<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper presented at *14th International Conference on Signal Processing and Communication Systems, ICSPCS 2020, Virtual, Adelaide, Australia, 14 December 2020 through 16 December 2020*.

Citation for the original published paper:

Elwardy, M., Hu, Y., Zepernick, H-J., Chu, T M., Sundstedt, V. (2020)
Comparison of ACR Methods for 360° Video Quality Assessment Subject to
Participants' Experience with Immersive Media
In: Wysocki T.A., Wysocki B.J. (ed.), *2020 14th International Conference on Signal
Processing and Communication Systems, ICSPCS 2020 - Proceedings*, 9310071
Institute of Electrical and Electronics Engineers Inc.
<https://doi.org/10.1109/ICSPCS50536.2020.9310071>

N.B. When citing this work, cite the original published paper.

© 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:bth-21062>

Comparison of ACR Methods for 360° Video Quality Assessment Subject to Participants' Experience with Immersive Media

Majed Elwardy, Yan Hu, Hans-Jürgen Zepernick, Thi My Chinh Chu, and Veronica Sundstedt
Blekinge Institute of Technology, SE-371 79 Karlskrona, Sweden
E-mail: majed.elwardy@bth.se

Abstract—Watching 360° videos on head-mounted displays (HMDs) has become more popular in recent years. As such, subjective quality assessment of 360° videos on HMDs is needed to obtain a ground truth on the quality as perceived by users. Due to the lack of standardized methods, the absolute category rating (ACR) with hidden reference (ACR-HR) method for conventional videos has also been used for subjective quality assessment of 360° videos on HMDs. A modified ACR (M-ACR) method tailored for assessing 360° videos on HMDs has recently been proposed. In this paper, we compare the ACR-HR and M-ACR with hidden reference (M-ACR-HR) method regarding subjective quality assessment of 360° videos on HMDs. The comparison is conducted subject to participants' experience with watching immersive media on HMDs. The findings include: 1) Average rating times to cast opinion scores is lower for the ACR-HR method irrespective of participants' experience, 2) Mean opinion scores are similar for both methods irrespective of participants' experience, 3) The M-ACR-HR method is more reliable, especially with experts and participants with no experience, 4) Discomfort is significantly lower for the ACR-HR method, especially, for participants with no experience.

I. INTRODUCTION

Immersive media such as virtual reality (VR) content shown on head-mounted displays (HMDs) have become more popular among consumers. Enablers and key drivers of this trend are the large advances in wearable devices and interaction technologies, hardware and software components, and increased connectivity in terms of bandwidth supported by the upcoming 5G and future 6G mobile communication systems [1], [2]. Because humans are the final judges of the perceptual quality of immersive media applications, it is essential to engage statistically meaningful ground truths from subjective experiments into the design of immersive media systems.

In particular, 360° videos on HMDs and their subjective and objective quality assessment have seen an increased interest in recent years. In contrast to 2D videos shown on conventional displays, 360° videos presented on HMDs provide a higher field of view (FOV), three degrees of freedom of unlimited rotational movements and limited translational movements related to head movements. Degradations when watching 360° videos on HMDs are more likely to impact on the subjective quality perceived by the viewers due to a higher FOV, resolution, and compression. On the other hand, standardized methodologies for 360° videos on HMDs are not yet available causing the research on related quality assessment resorting

to methodologies from 2D video quality assessment. To close this gap, the International Telecommunications Union (ITU) in ITU-T Study Group 12 and the Video Quality Experts Group (VQEG) jointly commenced to develop Recommendation ITU-T P-360-VR on subjective test methodologies for 360° video on HMDs [3]. The set of methodologies assessed by ITU-T Study Group 12 have been implemented in the tool reported in [4] for Android-based HMDs.

This paper contributes to the evaluation of subjective quality assessment methods for immersive media by providing a comparison between two absolute category rating (ACR) methods for 360° video quality assessment subject to participants' experience with immersive media on HMDs. In particular, the ACR with hidden reference (ACR-HR) method known from 2D video quality assessment and the modified ACR-HR (M-ACR-HR) method [5] that was introduced to account for users having little experience with immersive media on HMDs are compared. A detailed statistical analysis of the mean opinion scores (MOSs) gathered in a campaign of subjective experiments is provided to reveal the differences among the two methods and to shed more light on their contribution to 360° video quality assessment.

A. Related Work

Several subjective video quality assessment methods have been standardized by the ITU including the ACR, ACR-HR, and degradation category rating (DCR) method described in ITU-T P.910 [6] and ITU-T P.915 [7]. The DCR method is also known as the DSIS method [8]. In contrast to the ACR and ACR-HR method which use a quality scale, the DSIS method uses an impairment scale. The ACR-HR method inherits the benefits of the ACR method but implicitly shows the reference videos which allows to remove bias that may be induced by participants' likes or dislikes of scene contents.

Other prominent methods include the subjective assessment of multimedia video quality (SAMVIQ) method [9], subjective assessment of multimedia panoramic video quality (SAM-PVIQ) method [10], and the pair comparison (PC) method [6]. In the SAMVIQ method, a multi-stimulus approach with random access is used where the participants are allowed to view each test stimulus including explicit and hidden references multiple times as desired and to modify the opinion score each time. Further, the SAMVIQ method uses a continuous rating

scale while the ACR and DSIS method use five-level rating scales. With the SAMPVIQ method, which is more tailored to 360° videos, each set comprising of a particular reference video of a scene and the corresponding test videos including the hidden reference is divided into several non-overlapping groups. Each group contains the explicit reference video of the considered scene and three test videos. The stimuli are presented group-by-group and scene-by-scene. This approach is sought to reduce the impact of participants' fatigue on their opinion scores which may be a larger issue with 360° videos than with conventional 2D videos. In the PC method, two distorted stimuli are presented side-by-side and the participants are asked to choose the stimulus with the highest quality. Due to the large number of pairs to be compared when grouping all possible stimuli in the PC method, this method may be used after a subjective experiment with the ACR and DSIS method on a subset of test videos that require higher resolving power with respect to distortions.

