ŠTEFICA MRVELJ, Ph.D.[1]
E-mail: smrvelj@fpz.unizg.hr
MARKO MATULIN, Ph.D.[1]
(Corresponding author)
E-mail: mmatulin@fpz.unizg.hr
SERGO MARTIROSOV, Ph.D. Student[2]
E-mail: martiros@rti.zcu.cz
[1] University of Zagreb
Faculty of Transport and Traffic Sciences
Vukelićeva 4, 10000 Zagreb, Croatia
[2] University of West Bohemia
Univerzitni 2732/8, 30100 Pilsen, Czech Republic

Information and Communication Technology
Original Scientific Paper
Submitted: 21 Dec. 2019
Accepted: 10 Apr. 2020

SUBJECTIVE EVALUATION OF USER QUALITY OF EXPERIENCE FOR OMNIDIRECTIONAL VIDEO STREAMING

ABSTRACT

This paper reports on the results of subjective testing of user Quality of Experience (QoE) for omnidirectional video (ODV) streaming quality. The test was conducted among 20 test subjects who watched three ODVs using a Head Mounted Display (HMD) system. The length of the videos was between two and three minutes. The first video was used for training purposes and contained no quality degradations. The quality of the other two ODVs was degraded by manipulating the resolution or by introducing different frame drop patterns. While watching the pre-prepared videos the subjects indicated if they noticed the changes in the quality and then rated it. After watching each video, the subjects completed a separate questionnaire, which evaluated their level of enjoyment and discomfort with the video. The results showed that the degradation of both objective parameters (video resolution and frame rate) impacted the subjects' perception of quality; however, the impact was somewhat alleviated in ODV which contained dynamic scenes and fast camera movements.

KEY WORDS

360-degree video; streaming; video resolution; frame drop; quality of experience;

1. INTRODUCTION

In recent years reliable multimedia content delivery over heterogeneous networks has become a paramount objective that network operators and service providers must achieve if they want to survive on the competing markets. Ensuring the level of service quality that can cope with the increasing user demands is crucial since a sheer number of service providers enables end-users to change their provider of choice with ease. This is true for both mobile and fixed networks. Being at the forefront of the multimedia content usage, video streaming and online gaming services are pushing the development of these networks. For instance, the fifth-generation mobile network (5G) is currently being developed and piloted as an answer to user demands who want reliable high-definition and ultra-high-definition video streaming services and online gaming with hundreds of other players in real-time, both of which require a significant amount of network resources. The support for this statement can be found in [1] where Ericsson reports how users expect the 5G network to (a) provide them with more mobility through a more stable, faster and high-bandwidth network, (b) increase the network efficiency, thus, prolonging the battery life of the mobile devices, (c) improve their virtual reality (VR) experience due to higher resolution which will be supported, (d) make VR more accessible through cheaper data plans, and (e) decrease the feelings of nausea and sickness in VR by providing high-bandwidth and less-lag network.

This 'user-oriented' context of service provision means that the quality of the delivered service must be holistically evaluated so that different influential factors, as well as the significance of their impact on the user perception, could be disclosed. To this end, services are normally evaluated using objective (network dependent) and subjective (user dependent) quality measures. In this paper, the focus is on the subjective evaluation of omnidirectional video

(ODV) streaming quality, i.e. the videos which surround the viewing angle of a user. The paper represents a) the continuation of our previous work [2, 3] where extensive subjective evaluation for video streaming service was conducted and the results were used for modelling the user perception based on a number of objective input parameters, and b) one of the outputs of the Quality of Experience for Virtual Reality Applications (QoE4VR) project [4].

What motivated us for this work is the increasing popularity of ODV streaming service together with an increasing market penetration of devices capable of presenting different VR contents such as videos, games, or learning applications. For instance, Cisco in [5] forecasts that the usage of VR applications (mainly online games and video streaming) will increase 12-fold between 2017 and 2022, reaching 4.02 Exabyte per month of data traffic. Additionally, our decision to step into this research path was highly motivated by the increasing interest of the research community from this domain that is currently beginning to analyse QoE for ODV streaming. The contribution of the paper can be outlined as follows.

1) An overview of current approaches to the objective and subjective evaluation of ODV quality is given.
2) A methodology for conducting the subjective evaluation of ODV streaming quality is proposed.
3) The impact of two objective parameters (ODV resolution and frame drop) on user perception is investigated and the obtained results are presented and discussed.
4) An appraisal of the implemented research methodology is provided.

The paper is organized as follows. In Section 2 related work is discussed, focusing on objective and subjective evaluation of ODV quality. The description of the experiment conducted in our research can be found in Section 3, while the obtained results and discussion are presented in Section 4. The concluding remarks and future work are outlined in Section 5.

2. RELATED WORK

The analysis of the user QoE for ODV streaming as well as understanding how it changes in relation to different network and user environment conditions is still in its infancy and represents a major challenge for the scientific community. The size of the challenge is influenced by the new characteristics of this type of content and transfer technologies that are used to deliver it to the end-users. However, prior research achievements within the domain of evaluating 2D and 3D video quality can be significantly used for the analysis of ODV quality. Current state-of-the-art research about the ODV quality analysis is focused on describing and developing the coding techniques and projection methods, capacity planning of the networks that are used for ODV transfer as well as defining the new objective and subjective methods for the evaluation of the user QoE; the latter being the focus of this literature review.

As mentioned previously, the investigations by different researchers about the quality of ODV presentation are based on the already achieved know-how, i.e. the knowledge obtained from objective and subjective evaluation of 2D and 3D video quality. The objective evaluation of ODV quality originates from different upgrades of developed video quality metrics, with the main purpose of making the metrics more suitable for the evaluation of ODVs. For instance, the authors in [6] define the framework for evaluation of ODV quality which is based on the well-known and frequently used Peak Signal to Noise Ratio (PSNR) metrics. Similar work has been completed by the group of authors in [7] who modified the PSNR metrics, creating the weighted PSNR, thus, making it applicable for the ODV quality evaluation. Furthermore, in [8] the authors developed WS-PSNR (Weighted to Spherically uniform PSNR) metrics which calculates the difference between the pixels of the original and analysed image, by assigning different weights to the specific pixels depending on their position in the spherical image. Further upgrades of PSNR metrics include [9, 10] where S-PSNR-NN (Spherical PSNR Nearest Neighbor), S-PSNR-I (Spherical PSNR Interpolation) and CPP-PSNR (Crasters Parabolic Projection PSNR) metrics are developed, suitable for comparing the two images of different resolutions and/or projections. The Resized-PSNR (R-PSNR) metric is developed in [11]. The authors found that the traditional video quality metrics are not suitable for measuring the quality of ODV since they will take the redundant pixels into account, so they developed R-PSNR metrics which considers only, as the authors call it, the "meaningful" pixels for quality measurement while the redundant pixels are discarded.

