Subjective and Objective Evaluation of HDR Video Coding Technologies

Philippe Hanhart, Martin Řeřábek, and Touradj Ebrahimi
Multimedia Signal Processing Group
EPFL, Lausanne, Switzerland

***Abstract*—This paper reports the details and results of a subjective and objective quality evaluation assessing responses to an MPEG call for evidence (CfE) on high dynamic range (HDR) and wide color gamut video coding. Five HDR video contents, compressed at four bit rates by each proponent responding to the CfE, were used in the subjective assessments. To be able to evaluate the performance of objective quality metrics, the double stimulus impairment scale (DSIS) method was used for subjective assessments instead of previously published paired comparison to an anchor. Subjective results show evidence that coding efficiency can be improved in a statistically noticeable way over the HEVC anchor in terms of perceived quality. However, when compared to paired comparison, less statistically significant differences are observed because of the lower discrimination power of the DSIS method. The collected subjective scores were used as a ground truth to benchmark and analyze the performance of objective metrics. Results show that HDR-VDP-2 and PQ2VIFP have the highest correlation with subjective scores and outperform other investigated metrics.**

Keywords*—*High dynamic range video, subjective evaluation, video coding, video compression.

I. Introduction

The purpose of the call for evidence (CfE) [1] released in February 2015 by the Moving Picture Experts Group (MPEG) was to explore whether the coding efficiency and/or the functionality of HEVC Main 10 and Scalable Main 10 profiles can be significantly improved for high dynamic range (HDR) and wide color gamut (WCG) content. Potential evidence could include among others new video compression algorithms and coding tools, new signal processing techniques, as well as different color spaces and transfer functions.

The CfE considers four different categories addressing various applications, including backward compatibility with existing standard dynamic range (SDR) content, with either normative or non-normative changes to existing HEVC profiles. However, only responses to two categories were tested in the formal subjective evaluations: Category 1, i.e., single layer solution for HDR, and Category 3a, i.e., non-normative changes to the existing HEVC Main 10 Profile. Therefore, the results reported in this contribution cover only the five Category 1 submissions (P11, P12, P13, P14, and P22) and the four Category 3a submissions (P31, P32, P33, and P34).

This work was performed in the framework of ImmersiaTV under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 688619) and funded by Swiss State Secretariat for Education, Research and Innovation SERI.

978-1-5090-0354-9/16/$31.00 ©2016 IEEE

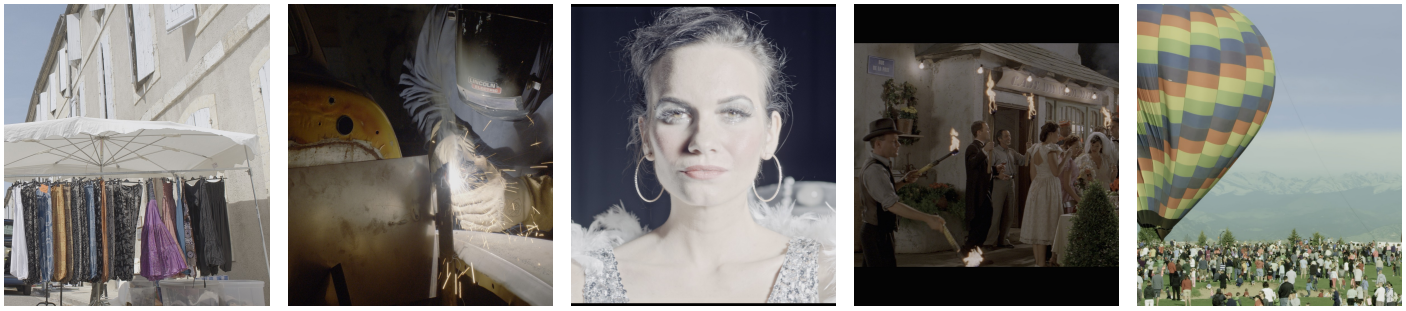
To evaluate responses to the CfE for HDR and WCG video coding, we initially used a partial paired comparison method to have a direct answer to the question whether a proponent can achieve better visual quality when compared to an anchor at a similar bit rate [2]. However, the drawback of this method is that a direct comparison of different proponents cannot be made in a reliable way. For this purpose, a full paired comparison would be necessary, but it requires tremendous efforts for the subjective evaluation. Another drawback is that the paired comparison results could not be used to measure the correlation between objective quality metrics and perceived visual quality. Thus, we have performed a subjective quality assessment on the CfE material using the DSIS method to obtain mean opinion score (MOS) values for all test stimuli. To be able to compare the results obtained with the DSIS method to those of our previous study using the partial paired comparison method [2], as little changes as possible were made to the evaluation methodology and viewing environments.

