https://doi.org/10.2352/EI.2023.35.8.IQSP-263
© 2023, Society for Imaging Science and Technology

Subjective video quality for 4K HDR-WCG content using a browser-based approach for "at-home" testing

Lukáš Krasula; Netflix; Los Gatos, USA
Anustup Choudhury; Dolby Laboratories Inc.; Sunnyvale, USA
Scott Daly; Dolby Laboratories Inc.; Sunnyvale, USA
Zhi Li; Netflix; Los Gatos, USA
Robin Atkins; Dolby Laboratories Inc.; Sunnyvale, USA
Ludovic Malfait; Dolby Europe Ltd.; London, UK
Aditya Mavlankar; Netflix; Los Gatos, USA

Abstract

The paper describes a design of a subjective experiment for testing the video quality of High Dynamic Range, Wide Color gamut (HDR-WCG) content at 4K resolution. Due to Covid, testing could not use a lab, so an at-home test procedure was developed. To aim for calibration despite not fully controlling the conditions and settings, we limited subjects to those who had a specific TV model, which we had previously calibrated in our labs. Moreover, we performed the experiment in the Dolby Vision mode (where the various enhancements of the TV are turned OFF by default). A browser approach was used which took control of the TV, and ensure the content was viewed at the native resolution of the TV (e.g., dot-on-dot mode). In addition, we know that video imagery is not ergodic, and there is wide variability in types of low levels features (sharpness, noise, motion, color volume, etc.) that affect both TV and visual system performance. So, a large number of test clips was used (30) and the content was specifically chosen to stress key features. The obtained data is qualitatively similar to an in-lab study and is subsequently used to evaluate several existing objective quality metrics.

Introduction

There are numerous video quality metrics of useful performance (e.g., correlations of 0.88) but these are for SDR (Standard Dynamic Range) and mostly SDTV (Standard Definition TV) resolution. These are constantly being improved and tuned for different applications or goals. But today's consumer video ecosystem is now HDR-WCG (High Dynamic Range - Wide Color Gamut), while still co-existing with the older SDR and SDTV ecosystem. The newer ecosystem is also at the higher resolutions of HDTV (1920x1080) or higher (UHDTV, a.k.a. 4k = 3840x2160). The dynamic range increase in the newer system is several orders of magnitude, going from SDR to HDR, and the color gamut increase by approximately 2X from the SDTV/SDR gamut of Rec. 709 to that of the newer ITU-R spec of BT. 2100 [1]. Currently, most content for the BT. 2100 format only fills a P3 color gamut, which is less than the full specification. However, for HDR color aspects, it is known that color volume is a more important descriptor than color gamut, and that color volume improvement ratios exceed color gamut ratios. There are several datasets and models for predicting HDR-WCG still image quality, but none yet for HDR-WCG video quality. So, there is a need for both data and quality models for video HDR-WCG at the 4K resolution, to which this work is addressed.

The objective is to design a subjective experiment for testing the video quality of HDR WCG content at 4K resolution. We had an experiment designed to be played on a custom 1600 cd/m^2 65" OLED TV and bypassing its internal video processing. However, due to the Covid pandemic, testing could not bring subjects into a lab, so an at-home test procedure was developed. To aim for calibration despite not being in the lab, we limited subjects to those who had a specific TV model, which we had previously calibrated in our labs. In addition, we know that video imagery is not ergodic, and there is wide variability in types of low-level features (sharpness, noise, motion, color volume, etc.) that affect both TV and visual system performance. So, a large number of test clips was used (30) and the content was specifically chosen to stress key features, often referred to as corner cases. The key compression parameters of resolution and DCT Quantization Parameter (QP) were varied and their effects on quality were analyzed.

A browser-based at home study was developed to collect observers' opinions following ITU-T P.913 DCR methodology. The reference video and the distorted video were shown iteratively three times and then the user was queried to make a choice on how the distorted video looked in comparison to the original video. The ratings were done on a standard 5-point impairment scale. 4K resolution HDR-WCG content with 10 bits was displayed on a 4K HDR OLED TV capable of P3 color gamut, 10 bits RGB, 700 cd/m^2 maximum luminance, and < 0.005 cd/m^2 black level in the Dolby Vision format.

Key Problems with Crowdsourcing Image Quality

While there has been much work in characterizing image quality via crowdsourcing, initially starting with hard copy [2] and expanding to displays [3], there are key problems that can significantly affect the accuracy and utility of the results. The particulars of the business aiming to use such results will determine their utility. The early work with hard copy image quality was promising as there was limited variability in the way the test images were displayed. For example, the printed images were all the same size, resolution, color gamut, the illumination was at office levels or higher, the dynamic range of prints is relatively small ($<$ 2log10), and the viewing distances tended to be at arms' length. Once this concept moved to displays, further work needed to be done to regularize the displayed resolution and viewing distance, such as scaling the image to match a credit card [3], as well as using visual characterization to assess the subjects' displays' gamma [3, 4].