Apart from the PSNR metrics, some authors are redeveloping SSIM (Structural Similarity) index, making it applicable in the process of evaluating ODV quality. For instance, in [12] the authors modified the index by introducing the link between a 2D stereoscopic image and a spherical representation of that image. The authors called the developed method the S-SSIM (Spherical SSIM) index. Another modification of the index is made in [13] where the performance results of the MS-SSIM (Multi-Scale SSIM) index are presented; this index enables evaluation of the two images (original and analysed) of different resolutions. Furthermore, Oznicar et al. in [14] showed how the analysis of a user viewing angle can be beneficial in the process of network capacity planning since it is not necessary to transmit the segments of the ODV which are not in the current focus of the user. The same group of authors upgraded their method in [15].

A group of authors in [16] achieved a breakthrough in the evaluation of ODV presentation quality by developing both objective and subjective evaluation methods. The objective method is also based on the PSNR metric, while the results obtained with the subjective method consider the user area of focus in the entire spherical image. It is shown that most test subjects are focused on the front region of the video while watching the ODVs, which is an expected result. Birkbeck et al. in [17] analyse the perceived ODV quality in relation to the type of video projection (Equirectangular, Cubemap, and Equi-Angular Cubemap projection), showing how the analysis of ODV quality and user experience requires devoting more time to additional research activities since the quality of 2D videos can be evaluated without focusing on these aspects. The relationship between ODV projection and the user experience is also analysed in [18, 19].

In [20] the authors developed a system (i.e. a computer application) used for subjective evaluation of ODV quality; the authors claim that the application can be used for the analysis of quality using different methods. The authors also conducted a pilot project, showing the workflow of their application in a typical use case scenario. Similarly to [20], an Android VR application called MIRO360 was developed in [21]. This application is used to assess the subjective quality of ODVs using different methodologies that are being studied for the development of ITU-T P.360-VR recommendation.

The impact of different impairments during the ODV streaming session on the user perception is investigated in [22]. The authors showed that, if triggered in proper time periods, the virtual walls and slowdown impairments (that change the way a user can interact with the ODV) are better perceived by the users than visual quality degradation from video compression. A group of authors in [23] was focused on the development of a new subjective method for ODV quality evaluation. Specifically, the authors investigated the impact of frame freeze effect on the user perception of quality. This research is especially interesting since the authors in conclusion of their work highlight that the scientific community should, in the future, invest additional effort in (a) establishing a publicly available ODV database, (b) development and standardization of objective and subjective methods for ODV quality evaluation and (c) conducting the analysis and evaluation of quality of user experience on a larger number of test subjects.

An example of how to use the results of the subjective evaluation of ODV quality for the development of QoE models is shown in [24]. The authors investigated the impact of different projection schemes, bitrate, spatial and temporal video characteristics on the user Mean Opinion Scores (MOS) and found that the developed models performed reasonably well (0.71 and 0.77 in Pearson linear and Spearman rank-order correlation coefficient scores, respectively).

It is noteworthy to mention a third category of QoE assessment methodologies, which exist between the subjective and objective ones. The category is called hybrid evaluation methods because they employ an automatic objective quality estimator and combine the derived rating with prior available subjective scores. Machine Learning tools often serve as a base of these hybrid methods, while subjective test scores are used as input to train a QoE model. The model then maps network parameters (e.g. network delay, jitter, packet loss rate and/or other) to MOS values [25]. However, the development of such models for QoE evaluation for ODV streaming is still expected.

From the perspective of our paper and this review, it is useful to see that different authors investigate the user subjective opinions about various quality degradations which can be caused by network impairments. Namely, the focus is on the video bitrate, framerate, resolution, frame freeze effect

and discovering the significance of their impact on the user QoE. Yet, the need for further research is often highlighted since the ODV streaming service is just starting to gain its popularity. Additionally, we learned that frequently used MOS discrete scale is employed also for subjective data collection during ODV quality evaluation, as is the case in this research.

3. EXPERIMENT SETUP

3.1 Preparation of test materials

For our experiment, we pre-prepared three ODVs. To avoid boredom of the test subjects, the content of each ODV was different (instead of using the same video with different properties three times). As an HMD system, an HTC Vive was used. The device uses two OLED panels, one per eye, each having a display resolution of 1,080×1,200 pixels (2,160×1,200 combined pixels). Since its 1,080p max resolution, there was no point in using videos with higher resolution, as there would be no additional quality added to the video viewing experience. All ODVs were downloaded with an initial quality of 1,080p @30 frames per second (fps) from YouTube, a video hosting service. To avoid any discomfort for test participants, the duration of the videos did not exceed 3 minutes.

The content of the first video was polar night and the northern lights in increased playback speed. The properties of this video were left unchanged since it was used for training purposes. Due to the mostly static camera and only polar lights movement, the video did not exhibit any discomfort for the participants.

The second video was used to test how changes of ODV resolution during playback affect the user's perception of its quality. For this, we downloaded the same video four times, each time in a different resolution (1,080p, 720p, 480p, and 360p). Then, in Adobe Premiere Pro video editing software, we took the 1,080p video and inserted several short clips of lower resolution videos as indicated in *Figure 1*. The objective was to emulate the changing network conditions, in which video resolution can be downgraded and upgraded during the streaming. Additionally, we inserted one re-buffering event which lasted for two seconds (indicated by the red rectangle in *Figure 1*). During those two seconds, the playback of the video stopped, and then it continued at a lower resolution (360p). Other changes in the ODV resolution indicated in the Figure were implemented without stalling the playback. Note that the content of the video was a guide through Buckingham Palace and the pace of the video was slow to moderate. The video also contained an audio guide.

The third video was manipulated with the purpose of testing how different frame drop patterns affect user QoE. Due to audio frame drop being so obvious to the ear and because we did not want participants to easily notice changes in the video quality, the audio track of the video was left untouched, while we cut out frames from the video itself to emulate frame drop events. Namely, 1, 2, 3 or 4 consecutive frames were cut out (*Figure 2*) with the purpose of creating the segments of the video containing different frame drop patterns. Then, the segments were inserted into the original video in random timeline locations (*Figure 3*). The specified number of frames were first cut, then the previous frame was copied to the locations of the missing frame(s) so that a picture does not become black while the video player is on an empty frame timeline, but rather remains still until the next new untouched frame starts. The same tool, Adobe Premiere Pro, was used in these processes.