This paper describes the details and the results of the subjective and objective quality evaluations conducted to benchmark the potential HDR video coding technologies. The subjective tests were performed using the DSIS method and a side-by-side presentation. Overall, 30 naïve subjects participated in the subjective experiments.

An important issue for the development of future HDR video compression algorithms is related to the selection of the objective metrics used to measure quality. A few studies have investigated this problem, but they rely on rather limited databases to benchmark the metrics [3], [4]. Therefore, currently, there is no agreement on which metrics should be used, as there is not enough evidence than one metric outperforms significantly over others.

In this paper, subjective scores collected during the evaluations of the CfE were used to benchmark several objective metrics. In particular, we used the 196 MOS and corresponding confidence interval (CI) values to measure the correlation between objective and subjective scores using widely used performance indexes, i.e., the linearity, monotonicity, accuracy, and consistency of the estimation of MOS. The objective metrics were benchmarked and evaluated based on their correlation with the perceived visual quality.

The remainder of this paper is organized as follows. Section II presents details related to subjective evaluation, such as description of used contents, methodology, and performed statistical analysis together with summary of results. Section III describes results of objective measurements and their correlations with perceived video quality. Section IV concludes the paper.



(a) *Market3* (b) *AutoWelding* (c) *ShowGirl2* (d) *WarmNight* (e) *BalloonFestival*

Figure 1: Representative frames of the sequences used in the experiments. Tone-mapped versions are shown, since typical displays and printers are unable to reproduce higher dynamic range images.

II. SUBJECTIVE EVALUATION

This section describes the dataset and methodology used in the subjective quality evaluations, as well as the processing of the collected individual subjective scores.

A. Dataset

The same dataset as in [2] was used for the subjective evaluations and consisted of five HD resolution HDR video sequences, namely *Market3*, *AutoWelding*, *ShowGirl2*, *WarmNight*, and *BalloonFestival*. Figure 1 shows a typical frame example of each content. Each video sequence was cropped to 950×1080 pixels so that the video sequences were presented side-by-side with a 20-pixels separating black border. The same cropping window as in [2] was used for each video sequence. However, in this study, the video sequences were played at their native frame rate, whereas they were all played at 24 fps in [2] due to the Pulsar's fixed frame rate. In particular, in this study, the sequence *Market3* was played at 50 fps and all of its 400 frames were shown, whereas only the first 240 frames were shown in [2]. For the other sequences, the same frames as in [2] were selected. The coordinates of the cropping window and selected frames are given in Table I. The source material was in EXR format, but, for the experiments, the data was converted to the Sim2 packed format and stored in uncompressed 8-bit AVI files. The side-by-side video sequences were generated using the HDRMontage tool from the HDRTools software [5]. Then, the conversion was made for the Sim2 display using the Sim2Convert tool provided in the HDRTools software [5].

B. Methodology

The DSIS Variant I method with a five-grade impairment scale (*Very annoying*, *Annoying*, *Slightly annoying*, *Perceptible, but not annoying*, and *Imperceptible*) [6] was selected. Two video sequences were presented simultaneously in side-by-side fashion. One of the two video sequences was always the reference (unimpaired) video sequence. The other was the test video sequence, which was a reconstructed version of the reference. To reduce the effect of position of video sequences on the screen, the participants were divided into two groups: the left video sequence was always the reference video sequence for the first group, whereas the right video sequence was always the reference video sequence for the second group. After the presentation of each pair of video sequences, a six-second voting time followed. Subjects were asked to rate the impairments of the test video sequence in relation to the reference video sequence.