Display and TV Capability Differences

However, in the last 10 years there has been significant display capability increases, while the older capability still exists at the lower price points. As a result, there is an extremely wide range of capability, and hence visibility differences along multiple dimensions. Critical aspects include screen size, pixel resolution, dynamic range, color gamut, temporal response, AR coating, noise level, bit-depth and dithering, display processing, and viewing angle performance.

Screen size tends to affect field of view (FOV) which affects the mapping of a display's Nyquist to the visual cy/deg, which in turn affects visibility of high frequency aspects of encoding via the spatial CSF (Contrast Sensitivity Function) and viewing distance. The display's pixel resolution affects the visibility of compression distortions, especially if the display is required to do down-scaling, or up-scaling when edge sharpening algorithms are included. The dynamic range includes factors such as the overall range, but also the specifics of black level and maximum luminance. Current displays range from as low as 200:1 (2.3log10) for SDR displays to over 10,000:1 (4log10) for HDR displays. These factors affect the displayed contrast of signals and distortions, and thus their visibility. Color gamut can range from less than Rec. 709 to over P3. The smaller color gamuts can either reduce the color contrast, and thus affect visibility of chromatic sub-sampling and other spatio-chromatic distortions, or simply clip the wider gamut colors, completely removing visibility of distortions in that part of the color space. Temporal response can affect visibility of distortions by blurring motion, or cause judder distortion which may cause masking of motion artifacts. The AR (anti-reflection) coating affects reflectivity, which determines how much the ambient light elevates the black level or creates hot spots of glare on the screen. While most digital displays now can achieve extremely low noise levels, their video processing chips may introduce noise, or have noise reduction which can inadvertently affect visibility of low amplitude signal distortions. In addition, displays have their own processing chips which affect bit-depth and may use dithering. While excellent dithering algorithms can increase native bit-depths by up to 2 bits without any visibility of the dithering, there are lower quality dithering techniques, as well as low bit-depth line drivers. Further display processing may include sharpening algorithms, dynamic contrast, vivid modes, and other alterations of the intended image. Some display types have viewing angle effects, where perpendicular viewing gives the best performance, but the contrast and color saturation can reduce as the viewing angle moves away from perpendicular. In all these cases of display capability differences affecting the displayed signal, it can act in two ways. One way is to affect the visibility of high spatial frequencies, high temporal frequencies and low amplitude details of distortions, all of which are key factors in video compression. In addition, these factors obviously affect the best inherent quality a display can achieve. While compression quality assessment tries to be agnostic to the display maximum possible quality via full-reference methodology and use of degradation scales, it is unknown how the display's maximum capability may affect such a rating scale.

So, while in the crowdsourcing image quality databases of the past, [9, 10] most displays were SDR and CRT (Cathode Ray Tube) and the resolutions and dynamic ranges, contrast, viewing angle issues were essentially uniform, that situation does not apply to today's consumer ecosystem, as well as the more varied business questions around quality that exist today [5].

Ambient Illumination Differences

Due to the physics of the display screen surface, the illumination hitting the display can have strong effects on the displayed tone scale, which can significantly raise the black, lower the contrast/code value, and overall contrast. Further the luminance of the surround has strong effects on visibility within the screen [6]. These effects are shown in Figure 1.

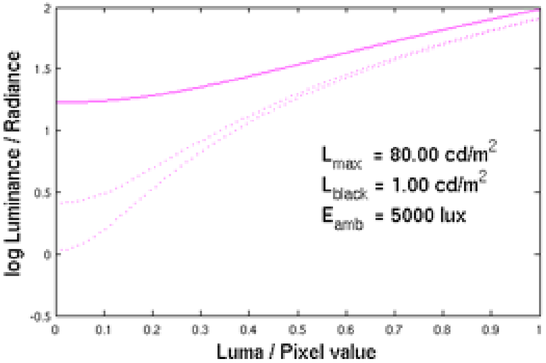

Figure 1. Effects of ambient light on a 100 cd/m^2 SDR display; solid curve shows the luminance displayed with an illumination of 5000 lux (is overcast daylight, or interior room with daylight through window) . The bottom curve shows the display response in the dark. The horizontal axis is normalized code value range (0-1024 for 10-bits)

The lower dashed curve shows the displays' performance (luminance vs code value signal level) at a very low ambient illumination level, and the middle curve shows how the displays range is reduced for an office lighting level. The upper solid curve shows the display performance in the ambient level corresponding to overcast daylight or daylight coming through a window of an indoor room. Not only is the dynamic range reduced (the delta from the minimum to the maximum of the curve) which affects overall quality, it is important to look at the slope at a given code value. This slope is the contrast/code value, which affects the contrast of small signals, which describes most compression distortions for the high to mid quality ranges. This figure is for an emissive display. Emissive displays, which is nearly all displays being used today (excepting a small percentage of reflective displays, e.g., electro-chromic, electro-phoretic, and hard copy), have their best performance in the dark. Ambient illumination reduces their performance, such that the display contrast reduces as the illumination is increased. The type of AR (anti-reflective) coating and geometry of the room lighting has effects as well. So, the unknown ambient lighting of most crowdsourcing experiments causes significant increase on the variance of MOS scores.

