This is the accepted version of a paper presented at 16th International Conference on Signal Processing and Communication System, ICSPCS 2023, Bydgoszcz, 6 Sept - 8 Sept 2023.

Citation for the original published paper:

ACR360: A Dataset on Subjective 360° Video Quality Assessment Using ACR Methods
https://doi.org/10.1109/ICSPCS58109.2023.10261151

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

©2023 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-25548
ACR360: A Dataset on Subjective 360° Video Quality Assessment Using ACR Methods

Majed Elwardy, Hans-Jürgen Zepernick, Yan Hu, Thi My Chinh Chu
Blekinge Institute of Technology, SE-371 79 Karlskrona, Sweden
E-mail: {mew, hjz, yhx, cch}@bth.se

Abstract—The recent advances in immersive technologies have been essential in the development of a wide range of novel standalone and networked immersive media applications. The concepts of virtual reality, augmented reality, and mixed reality relate to different compositions of real and computer-generated virtual objects. In this context, 360° video streaming has become increasingly popular offering improved immersive experiences when viewed on a head-mounted display (HMD). An important component in the development of novel immersive media systems are subjective tests in which participants assess the quality of experience of representative test stimuli. In this paper, the annotated ACR360 dataset is presented which is publicly available on GitHub. The ACR360 dataset contains a wide range of psychophysical and psychophysiological data that was collected in subjective tests on 360° video quality. The test stimuli were shown on an HMD and rated according to the absolute category rating (ACR) and modified ACR (MACR) methods. To support an easy exploration and utilization of the ACR360 dataset by the research community, its structure on GitHub is described and a comprehensive illustration of analysis options are provided for each data category. The ACR360 dataset may be used for conducting meta-analysis in combination with other datasets to improve precision and to pursue research questions that cannot be answered by an individual study.

Index Terms—Immersive media, 360° video, subjective test, quality assessment, absolute category rating, annotated dataset.

I. INTRODUCTION

Immersive media have gained significant attention in recent years due to their great potential of supporting novel applications ranging from entertainment and social interaction, to education, healthcare and wellness, and business [1]–[4]. The trend from extended realities to the metaverse includes the entire real-virtuality continuum of combining real objects and fully computer-generated objects. Apart from virtual reality (VR), augmented reality, and mixed reality, viewing 360° videos on head-mounted displays (HMDs) is foreseen to account for a large portion of VR content [5], [6]. In particular, 360° video, also referred to as omnidirectional video, spherical video, surround video, or immersive video, are recorded using an omnidirectional camera and may be shown on a conventional display or viewed on an HMD. While the user can control the viewing direction on conventional displays via a keyboard or hand-controller, a more natural and immersive experience is offered on HMDs allowing to control viewing directions through head movements.

An important component in the development of immersive media processing chains, networked immersive media systems, applications, and services, are subjective tests in which participants assess the quality of experience (QoE) of representative test stimuli. The recordings gathered during subjective tests, constituting a ground truth for system design, may include participants’ opinion scores, psychophysical and psychophysiological data, and other viewer behavior-related data. As such, as subjective tests are time consuming and resource expensive, there exists a high demand on publicly available annotated datasets on subjective experiences of immersive media. Furthermore, in light of the recent COVID-19 pandemic, the availability of annotated datasets is also considered as vital to ensure progress in experimental in-person research under opportunity limited conditions [7].

Regarding annotated dataset on the QoE of conventional image and video stimuli, a tremendous amount of data has been made publicly available to the research community. For example, the Qualinet database contains 226 annotated datasets [8] while surveys of public image and video datasets for quality assessment are provided, e.g., in [9], [10]. On the other hand, the portfolio of public annotated datasets on the various factors of QoE of immersive media is not as developed. Surveys on related public annotated datasets that may be used as ground truth in the area of 360° video quality assessment are provided in [11], [12]. These surveys demonstrate the need for additional annotated datasets covering a broader range of subjective aspects of immersive media. However, there exist only a few papers presenting public annotated datasets that focus on subjective quality assessment of 360° videos on HMDs, e.g., [13], [14]. Recently, the results of a large scale cross-lab test campaign involving 10 laboratories and engaging over 300 participants has been reported in [15] and made available to the public. This comprehensive study has been conducted by the Immersive Media Group of the Video Quality Experts Group (VQEG) and was instrumental for developing ITU-T Recommendation P.919 on subjective test methodologies for 360° videos on HMDs. The study validated that the absolute category rating (ACR) and degradation category rating (DCR) are indeed well suited for conducting subjective tests of 360° videos on HMDs.