Table I: HDR test sequences used in the subjective evaluations.

sequence	fps	window		frames		anchor bit rates (kbits/s)			
						R4	R3	R2	R1
Market3	50	970	1919	0	399	1248	2311	4224	7913
AutoWelding	24	600	1549	162	401	454	778	1383	3157
ShowGirl2	25	350	1299	94	333	574	971	1652	3316
WarmNight	24	100	1049	36	275	462	780	1328	2441
BalloonFestival	24	0	949	0	239	1276	2156	3767	6644

C. Test Environment

The tests were performed in the same environment as in [2]. In particular, the experiments were conducted in the Multimedia Signal Processing Group (MMSPG) test laboratory at EPFL, which fulfills the recommendations for subjective evaluation of visual data issued by ITU-R [6]. The test room is equipped with a controlled lighting system of a 6500 K color temperature. The color of all background walls and curtains in the room is mid grey. The laboratory setup is intended to ensure the reproducibility of the subjective tests results by avoiding unintended influence of external factors. In the experiments, the luminance of the background behind the monitor was about 20 $\mathrm{cd/m^2}$. The ambient illumination did not directly reflect off of the display. However, the test was performed on a full HD 47" Sim2 HDR47E S 4K HDR monitor instead of a Pulsar. Similarly to [2], three subjects assessed the displayed test video content simultaneously in every session. They were seated in an arc configuration, at a constant distance of about 3.2 times the picture height (measured from the middle of the screen), as suggested in recommendation ITU-R BT.2022 [7].

D. Test Planning

Before experiments, oral instructions were provided to explain the evaluation task and a consent form was handed to subjects for signature. A training session was organized to allow subjects to familiarize with the assessment procedure. The same contents were used in the training session as in

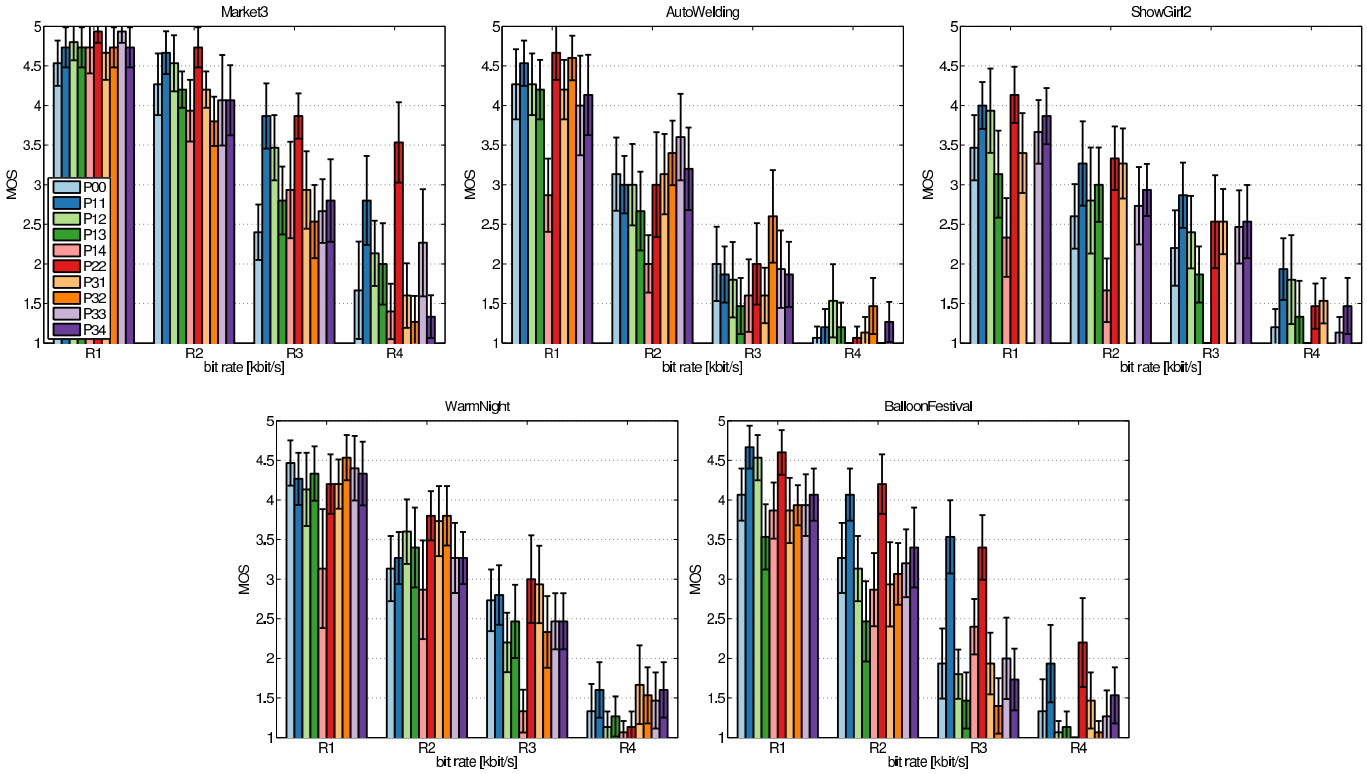


Figure 2: Subjective results (MOS and CI). For content ShowGirl2, dummy values are used for one proponent, as the decoded material was not provided for this sequence.