Viewer Differences

Viewer differences should also be accounted for, and there are three key factors. These are primarily the viewing distance, the viewer's acuity, and the viewer's engagement. In lab environments, it is common to have a fixed viewing distance, typically such that the display and image content Nyquist frequency results in 30 cy/deg on the retina to match the TV signal design specifications. That retinal frequency is the maximum that can be seen by an average viewer for light adaptation levels below around 600 cd/m^2. In addition to considering such maximum cut-off frequency, it is important to consider the entire spatial CSF (frequency response). The viewing distance affects the mapping of the digital frequencies displayed to the retinal frequencies. The CSF is band-pass, so that it has a maximum visibility near 4 cy/deg and decreases for both higher and lower frequencies. As the viewing distance gets closer, the display frequencies shift to lower retinal frequencies. This means that high frequency distortions get easier to see as the viewing distance gets closer. However, for very low frequency distortions, which typically result from tone scale and color display mapping algorithms, these become more difficult to see as the viewing distance decreases because their frequencies shift towards the very low spatial frequencies which are less sensitive due to the band-pass CSF. The CSF affects the visibility of content, image intrinsic noise and its masking effects, as well as the compression distortions. The second factor, viewer acuity, describes how high of a frequency one can see. It is typically described with the 20/20 terminology (or 6/6 in the EU). Most TV signal processing and formats are designed for a 20/20 viewer [7]. However, such a viewer is the average (including correction for glasses) which means there are those who can see higher spatial frequencies [8]. There is a strong effect of age on acuity, such that those under the age of 25 are often those with acuity better than 20/20.

The last key factor of viewer differences is a variability that can occur within a viewer/subject, and the umbrella term engagement is used. This describes how well the viewer is paying attention to the imagery/stimuli of the test, and how much they are trying in the task. A distracted or multi-tasking viewer is less able to fully pay attention to all parts of the image, and in the case of video may even miss key portions (groups of frames) of the video stimulus. There is also the type of viewer, particularly for paid crowdsourcing, that does not bother to actually do the task, and randomly makes responses. Analysis of one study found many observers whose responses to a paired comparison test simply alternated their responses. Of course, this can occur in lab experiments as well, but tends to be more common in crowdsourcing. Modern studies have developed techniques to weed out such lazy viewers, but this tends to reduce the data that can be used. Other approaches [12] try to weigh observers' contributions based on their reliability.

Browser-Based Approach to “At-Home” Testing

Conducting subjective tests outside of laboratories has shown increased interests since the Covid-19 pandemic as many labs were not accessible. There is a wide range of possibilities when it comes to testing outside of the lab, from conducting the test in a different, but rather controlled environment, to running an experiment on micro-tasks platforms over the internet [21],[22] . Conducting tests in subjective tests labs provides the highest degree of control in terms of environment, equipment, participants and administration. Deviating from this ideal environment adds uncertainty that experimenters should account for when planning their experiments.

Crowdsourcing often refers to test conducted on micro-tasks platforms such as Amazon Mechanical Turk (MTurk), involving anonymized subjects who typically participate in a subset of an experiment. This approach provides high scalability and fast execution at the expense of uncontrolled participants, equipment and environment. In between crowdsourcing and lab testing is what the Authors describe as ”remote testing”, where known participants attend a full-length experiment, as if they were sitting in the lab. Participants can typically be recruited through the usual methods (job agencies, colleagues, email lists, etc...). This approach provides a greater level of trust in the participants in terms of attention to the material being presented, but also in setting up the equipment and environment.

The authors chose the remote testing approach for this study. An advanced cloud-based subjective test platform was developed to administer subjective tests within a web browser. It is important to note that while web-browsers are very convenient endpoints for deploying applications, the lack of direct access to the device hardware requires special attention to ensure the correct presentation of the test material, especially in terms of rescaling. Our platform was developed with these aspects in mind and incorporate modules for detecting the native resolution of the display, managing full screen modes and video presentation.

Workflow for testing on smart TVs

The presentation devices for this study were LG OLED TVs. A custom-made WebOS TV application was developed to enhance the capabilities of our subjective test platform, enabling access to TV hardware information and Dolby Vision rendering. The subjective test platform was crafted to enable a simple process for the participants and the workflow is depicted in Figure 2.

Participants received an email providing instructions on how to install the WebOS application on their TV, and a unique URL that gave direct access to the subjective test. A click on this URL prompted the participants to pair their TV with the subjective test platform by typing a code displayed on the WebOS application. Once paired, the subjective test started automatically on the TV with a set of instructions for TV settings, room conditions and test procedure, followed by a demographics survey. Then the test started with a set of training material to get participants acquainted with the task, the rating scale and the range of distortions. Participants navigated the test and entered their responses using the TV remote control.

Test Setup and methodology

The test procedure followed typical procedure for in-lab testing, with additional steps for instructing participants to configure their TV and environment. A summary of the test properties is available in Table 2.