The content of the video was a Hong Kong tour from the air. The camera was moving at low and fast speeds; the video contained dynamic music and an audio narration explaining some of the landmarks of the city.

Hence, the videos used in this experiment were prepared in advance and played to the participants during the testing. As seen from *Figures 1* and *2*, the start and the end of the videos were always left untouched so that the subjects could immerse themselves in the video at the beginning of the screening and contemplate what they had experienced toward the end.

3.2 Participants and test procedure

The experiment was conducted among 20 participants (25% of whom were female) between 20 and 25 years of age. The participants were selected from the population of students for two reasons: a) according to [26], people between the ages of 18 and 24 can be considered as a typical video streaming service users, and b) this particular population was easily accessible for conducting such a survey (i.e., the convenience sampling method [27] was used).

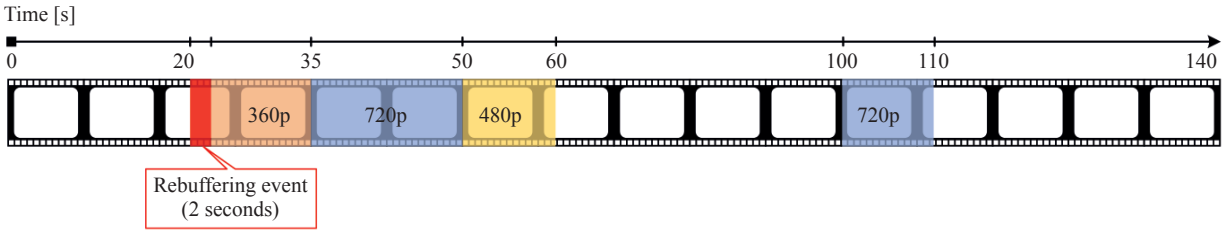

Figure 1 – Timeline of the video used for testing the effects of the resolution changes

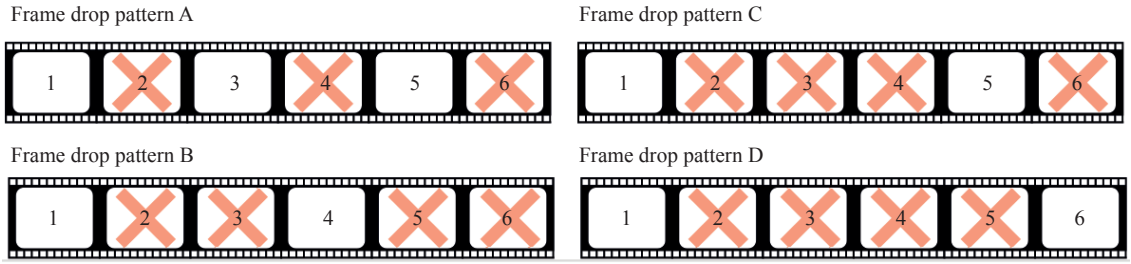

Figure 2 – Different frame drop patterns used in the simulation (X indicates the deleted frames)

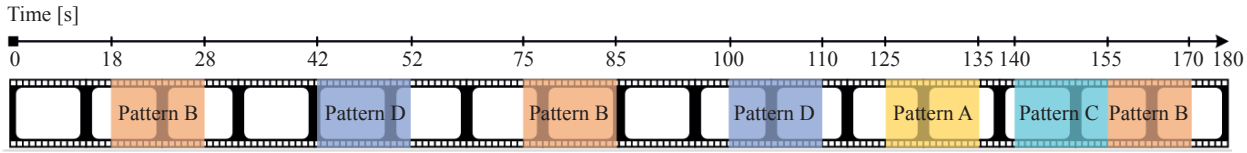

Figure 3 – Timeline of the video used for testing the effects of different frame drop patterns

Before the start of the test, the subjects were informed about the test; however, without revealing the quality degradations that would occur in the videos. They were instructed just to sit and relax and if during the video playback, they notice any changes in the video quality, state a grade between 1 (*bad quality*) and 5 (*excellent quality*), i.e. a typical MOS discrete scale was used for rating. During the playback, we were writing down all their subjective grades about the current level of quality in a simple timeline table that we prepared. Thus, the video playback was not interrupted due to the rating and we were able to collect user opinions for the specific video segment of each video. Note that some subjects stated more than one grade for the same segment of a video. This meant that either they could not decide between the two grades or they thought that the quality was changed when it remained the same. We recorded all the subjects' grades and used their median for further analysis in cases when more than one grade was given.

The tests were conducted in a well-isolated room with no outside sources of noise. The computer used for video playback had Intel Core i7 6800K processor, 16 GB of RAM and EVGA GeForce GTX 1080 Ti FTW3 GAMING video graphic card. As mentioned earlier, we used an HTC Vive HMD system. The participants were viewing the videos in a seated position on a chair that enabled them to rotate freely. The first video was shown to participants just to accustom them to the 360-degree perspective in VR and the HMD system used to render the ODVs, i.e. no rating of its quality was required. Apart from the first video, two other videos were shown to each participant in a different order, to eliminate any effects of one video type having on another. After watching each video, the subjects completed a short questionnaire about the video. The questionnaire contained the following questions:

1) Were you able to immerse yourself in the video? (available answers were: *Yes* | *Partially* | *No*).
2) Was the video interesting to you? (the answers were given on a discrete scale from 1 being *boring* to 5 being *very interesting*).
3) Did you enjoy watching the video? (the answers were given on a discrete scale from 1 being *no satisfaction* to 5 being *very satisfying*).
4) How stressful was it to watch the video? (the answers were given on a discrete scale from 1 being *not stressful* to 5 being *very stressful*).

5) Did you experience any sense of fear during watching? (the answers were given on a discrete scale from 1 being *no such feeling* to 5 being *a strong sense of that feeling*).
6) Did you experience any sense of balance loss during watching? (the answers were given on a discrete scale from 1 being *no such feeling* to 5 being *a strong sense of that feeling*).
7) Did you experience any sense of nausea during watching? (the answers were given on a discrete scale from 1 being *no such feeling* to 5 being *a strong sense of that feeling*).
8) Considering your answers to questions 5 to 7, are the provided answers impacted by the camera position and movement in the video? (the answers were given on a discrete scale from 1 being *no impact* to 5 being *very strongly impacted*).
9) Please evaluate the overall quality of the video presentation and your watching experience. (the answers were given on a discrete scale from 1 being *bad* to 5 being *excellent*).