the test session to highlight the areas where distortions can be visible. Five training samples were manually selected by expert viewers, one for each level of the impairment scale and a different content for each sample. The samples were presented in the following order: *Imperceptible* (*Market3*), *Very annoying* (*AutoWelding*), *Annoying* (*WarmNight*), *Slightly annoying* (*ShowGirl2*), and *Perceptible, but not annoying* (*BalloonFestival*). The training materials were presented to subjects exactly as for the test materials, thus in side-by-side fashion. The overall experiment was split into four test sessions. Each test session was composed of 50 basic test cells, corresponding to approximately 14 minutes each. To reduce contextual effects, the stimuli orders of display were randomized applying different permutation for each group of subjects, whereas the same content was never shown consecutively. The test material was randomly distributed over the four test sessions. Each subject took part in exactly two sessions. One dummy pair, whose score was not included in the results, was included at the beginning of the each session to stabilize the subjects' ratings. Between the sessions, the subjects took a five-minute break. A total of 30 naïve subjects (3 females and 27 males) took part in experiments, leading to a total of 15 ratings per test sample. Subjects were between 20 and 26 years old with an average and median of 22.9 and 23 years of age, respectively. All subjects were screened for correct visual acuity and color vision using Snellen and Ishihara charts, respectively.

E. Statistical Analysis

To detect and remove subjects whose scores appear to deviate strongly from the other scores in a session, outlier detection was performed. The boxplot inspired outlier detection technique proposed in [8] was used. In this study, no outlier subjects were detected. Then, the MOS values were computed for each test stimulus as the mean across the rates of the valid subjects, as well as associated 95% CIs, assuming a Student's t-distribution of the scores. To understand whether the difference between two MOS values corresponding to two different compression algorithms is statistically significant, a multiple comparison significance procedure was applied to the data [9]. Particularly, for each bit rate and content, a one-way ANOVA test was conducted to compare all compression algorithms pairwise to understand whether the differences of their means were statistically significant [9].

F. Results and Discussions

Figure 2 shows the resulting MOS/CI plots for different contents. As it can be observed, the CIs of the different proponents overlap in most cases, meaning that there are few cases where there is a statistically significant difference in visual quality. Nevertheless, improvements can still be observed, especially for proponents P11 and P22. Because the improvements over the anchor (P00) are rather limited in many cases, they are harder to distinguish with an indirect comparison, e.g., DSIS, than with a direct comparison, e.g., paired comparison. Surprisingly, the MOS for content ShowGirl2 are all below 4, except for Proponent P22 at R1. We believe that the scores for this content are lower because of the color differences induced by the viewing angle dependency of the Sim2 monitor and because subjects are more sensitive to color differences in human skin than for other regions. These results

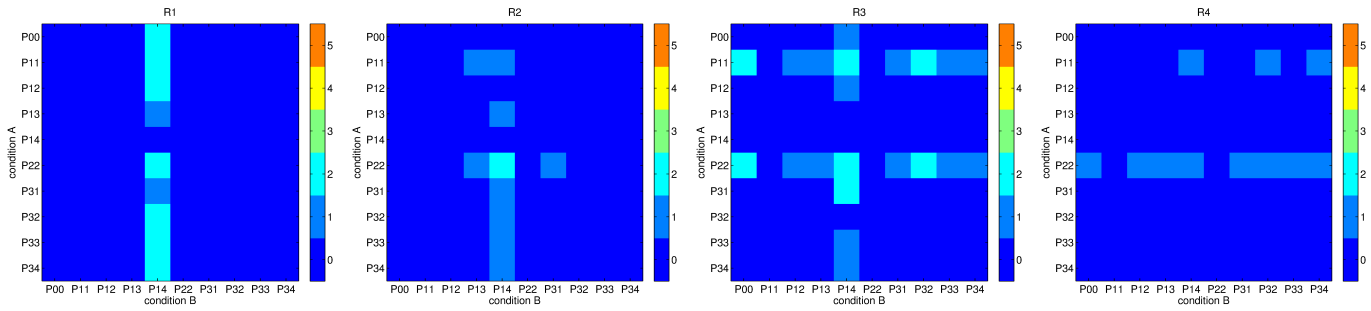


Figure 3: Results of the multiple comparison tests for different test conditions, i.e., combination of algorithm and bit rate (R1 to R4). In each plot, the color of each square shows the number of times (i.e., for how many contents) the MOS corresponding to condition A is statistically significantly better than the MOS corresponding to condition B.