The device requirement was an LG 4K OLED TV B8/C8/C9/CX and the set up instructions were specifically written for these TV models. Participants were requested to set up their TV in *Dolby Vision Cinema* mode and to disable AI enhancements. This configuration ensures appropriate screen calibration and minimal post-processing.

Participants were further asked to set themselves into a dark

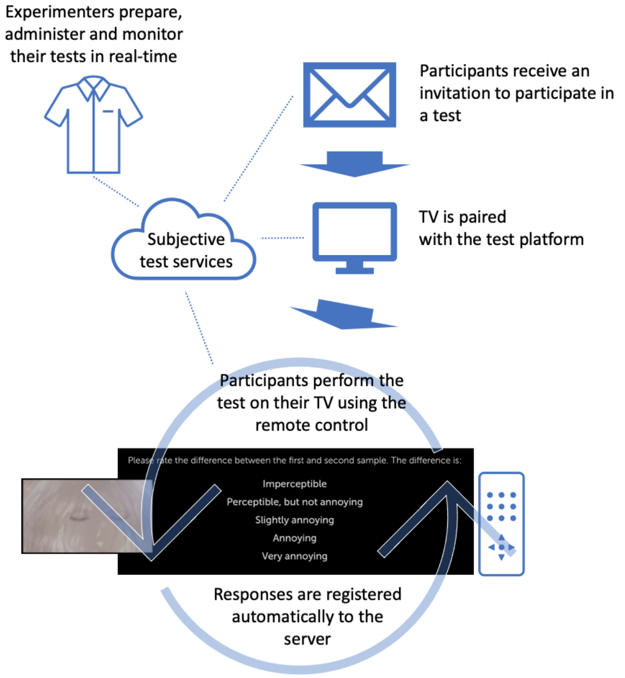
Experimenters prepare, administer and monitor their tests in real-time
Participants receive an invitation to participate in a test
Subjective test services
TV is paired with the test platform
Participants perform the test on their TV using the remote control
Please rate the difference between the first and second sample. The difference is:
Imperceptible
Perceptible, but not annoying
Slightly annoying
Annoying
Very annoying
Responses are registered automatically to the server

Figure 2. *Simplified workflow for conducting subjective tests on smart TVs.*

home theater environment, and to sit at 1.6 times the height the screen.

The test methodology was the ITU-T P.913 Degradation Category Rating (DCR) procedure, also known as Double Stimulus Impairment Scale (DSIS), in which the reference and the test stimuli are presented in pairs, sequentially, with the reference always first [11]. In this test, the paired stimuli were presented up to three times unless the participant interrupted the repetition by pressing a key on their remote. After the presentation of the stimuli, participants were requested to *rate the difference between the first and the second sample* on the scale following 5-point discrete scale, mapped from 5 (Imperceptible) to 1 (Very annoying):

Table 2: Degradation Category Rating scale (DCR)

Imperceptible	5
Perceptible, but not annoying	4
Slightly annoying	3
Annoying	2
Very annoying	1

Test Sequences

We used 30 different clips, each spanning approximately 5 seconds. These clips were content from popular shows on Netflix and some trailer/movie content from Dolby. The clips were chosen such that it covers a large variety of dynamic range, color gamut, sharpness and motion complexity. Along with sequences that contain natural imagery, our test set also contained animation sequences. We specifically included challenging cases for the display behavior as well as for human vision. For example we included scenes that had complex motion effects on eye tracking such as (say) scenes having snow falling along with some camera motion. Several of these clips represented highest possible quality with the state-of-the-art technology.

Table 1: Test properties

Devices	LG 4K OLED TV B8/C8/C9/CX in Dolby Vision Cinema mode
Environment	At home remote testing, dark home theater
Viewing distance	1.6 times the height of the screen
Methodology	ITU-T Rec. P.913 DCR (DSIS) Up to three presentations per trial
Software	Proprietary web platform (DSVLab)
Sources	30 5-second clips, with a variety of dynamic range, color gamut, sharpness and motion complexity
Encoding	HEVC Dolby Vision Profile 5
Test conditions	Resolutions: 2160p, 1080p, 720p and 540p Quantization Parameter (QP)s: 18, 22, 26, 30 and 34
Test sequences	250 sequences 2160p clips with QP18 used as reference 540p clips only with QP22
Participants	25 Netflix and Dolby employees and family
Test sessions	3 separate sessions of 30 minutes 3 blocks per session

The videos were also encoded using Dolby Vision Profile 5. The test sequences contained content with various different resolutions such as 2160p (3840 x 2160), 1080p (1920 x 1080), 720p (1280 x 720) and 540p (960 x 540). While encoding the content, we selected various different Quantization Parameters (QP) such as QP18, QP22, QP26, QP30 and QP34. Please note that the lowest QP value (QP18) in our test sequence has the highest quality. Likewise, the highest QP value (QP34) has the lowest quality.

Including all possible combinations of QPs and resolutions would have made our testing process quite time-consuming. We therefore sampled the data such that we had 250 different sequences. One such sampling was that we only had 540p clips with QP22. While the content was being displayed during testing, we used the 2160p content at QP18 as our ‘Reference’.