Although developed for 2D and 3D video systems, the aforementioned subjective quality assessment methods have also been used for evaluating the quality of immersive media such as 360° videos, digital games, and 3D graphics shown in a VR environment. In particular, quality assessment of 360° videos on HMDs using the ACR method can be found in, e.g., [11]–[16], applications of the ACR-HR method are contained in [12], [17], the DSIS method is used in [18], [19], the SAMVIQ method is applied in [10], and the SAMPVIQ method is used in [10], [20].

In order to address the issue that immersive media are rather new in the consumer markets and users may be still unfamiliar with related services, a modified ACR (M-ACR) method has been proposed in [5] for subjective quality assessment of 360° videos. In the M-ACR method, each test video is shown twice for the duration of 10 s with a 3 s mid-gray screen in between the videos. In particular, in [5], the M-ACR method was used to evaluate the subjective quality of HEVC/H.265-encoded 360° videos at different bitrates for ultra high-definition (UHD, i.e., 3840×2160 p) and full high-definition (FHD, i.e., 1920×1080 p) resolution. The test videos were shown on an Oculus Rift HMD with a resolution of 2160×1200 p. The participants were asked to provide their ratings verbally for each test video which then was written down by the experimenter. The results provide insights that may guide the choice of coding and resolution settings subject to given bitrate constraints. As with the ACR and ACR-HR method, the M-ACR and M-ACR-HR use the same quality scale but approximately doubles the time needed to conduct a test session due to showing each stimulus twice.

In [21], the impact that the experience of participants with immersive media on HMDs has on 360° video quality assessment was studied using the M-ACR-HR method. It was shown that screening of the distribution of opinion scores detects many outliers for participants with no experience in watching immersive media on HMDs. Further, participants with no experience provided lower opinion scores on average compared to experts or participants with some experience with

immersive media. Lower average rating times were observed in Session 2 compared to Session 1 of the subjective experiment irrespective of participants' experience. The reduction in average rating times in Session 2 was more pronounced for the experts which indicates that the quality assessment task has been learned quite well in Session 1.

The work in [22] studied the number of participants that are needed in subjective experiments for 360° videos using the M-ACR-HR method such that a statistically meaningful subjective quality assessment is supported. A total of 120 360° videos of 10 s duration comprising four different scenes with five resolutions, five quantization parameters (QPs), and reference videos with near to lossless encoding were shown on an HTC Vive Pro HMD with an integrated eye-tracker. Based on the differential margin of error (MoE) of the differential mean opinion score (DMOS), it was shown that 21 to 26 valid participants would be sufficient to drive the differential MoE below given thresholds.

In [23], the ACR-HR and DSIS methods were compared to address the question if an explicit reference is needed in subjective experiments for 3D graphics when shown in a virtual environment on an HTC Vive Pro HMD. A motivation of this work was to reveal if these subjective methods are valid and accurate in view of participants' lack of prior knowledge of 3D graphics data. The quantitative analysis of the opinion scores obtained in a subjective experiment suggests that the DSIS method is more accurate and requires less participants compared to the ACR-HR method. The results also provide evidence that participants with no experience watching 3D graphics had difficulties to differentiate among all distortions.

A comparison of the DSIS and M-ACR method for subjective quality evaluation of 360° videos is provided in [19]. Six scenes of 10 s duration were chosen with each scene given in FHD and UHD, and five different bitrates up to a maximum of 15 Mbps. The 60 test videos were shown on an Oculus Rift HMD. The results indicate that the M-ACR method provides slightly higher MOS for lower bitrates while the DSIS method gives slightly higher MOS for bitrates above 4 Mbps in UHD. In general, a strong correlation between the scores obtained with the M-ACR and DSIS method is observed. In the M-ACR method, participants were asked to rate the quality of the test videos while an impairment scale is used with the DSIS method to rate the degradation of the test videos. As such, the impact of the actual task to be performed on the given scores remains to be considered.

Building on [19], the work reported in [24] compares the ACR, M-ACR, and DSIS method for subjective quality assessment of 360° videos. In particular, five different scenes of 10 s duration were selected with 4K, 6K, and 8K resolution (2D width in equirectangular projection) and four different target bitrates up to a maximum of 25 Mbps. The 60 test videos were shown on an HTC Vive Pro HMD with 2880×1600 p resolution. From the MOS deduced from the subjective experiments, it was conjectured that the DSIS method is statistically more reliable in terms of normalized mean confidence interval compared to the M-ACR method which is followed by the ACR

method. Further, it is recommended to use the DSIS method for high resolution stimuli where resolving finer degradation levels is of main interest. The ACR method is recommended when less resolving power is acceptable but the duration of the subjective experiment is of main concern.

In this paper, we compare the ACR-HR and M-ACR-HR method for subjective quality assessment of 360° videos subject to the participants' experience with immersive media on HMDs. Although the M-ACR-HR method shall cope with participants that have little experience with watching immersive media on HMDs, to the best of our knowledge, there have been no studies that compare the ACR-HR and M-ACR-HR method subject to the impact of the participants' experience with 360° videos on HMDs. Given that the DSIS method differs from the ACR-HR and M-ACR-HR method in posing a video degradation assessment task to the participants compared to a video quality assessment task, we focus on the ACR-HR and M-ACR-HR method to remove potential bias due to different assessment tasks.