The time required to complete this questionnaire also served to the subjects as a short rest from the HMD system and allowed us to prepare the next test sequence for rating. After watching all three videos and completing the questionnaires, the subjects were asked to state how often they consume VR content and to compare ODV playback in VR with normal 2D video playback on TV or computer screens. Lastly, the subjects were also asked not to discuss the purpose of the experiment nor the quality degradation of the videos with anybody during the test period which lasted for two weeks. Since we were in the room with the subjects during the screening and rating, we were able to provide further explanation to the subjects if needed. Hence, we did not have to discard any data collected in this manner due to the misinterpretation of the questions. The duration of the whole test was between 25 and 30 minutes per subject.

4. RESULTS AND DISCUSSION

4.1 Background data

Prior to presenting the results of the analysis on how video resolution and different frame drop patterns affected the participants' QoE, it is important to understand the role of other influential factors. Hence, in this section, we first analyse the data collected with the questionnaires which were completed by the subjects after watching each ODV. These results will serve as an additional source of information that will be useful for data interpretation and discussion which will follow. The data are presented in *Table 1*. Note that the analysis of the first question from the questionnaire is not included in the table since the answers to that question were not quantifiable. Furthermore, the first video was used for training purposes only, hence, it is left out of the table as well.

The subjects found the Hong Kong air tour video more interesting than the Buckingham Palace video, with an average rating (Avg) of 4.15; the standard deviation (StDev) equalled 0.81. When asked if

Table 1 – Subjects' responses to the questions asked after watching each video

Question (the available answers to the questions can be found in Section 3.2)	The second video (Buckingham Palace)		The third video (Hong Kong)	
	Avg	StDev	Avg	StDev
2. Was the video interesting to you?	3.55	0.94	4.15	0.81
3. Did you enjoy watching the video?	3.20	1.06	3.90	0.72
4. How stressful was it to watch the video?	1.30	0.66	1.25	0.44
5. Did you experience any sense of fear during watching?	1.15	0.37	1.30	0.57
6. Did you experience any sense of balance loss during watching?	1.60	1.14	1.45	0.83
7. Did you experience any sense of nausea during watching?	1.50	1.15	1.20	0.89
8. Considering your answers to questions 5 to 7, are the provided answers impacted by the camera position and movement in the video?	2.10	1.48	2.20	1.51
9. Please evaluate the overall quality of the video presentation and your watching experience.	2.90	0.91	3.55	0.83

they felt immersed in that video, 50% of the subjects responded *Yes* and the other 50% responded *Partially*. The subjects also found the third video more enjoyable; with an average rating of 3.9 (St-Dev equalled 0.72). Out of the three ODVs used in the experiment, this video contained the most dynamic scenes and fast camera movements, probably making it the most interesting to the subjects.

We can also observe how, on average, the videos did not inflict any significant discomfort to the subjects (stress, fear, loss of balance, or nausea), i.e. the average ratings to questions 4-7 are always between 1 and 2 (higher grade interpreted as the stronger feeling). However, while watching the Hong Kong air tour, two test subjects (numbers 12 and 13) experienced low to moderate discomfort in terms of losing the sense of balance, while subject number 13 also reported a high level of nausea. Both those subjects gave lower grades when asked about the overall quality of the Hong Kong ODV presentation and their watching experience (the grades are 3 and 2 for the subject number 12 and 13, respectively). Yet, the same subjects gave a high grade to the same question regarding the first (training) video. In that video the camera was static, and the subjects did not experience any discomfort; they were enjoying the polar night scenes and found it to be *very interesting*, thus, confirming how camera position and the position of the video can have a significant impact on user QoE.

The average overall quality of the video presentation and user watching experience was 2.9 (i.e. close to *fair*) and 3.55 (i.e. between *fair* and *good*) for the second and third video, respectively (question number 9 in *Table 1*). Note that only one test subject reported that he uses VR technology *occasionally*, while 11 and 8 of them reported that they use it *rarely* and *never before*, respectively. Due to this infrequent usage of VR among our test participants, it could be expected that they would grade their experience higher because of the "wow factor" originating from, mostly, the first time use of ODV in VR. However, this did not happen. Clearly, the visual impairments inserted into the videos impacted the subjects' perception, causing degradation of their watching experience.

4.2 The effect of video resolution changes

As explained in Section 3.1, while watching the second ODV, the subjects experienced the resolution changes at arbitrarily chosen video timeline locations, thus, emulating the adaptive video streaming session when the resolution may be downgraded/upgraded, depending on the network conditions and video buffer state. The subjects were instructed to rate the quality of the ODV each time when they noticed the change in video quality by stating a grade ranging between 1 (which meant *bad quality*) and 5 (which meant *excellent quality*). The medians of the subjects' scores, as well as 95% confidence intervals for the medians, can be found in *Figure 4* while *Table 2* contains the results of the Wilcoxon test for paired samples indicating statistically significant differences between the samples. Note that the p values in the shaded fields in *Table 2* indicate that the significant differences between the two samples were observed (i.e. $p < 0.05$). After testing the data normality and proving that the distribution of the samples is not

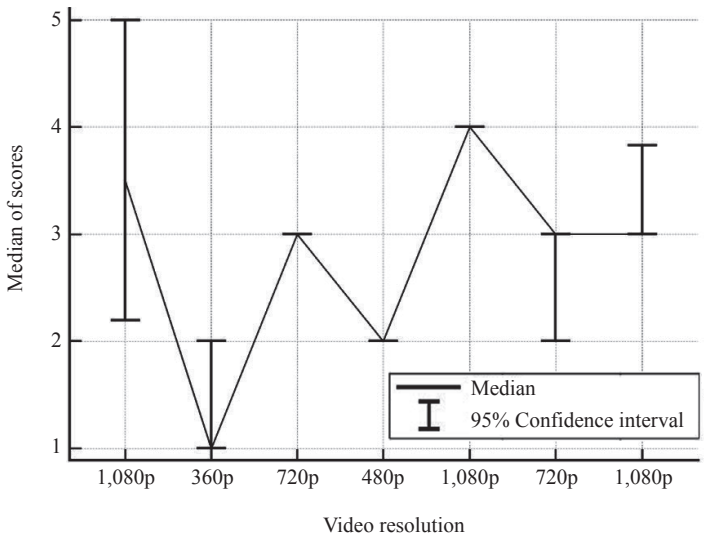

Figure 4 – The medians of the scores for ODV used to test the effect of the video resolution changes