Due to copyright issues, we can’t demonstrate sample frames from the sequences but the titles of the sequences can be seen in Figure 6.

Subjective Results

We managed to recruit 25 participants able to create viewing conditions suitable for the test in their homes. Each participant was asked to complete 3 separate sessions of 20-30 minutes each with a break of at least a few hours in between the sittings. The majority of the subjects managed to finish all of the sessions leading to approximately 20 individual votes per test sequence.

The overall result for this experiment, across the 30 sources, is depicted in Figure 3. The figure shows the average of all scores collected for a given combination of resolution and QP, along with

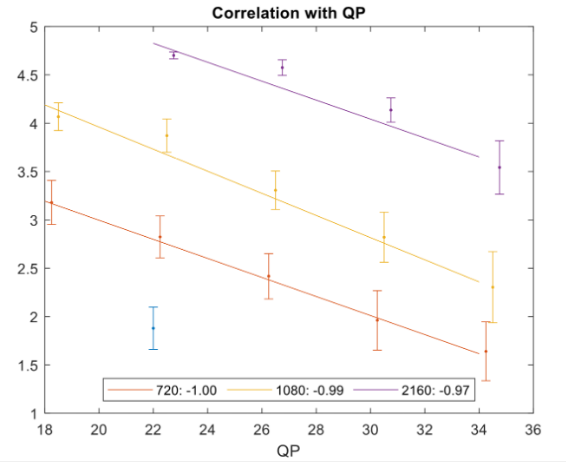

Figure 3. *Average Subjective scores per QP and resolution*

their 95% confidence interval. The results agree with general expectations, with the subjective scores decreasing as QP increases or resolution decreases.

Within each resolution, the mean opinion scores highly correlate with QP, with a strong linear relationship for 720p, 1080p and 2160p. Note that a saturation of the rating scale can be observed for 2160p at QP 22.

Within each QP values, mean opinion scores decrease significantly as the resolution decreases. Note that the results indicates that lower resolution at low QP may be preferred over higher resolution at high QP. For example, 1080p QP18 was rated higher that 2160p QP34. The same applies for 720p QP18 and 1080p QP34.

Subjective Results post-processing

To recover the overall quality scores, we employed the method from section 12.6 of the ITU-T Recommendation P.913 [11], originally described in [12]. This technique uses Maximum Likelihood Estimation (MLE) to explain the individual scores by modeling each subject's bias and inconsistency. These are then utilized to extract the most relevant information about the overall quality from each participant.

The recovered quality scores for each sequence are shown in Figure 4. The scores are sorted to showcase well-balanced coverage of the quality scale. This makes the data well-suited for training and/or testing of objective quality metrics.

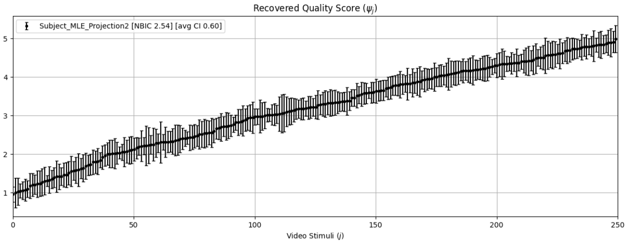

Figure 4. *Recovered quality scores with 95% confidence intervals.*

Bias and Inconsistency Analysis

To better understand the observers' behavior, we look deeper into bias and inconsistency obtained by the above-described subjective score recovery technique.

Bias is a global shift in an observer's scoring compared to the others. A subject with a high positive bias is "more forgiving", i.e. giving even the very low-quality sequences higher scores, typically someone less sensitive to quality drops. A negative bias, on the other hand, means the observer is highly sensitive and their overall scores are lower. Such behavior is typical for experts or people used to watching the highest quality content on premium devices. Even though bias provides an interesting view into each subject's scoring process, it usually does not say anything about their reliability or attentiveness unless also paired with high inconsistency.

Inconsistency shows how the participant's votes agree with the rest of the voters. High levels of inconsistency are a sign of unusual behavior, very often connected to less careful observations and/or misunderstanding of the task. It has been shown [12] that weighting each subject's contribution towards the overall quality scores by their consistency is beneficial over subjects rejection mechanisms. Each subject's bias and inconsistency with the corresponding 95% confidence intervals are depicted in Figure 5.

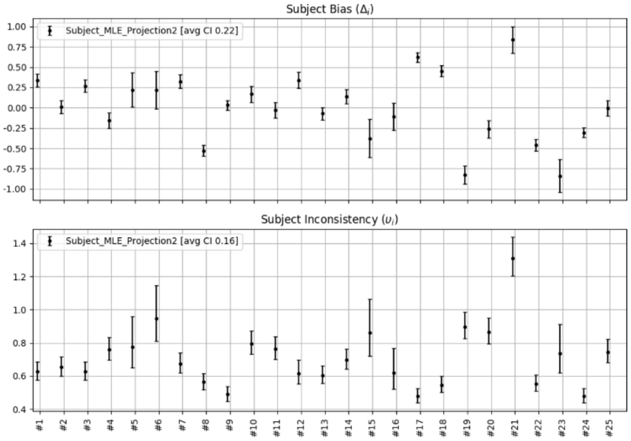

Figure 5. *Each subject's bias and inconsistency with 95% confidence intervals.*

The most interesting observation is that the subject inconsistency is on the levels typical for controlled in-lab studies rather than remote (crowdsourcing) tests. This suggests that our approach is valid and high-quality subjective data can also be obtained outside a lab if certain precautions are taken care of. The only exception is subject #21 who seemed to disagree with the others more than is typical and, combined with the strong positive bias, we conclude that there was either something wrong with the observer's setup or they did not follow the test carefully. Note that larger confidence intervals for certain subjects are caused by the completion of only a subset of the test sessions and do not indicate non-standard behavior.