Motivated by all of the above, in this paper, the annotated ACR360 dataset is presented which is publicly available on GitHub [16]. The ACR360 dataset contains the psychophysical and psychophysiological data that was recorded during subjective tests on 360° video quality assessment on an HMD. The quality of the visual stimuli was assessed according to the ACR and modified ACR (MACR) methods. The structure of
the ACR360 dataset is provided to allow an easy exploration and utilization of the data by the research community. Comprehensive examples on the analysis options are provided for each data category illustrating potential research questions that may be pursued with the help of the ACR360 dataset. The ACR360 dataset may also be used for conducting meta-analysis in combination with other datasets to improve precision and to pursue research questions that cannot be answered by an individual study.

The reminder of this paper is organized as follows. Section II describes the experimental setup used throughout our various subjective test campaigns conducted over the years as needed for the understanding of the ACR360 dataset structure and its content. In Section III, the ACR360 dataset structure is provided, i.e., file directory and folder content on GitHub. Section IV discusses and visualizes the opinion scores, rating times, head movements, pupil dilations, galvanic skin responses (GSRs), and simulator sickness questionnaire (SSQ) scores recorded for the ACR and MACR methods. Finally, conclusions are drawn in Section V.

II. EXPERIMENTAL SETUP

This section describes the experimental setup used in our subjective test campaigns on 360° video quality assessment on an HMD [17]–[23] to the extent needed for the understanding of the ACR360 dataset. More information about test procedures, 360° video processing, generation of the test stimuli, and specifications of the software suites and hardware equipment used in the various subjective tests can be found in [17].

A. Stimuli, Software, and Equipment

The 360° reference videos are based on four 8K resolution natural scenes taken from the VQA-ODV dataset [24], [25], are cut to 10 s duration, and played out with a frame rate of 29.97 frames per second (fps). Fig. 1 shows sample frames of the four selected scenes. A summary of the 360° reference videos, 360° test videos, software suites, and hardware equipment used in the various subjective tests is provided in Table I. Note that the test stimuli were generated by compressing the reference stimuli using different quantization parameters (QPs) [26], [27].

B. Test Methods

The subjective quality assessment of the 360° video stimuli was performed using a five-level rating scale: (5) Excellent, (4) Good, (3) Fair, (2) Poor, (1) Bad. On this basis, a category judgement of the test videos was conducted according to the following two test methods:

- ACR method: Each stimulus is shown once followed by rating its quality on the above five-level rating scale [29].
- MACR method: Each stimulus is shown twice with a 3-second mid-grey screen between the back-to-back presentations [30]. The quality of each video is assessed after its second presentation has finished using the above five-level rating scale.

It is noted that the ACR method has been widely used for conventional video quality assessment [31]–[33] and has recently also been recommended in the context of subjective tests for 360° videos on HMDs [29]. The MACR method was suggested in [30] to account for participants that have little experience with viewing immersive media on HMDs.