B. Contributions and Paper Structure

Contributions of this paper on comparing the performance of the ACR-HR and M-ACR-HR method for 360° video quality assessment include the following:

- The results of subjective experiments for 360° video quality assessment shown on an HTC Vive Pro HMD using the ACR-HR and M-ACR-HR method are compared.
- Three classes of experience levels with immersive media on HMDs are considered in the comparison of the experimental data: Experts, sometimes used, never used.
- A detailed statistical analysis of the data obtained from subjective experiments for the ACR-HR and M-ACR-HR method is provided supporting the comparison in terms of average rating times, MOS and standard deviation (STD), t-tests between opinion scores of both methods, and answers to a simulation sickness questionnaire (SSQ).
- The aforementioned statistics are provided and compared irrespective of as well as subject to participants' experience with immersive media on HMDs.

Main findings of this work may be summarized as follows:

- Average rating times to cast an opinion score are lower for the ACR-HR method irrespective of the scene shown and participants' experience with immersive media on HMDs.
- The progression of MOS values for different resolutions versus QPs is similar for both methods irrespective of participants' experience. This finding is also supported by the performed t-tests.
- A Bonferroni-corrected t-test revealed that the M-ACR-HR method is, in general, more reliable compared to the ACR-HR method. This finding applies in particular to the experts and the participants that have never used immersive media on HMDs.
- Analysis of the SSQs shows that discomfort is kept significantly lower for the ACR-HR method while simulator sickness symptoms are reported much stronger for the M-ACR-HR method.

The rest of this paper is organized as follows. In Section II, the experimental design is summarized. The statistical analysis of the data obtained from the subjective experiments is provided and discussed in Section III. This includes average rating times, MOS, STDs, t-tests, and an analysis of the SSQs. Finally, conclusions are given in Section IV.

II. EXPERIMENTAL DESIGN

In this section, the experimental setup used throughout our campaigns of subjective experiments on aspects of 360° video quality assessment is summarized to the extent needed for the understanding of the work reported in this paper. More details about the experimental design can be found in [21].

A. 360° Reference and Test Videos

Four 360° video scenes, i.e., “Alcatraz”, “BloomingAppleOrchards” (Blooming), “FormationPace” (Formation), and “PandaBaseChengdu” (Panda), with different contents in 7680 × 3840p (8K) resolution, equirectangular projection (ERP), and frame rate of 29.97 frames per second (fps) were selected from the VQA-ODV database [25], [26] to produce 360° reference videos of 10 s duration. These scenes were selected to reduce the risk of simulator sickness, which otherwise could affect the quality assessment task. By down-sampling the 8K reference videos with the bi-cubic scaling algorithm, additional reference videos with 6K, 4K, 2K and optimal resolution (OR) of 3600 × 1800p recommended in [17] for the HTC Vive Pro HMD were generated. The libx265 tool [27], [28] was used to produce the test videos by encoding the reference videos into HEVC/H.265 format with five different quantization parameters (QPs), i.e., QP = 22, 27, 32, 37, and 42. Given the high bitrates of the reference videos ranging between 844.65 Mbps and 52.33 Mbps, the multiple rate control algorithm of the libx265 encoder was used with a constant rate factor (CRF) of CRF = 10 (perceptually lossless encoding) to constrain the bitrates to the range between 226.79 Mbps and 15.48 Mbps. A total of 120 360° videos were generated including the reference videos with different resolutions and the corresponding test videos with different QPs.

B. Software Suite and Technical Equipment

The test platform was developed using the Unity 3D game engine Version 2018.3.11f1 and Visual Studio 2017. The test platform includes graphical user interfaces for giving the opinion scores to the stimuli according to the ACR-HR and M-ACR-HR method. The Unity build-in random function was used to randomize the video presentation. The 360° videos were shown on an HTC Vive Pro HMD with gaze data output at a rate of 120 Hz obtained from a Tobii Pro prototype eye-tracker integrated into the HMD. A Corsair One i160 Gaming-PC with an Intel I9-9900K Processor of 3.6GHz clock rate was used as hardware platform. This computer is equipped with an NVIDIA GeForce RTX 2080 TI graphics card comprising of 4352 CUDA cores and allows a maximum digital resolution of 7680 × 4320p.

C. Procedures and Test Methods

The common components of the considered two test methods include the following. An introduction to the tasks of the experiment, safety instructions, and risks with using HMDs were provided to the participants. Then, a visual acuity and color vision test were conducted. The participants that passed the vision tests and signed the consent form commenced with a training session. Once a participant feels confident with the equipment, procedures, and tasks, a session of the subjective experiment starts with eye-tracker calibration followed by the presentation of the stimuli. Given that the M-ACR-HR method shows each test video twice back-to-back and hence approximately doubles the duration of the subjective experiment compared to the ACR-HR method, the 120 360° videos were shown in two test sessions while one test session for the ACR-HR method accommodated all videos. All test sessions were concluded with answering an SSQ and general comments of the participants to the experiments were collected. Further, an SSQ was conducted with the ACR-HR method prior to the subjective experiment and an SSQ was implemented prior to the second session of the M-ACR-HR method.

1) *ACR-HR method*: Each hidden reference and test video is shown once and its quality is then directly rated as recommended in [6], [7]. A five-level quality rating scale is used: (5) Excellent, (4) Good, (3) Fair, (2) Poor, (1) Bad.

2) *M-ACR-HR method*: Each hidden reference and test video is shown twice separated by a 3 s mid-grey screen between the two presentations. Then, the participants rate the quality of the shown video using the same five-level quality scale as with the ACR-HR method.

