”Discriminability-Experimental Cost” tradeoff in subjective video quality assessment of codec: DCR with EVP rating scale versus ACR-HR

Andréas Pastor, Pierre David, Ioannis Katsavounidis, Lukas Krasula, Hassene Tmar, Patrick Le Callet

"Discriminability–Experimental Cost" tradeoff in subjective video quality assessment of codec: DCR with EVP rating scale versus ACR–HR

Andréas Pastor[1], Pierre David[1, 2], Ioannis Katsavounidis[3], Lukáš Krasula[4], Hassene Tmar[3], Patrick Le Callet[1, 5]
[1]*Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France*
[2]*CAPACITÉS SAS* [3]*Meta, USA* [4]*Netflix Inc., USA* [5]*Institut universitaire de France (IUF)*

***Abstract*—This work compares two subjective studies conducted in a controlled laboratory environment on SDR HD, UHD, and HDR UHD contents using naive observers. The goal of these tests is to compare the precision and accuracy of a modified Degradation Category Rating (DCR) and Absolute Category Rating with Hidden Reference (ACR–HR) subjective methods for video quality assessment. The modified version of the DCR method includes a repetition of both reference and distorted stimuli; and utilizes an 11–grade rating scale from Expert Viewing Protocol (EVP) of ITU–R BT.500-15 standards. In the second subjective protocol, ACR–HR operates without repetition and with the 5–grade quality scale from ITU standards.**

We perform an extensive analysis of the scale usage and compare Mean Opinion Scores (MOS) scores discriminability in both subjective studies. We show that both methods can retrieve accurate MOS. However, the ACR–HR method achieves better discriminability among MOS than DCR with EVP rating scale, while reducing by a factor of two the experimental effort, i.e., the cost of the experiment.

The findings of this work give new insight into how to perform cost-efficient subjective tests for video quality estimation with naive observers, and how to retrieve good MOS estimates.

***Index Terms*—subjective methodologies, video quality assessment, modern video CODEC, system validation**

I. Introduction

Conducting subjective video quality assessment experiments is necessary to obtain reliable quality scores and validate objective quality metrics, encoding pipelines, and systems.

Several methodologies are available and defined in International Telecommunication Union (ITU) standards [1]–[3]. However, these methodologies can be considered time-consuming and expensive to run. They require recruiting observers and inviting them to in-lab experiments lasting, on average, between 30 to 60 minutes. It is then critical to optimize the trade-off between the quantity and quality of the collected data from a given panel of observers. From an available budget, how many video sequences can we afford to test knowing the efficiency of a subjective quality assessment protocol and its required annotation time per observer?

This work investigates and compares two quality assessment methodologies in the context of video quality assessment for High Definition (HD) videos, and Ultra High Definition (UHD) Standard Dynamic Range (SDR) and High Dynamic Range (HDR) videos. Several works have focused on subjective methodology comparison for different multimedia systems. For 2D videos, [4] compared Subjective Assessment Methodology for Video Quality (SAMVIQ) and ACR–HR on HDTV, VGA, and QVGA sequences. In [5], multiple ACR–HR, DCR, and SAMVIQ versions with differing rating scales are compared for mobile videos and 3D videos in [6]. Results presented in [7], [8] compare the M–ACR method proposed in [9] for 360 video quality evaluation with other subjective methodologies: ACR and DCR. It is concluded that DCR is statistically more reliable for 360 video quality, as reported in ITU-T Rec. P.919 [10]. Other works have explored subjective methodologies for 3D graphics quality evaluation [11], for 360 audiovisual with spatial audio contents using expert or naive assessors [12], and to compare pairwise, triplet, and quadruplet-based methods for small videos [13].

The first protocol tested in this work is the ACR–HR methodology. It is a category judgment method where the test sequences are presented one at a time and rated independently on a category scale by a single observer. This methodology, well-known for its simplicity and efficiency, allows assessing many sequences in a session. This efficiency is balanced by precision, as ACR can require more observers than other methodologies. ITU standards recommend the use of at least 24 naive observers.