<table>
<thead>
<tr>
<th>360° video stimuli</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref. stimuli</td>
</tr>
<tr>
<td>Test stimuli</td>
</tr>
<tr>
<td>QP: 22, 27, 32, 37, 42</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Software suites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test platform development</td>
</tr>
<tr>
<td>Unity 3D Version 2018.3.11f1 and</td>
</tr>
<tr>
<td>Visual Studio 2017</td>
</tr>
<tr>
<td>Bio recordings</td>
</tr>
<tr>
<td>iMotion Software Version 7.1</td>
</tr>
<tr>
<td>SSQ</td>
</tr>
<tr>
<td>iMotion Software Version 7.1 based GUI</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Human-machine interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMD</td>
</tr>
<tr>
<td>HTC Vive Pro with integrated eye-tracker</td>
</tr>
<tr>
<td>Interaction</td>
</tr>
<tr>
<td>HTC Vive controller:</td>
</tr>
<tr>
<td>(1) Execute calibration instructions</td>
</tr>
<tr>
<td>(2) Cast quality ratings</td>
</tr>
<tr>
<td>Sensors</td>
</tr>
<tr>
<td>Shimmer GSR biosensor:</td>
</tr>
<tr>
<td>(1) Galvanic skin response</td>
</tr>
<tr>
<td>(2) Heart rate</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>High-performance computing platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
</tr>
<tr>
<td>Corsair One i160 Gaming PC with:</td>
</tr>
<tr>
<td>(1) Intel I9-9900K processor</td>
</tr>
<tr>
<td>(2) NVIDIA GeForce RTX 2080 TI graphics card</td>
</tr>
</tbody>
</table>

Fig. 1. Sample frames of the 360° video scenes [24], [25].
of events. The actual data files in these sub-folders are in .csv format while further details about the file structure can be found in the Readme.md files. A data file containing information about the participants is also provided for each test method. The Test_Scenes folder provides information about the 360° reference videos and 360° test videos along with instructions for downloading these videos.

The content of the folders and details of the data files in each sub-folder are provided according to the following conventions for the ACR and MACR methods. Note that the notation $X \in \{\text{ACR, MACR}\}$ is used in the sub-folder names:

- **Data_X_MOS**: This folder contains the opinion scores given by each participant to the video stimuli along with their rating times. Four different .csv files are provided for each participant, one for each of the four scenes.
- **Data_X_HMD**: This folder contains recordings of the participants’ head movements during HMD exposure, i.e., head positions and head rotations.
- **Data_X_EYE**: The eye-tracking data of each participant is organized into three files containing the recorded gaze left, gaze right, and gaze combined data.
- **Data_X_GSR**: This folder contains the GSR files of the recorded GSR signals in micro-Siemens ($\mu$S), GSR quality (valid/not-valid), and heart rate in beats per minute.
- **Data_X_SSQ**: This folder reports the symptom variability scores given by each participant to the 16 sickness symptoms of the SSQ. For the ACR method, the Pre-SSQ and Post-SSQ files contain the SSQ scores given before and after exposure to the HMD, respectively. For the MACR method, the Post-SSQ-S1 and Post-SSQ-S2 files contain the SSQ scores given

### III. DATASET STRUCTURE

The ACR360 dataset directory on GitHub [16] is organized into five main folders, i.e., two folders for each test method and one folder containing the test sequences, as shown in Figure 4. Six sub-folders are used for each test method providing the respective sets of opinion scores, head movement data, eye-tracking data, GSR recordings, SSQ scores, and timestamps for each participant to record skin conductivity during HMD exposure. In addition, the SSQ and sickness symptoms as introduced in [34] were explained as well as the symptom variable score used to rate the presence of each symptom: (0) None, (1) Slight, (2) Moderate, (3) Severe. In the ACR method, a pre-session SSQ (Pre-SSQ) and post-session SSQ (Post-SSQ) were file before and after the quality assessment task on the HMD. In the MACR method, the participants completed a post-session SSQ after the first session (Post-SSQ-S1), followed by completing a post-session SSQ after the second session (Post-SSQ-S2).

### C. Test Procedures

Figs. 2-3 show the schedules of components and stimuli presentation for the test sessions using the ACR and MACR method, respectively. Before starting the actual subjective tests, each participant was given an explanation of the quality assessment task, instructions on the equipment usage, and a briefing on safety precautions regarding the risks of using HMDs. The GSR sensor electrodes were set up on the participant to record skin conductivity during HMD exposure. In addition, the SSQ and sickness symptoms as introduced in [34] were explained as well as the symptom variable score used to rate the presence of each symptom: (0) None, (1) Slight, (2) Moderate, (3) Severe. In the ACR method, a pre-session SSQ (Pre-SSQ) and post-session SSQ (Post-SSQ) were file before and after the quality assessment task on the HMD. In the MACR method, the participants completed a post-session SSQ after the first session (Post-SSQ-S1), followed by completing a post-session SSQ after the second session (Post-SSQ-S2).