D. Participants

A total of 33 volunteers agreed to take part in the M-ACR-HR method of which three failed the vision tests. The ages of the remaining 30 participants (9 females and 21 males) ranged from 23 to 46 years with an average age of 29.53 years. The participants' experience with immersive media on HMDs was self-reported as: 0 "often used" (weekly or monthly), 13 "sometimes used" (few times a year), and 17 "never used".

For the ACR-HR method, 32 volunteers were recruited with two of them failing the vision tests. The ages of the remaining 30 participants (7 females and 23 males) ranged from 20 to 36 years with an average age of 25.37 years. The participants' experience with immersive media on HMDs was self-reported as: 0 "often used", 17 "sometimes used", and 13 "never used".

A pilot study was also conducted for both methods engaging 5 experts that were familiar with immersive media (2 females and 3 males). The ages of these experts span between 30 to 60 years with an average age of 38.2 years.

III. EXPERIMENTAL RESULTS

A participant screening was conducted regarding consistency of opinion scores in order to remove outliers from the subsequent statistical analysis of the data obtained in the subjective experiments. The method recommended by VQEG in [29] for ACR tests and the Matlab code made available

by VQEG in [30] were used. After screening, the numbers of valid participants were obtained as shown in Table I.

TABLE I
SCREENING RESULTS FOR CONSISTENCY OF OPINION SCORES

Method	Number of Valid Participants		
	Sometimes used	Never used	Total
ACR-HR	15	11	26
M-ACR-HR	13	13	26

A. Average Rating Time

The average rating times consumed by the participants subject to the different levels of experience with immersive media on HMDs are shown in Figs. 1(a)-(d). In particular, a comparison of average rating times for each of the four scenes and averaged over all four scenes for the ACR-HR and M-ACR-HR method are provided for each class of participants.

Fig. 1(a) shows the average rating time needed by the experts to cast an opinion score in the ACR-HR and M-ACR-HR method. The average rating times when using the ACR-HR method are shorter compared to the M-ACR-HR method irrespective of the considered scene. Apart from the scene "Blooming", the difference of average rating times between the two methods is rather large for the scenes "Alcatraz", "Formation", and "Panda". This result suggests that the faster scheduling of the test videos in the ACR-HR method (each stimulus is shown only once) compared to the slower progression of test videos in the M-ACR-HR method (each stimulus is shown twice with a mid-gray screen in between) also carries over to the rating time of the experts.

Fig. 1(b) presents the average rating time needed by the participants that have some experience with immersive media on HMDs. Similar as with the experts, this class of participants also rates the quality of the test videos for each of the four scenes in average faster in the ACR-HR method compared to the M-ACR-HR method. However, the difference of the average rating times consumed between the two methods is not as pronounced as observed for the experts. Further, the average rating times take generally longer for the different four scenes and two methods compared to those for the experts.

Fig. 1(c) depicts the results for the participants that had no experience with immersive media on HMDs. Again, average rating times are shorter for the ACR-HR method compared to the M-ACR-HR method although not as different as for the experts but more similar to those results for the participant that have some experience with immersive media on HMDs.

Finally, Fig. 1(d) compares the average rating times for the ACR-HR and M-ACR-HR method and the three classes of participants irrespective of the actual scene. The results indicate that the experts for both methods rate faster compared to the participants with some or no experience with immersive media on HMDs. Further, the average rating time consumed by the experts when using the ACR-HR method is significantly shorter compared to the M-ACR-HR method while this difference is not as large for the other two classes of participants.

Fig. 1. Average rating times and 95% CIs for the ACR-HR and M-ACR-HR method subject to participants' experience with immersive media on HMDs.

Overall, the average rating times for the four scenes, two methods, and three classes of participants turn out to be less than 2.5 s. This finding may assist in designing subjective experiments such that the total duration of participants' exposure to immersive media on HMDs is kept within reasonable limits.

B. MOS and Standard Deviation

In this section, the MOS and STD obtained from the opinion scores given by the participants to the 120 360° videos in both methods are discussed.

1) Results irrespective of participants' experience:

Figs. 2(a)-(b) show the MOS averaged over the four scenes and irrespective of participants' experiences with immersive media on HMDs. Comparing the MOS obtained using the ACR-HR method (Fig. 2(a)) with those using the M-ACR-HR method (Fig. 2(b)) shows a similar progression of decreasing MOS with increasing QP for the considered resolutions. However, the MOS values for the ACR-HR method are slightly higher compared to the M-ACR-HR method (see MOS values for Ref. and QP=42, and fixed resolution). As for the optimal resolution (OP of around 3K) suggested in [17], the MOS progression is similar to that for 4K resolution for both methods.

Fig. 2(c) shows the average STD over the four different scenes associated with the respective MOS in increasing order from 1 (bad) to 5 (excellent). In addition, the polynomial fit to the average STD data points are provided which were obtained using the Matlab curve fitting toolbox. Similar as for conventional videos, the uncertainty about the MOS values is lower for the low and high end of the quality scale while the uncertainty becomes higher towards the mid-quality range for

both methods. Further, the M-ACR-HR method gives lower average STDs at the lower and higher end of the quality scale as compared to the ACR-HR method. This finding suggests that the participants are in closer agreement with their ratings in the M-ACR-HR method when it comes to test videos towards bad and excellent quality. For the mid-range quality ratings, in contrast, the ACR-HR method provides slightly lower average STDs compared to the M-ACR-HR method. In other words, showing each test video twice with the M-ACR-HR method results in higher agreement (lower average STDs) among the ratings of the participants at the lower and higher end of the quality scale compared to the ACR-HR method while lower agreement (higher average STDs) is observed in the mid-range quality regime.