SOS Analysis

We further conducted a Standard Deviation of Scores (SOS) analysis described in [13]. This method allows comparison to other studies based on analyzing standard deviations of scores for

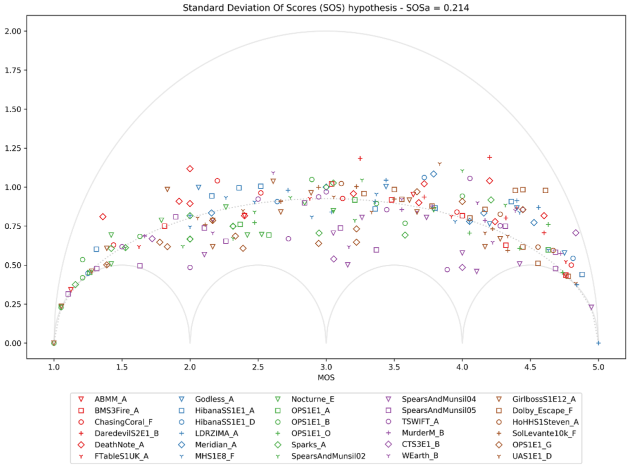

Figure 6. *SOS analysis plot.*

each sequence. SOS hypothesis assumes a square relationship between standard deviations and means which is parameterized by a parameter a. The value of a obtained from the raw scores is then compared to typical values for a given application.

Our experiment resulted in $a = 0.214$ which is similar to other in-lab tests in video compression and streaming. This further validates the quality of the obtained data. The plot of standard deviations against their respective means (also known as Mean Opinion Scores – MOS) is depicted in Figure 6.

The plot also serves as an indicator of the content difficulty as the standard deviation is directly linked to the inter-observer agreement. Larger values indicate cases where observers' opinions differed more. We can see that a few contents stand out. For example "DeathNote_A" is a dark scene that contains a scene cut which could lead to different observers focusing on a different portion of the content. "DareDevilS2E1_B", on the other hand, included a camera panning leading to slight flickering on the OLED screen which some observers could consider to be an encoding artifact.

Objective Metrics Performance

In order to benchmark existing objective image and video quality metrics on the above-described dataset, we first establish a baseline performance achievable by a combination of known encoding parameters, namely the used quantization parameter (QP) and encoding resolution. When we average the subjective scores (MOS) across all sequences of the specific combination of QP and resolution, we can find a good fit by optimizing parameters a_1, a_2, and a_3 in the following equation:

$$MOS = a_1 \times Q1 + a_2 \times \log_2(resolution) + a_3. \qquad (1)$$

Setting $a_1 = -0.1027$, $a_2 = 1.3464$, and $a3 = -7.7294$ leads to a very good correlation of 0.99. The resulting 2D plane can be found in Figure 7.

If we try to use this combination to estimate the quality of the individual clips, we reach the Pearson Linear Correlation Coefficient (PLCC) of 0.83. In order for a quality metric to be useful for 4K HDR WCG content, it should be able to reach significantly higher PLCC on our dataset.

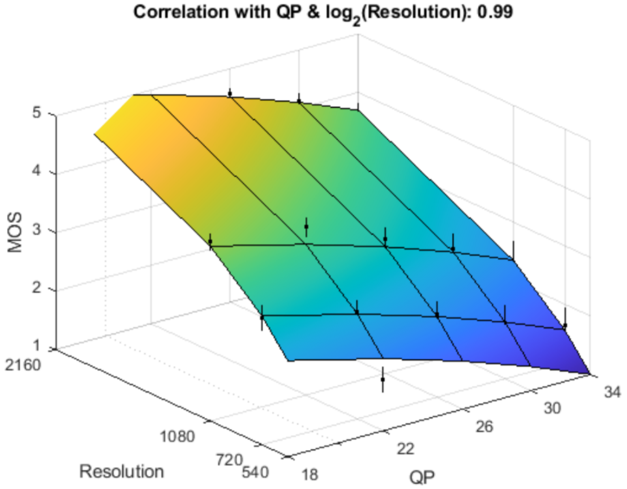

Figure 7. *Estimating MOS (averaged across all clips) by combining QP and encoding resolution.*

We selected 10 popular, publicly available image and video quality metrics for benchmarking – Peak-Signal-To-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) and Multi-Scale Structural Similarity Index Measure (MS-SSIM) [14], Visual Information Fidelity (VIFp) [15], Additive Distortion Metric (ADM) [16], Video Multi-method Assessment Fusion (VMAF) [17], HDR Video Quality Measure (HDR-VQM) [18], HDR Visual Difference Predictor (HDR-VDP 2.2) [19], HDR-VDP 3 [20], and its extension into temporal domain denoted as HDR-VDP 3 Flicker. All metrics were computed directly on the I channel of the videos that were natively in ICtCp color space with the exception of HDR-VDP versions which require the data to be linearized with electro-optical transfer function (EOTF) – in this case inverse PQ.