### D. Participants

A pilot test was conducted for both test methods before conducting the full subjective tests with more participants. Table II summarizes demographic information about the panels of participants. The table also provides participants’ self-reported experience with watching immersive media on HMDs.

#### TABLE II DEMOGRAPHIC INFORMATION

<table>
<thead>
<tr>
<th>Test Method</th>
<th>Gender</th>
<th>Age Range</th>
<th>Average Age</th>
<th>Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilot tests with ACR and MACR</td>
<td>2 females, 3 males</td>
<td>30-60 years, average of 38.2 years</td>
<td>5 experts familiar with immersive media</td>
<td></td>
</tr>
<tr>
<td>Subjective tests with ACR</td>
<td>7 females, 23 males</td>
<td>20-36 years, average age of 25.37 years</td>
<td>Often used: 0, Sometimes used: 17, Never used: 13</td>
<td></td>
</tr>
<tr>
<td>Subjective tests with MACR</td>
<td>9 females, 21 males</td>
<td>23-46 years, average age of 29.53 years</td>
<td>Often used: 0, Sometimes used: 13, Never used: 17</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2. Procedure of the subjective tests using the ACR method ©[2019]IEEE. Reprint, with permission, from [17].

Fig. 3. Procedure of the subjective tests using the MACR method which is split into Session 1 (S1) and Session 2 (S2).
files hold the SSQ scores given after the first and after the second session, respectively.

- **Data_X_TimeStamp**: Timestamps that mark the start and end of each event such as calibration, stimuli presentation, quality rating, and SSQ, are provided in this folder (see also Figs. 2-3).

- **Participants_X_Information.csv**: Age, gender, occupation, and level of experience with immersive media of each participant are provided in this folder.

### IV. Recorded Data

In this section, we illustrate the recorded data for both the ACR and MACR method. A detailed comparison of the ACR and MACR methods are available in [35]. For the purpose of illustrating the use of the ACR360 dataset, we selected the first four participants P1-P4 for the ACR method and participants P20-P23 for the MACR method. To prevent any ambiguities in the illustrations, the participants selected for each method are distinct such that they do not overlap.

#### A. Opinion Scores

Opinion scores measure the participants’ subjective perception of the stimuli and are averaged over the number of participants to obtained mean opinion scores (MOS). Figs. 5(a)-(e) show the average MOS over all four scenes for different resolution-QP pairs that were obtained for both methods. A consistent trend of decreasing average MOS with increasing QP is observed for each resolution. The results also highlight the impact of resolution on video quality, with 8K reference videos providing good quality which decreases to fair quality for increased compression with QP = 42. In contrast, the 2K reference videos have already only fair quality without compression which rapidly decreases with increasing compression toward QP = 42.

#### B. Rating Times

Rating times may provide insights into the difficulty of giving a quality score to a presented video stimulus. Fig. 6 presents violin plots of the rating times for both methods, representing the time it takes the participants to provide their opinion scores. The violin plots comprise of a box plot and a kernel density plot of the respective data. The ordinate represents the rating times and the abscissa organizes the results into the four video scenes, and results averaged over the four scenes.

It is observed that wider sections of the rating time distribution are located closer to 1 s for the ACR method and 2 s for the MACR method. The rating times are focused around the first quartile below the median of the box plots for both methods. These results support the conjecture that participants provide their ratings faster for the ACR method compared to the MACR method. This finding is underpinned by the non-overlapping notches of the pairs box plots with those for the ACR method being located lower compared to the MACR method. The same applies to the mean rating times (black marker) which are lower for the ACR method compared to the MACR method.