Fig. 2(d) shows the STD of the MOS values for each of the 120 resolution-QP pairs and the respective polynomial fits for the ACR-HR and M-ACR-HR method. The same conclusions as for the progression of average STDs can be drawn for the case of STDs without averaging over the four different scenes.

2) Results subject to participants' experience: Figs. 3(a)-(c) and Figs. 4(a)-(c) show the progression of MOS versus QP for the ACR-HR and M-ACR-HR method, respectively, obtained for the experts, and participants with some and no experience with immersive media on HMDs. In general, similar progressions of MOS for the different resolution-QP pairs can be observed for both methods. Also, for both methods, the experts are able to better differentiate among the 5 resolutions of the 360° reference videos compared to participants with some or no prior experience. Further, fair quality (MOS=3) is rated in the ACR-HR method by all participant classes for 8K

Fig. 2. Overall statistics without differentiation among participants' experience with immersive media on HMDs: (a) Average MOS and 95% CI over the four different scenes for the ACR-HR method, (b) Average MOS and 95% CI over the four different scenes for the M-ACR-HR method, (c) Average standard deviation and polynomial fit over the four different scenes (30 videos per scene), (d) Standard deviation and polynomial fit for the 120 videos.

resolution even when using the high quantization of QP=42. In the M-ACR-HR method, which shows each stimulus twice, the quality is rated below fair for the 8K resolution at QP=42. To also keep the quality for the resolutions of OR, 4K, and 6K at least on a fair level, quantization should not exceed QP=32. On the other hand, for the 2K resolution, already the reference video is given only fair quality and decreases further with an increase of the QP. These results may assist the operation of immersive media systems such that targeted quality levels are reached by choosing suitable resolutions and/or QPs.

Figs. 5(a)-(c) show the STD averaged over the four different scenes for the ACR-HR and M-ACR-HR method along with polynomial fits subject to the three considered classes of participants. Similar results are obtained for the experts and participants having some experience with immersive media on HMDs. Specifically, the M-ACR-HR method provides higher certainty (lower average STDs) about the MOS values compared to the ACR-HR method towards the lower and higher end of the quality scale. In the mid-quality range, the ACR-HR method supports higher certainty in the ratings than the M-ACR-HR method. These characteristics are less pronounced and more asymmetric for the experts compared to the participants with some experience. For the participants that have never used immersive media on HMDs, the progression of the average STD is virtually the same for both methods in the quality range from 1 (bad) to 3 (fair). For MOS values from 3 (fair) to 5 (excellent), the certainty increases with the

M-ACR-HR method compared to the ACR-HR method.

Figs. 6(a)-(c) present the STD and their polynomial fits with respect to the total of 120 360° videos. Similar progressions as for the average STD is observed for the participants having some and no experience with immersive media on HMDs. However, for the experts, the progression of the STD becomes more asymmetric and indicates increased certainty (lower standard deviation) for the M-ACR-HR method towards the higher end of the quality scale.

C. t-tests

The MOS results shown in Section III-B suggest rather small differences between the ACR-HR and M-ACR-HR method. To shed more light on potential differences, inferential statistics in terms of t-tests are presented in this section. In particular, the following steps have been used to deduce whether differences among the MOS values of the two methods are statistically significant:

- **Step 1:** Calculate the average MOS values over all four scenes for each resolution-QP pair and participant for the ACR-HR method. This will result in a set of 31 average MOS values for each resolution-QP pair (5 experts, 15 sometimes used, 11 never used).
- **Step 2:** Calculate the average MOS values over all four scenes for each resolution-QP pair and participant for the M-ACR-HR method. This will result in a set of 31 average MOS values for each resolution-QP pair (5 experts, 13 sometimes used, 13 never used).

Fig. 3. Average MOS and 95% CI over the four scenes for the ACR-HR method subject to participants' experience.

Fig. 4. Average MOS and 95% CI over the four scenes for the M-ACR-HR method subject to participants' experience.

Fig. 5. Average standard deviation and polynomial fit over the four different scenes (30 videos per scene).

Fig. 6. Standard deviation and polynomial fit for the 120 videos.

- **Step 3:** Perform t-tests among the sets of average MOS values of the ACR-HR and M-ACR-HR method for each resolution-QP pair with significance level set to $\alpha = 0.05$. The results of the t-tests are presented in Tables II-V which

contain the p -values for the different experience classes of participants, i.e., overall, experts, sometimes used, and never used immersive media on HMDs. The MOS values for the resolution-QP pairs are considered as significantly different in

a statistical sense if the p -value is lower than the significance level of 0.05 (see bold numbers in the tables):

- **Overall:** The MOS for 10 of the 25 resolution-QP pairs turn out to be significantly different. These differences tend to be in the low resolution and high QP regime, i.e., when the quality of the videos is in range from poor to bad. This result is consistent with the MOS progression in Figs. 2(a)-(b) where the MOS values obtained with the ACR-HR method are noticeably higher for QP= 42 compared to those of the M-ACR-HR method.
- **Experts, sometimes used, and never used:** MOS values of only 1 (experts), 3 (sometimes used), and 1 (never used) resolution-QP pairs turn out to be statistically different. As such, both methods seem to provide a rather similar MOS progression when applied individually to the three participant classes.

In addition, a Bonferroni-corrected t-test was performed between each resolution-QP pair constituting a post-hoc multiple comparison for each of the two methods. These comparisons give insights into a methods' ability to differentiate among the quality associated with the different resolution-QP pairs. In other words, the reliability of a method is revealed and as such relates to the STD (see also Figs. 6(a)-(c)). Given the 120 different 360° videos, there exist $120 \times 119/2 = 7140$ possible combinations of resolution-QP pairs.