Table 2 – Descriptive statistics and results of the Wilcoxon test for paired samples for the second video

Wilcoxon test (paired samples)	1,080p	360p	720p	480p	1,080p	720p	1,080p
Lowest value	2	1	2	1	2	1	2
Highest value	5	3	5	4	5	5	5
Median	3.5	1	3	2	4	3	3
95% Confidence interval for the median	2.2 - 5	1 - 2	3 - 3	2 - 2	4 - 4	2 - 3	3 - 3.8
Two-tailed probability (p-value)		0.0313	<0.0001	0.0001	<0.0001	0.0001	0.0625
Sample size	6	20	20	20	20	20	20

Normal, we decided to go forward with the Wilcoxon test to test statistical significance between the ratings of the two neighbouring video segments. The Wilcoxon test ranks the absolute values of the differences between the paired observations in the samples and calculates the statistics on the number of negative and positive differences. It is usually used to compare two sets of scores that come from the same participants, which was beneficial to us in this experiment.

The considerable drop in the perceived quality, i.e. the median of scores, has been recorded between the first two video segments. The median of scores dropped from 3.5 to 1, capturing the change in the subjects' perception when the resolution degraded from the original 1,080p to 360p. Remember that during this resolution shift, the video playback was interrupted by the 2-second re-buffering event which caused all subjects to notice the change. As seen from *Table 2*, the changes in ODV resolution have a significant impact on the subjects' perception of quality, i.e. differences between the medians of scores are statistically significant. This is true for all video segments except for the last one which can be explained by arguing that the shift from 720p to 1,080p resolution did not elicit enough improvement of the quality to be recognized by the subjects. Had the ODV been longer, perhaps the subjects' perception would have been further improved; however, in this case when the video lasted for 140 seconds it remained at the *fair quality* (the linguistic meaning of the last median shown in the figure and the table). From this, we can infer that it is hard to recover the perception of quality once when it has been downgraded, i.e. when it is disturbed beyond a certain threshold (the concept of thresholds was first introduced by Fiedler et al. [28]).

Apart from testing the statistical significance between the ratings of the two neighbouring video segments, we also tested the significance of the segments of the same quality. We found that there are no statistically significant differences between the medians of scores apart from the last two 1,080p video segments (in the second half of the video). Here, the difference is significant because most test subjects did not perceive the improvement of quality from 720p to 1,080p, as we discussed earlier.

In [2] we showed how the user QoE can be redeemed if the video content is perceived as interesting to the user, i.e. entertained users who enjoy the content can become more forgiving to the occasional occurrence of quality degradations. From *Table 1* we know that the average level of interest and enjoyment for this video equalled 3.55 and 3.2, respectively, i.e. on average the subjects were only moderately entertained with the guide through the Palace. Together with the quality degradations inserted into the video, this resulted in a lower level of user experience. A detailed analysis of the relationship between the level of interest and the enjoyment of the subjects and their subjective feeling about the presentation quality and watching experience (*Figure 5*) shows how the subjects who experienced higher levels of interest and enjoyment also exhibited a higher level of watching experience and vice versa. Thus, we can argue that the level of QoE was restored for those subjects who found this ODV interesting and entertaining. The only exception is subject number 9 who reported the highest level of interest and enjoyment but his perception of quality and watching experience remained low.

4.3 Frame drop patterns and user QoE

The impact of four different frame drop patterns on the user QoE was also tested in this experiment. Namely, frame drop patterns A, B, C, and D (indicating the drop of 1, 2, 3 or 4 consecutive frames, respectively) were inserted into the original ODV with the purpose of testing the user perception when

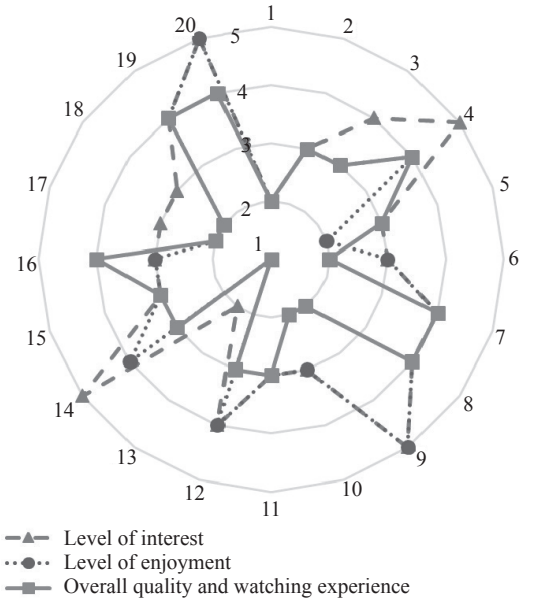

Figure 5 – Relation between the level of interest and enjoyment and the overall video quality and watching experience of the subjects

the video becomes choppy. The results are presented in *Figure 6*. As before, the descriptive statistics for each video segment can be found in *Table 3* (the p values in the shaded fields indicate that significant differences were observed between the two samples). Note that the Wilcoxon test was again used to test significant differences. Additionally, in this part of the analysis, we disregarded the ratings of three test subjects because they stated the same grade three times or less during the ODV screening.

In this test, the lower medians of scores were recorded toward the end of the video, during the playback of frame drop patterns C and B (the medians equalled 2). Another low median of scores is derived for the fourth video segment (for pattern D) when it also equalled 2, which is an expected result. However, during the middle of the video, the medians of scores remained relatively high despite the patterns which were inserted. Again, this result can be related with the previous finding of how this video was the most interesting and entertaining to the test subjects. The fast pace of the video, dynamic camera movement from air accompanied by energetic music clearly impacted the subjects' perception. We can also support this statement by analysing the statistical significance of the obtained results. As can be seen from *Table 3*, differences between the medians of scores for the several frame drop patterns are insignificant, especially in the middle of the video. For instance, pattern D, which includes a drop of 4 consecutive frames and causes high disturbances in the video playback, was inserted two times in this ODV. Note that both patterns lasted for 10 seconds, which was more than enough time for patterns to get noticed. However, the second time when the pattern was shown there was no statistically significant difference recorded compared to the previous segment when there was no pattern inserted. Note that the Wilcoxon test showed that there were no statistically significant differences between the samples of the same quality in cases when there was no video degradation (i.e. for NP segments shown in *Figure 6*). However, the differences are significant between the segments of the same quality in cases when the patterns were inserted. This proves how the same patterns can induce contrasting impact on the user perception depending on the video scenes when they appear.