#### C. Head Movements

The head movement data recorded during HMD exposure can provide insights into participants’ test stimuli exploration behaviors. Figs. 7(a)-(d) and Figs. 8(a)-(d) depict the cumulative distribution functions (CDFs) of rotational head movements in terms of yaw, pitch, and roll angles, for participants P1-P4 in the ACR method and P20-P23 in the MACR method, respectively. The results suggest that the participants explore
D. Pupil Dilation

The pupil dilation data, recorded from eye-tracking measurements, can provide objective measures of participants’
erance or cognitive load in response to the test stimuli. Figs. 11-12 show the eye pupil diameters for the left and right eye of participants P1-P4 in the ACR method and P20-P23 for the two sessions in the MACR method, respectively.

These results indicate that eye pupil diameters may be significantly different between participants for the ACR method while these differences are not as pronounced for the MACR method. This observation may lead to the conjecture, subject to further study, that the faster progression of test stimuli in the ACR method may cause higher cognitive load to some participants.

Figs. 13-14 show the GSR amplitudes measured for participants P1-P4 during HMD exposure for the ACR method and P20-P23 during the two sessions for the MACR method, respectively. It is noted that the GSR amplitudes tend to increase or fluctuate over time based on the mental load associated with the duration of HMD exposure for both methods. In addition, it is observed that these characteristics may differ significantly among participants which indicates individual variability in physiological responses. The recorded GSR signals can also be used to calculate more aggregated metrics associated with the progression of the GSR such as GSR peaks and peaks per minute. These metrics can provide further insights into the participants’ physiological responses and emotional arousal during HMD exposure.

**E. Galvanic Skin Response**

GSR measures changes in the electrical conductance of the skin which can reflect emotional arousal or stress levels.

**F. SSQ Scores**

The SSQ scores represent the self-reported symptoms of simulator sickness experienced by the participants during the HMD exposure. SSQs are used to reveal the level of
discomfort or adverse effects on participants in given test scenarios.

Fig. 15 shows the mean Pre-SSQ and Post-SSQ scores for each of the 16 symptoms for the ACR method, aggregated over all participants. The results reveal that only minor levels of selected symptoms were reported in the Pre-SSQ by some participants prior to exposure to the 360° videos on the HMD. Furthermore, the Post-SSQ scores reported after a test session slightly increased compared to the Pre-SSQ scores given before a test session.

Fig. 16 provides the average Post-SSQ scores over all participants after performing the quality assessment task on the HMD for the MACR method. Clearly, the Post-SSQ-S1 and Post-SSQ-S2 scores for the MACR method are higher compared to the Post-SSQ scores for the ACR method. This finding suggests that the MACR method induces a potentially higher level of simulator sickness than the ACR method. However, the mean SSQ scores obtained for each of the 16 individual sickness symptoms are very low for both methods compared to the score of 20, in which case a test setting would be considered as problematic.

V. CONCLUSIONS

In this paper, we have presented the annotated ACR360 dataset which has been made publicly available on GitHub. The ACR360 dataset provides the psychophysical and psychophysiological data that were recorded during subjective tests on 360° video quality assessment on an HMD using the ACR and MACR methods. The data recorded covers opinion scores, rating times, head movements, pupil dilation, GSR, and SSQ scores for each participant. The discussion and visualization of each data category shows the vast scope of the ACR360 dataset as a ground truth for designing 360° video systems, benchmarking algorithms of 360° video processing chains, and conducting meta-analysis. The ACR360 dataset also opens up opportunities for further research in understanding human behavior in immersive media through the use of artificial intelligence and data-driven objective quality modeling. Future research efforts may be directed toward: 1) performing meta-analysis in combination with other datasets to improve the precision of results and to gain insight into other QoE related metrics not covered so far, 2) developing new objective models and metrics that can predict and quantify the quality of 360° video applications, 3) expanding subjective tests to other types of immersive media along with making related dataset available to the research community.

VI. ACKNOWLEDGMENTS

This work was supported in part by the Knowledge Foundation, Sweden, through the ViaTecH and HINTS projects under contract numbers 20170056 and 20220068, respectively. The studies were granted ethical approval (Dnr. 2018/863). We would like to thank Francisco Lopez Luro and Diego Navarro for their advice given to the development of the test platform, and Veronica Sundstedt for valuable discussions in the course of this work. We thank all volunteers who generously shared their time to participate in the subjective tests.
REFERENCES