The performances of the tested metrics, expressed in terms of PLCC, are plotted in Figure 8. The red dashed line indicates the baseline correlation obtained from the combination of QP & $\log_2$(Resolution) described above. The plot indicates that VMAF is the only tested metric able to estimate quality significantly better than the baseline predictions ($PLCC = 0.89$). ADM and the two versions of HDR-VDP 3 are on par with the baseline while the rest of the metrics only achieve much lower correlations. While the state-of-the-art techniques such as HDR-VDP achieve similar performance to the baseline model, please note that the baseline model has been over-fit to our data. It remains to be seen how well the performance can extrapolate to sequences outside of this dataset. On the other hand, HDR-VDP has been known to have good performance across a wide variety of datasets. This clearly demonstrates the challenge that 4K HDR WCG content brings to the objective quality assessment with even the best performing metric still leaving a lot of room for improvement.

Conclusion

We developed a one-of-its-kind dataset of 4K HDR WCG videos encoded with Dolby Vision Profile5 and described an "at home" experiment conducted to annotate the database with subjective opinion scores. We further demonstrated that it is possible to obtain the quality of data in an "at-home" environment

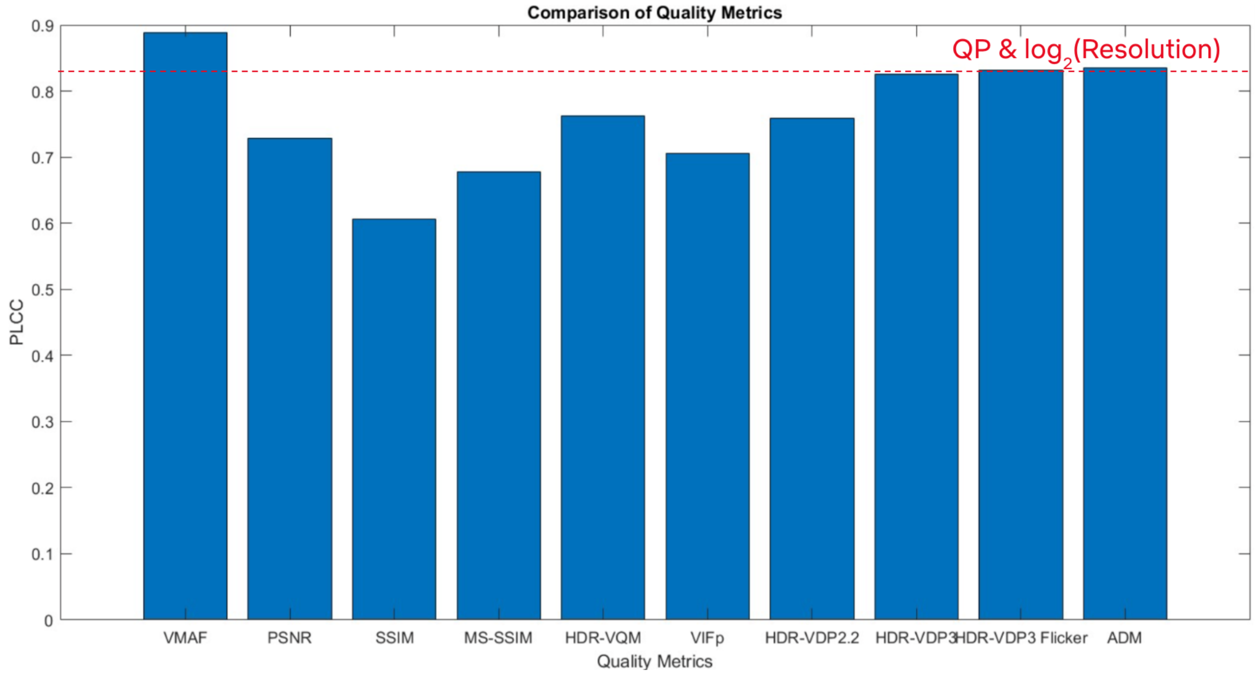

Figure 8. *Pearson Linear Correlation Coefficient for the tested objective metrics. The red dashed line indicates the baseline correlation obtained from the combination of QP & $\log_2$(Resolution).*

comparable to “in-lab” tests, if specific precautions are taken into account. In an objective analysis, we found that most standard image and video quality metrics (both SDR and HDR) do not correlate well with observer opinions in this challenging application. The best performing metric was VMAF, reaching the Pearson Linear Correlation Coefficient of 0.89. It was the only tested metric able to significantly outperform combination of QP and encoding resolution as a predictor of quality.