The second one explored is modified from the DCR method. DCR is an impairment assessment methodology using a discrete annotation scale. In a session, all the impaired sequences are compared with explicit references. Due to the presence of explicit references, the annotation speed is reduced, except when the explicit reference can be displayed side by side with the impaired sequences or on a second screen. An explicit reference reduces the observer's cognitive load, facilitates voting, and consequently increases the Mean Opinion Score (MOS) precision obtained from a fixed number of observers.

Two significant differences between ACR–HR and DCR exist. The first one is the type of scale. ACR uses a discrete quality scale, while DCR uses a discrete impairment scale. The labels are not the same and not necessarily in the same amount. The second distinction is the presence of the explicit reference for the DCR method. It modifies the observers' task toward a fidelity task, as they have to construct their judgment against an explicit reference and rate how annoying the impairment is. ACR–HR observers assess sequences' absolute quality.

The presentation of the work starts with section II explaining the DCR with EVP rating scale and ACR–HR methods,

TABLE I: 11–grade EVP rating scale

Scores	Impairment items	Levels
10	Imperceptible	
9	Slightly perceptible	Somewhere
8		Everywhere
7	Perceptible	Somewhere
6		Everywhere
5	Clearly perceptible	Somewhere
4		Everywhere
3	Annoying	Somewhere
2		Everywhere
1	Severely annoying	Somewhere
0		Everywhere

the test environment, and the sequences under test. Section III presents an analysis of Mean Opinion Scores and observers' usage of the rating scales in III-A. In III-B, we investigate sequences where the two methodologies agree or disagree on estimated MOS. Section III-C focuses on evaluating the MOS precision with an increasing number of participants and increasing experimental effort. Lastly, section IV summarizes our findings and concludes this work.

II. Subjective Tests Design overview

A. Video Test Sequences:

In this work, we consider three viewing scenarios:

1) Standard Dynamic Range (SDR) High Definition (HD) video sequences; 10 secs long, 1920x1080 in resolution at 60 fps, and in BT709 Y'CbCr 10–bit color format.
2) SDR Ultra High Definition (UHD) sequences; 10 seconds long, 3840x2160 in resolution, with one sequence at 60 fps, others at 30 fps, and in BT709 Y'CbCr 10–bit.
3) High Dynamic Range (HDR) UHD sequences; 10 seconds long, BT2020 Y'CbCr 10–bit color format, 60fps.

The sequences are encoded with the Random Access (RA) mode of modern video encoding implementations. The test sequences for the three scenarios are evaluated in separate viewing sessions. First, the SDR–HD scenario contains four sequences evaluated with 10 Hypothetical Reference Circuits (HRC). Hence, there are 44 Processed Video Sequences (PVS) to annotate in one viewing session. In the SDR–UHD scenario, five sequences are considered under 10 HRCs. For the DCR with EVP rating scale subjective test, the 55 PVS are split into two sessions, as it was too long to annotate them into one viewing session. This decision has the purpose of reducing observer fatigue. For the ACR–HR test, one viewing session could fit all the PVS. Finally, in the HDR–UHD scenario, five sequences are encoded, resulting in 55 PVS. Similarly to UHD SDR, the PVS are split into two viewing sessions for the DCR with EVP rating scale experiment, but not for the ACR–HR experiment.

B. The DCR with EVP rating scale experiment setup

1) DCR with EVP rating scale subjective methodology and rating scale: The test procedure is the Degradation Category Rating (DCR) method specified in ITU–T Rec. P.910 [2], modified with a repetition. Both the source and the coded sequence are shown twice, as follows:

$$| \, Source \, | \, Coded \; sequence \; A \, | \, Source \, | \, Coded \; sequence \; A \, |$$

TABLE II: 5–grade ACR–HR scale

Scores	Quality items
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

The transitions are 2–second pauses on a middle gray screen. The Expert Viewing Protocol (EVP) scale utilized in this test is described in Table I. It is an 11–grade scale specified in ITU–R BT.500-14 [1] "Annex 8 to Part 2", the annex on expert viewing protocol. The scale ranges from "0" (lowest quality) to "10" (highest quality). Observers evaluate the impairment by choosing between 6 impairment items + the location of this impairment: global or local. These two queries under a single question, and both video sequence repetitions, are assumed to help refine the subjective opinion one can have. For the rest of the paper, we will refer to this DCR with EVP rating scale as "DCR–EVP".