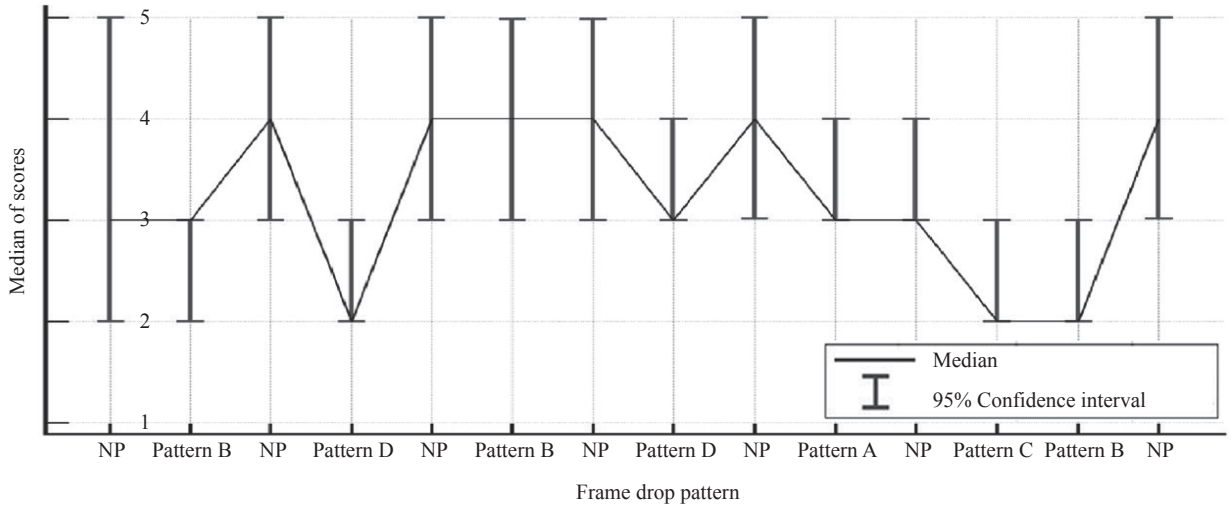

Figure 6 – Medians of the scores for ODV used to test the effect of the frame drop patterns (NP = no pattern)

Table 3 – Descriptive statistics and results of the Wilcoxon test for paired samples for the third video

Wilcoxon test (paired samples)	NP	Pattern A	NP	Pattern D	NP	Pattern B	NP	Pattern D	NP	Pattern A	NP	Pattern C	Pattern B	NP
Lowest value	2	2	2	1	2	2	2	1	2	2	2	1	1	2
Highest value	5	4	5	4	5	5	5	5	5	5	5	3	4	5
Median	3	3	4	2	4	4	4	3	4	3	3	2	2	4
95% Confidence interval for the median	2 - 5	2 - 3	3 - 5	2 - 3	3 - 5	3 - 5	3 - 5	3 - 4	3 - 5	3 - 4	3 - 4	2 - 3	2 - 3	3 - 5
Two-tailed probability (p-value)		0.3750	0.0002	0.0001	0.0002	0.6250	0.6406	0.1641	0.0093	0.0078	>0.05	0.0001	0.6406	0.0001
Sample size	6	17	17	17	17	17	17	17	17	17	17	17	17	17

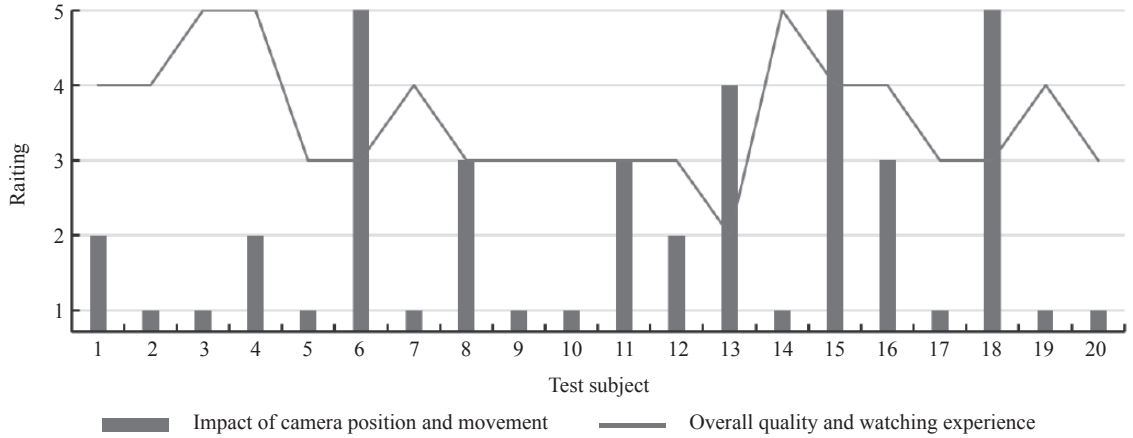

Figure 7 – The relation between the camera position and movement and the overall video quality and watching experience of the subjects

Due to the fast pace of the video and aerial viewing perspective, it is worthy to investigate the relation between the camera position and movement and the user perception about the overall video quality and watching experience (depicted in *Figure 7*). Note that the subjects rated the *impact of camera position and movement* on a scale from 1 to 5; a higher rating was interpreted as a higher impact on the subjects' feelings of discomfort. Hence, we observed that the subjects who reported high impact also reported sensing some kind of discomfort (e.g. feelings of nausea or fear or they lost the sense of balance). The figure depicts how subjects who reported that the camera movement impacted their level of discomfort also experienced a lower level of overall quality and watching experience, except for subject number 15. Notwithstanding, the choppiness of the video caused by the frame drops also contributed to the subjects' discomfort. In future research, the same frame drop patterns could be applied to two different paced videos, with different camera dynamics, with the purpose of distinguishing the effect of each individual parameter on subjects' level of discomfort and QoE.

5. CONCLUSION

The experience gained through the implementation of this experiment provokes us to critically think about the methodology used for testing and to underline specific cognitions that arose in this process.