Our future work includes using the dataset for training the existing, as well as new video quality metrics.

References

[1] Recommendation ITU-R BT. 2100, ”Image parameter values for high dynamic range television for use in production and international programme exchange,” 7/2018.

[2] N. Moroney, ”Unconstrained web-based color naming experiment”, Proc. SPIE 5008, pp. 36-46 (2003).

[3] N. Moroney and G. Beretta, ”The world wide “gamma””, IS&T Color Imaging Conference 2010.

[4] J. Gille and J. Larimer, Using the Human Eye to Characterize Displays, Human Vision and Electronic Imaging VI, Proc. SPIE, 4299, pp. 439-454 (2001).

[5] Robert S. Allison, Kjell Brunnström, Damon M. Chandler, Hannah R. Colett, Philip J. Corriveau, Scott Daly, James Goel, Juliana Y. Long, Laurie M. Wilcox, Yusizwan M. Yaacob, Shun-nan Yang, Yi Zhang, ”Perspectives on the definition of visually lossless quality for mobile and large format displays,” J. Electron. Imaging 27(5), 053035 (2018), Doi: 10.1117/1.JEI.27.5.053035.

[6] Scott Daly, Pavel Korshunov, Touradj Ebrahimi, Timo Kunkel, and Robert Wanat (2019) “Black level visibility as a function of ambient illumination.” SMPTE Motion Imaging Journal, May 2019. https://ieeexplore.ieee.org/document/8700636

[7] Sean T. McCarthy, Scott Daly, and Timo Kunkel (2019) “Frame work for evaluating display resolution and size in the context of video compression and visual acuity”. SID Display Week. Paper 79-1.

[8] L. Frisen and M. Frisen (1981) “How good is normal visual acuity? A study of letter acuity thresholds as a function of age ”Archives Clinical Experimental Ophthalmology , DOI: 10.1007/BF00413146

[9] Larson, E.C.; Chandler, D.M., ”Most apparent distortion: full-reference image quality assessment and the role of strategy,” J. Electron. Imaging 2010, 19, 011006.

[10] Ponomarenko, N.; Jin, L.; Ieremeiev, O.; Lukin, V.; Egiazarian, K.; Astola, J.; Vozel, B.; Chehdi, K.; Carli, M.; Battisti, F., ”Image database TID2013: Peculiarities, results and perspectives,” Signal Process. Image Commun. 2015, 30, 57–77.

[11] Recommendation ITU-T 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).

[12] Z. Li, C. G. Bampis, L. Krasula, L. Janowski, I. Katsavounidis, ”A Simple Model for Subject Behavior in Subjective Experiments,” arXiv:2004.02067, doi: 10.48550/arXiv.2004.02067.

[13] T. Hoßfeld, R. Schatz and S. Egger, ”SOS: The MOS is not enough!,” 2011 Third International Workshop on Quality of Multimedia Experience, Mechelen, 2011, pp. 131-136.

[14] A. K. Venkataramanan, C. Wu, A. C. Bovik, I. Katsavounidis and Z. Shahid, ”A Hitchhiker’s Guide to Structural Similarity,” in IEEE Access, vol. 9, pp. 28872-28896, 2021, doi: 10.1109/ACCESS.2021.3056504

[15] H.R. Sheikh.and A.C. Bovik, ”Image information and visual quality,” IEEE Transactions on Image Processing, vol.15, no.2, pp. 430-444, Feb. 2006.

[16] S. Li, F. Zhang, L. Ma and K. N. Ngan, ”Image Quality Assessment by Separately Evaluating Detail Losses and Additive Impairments,”

in IEEE Transactions on Multimedia, vol. 13, no. 5, pp. 935-949, Oct. 2011, doi: 10.1109/TMM.2011.2152382.

[17] Z. Li, K. Swanson, C. G. Bampis, L. Krasula, A. Aaron, ”Toward a Better Quality Metric for the Video Community,” Netflix Technology Blog, [online] `https://tinyurl.com/2p8fce64`.

[18] M. Narwaria, M. Perreira da Silva, P. Le Callet, ”HDR-VQM: An Objective Quality Measure for High Dynamic Range Video,” Signal Processing: Image Communication, 2015, 35, pp.46-60.

[19] M. Narwaria, R. K. Mantiuk, M. Perreira Da Silva, P. Le Callet, ”HDR-VDP-2.2: A Calibrated Method for Objective Quality Prediction of High Dynamic Range and Standard Images,” Journal of Electronic Imaging, 24(1), 2015.

[20] R. Mantiuk, K. J. Kim, A. G. Rempel, W. Heidrich, ”HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions,” ACM Transactions on Graphics, 30(4), 40:1–40:14, 2011, doi: 10.1145/2010324.1964935.

[21] ITU-T Technical Report PSTR-CROWDS, ”Subjective evaluation of media quality using a crowdsourcing approach”, May 2018.

[22] Recommendation ITU-T P.808, ”ubjective evaluation of speech quality with a crowdsourcing approach”, June 2021.