Table VI shows the results of the Bonferroni-corrected t-test for 7140 pairs. In general, a larger number of pairs can be differentiated by the M-ACR-HR method compared to the ACR-HR method which translates to a higher reliability of the M-ACR-HR method. This finding applies to the overall results irrespective of participants' experiences and is even more pronounced for the experts (see also Fig. 6(a)) and participants that have never watched immersive media on HMDs (see also Fig. 6(c)). In contrast, for the participants with some experience, the ACR-HR method is slightly more reliable with a larger number of pairs being differentiated compared to the M-ACR-HR method (see also Fig. 6(b)).

D. Analysis of the Simulator Sickness Questionnaire

The SSQ proposed in [31] accounting for 16 symptoms has been used here for the ACR-HR and M-ACR-HR method. Accordingly, symptoms are rated through a symptom variable score: (0) None, (1) Slight, (2) Moderate, (3) Severe. The 16 symptoms can be organized into three symptom clusters [31]:

- **Nausea (N):** General discomfort, increased salivation, sweating, nausea, difficulty concentrating, stomach awareness, burping.
- **Oculomotor (O):** General discomfort, fatigue, headache, eyestrain, difficulty focusing, difficulty concentrating, blurred vision.
- **Disorientation (D):** Difficulty focusing, nausea, fullness of head, blurred vision, dizziness (eyes open), dizziness (eyes closed), vertigo.

Details on the calculation of the SSQ values for the symptom clusters N, O, and D from the respective 16 individual symptoms and the total score (TS) are given in [31].

TABLE II
T-TEST BETWEEN RESOLUTIONS-QP PAIRS FOR THE ACR-HR AND M-ACR-HR METHOD: OVERALL

	Ref.	22	27	32	37	42
8K	0.4972	0.3231	0.1296	0.2342	0.0964	0.1147
6K	0.5467	0.1242	0.5726	0.3363	0.0439	0.5909
4K	0.3234	0.1274	0.0281	0.0466	0.0146	0.0108
OR	0.5785	0.1315	0.0654	0.1000	0.0262	0.0263
2K	0.1613	0.2968	0.0319	0.0437	0.0444	0.1554

TABLE III
T-TEST BETWEEN RESOLUTIONS-QP PAIRS FOR THE ACR-HR AND M-ACR-HR METHOD: EXPERTS

	Ref.	22	27	32	37	42
8K	0.6103	0.8709	0.1950	0.6103	0.8494	0.4492
6K	0.6883	0.3559	0.5319	1.0000	0.8743	0.8494
4K	0.3175	0.7585	0.5077	0.7924	0.6103	0.7078
OR	0.4021	0.5212	1.0000	0.5319	0.3122	0.3016
2K	0.7103	0.8651	0.4236	0.8196	0.0415	1.0000

TABLE IV
T-TEST BETWEEN RESOLUTIONS-QP PAIRS FOR THE ACR-HR AND M-ACR-HR METHOD: SOMETIMES USED

	Ref.	22	27	32	37	42
8K	0.7439	0.7439	0.9702	0.4913	0.2538	0.4173
6K	0.8026	0.6087	0.9120	0.9552	0.0835	0.9356
4K	0.4053	0.5078	0.1785	0.1282	0.0232	0.0057
OR	0.6316	0.6644	0.2345	0.2608	0.2019	0.0952
2K	0.3603	0.7088	0.0759	0.0862	0.0128	0.1690

TABLE V
T-TEST BETWEEN RESOLUTIONS-QP PAIRS FOR THE ACR-HR AND M-ACR-HR METHOD: NEVER USED

	Ref.	22	27	32	37	42
8K	0.7556	0.2805	0.1087	0.2119	0.2866	0.2764
6K	0.4318	0.2991	0.8554	0.2430	0.2147	0.4399
4K	0.9332	0.1966	0.1455	0.1939	0.2538	0.2899
OR	0.7148	0.0307	0.1340	0.1210	0.1736	0.2761
2K	0.1789	0.2386	0.3865	0.1727	0.9246	0.5908

TABLE VI
BONFERRONI-CORRECTED T-TEST FOR 7140 POSSIBLE RESOLUTION-QP PAIRS AND SIGNIFICANCE LEVEL $\alpha = 0.05$

	Overall	Experts	Sometimes used	Never used
ACR-HR	3427	1183	2536	1513
M-ACR-HR	3512	1850	2444	1967

The results presented in the sequel extend our work reported in [32] on evaluating simulator sickness for 360° videos on an HMD for the M-ACR-HR method by providing a comparison to the ACR-HR method.

Figs. 7(a)-(d) show the mean weighted SSQ scores and 95% CI obtained from the post-session SSQs for the overall participants and subject to their experience with immersive media on HMDs. Because the M-ACR-HR method shows each stimulus twice back-to-back, the 120 360° videos were shown in two sessions each presenting 60 stimuli. In this way, a similar session duration as for the ACR-HR method is maintained in each session of the M-ACR-HR method.

Fig. 7(a) provides the overall results irrespective of partic-

Fig. 7. Statistical analysis of symptom cluster scores in terms of mean weighted SSQ scores and their 95% CI.

ipants' experience. The results show that the mean weighted SSQ scores obtained for all three symptom clusters N, O, D, and the total score TS (also referred to as total severity) are significantly lower for the ACR-HR method compared to the M-ACR-HR method. Nausea related symptoms were perceived less severe compared to oculomotor and disorientation related symptoms. As for the M-ACR-HR method, it is observed that the mean weighted SSQ scores are slightly lower in Session 2. It should be mentioned that a similar order was observed for the average rating times reported in [21] for the M-ACR-HR method, with opinion scores given faster in Session 2. These characteristics may be attributed to a learning effect that took place in Session 1 to the benefit of Session 2.