Moreover, a stabilization phase of three sequences is included at the beginning of the test: a no-impairment example, one with clearly perceptible impairment, and one with severely annoying impairment. This phase lets observers get used to the testing methodology scale and voting interface. All the collected stabilization scores are discarded during later analysis. The video source (SRC) used to generate the calibration sequences differs from the SRCs evaluated during the second part of the subjective tests.

Before the start of the test, we check and ensure every observer's vision. More specifically, we estimate their visual acuity with Snellen charts and their color perception with Ishihara plates. Observers who do not meet the requirements (normal or corrected-to-normal acuity + normal color vision) are rejected.

For some analyses conducted later in this work, DCR–EVP MOS are scaled to a 1–5 range. The mapping function is:

$$MOS^{j}_{1to5} = 4 \times \frac{MOS^{j} - minScale}{maxScale - minScale} + 1 \qquad (1)$$

where j is the index of a PVS, $minScale = 0$ and $maxScale = 10$. This scaling is applied notably to compare the Confidence Intervals (CIs) size across both subjective tests.

2) DCR–EVP Observers: We recruited naive observers from our panel and invited them to our lab for the experiment: they were compensated with gift cards. Ages of observers ranged from 19 to 62 years, with a good representation of nationalities and educational backgrounds. We collected over 24 observers' Opinion Scores (OS) for each PVS.

C. The ACR–HR experiment setup

1) ACR–HR subjective methodology and rating scale: We used the Absolute Category Rating with Hidden Reference (ACR–HR) method, specified in ITU–T Rec. P.910 [2]. It is a category judgment method where test sequences are presented individually, without repetition, and rated with a category scale. The test includes viewing and evaluating the reference sequences without an explicit signal to observers that they are rating it, thus the term "hidden reference". The rating of these

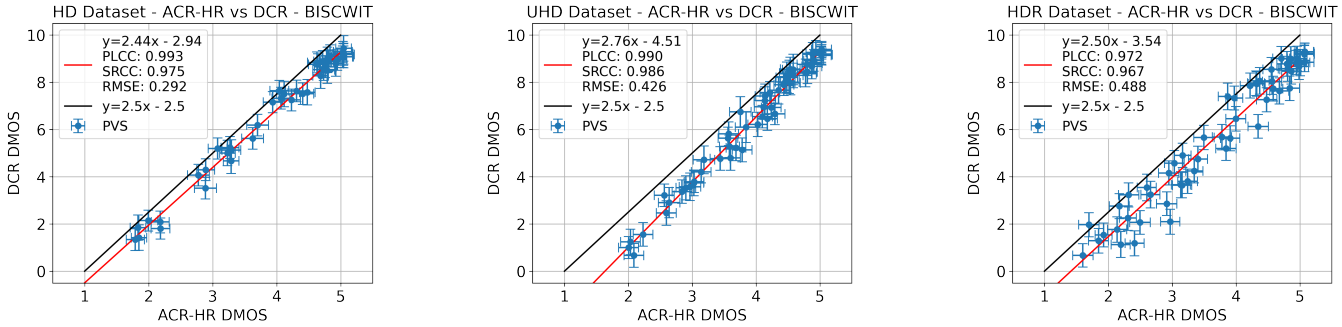

Fig. 1: Scatter plot of the BISCWIT MOS from DCR–EVP and ACR–HR subjective tests.

hidden reference conditions is identical to other sequences. The category scale is a 5–grade scale to rate the overall quality; see Table II. For the analysis, the Mean Opinion Scores (MOS) are converted to Differential Mean Opinion Scores (DMOS), as specified in ITU–T Rec. P.910 [2]:

$$DMOS_o^j = 5 - (MOS_o^{ref} - MOS_o^j) \tag{2}$$

$$DMOS^j = \frac{1}{N}\sum_{o=1}^{N} DMOS_o^j \tag{3}$$

where o and j are indexes of observers and PVS, respectively. N is the observers' total number for the PVS.