We would first like to draw attention to one of the obvious features of this kind of experiment: the considerable amount of time required to conduct the subjective tests. This is the hampering characteristic of almost every subjective test, but it is especially emphasised for testing the quality of services that are used in VR. This is because it is difficult to use, for instance, crowd testing platforms which are powerful tools for reaching the targeted population for the experiment. Thus, it may be a challenging task to conduct the test on a large number of test subjects, keeping in mind that some potential test participants have to be discarded due to their sensitiveness to a 360-degree perspective. The perspective itself may inflict a high level of discomfort, making individuals unwilling to participate in the experiment. Moreover, due to its innovativeness, it is not advisable to rely on the test subjects' knowledge on how to use the technology and services which are provided in VR. Hence, researcher(s) must devote additional time to each test participant (e.g. for training purposes on how to use an HMD system, how to interact with the video, etc.) to ensure the validity of the test and the collected data. Another approach would be to use experienced VR users; however, there are not many of them yet, at least in our environment; in turn, this opens new opportunities for our future research. Another limitation of the current experiment is the fact that a subjective evaluation was conducted among the student population. Hence, we cannot claim that the results strongly correlate with the perception of a more versatile panel of users; this will have to be addressed in our upcoming experiments.

As expected, and confirmed in the past, the results of subjective tests of video quality depend on the type of content which is used. In our experiment, we used ODVs of different contents to avoid boredom of the subjects and to be able to test the impact of the camera position and movement on the user QoE. However, we can also argue that a publicly available ODV database, containing videos of different contents and lengths, viewing perspective, and spatial-temporal characteristics, would be most helpful in this kind of research and it would enable the comparative analysis of the obtained results.

Out of the two objective parameters tested we showed that the changes in ODV resolution produced the most impact on the subjects' QoE. Moreover, we witnessed how stalling the video playback can further degrade the perception of quality, which is then difficult to recover, especially if shorter test sequences are used for the experiment. The importance of ODV pace and camera dynamics was also disclosed when evaluating the impact of different frame drop patterns. In our case, even the choppiest video segments remained unnoticed by some of our test subjects owing to the fact that the ODV used in this test was evaluated as the most interesting and entertaining by the subjects.

In our future research we are planning to address some of the issues highlighted in this paper, namely, work on improvement of test methodology, conducting tests on a larger target group of, perhaps, more experienced and age-versatile VR users, using the same ODV for testing different objective parameters, and employing different subjective methodology (e.g. double-stimulus) for testing the user QoE. The work towards establishing the ODV database is also one of the possible future research paths.

ACKNOWLEDGEMENT

The results presented in this paper originate from the Quality of Experience for Virtual Reality Applications (QoE4VR) project activities. The project is funded by the University of Zagreb under *Short-term Financial Support for Researchers* program in 2017.

Dr. sc. **ŠTEFICA MRVELJ**[1]
E-mail: smrvelj@fpz.unizg.hr
Dr. sc. **MARKO MATULIN**[1]
(Corresponding author)
E-mail: mmatulin@fpz.unizg.hr
SERGO MARTIROSOV, doktorand[2]
E-mail: martiros@rti.zcu.cz
[1] Sveučilište u Zagrebu, Fakultet prometnih znanosti
Vukelićeva 4, 10000 Zagreb, Hrvatska
[2] Západočeská univerzita v Plzni
Univerzitni 2732/8, 30100 Plzeň, Česká republika

SUBJEKTIVNA EVALUACIJA ISKUSTVENE KVALITETE USLUGE STRUJANJA OMNIDIREKCIONALNIH VIDEOSADRŽAJA

SAŽETAK

Ovaj rad izvještava o rezultatima subjektivnih testiranja iskustvene kvalitete usluge (Quality of Experience - QoE) prijenosa omnidirekcionalnih videosadržaja (ODV) tehnologijom strujanja. Test je proveden na 20 ispitanika koji su gledali tri ODV-a koristeći HMD (Head Mounted Display) sustav. Trajanje videosadržaja bilo je između 2 i 3 minute. Prvi video upotrijebljen je u svrhu upoznavanja ispitanika s ODV-ima te nije sadržavao degradacije u kvaliteti prikaza. Kvaliteta drugog i trećeg ODV-a bila je degradirana na način da su uvedene promjene u

videorezoluciji kao i različiti obrasci gubitka sličica tijekom prikazivanja. Dok su gledali tako unaprijed pripremljene videosadržaje, ispitanici su se izjašnjavali primjećuju li degradacije u kvaliteti prikaza te su ujedno i ocjenjivali tu kvalitetu. Nakon što su odgledali pojedini ODV, ispitanici su popunjavali i zaseban upitnik u kojem su ocjenjivali razinu svoga zadovoljstva i nelagode s ODV-om. Rezultati su pokazali kako degradacije obaju objektivnih parametara videokvalitete koja su testirana (rezolucija i broj sličica) utječe na percepciju ispitanika o kvaliteti ODV-a. Ipak, taj utjecaj je donekle ublažen u onom ODV-u koji je sadržavao više dinamičnih scena i brze pokrete kamere.

KLJUČNE RIJEČI

360° video; strujanje; video rezolucija; gubitak sličica; iskustvena kvaliteta;

REFERENCES

[1] Ericsson Consumerlab. *Merged Reality: Understanding How Virtual and Augmented Realities Could Transform Everyday Reality*. Ericsson; 2017. Available from: https://www.ericsson.com/en/trends-and-insights/consumerlab/consumer-insights/reports/merged-reality [Accessed June 2019].

[2] Mrvelj Š, Matulin M. Impact of Packet Loss on the Perceived Quality of UDP-based Multimedia Streaming: A Study of User Quality of Experience in Real-life Environments. *Multimedia Systems*. 2018;24(1): 33-53. Available from: doi:10.1007/s00530-016-0531-8

[3] Matulin M, Mrvelj Š. Modelling User Quality of Experience from Objective and Subjective Data Sets using Fuzzy Logic. *Multimedia Systems*. 2018;24(6): 645-667 Available from: doi:10.1007/s00530-018-0590-0

[4] https://www.fpz.unizg.hr/qoe4vr/ [Accessed September 2019].

[5] Cisco. *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017-2022*. White Paper. Cisco; 2019. Available from: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.html [Accessed June 2019].