Fig. 7(b) presents the mean weighted SSQ scores for the experts supporting the same finding as for the overall results, i.e., the ACR-HR method induces lower levels of sickness symptoms than the M-ACR-HR method, for nausea N, oculomotor O, and total score TS. In contrast, for the limited number of five experts, the M-ACR-HR method outperforms the ACR-HR method for disorientation related symptoms.

Fig. 7(c) shows the results obtained for participants having some experience with immersive media on HMDs. The same conclusions as for the overall results shown in Fig. 7(a) can be drawn, i.e., nausea related symptoms appear less severe compared to oculomotor and disorientation related symptoms. The aforementioned learning effect related to casting an opinion score and improved comfort to watching stimuli on HMDs is also supported for this class of participants. Again, the ACR-HR method causes significantly lower discomfort compared to the M-ACR-HR method.

Fig. 7(d) shows rather different results for the mean weighted SSQ scores for the participants which have never watched immersive media on HMDs compared to the overall results and those for the experts and participants with some experience. In particular, for the M-ACR-HR method, scores related to the nausea cluster are the lowest, followed by symptoms related to the oculomotor cluster with higher scores, and the disorientation cluster causing the highest discomfort. Further, the mean weighted SSQ scores for each symptom cluster and the accumulated total score TS increase from Session 1 to Session 2. On the other hand, discomfort is kept significantly lower for the ACR-HR method compared to the M-ACR-HR method irrespective of the symptom cluster.

IV. CONCLUSIONS

In this paper, we have compared the performance of the ACR-HR and M-ACR-HR method in the context of subjective quality assessment of 360° videos subject to participants experience with immersive media on HMDs. The analysis of the average rating times required to cast an opinion score reveals that this is lower for the ACR-HR method irrespective of the scene and participants' experience. Given that the ACR-HR method shows each stimulus only once while the M-ACR-HR presents each stimulus twice back-to-back before rating, the overall time consumed to conduct a subjective experiment is significantly lower for the ACR-HR method. As such, a much larger set of test conditions can be assessed with the ACR-HR method in a shorter time. On the other hand, the progression of the MOS values versus QPs for the different resolutions is similar for both methods irrespective of the

participants' experience. However, in general, the M-ACR-HR method is more reliable when it comes to differentiating among quality levels and certainty of the given ratings. This holds specifically for the experts and the participants that have never used immersive media on HMDs. The analysis of the SSQs has shown that significantly lower discomfort is induced by the ACR-HR method while simulator sickness symptoms are much stronger for the M-ACR-HR method.

ACKNOWLEDGMENT

This work was supported in part by The Knowledge Foundation, Sweden, through the ViaTecH project (Contract 20170056). The study has been granted ethical approval by the Regional Ethics Review Board in Lund, Sweden (Dnr. 2018/863). We thank all volunteers who generously shared their time to participate in the subjective experiments. Special thanks go to Francisco Lopez Luro and Diego Navarro for their advice given to the development of the test platform.