Similarly to the DCR–EVP experiment, a stabilization phase over three stimuli was included for high, middle, and low qualities. Likewise, the observer's vision was tested to ensure normal and corrected-to-normal vision.

Unlike the DCR–EVP experiment, SDR–UHD and HDR–UHD PVS are not split into two viewing sessions and combined into a single one. Annotation speed for ACR–HR is drastically increased by roughly a factor of 4, without the repetition and the presence of an explicit reference required by the DCR methodology.

2) ACR–HR Observers: We collected 90 observers' opinion scores for each PVS. Observers were pooled from our observer's panel while avoiding as much as possible inviting people who had already participated in the DCR–EVP experiment. Nevertheless, the DCR–EVP and ACR–HR experiments were three months apart to avoid any bias from the first study.

In this ACR–HR viewing configuration, observers were invited to the lab for 30-minute sessions and had the time to annotate all HD and UHD sequences, with a 3–minute break in between. Half of the observers annotated HD, then UHD contents, and the other half first UHD, then HD contents. For the HDR scenario, 45 observers were invited for 45–minute sessions. They were instructed to annotate all HDR PVS in a first viewing session A, followed by a 3–minute break, and then in a second viewing session B, observers annotated a second time all HDR PVS. PVS in sessions A and B were the same, but this information was not disclosed to observers.

D. Test Environment

1) For SDR–HD and SDR–UHD scenarios: The testing environments are two laboratory rooms, calm with controlled lighting conditions as stipulated in ITU–T Rec. BT.500 [1]. The observers are placed at 1.6 times the screen's height. In each experiment, we present the sequences at native resolution: HD videos are padded with neutral gray on our UHD displays. Hence, we achieved an effective viewing distance that was the recommended 3.2 times the displayed 1080p portion of the display height for the HD contents. All PVS were played from Y4M decoded streams. We calibrated the TVs with a color probe over 461 color references, and we tuned to D65 white at 120 cd/m².

2) For HDR–UHD scenario: The testing environment is one laboratory room, calm with controlled lighting as stipulated in ITU–R Rec. BT.2100 [14]: 5 cd/m² for the luminance of the surround. The observers are placed at 1.6 times the height of the screen. In each experiment, the sequences are presented at native resolution. All PVS are played from Y4M decoded streams. The TV is calibrated with a color probe and tuned to display D65 white at 950 cd/m². Color calibration is performed using Calman Home for Sony[1].

For the SDR–HD and UHD sessions, we use two 55" Sony TVs with LCD with LED backlight displays. We operate a 65" QD–OLED display for the HDR–UHD sessions with a measured peak luminance after calibration at 965 nits.

III. Results and Analysis

To analyze the results of the DCR–EVP and ACR–HR experiments, two MOS computed for each PVS: ***regular MOS*** (opinion scores average), and ***BISCWIT MOS*** calculated using MLE Content Oblivious Alternative Projection from SUREAL package[2] and detailed in ITU P.913-12.6.

A. Analysis on Opinion Scores and rating scale usage

In this section, we present an analysis of the Opinion Scores obtained during the two subjective experiments.

In figure 1, the scatter plot illustrates the relationship between the DCR–EVP MOS and ACR–HR DMOS scores. We fit a linear function (in red) and extract coefficients a and b, the slope, and the intercept. We analyze these coefficients to see how differently observers use the rating scales.

We performed the same analysis with the BISCWIT MOS aggregation technique presented above: see figure 1. The black line translates the "one-to-one" relationship.

For all the plots, the fitted line, in red, is below the black line. For the high-quality range, it indicates that the DCR–EVP small perceived impairments, scores around 8–9 out of

[1]Calman Home for Sony: https://store.portrait.com/consumer-software/calman-home-for-sony.html