[6] Yu M, Lakshman H, Girod B. A Framework to Evaluate Omnidirectional Video Coding Schemes, Mixed and Augmented Reality. *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR), 29 September – 3 October 2015, Fukuoka, Japan*; 2015. p. 31-36. Available from: doi:10.1109/ISMAR.2015.12

[7] Sun Y, Lu A, Yu L. WS-PSNR for 360 Video Objective Quality Evaluation. *MPEG Joint Video Exploration Team*. Vol. 116. Chengdu; 2016.

[8] Sun Y, Lu A, Yu L. AHG8: WS-PSNR for 360 Video Objective Quality Evaluation. *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0040, 4th Meeting*. Chengdu; 2016.

[9] He Y, Vishwanath B, Xiu X, Ye Y. AHG8: InterDigital's Projection Format Conversion Tool. *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0021, 4th Meeting*. Chengdu; 2016.

[10] Zakharchenko V, Alshina E, Singh A, Dsouza A. AHG8: Suggested Testing Procedure for 360-degree Video. *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0027, 4th Meeting*. Chengdu; 2016.

[11] Wu S, Chen X, Fu J, Chen Z. Efficient VR Video Representation and Quality Assessment. *Journal of Visual Communication and Image Representation*. 2018;57: 107-117. Available from: doi:10.1016/j.jvcir.2018.10.018

[12] Chen S, Zhang Y, Li Y, Chen Z, Wang Z. Spherical Structural Similarity Index for Objective Omnidirectional Video Quality Assessment. *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), 23-27 July 2018, San Diego, United States*; 2018. p. 1-6. Available from: doi:10.1109/ICME.2018.8486584

[13] Wang Z, Simoncelli EP, Bovik AC. Multiscale Structural Similarity for Image Quality Assessment. *Proceedings of the 37th Asilomar Conference on Signals, Systems & Computers, 9-12 November 2003, Pacific Grove, United States*; 2003. p. 1398-1402. Available from: doi:10.1109/ACSSC.2003.1292216

[14] Ozcinar C, Cabrera J, Smolic A. Omnidirectional Video Streaming Using Visual Attention-Driven Dynamic Tiling for VR. *Proceedings of the IEEE International Conference on Visual Communications and Image Processing (VCIP) 2018, 9-12 December 2018, Taichung, Taiwan*; 2018. p. 1-4. Available from: doi:10.1109/VCIP.2018.8698638

[15] Ozcinar C, Cabrera J, Smolic A. Visual Attention-Aware Omnidirectional Video Streaming Using Optimal Tiles for Virtual Reality. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*. 2019;9(1): 217-230. Available from: doi:10.1109/JETCAS.2019.2895096

[16] Xu M, Li C, Chen Z, Wang Z, Guan Z. Assessing Visual Quality of Omnidirectional Videos. *IEEE Transactions on Circuits and Systems for Video Technology*. 2019. Available from: doi:10.1109/TCSVT.2018.2886277

[17] Birkbeck N, Brown C, Suderman R. Quantitative Evaluation of Omnidirectional Video Quality. *Proceedings of the 9th International Conference on Quality of Multimedia Experience (QoMEX), 31 May - 2 June 2017, Erfurt, Germany*; 2017. Available from: doi:10.1109/QoMEX.2017.7965684

[18] Zakharchenko V, Pyo Choi K, Alshina E, Hoon Park J. Omnidirectional Video Quality Metrics and Evaluation Process. *Proceedings of the Data Compression Conference (DCC), 4-7 April 2017, Snowbird, United States*; 2017. p. 472-472. Available from: doi:10.1109/DCC.2017.90

[19] Chen Z, Li Y, Zhang Y. Recent Advances in Omnidirectional Video Coding for virtual Reality: Projection and Evaluation. *Signal Processing*. 2018;146: 66-78. Available from: doi:10.1016/j.sigpro.2018.01.004

[20] Upenik E, Řeřábek M, Ebrahimi T. Testbed for Subjective Evaluation of Omnidirectional Visual Content. *Picture Coding Symposium (PCS), 4-7 December 2016, Nuremberg, Germany*; 2016. p. 1-5. Available from: doi:10.1109/PCS.2016.7906378

[21] Pérez P, Escobar J. MIRO360: A Tool for Subjective Assessment of 360 Degree Video for ITU-T P.360-VR.

Proceedings of the 11th International Conference on Quality of Multimedia Experience (QoMEX), 5-7 June 2019, Berlin, Germany; 2019. Available from: doi:10.1109/QoMEX.2019.8743216

[22] Sassatelli L, Winckler M, Fisichella T, Dezarnaud A, Lemaire J, Aparicio-Pardo R, Trevisan D. New Interactive Strategies for Virtual Reality Streaming in Degraded Context of Use. *Computers & Graphics*. 2019. Available from: doi:10.1016/j.cag.2019.10.005

[23] Schatz R, Sackl A, Timmerer C, Gardlo B. Towards Subjective Quality of Experience Assessment for Omnidirectional Video Streaming. *Proceedings of the 9th International Conference on Quality of Multimedia Experience (QoMEX), 31 May - 2 June 2017, Erfurt, Germany*; 2017. Available from: doi:10.1109/QoMEX.2017.7965657

[24] Yao S-H, Fan C-L, Hsu C-H. Towards Quality-of-Experience Models for Watching 360° Videos in Head-Mounted Virtual Reality. *Proceedings of the 11th International Conference on Quality of Multimedia Experience (QoMEX), 5-7 June 2019, Berlin, Germany*; 2019. Available from: doi:10.1109/QoMEX.2019.8743198

[25] Liotou E, Tsolkas D, Passas N. A roadmap on QoE metrics and models. *Proceedings of the 23rd International Conference on Telecommunications (ICT), 16-18 May 2016, Thessaloniki, Greece*; 2016. Available from: doi: 10.1109/ICT.2016.7500363

[26] Datta P, Izdebski L, Kumar N, Suh K. *"It came to me in a stream…" The upward arc of online video, driven by consumers.* Cisco white paper. Available from: https://www.cisco.com/c/dam/en_us/about/ac79/docs/sp/Online-Video-Consumption_Consumers.pdf [Accessed 22nd Feb 2020].

[27] Farrokhi F, Mahmoudi-Hamidabad A. Rethinking convenience sampling: defining quality criteria. *Theory and Practice in Language Studies*. 2012;2(4): 784-792. Available from: doi:10.4304/tpls.2.4.784-792

[28] Fiedler M, Hossfeld T, Tran-Gia P. A Generic Quantitative Relationship between Quality of Experience and Quality of Service. *IEEE Network*. 2010;24(2): 36-41. Available from: doi:10.1109/MNET.2010.5430142