REFERENCES

- [1] 3GPP TR 26.918 V15.2.0, *Virtual Reality (VR) Media Services Over 3GPP (Release 15)*, 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects, Mar. 2018.
- [2] M. Latva-aho and K. Leppänen (eds.), *Key Drivers and Research Challenges for 6G Ubiquitous Wireless Intelligence, 6G Research Visions 1*, 6G Flagship, University of Oulu, Finland, Sep. 2019.
- [3] ITU-T P.360-VR, *Subjective Test Methodologies for 360 Degree Video on HMD*, Int. Telecommun. Union, (for consent) 2020.
- [4] P. Pérez and J. Escobar, "MIRO360: A Tool for Subjective Assessment of 360 Degree Video for ITU-T P.360-VR," in *Proc. Int. Conf. on Quality of Multimedia Experience*, Berlin, Germany, Jun. 2019, p. 1–3.
- [5] A. Singla, S. Fremerey, W. Robitza, P. Lebreton, and A. Raake, "Comparison of Subjective Quality Evaluation for HEVC Encoded Omnidirectional Videos at Different Bit-rates for UHD and FHD Resolution," in *Proc. Thematic Workshops of ACM Multimedia*, ser. Thematic Workshops '17, Mountain View, CA, USA, Oct. 2017, pp. 511–519.
- [6] ITU-T P.910, *Subjective Video Quality Assessment Methods for Multimedia Applications*, Int. Telecommun. Union, Apr. 2008.
- [7] ITU-T P.915, *Subjective Assessment Methods for 3D Video Quality*, Int. Telecommun. Union, Mar. 2016.
- [8] ITU-R BT.500-13, *Methodology for the Subjective Assessment of the Quality of Television Pictures*, Int. Telecommun. Union, Jan. 2012.
- [9] ITU-T BT.1788, *Methodology for the Subjective Assessment of Video Quality in Multimedia Applications*, Int. Telecommun. Union, Feb. 2007.
- [10] B. Zhang, J. Zhao, S. Yang, Y. Zhang, J. Wang, and Z. Fei, "Subjective and Objective Quality Assessment of Panoramic Videos in Virtual Reality Environment," in *Proc. IEEE Int. Conf. on Multimedia Expo Workshops*, Hong Kong, China, Jul. 2017, pp. 163–168.
- [11] H. T. T. Tran, N. P. Ngoc, C. T. Pham, and Y. J. Jung, "A Subjective Study on QoE of 360 Video for VR Communication," in *Proc. IEEE Int. Workshop on Multimedia Signal Process.*, Luton, UK, Oct. 2017, pp. 1–6.
- [12] F. Lopes, J. Ascenso, A. Rodrigues, and M. P. Queluz, "Subjective and Objective Quality Assessment of Omnidirectional Video," in *Proc. SPIE Optical Eng. + Appl.*, San Diego, CA, USA, Sep. 2018, pp. 1–17.
- [13] A. Singla, S. Fremerey, W. Robitza, and A. Raake, "Measuring and Comparing QoE and Simulator Sickness of Omnidirectional Videos in Different Head Mounted Displays," in *Proc. IEEE Int. Conf. on Quality of Multimedia Experience*, Erfurt, Germany, June 2017, pp. 1–6.
- [14] M. Xu, C. Li, Y. Liu, X. Deng, and J. Lu, "A Subjective Visual Quality Assessment Method of Panoramic Videos," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, Hong Kong, China, Jul. 2017, pp. 517–522.
- [15] H. Duan, G. Zhai, X. Yang, D. Li, and W. Zhu, "IVQAD 2017: An Immersive Video Quality Assessment Database," in *Proc. Int. Conf. on Syst., Signals and Image Process.*, Poznan, Poland, May 2017, pp. 1–5.
- [16] S.-H. Yao, C.-L. Fan, and C.-H. Hsu, "Towards Quality-of-Experience Models for Watching 360° Videos in Head-Mounted Virtual Reality," in *Proc. Int. Conf. on Quality of Multimedia Experience*, Berlin, Germany, Jun. 2019, p. 1–3.
- [17] Y. Zhang et al., "Subjective Panoramic Video Quality Assessment Database for Coding Applications," *IEEE Trans. on Broadcast.*, vol. 64, no. 2, pp. 461–473, Jun. 2018.
- [18] E. Alexiou, E. Upenik, and T. Ebrahimi, "Towards Subjective Quality Assessment of Point Cloud Imaging in Augmented Reality," in *Proc. IEEE Int. Workshop on Multimedia Signal Process.*, Luton, UK, Oct. 2017, pp. 1–6.
- [19] A. Singla, W. Robitza, and A. Raake, "Comparison of Subjective Quality Evaluation Methods for Omnidirectional Videos with DSIS and Modified ACR," in *Proc. Human Vision and Electronic Imaging*, Burlingame, CA, USA, Jan. 2018, pp. 1–6.
- [20] B. Zhang, Z. Yan, J. Wang, Y. Luo, S. Yang, and Z. Fei, "An Audio-Visual Quality Assessment Methodology in Virtual Reality Environment," in *Proc. IEEE Int. Conf. on Multimedia & Expo Workshops*, San Diego, CA, USA, Jul. 2018, pp. 1–6.
- [21] M. Elwardy, H.-J. Zepernick, V. Sundstedt, and Y. Hu, "Impact of Participants' Experiences with Immersive Multimedia on 360° Video Quality Assessment," in *Proc. Int. Conf. on Signal Process. and Commun. Syst.*, Gold Coast, Australia, Dec. 2019, pp. 40–49.
- [22] H.-J. Zepernick, M. Elwardy, Y. Hu, and V. Sundstedt, "On the Number of Participants Needed for Subjective Quality Assessment of 360° Videos," in *Proc. Int. Conf. on Signal Process. and Commun. Syst.*, Gold Coast, Australia, Dec. 2019, pp. 50–59.
- [23] Y. Nehmé, J.-P. Farrugia, F. Dupont, P. L. Callet, and G. Lavoué, "Comparison of Subjective Methods, With and Without Explicit Reference, for Quality Assessment of 3D Graphics," in *Proc. ACM Symp. on Applied Perception*, Barcelona, Spain, Sep. 2019, pp. 17:1–17:9.
- [24] A. Singla, W. Robitza, and A. Raake, "Comparison of Subjective Quality Test Methods for Omnidirectional Video Quality Evaluation," in *Proc. IEEE Int. Workshop on Multimedia Signal Process.*, Kuala Lumpur, Malaysia, Sep. 2019, pp. 1–7.
- [25] C. Li, M. Xu, X. Du, and Z. Wang, "Bridge the Gap Between VQA and Human Behavior on Omnidirectional Video: A Large-Scale Dataset and a Deep Learning Model," in *Proc. ACM Int. Conf. on Multimedia*, Seoul, Republic of Korea, Oct. 2018, pp. 932–940.
- [26] VQA-ODV, Beihang University, School of Electronic and Information Engineering, Beijing, China, 2017 (accessed Apr. 27, 2019), <https://github.com/Archer-Tatsu/VQA-ODV>.
- [27] *H.264 Video Encoding Guide*, FFmpeg, 2018 (accessed July 4, 2019). [Online]. Available: <https://trac.ffmpeg.org/wiki/Encode/H.264crf>
- [28] *FFmpeg and H.265 Encoding Guide*, FFmpeg, 2018 (accessed July 4, 2019). [Online]. Available: <https://trac.ffmpeg.org/wiki/Encode/H.265>
- [29] *VQEG Report on the Validation of Video Quality Models for High Definition Video Content*, Video Quality Experts Group, 2010 (accessed June 28, 2019). [Online]. Available: <http://www.vqeg.org>
- [30] Video Quality Experts Group.
- [31] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness," *The Int. J. of Aviation Psychology*, vol. 3, no. 3, pp. 203–220, 1993.
- [32] M. Elwardy, H.-J. Zepernick, Y. Hu, T. M. C. Chu, and V. Sundstedt, "Evaluation of Simulator Sickness for 360° Videos on an HMD Subject to Participants' Experience with Virtual Reality," in *Proc. IEEE Conf. on Virtual Reality and 3D User Interfaces Abstracts and Workshops*, Atlanta, GA, USA, Mar. 2020, pp. 477–484.