[2]SUREAL: https://github.com/Netflix/sureal

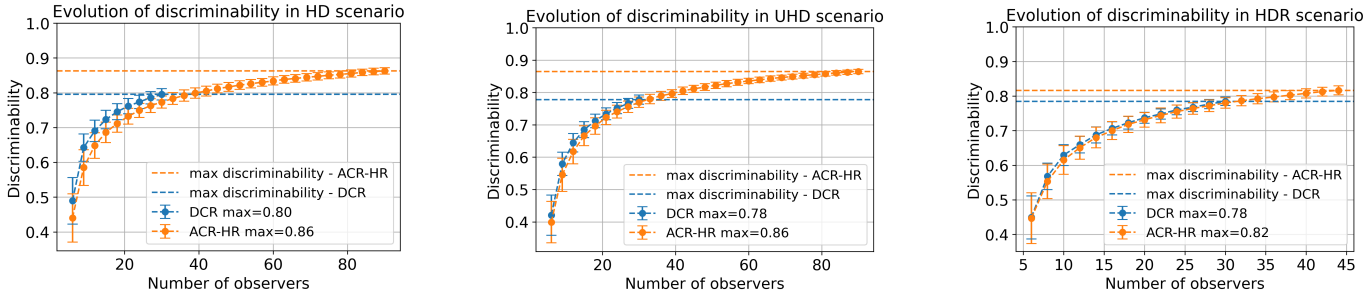

Fig. 2: Evolution of discriminability in SDR–HD, SDR–UHD, and HDR–UHD scenarios, when evaluated with ACR–HR or DCR–EVP subjective methodologies, and as a function of the number of observers.

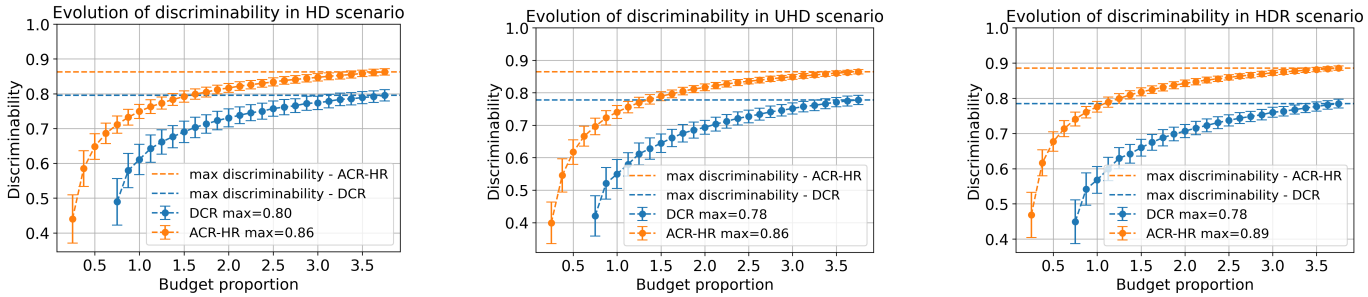

Fig. 3: Evolution of discriminability in SDR–HD, SDR–UHD, and HDR–UHD scenarios, when evaluated with ACR–HR or DCR–EVP subjective methodologies, and as a function of the budget proportion spent to collect the data.

TABLE III: Significance test on ACR–HR and DCR–EVP data.

Scenarios	Condition 1	Condition 2	Condition 3
SDR–HD	0 / 154	13 / 167	0
SDR–UHD	2 / 196	23 / 217	0
HDR–UHD	4 / 196	38 / 230	0

10, are mapped to scores of 4.5–5 on the ACR–HR DMOS scale. A DMOS of 5 for ACR–HR experiments implies the same perceived quality as the evaluated Hidden–Reference.

The slope of the red line is around 2.5, similar to the black line, implying that the annotation scale range in both subjective studies is relatively similar. The intercept of the red line suggests that the scores are, on average, relatively smaller in the DCR–EVP experiment than in the ACR–HR experiment, as we observe in the high-quality range of the scales. A larger slope, 2.76, in the UHD scenario, reveals that the observers used a larger scale range during the DCR–EVP experiment than in the ACR–HR experiment.

We can conclude from this analysis that naive observers are using both scale ranges to similar extents, with a slight benefit over UHD contents for the DCR–EVP scale.

B. Methodologies Agreement from an intra-content analysis

To show whether both methodologies agree on the significance of MOS difference for pairs of PVS and their ranking, we propose the following analysis.

