pp. 583-586.

[14] X. Corbillon, G. Simon, A. Devlic and J. Chakareski, "Viewport-adaptive navigable 360-degree video delivery," in *2017 IEEE International Conference on Communications (ICC)*, Paris, 2017, pp. 1-7. doi: 10.1109/ICC.2017.7996611.

[15] A. D. Aladagli, E. Ekmekcioglu, D. Jarnikov and A. Kondoz, "Predicting head trajectories in 360° virtual reality videos," in *2017 International Conference on 3D Immersion (IC3D)*, Brussels, 2017, pp. 1-6.

[16] Y. Bao, H. Wu, T. Zhang, A. A. Ramli and X. Liu, "Shooting a moving target: Motion-prediction-based transmission for 360-degree videos," in *2016 IEEE International Conference on Big Data (Big Data)*, Washington, DC, 2016, pp. 1161-1170.

[17] White Paper: "Cisco Visual Networking Index: Forecast and Methodology", Cisco Visual Netw. Index, San Jose, CA, USA, pp. 2017–2022, Feb. 18, 2019.

[18] A. Khatiullin, M. Erofeev and D. Vatolin, "Fast occlusion filling method for multiview video generation," in *2018 - 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Helsinki, 2018, pp. 1-4.

[19] C. Zhu and S. Li, "Depth image based view synthesis: New insights and perspectives on hole generation and filling," *IEEE Transactions on Broadcasting*, vol. 62, no. 1, pp. 82-93, March 2016.

[20] G. Luo and Y. Zhu, "Foreground removal approach for hole filling in 3D video and FVV synthesis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 10, pp. 2118-2131, Oct. 2017.

[21] O. Stankiewicz, M. Domański, A. Dziembowski, A. Grzelka, D. Mieloch and J. Samelak, "A free-viewpoint television system for horizontal virtual navigation," in *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2182-2195, Aug. 2018.