In each scenario and for each SRC, we form all the possible MOS pairs and test which ones are significantly different according to the DCR–EVP test data. The significance test employs a T-Test with a p_{value} of 0.01. We then report if ACR–HR MOS for these "DCR significantly different pairs" are also significantly different or not. In the case of ACR–HR doesn't provide significant differences on these pairs, we could conclude that the DCR test provides a better testing method to discriminate these stimuli pairs. The results of this analysis are detailed in table III in column **Condition 1**. For example, we can read for the SDR–HD scenario that among the 154 significantly different pairs according to the DCR test, all are also significantly different from the ACR–HR data point of view: translating that all the pairs detected as significantly different by DCR test are also detected as significantly different by ACR–HR method. For SDR–UHD and HDR–UHD scenarios, 2 and 4 pairs, respectively, are not significantly different from ACR–HR data.

Conversely, we can analyze when the ACR–HR test yields significantly different pairs if they are as well significantly different from the DCR test perspective. Results are provided under **Condition 2** in table III. A first observation is that more pairs statistically differ from ACR–HR collected data. For example, in the SDR–HD scenario, 154 pairs are statistically different according to the DCR test and 167 according to the ACR–HR test. This implies that more pairs are detected as significantly different by ACR–HR than DCR: 13 more, for example, for the SDR–HD scenario.

The last analysis, on ranking analysis, in column **Condition 3** of table III, we check how many pairs (A, B) are rated as stimuli A **significantly better than** B by DCR test, and stimuli A **significantly lower than** B by ACR–HR. Here, we want to showcase if there is any disagreement between the methods. As the reader can see, none of the scenarios exhibit such pairs.

C. Discriminability evolution of MOS

As suggested in [15], [16], we can investigate the discriminability evolution of MOS with an increasing number of observers. A two-sample Wilcoxon test is applied on all the possible pairs of MOS, and a p_{value} of 0.05 is employed to

compute the percentage of pairs significantly different. The statistical test is applied between two estimated MOS. An estimated MOS is computed from K out of N observers randomly selected. The number of possible pairs for HD, UHD, and HDR experiments is 946 pairs, 1485 pairs, and 1485 pairs, respectively. Results are in figure 2 with a growing value for K and 63% Confidence Intervals (CIs) bootstrapped over 1,000 simulations.

For the HD scenario, DCR achieves slightly greater discriminability than ACR–HR for a fixed number of votes per stimuli. However, for UHD and HDR scenarios, the difference is reduced to zero, as the two curves overlap in their uncertainty estimates. This result partly invalidates the argument that at the same number of observers, DCR provides more accurate votes than ACR.

In figure 3, We present a similar analysis based on increasing budget proportion. Here, budget proportion $B_{prop}^{k,method}$ refers to the ratio between the cost C_k^{method} to recruit K participants to perform DCR or ACR–HR viewing sessions, and the constant cost C_{24}^{ACR-HR} to recruit 24 observers for ACR–HR sessions. We selected 24 since it is recommended by ITU standards.

$$B_{prop}^{k,method} = \frac{C_k^{method}}{C_{24}^{ACR-HR}} \tag{4}$$

As an example, a budget ratio of 2 corresponds to recruiting, from our panel, 48 observers for an ACR–HR viewing sessions or 16 observers for a DCR viewing session, as DCR test observers are compensated more since the protocol displays twice the reference and the distorted sequences.

The general trend we observe in figure 3 is that spending more budget improves the discriminability between stimuli inside the dataset. For HD, UHD, and HDR datasets, the highest discriminability is achieved using ACR–HR subjective methodology. Moreover, we can see that one can achieve the maximum discriminability of the DCR experiment with less than half the budget, by using the ACR–HR experiment. Additionally, the increasing rate of discriminability slows down drastically on each curve after 60–80% of the budget is spent. This effect means that recruiting more observers in these studies will provide marginal gains in MOS accuracy and precision, which aligns with ITU standards recommendations.

IV. Conclusion

This work compares two subjective studies conducted on SDR–HD, SDR–UHD, and HDR–UHD contents. The first subjective study is with the DCR 11–grade rating scale proposed in ITU–R BT.500-14 with a repetition of both the explicit reference and distorted stimuli. In the second subjective study, we use the ACR–HR methodology with a 5–grade quality scale from ITU–T Rec. P.910 and without repetition. With extensive analysis of the scale usage by naive observers and the comparison of MOS scores discriminability in both subjective studies, we show that both methods can retrieve accurate estimates, and at a fixed number of observers, DCR–EVP achieves similar or slightly better discriminability than ACR–HR. However, the results obtained from ACR–HR tests are superior when compared at fixed experimental efforts on all the figures of merits. Moreover, this analysis gives new insight into how to perform experiments for modern codec comparison and validating dataset quality through MOS discriminability computation.

References

[1] ITU Rec. BT.500-15, "Methodologies for the Subjective Assessment of the Quality of Television Images," 2023.

[2] ITU-T Rec. P.910, "Subjective video quality assessment methods for multimedia applications," 2022.

[3] ITU-T Rec. P.913, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment," 2021.

[4] Stéphane Péchard, Romuald Pépion, and Patrick Le Callet, "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm," in *International Workshop on Image Media Quality and its Applications, IMQA2008*, 2008, p. 6.

[5] Toshiko Tominaga, Takanori Hayashi, Jun Okamoto, and Akira Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video," in *2010 Second international workshop on quality of multimedia experience (QoMEX)*. IEEE, 2010, pp. 82–87.

[6] Taichi Kawano, Kazuhisa Yamagishi, and Takanori Hayashi, "Performance comparison of subjective assessment methods for 3d video quality," in *2012 Fourth International Workshop on Quality of Multimedia Experience*, 2012, pp. 218–223.

[7] Ashutosh Singla, Werner Robitza, and Alexander Raake, "Comparison of subjective quality test methods for omnidirectional video quality evaluation," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 2019, pp. 1–6.

[8] Majed Elwardy, Yan Hu, Hans-Jürgen Zepernick, Thi My Chinh Chu, and Veronica Sundstedt, "Comparison of acr methods for 360° video quality assessment subject to participants' experience with immersive media," in *2020 14th International Conference on Signal Processing and Communication Systems (ICSPCS)*. IEEE, 2020, pp. 1–10.

[9] Ashutosh Singla, Stephan Fremerey, Werner Robitza, Pierre Lebreton, and Alexander Raake, "Comparison of subjective quality evaluation for hevc encoded omnidirectional videos at different bit-rates for uhd and fhd resolution," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 2017, pp. 511–519.

[10] Jesus Gutierrez, Pablo Perez, Marta Orduna, Ashutosh Singla, Carlos Cortes, Pramit Mazumdar, Irene Viola, Kjell Brunnström, Federica Battisti, Natalia Cieplińska, et al., "Subjective evaluation of visual quality and simulator sickness of short 360 videos: Itu-t rec. p. 919," *IEEE transactions on multimedia*, vol. 24, pp. 3087–3100, 2021.

[11] Yana Nehmé, Jean-Philippe Farrugia, Florent Dupont, Patrick Le Callet, and Guillaume Lavoué, "Comparison of subjective methods for quality assessment of 3d graphics in virtual reality," *ACM Transactions on Applied Perception (TAP)*, vol. 18, no. 1, pp. 1–23, 2020.

[12] Andreas Pastor, Pierre Lebreton, Toinon Vigier, and Patrick Le Callet, "Comparison of conditions for omnidirectional video with spatial audio in terms of subjective quality and impacts on objective metrics resolving power," working paper or preprint, Oct. 2023.

[13] Andréas Pastor and Patrick Le Callet, "Perceptual annotation of local distortions in videos: tools and datasets," in *Proceedings of the 14th Conference on ACM Multimedia Systems*, 2023, pp. 458–462.

[14] ITU-R Rec. BT.2100, "Image parameter values for high dynamic range television for use in production and international programme exchange," 2018.

[15] Randy F Fela, Andréas Pastor, Patrick Le Callet, Nick Zacharov, Toinon Vigier, and Soren Forchhammer, "Perceptual evaluation on audio-visual dataset of 360 content," in *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2022, pp. 1–6.

[16] Andréas Pastor and Patrick Le Callet, "Towards guidelines for subjective haptic quality assessment: A case study on quality assessment of compressed haptic signals," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2023, pp. 1667–1672.