show that a simultaneous side-by-side presentation on Sim2 might not be suitable and that a DSIS temporal presentation may be considered as an alternative to direct comparison methods, such as paired comparison. However, the temporal presentation relies more on short-term memory and has a lower discrimination power than simultaneous presentation. The Variant II with repetition of each stimulus could be considered to compensate this effect.

Figure 3 shows the results comparing all possible pairs, for each bit rate separately. These results confirm that there are few cases with significant visual differences. In particular, Proponents P11 and P22 show significant improvements, especially at lower bit rates, whereas Proponent P14 is outperformed by most proponents on some contents.

III. OBJECTIVE EVALUATION

This session describes the results of correlations between perceived video quality and objective measurements. The following objective quality metrics were selected and benchmarked

- Metrics computed in linear domain
  1) PSNR-x: PSNR computed on x component,
  2) PSNR-DEx: PSNR of mean of absolute value of deltaE2000 metric, derived with x as reference luminance value,
  3) PSNR-Lx: PSNR of mean square error of L component of the CIELab color space used for the deltaE2000 metric, derived with x as reference luminance value,
  4) avLumaErr: average error in Barten Steps [10], [11],
  5) avLumaPSNR: PSNR of average error in Barten Steps [10], [11],
  6) avColorErr: average color error [10], [11],
  7) avColorPSNR: PSNR of average color error [10], [11],
  8) HDR-VDP-2 [12], and
  9) HDR-VQM [13].
- Metrics computed in PQ-TF domain [14]
  10) tPSNR-x: PSNR computed on x component,
  11) PQ2SSIM,
  12) PQ2MS-SSIM, and
  13) PQ2VIFP: VIF pixel based version.
- Metrics computed using multi-exposure [15]
  14) mPSNR.

SSIM, MS-SSIM, and VIFP values were computed using MeTriX MuX Visual Quality Assessment Package. For these three metrics, the luminance information was extracted from the RGB values, clipped to the range $[0.005, 4000]$ $\mathrm{cd/m^2}$, transformed using the PQ EOTF, and normalized to the interval $[0, 255]$ before computing the metric. The MATLAB implementations of HDR-VDP-2 and HDR-VQM were used. The remaining metrics were computed using the HDRTool software (v0.9) modified by Philips to implement their Luvstar metric [10], [11]. For contents *ShowGirl2* and *WarmNight*, the top and bottom black borders were discarded when computing the metrics. For content *Market3*, the metrics were computed on the first 240 frames that were used in [2]. We assumed that there would not be much differences in the remaining 150 frames, as the cropped part contains rather constant/similar motion between the beginning and the end of the sequence.

Please note that for better orientation in the results, the tPSNR-Y metric corresponding to PSNR after PQ-TF on RGB and YUV conversion was renamed to tPSNR-Yyuv and the tPSNR-Y metric corresponding to PSNR on XYZ after conversion from RGB to XYZ and PQ-TF on XYZ was renamed to tPSNR-Yxyz. Similarly, the tPSNR-Y metric corresponding to PSNR after PQ-TF on RGB and Yu'v' conversion was renamed to tPSNR-Yyupvp.

A. Performance Evaluation

The results of subjective visual experiments are considered as ground truth to evaluate how well an objective quality metric estimates perceived quality. The result of execution of a particular objective metric is an objective quality rating (OQR), which is expected to be the estimation of the MOS corresponding to a video sequence.

A regression was fitted to the [OQR, MOS] data set to map the objective scores to the subjective ratings. Note that different objective metrics typically have different range of values, so the mapping to a common scale also facilitates the comparison of different models. To consider the intrinsic nature of bounded rating scales, as well as nonlinearities and saturation effects of the human visual system, a non-linear mapping function was used

$$\tilde{M} = a + \frac{b}{1 + \exp[-c(O - d)]} \tag{1}$$

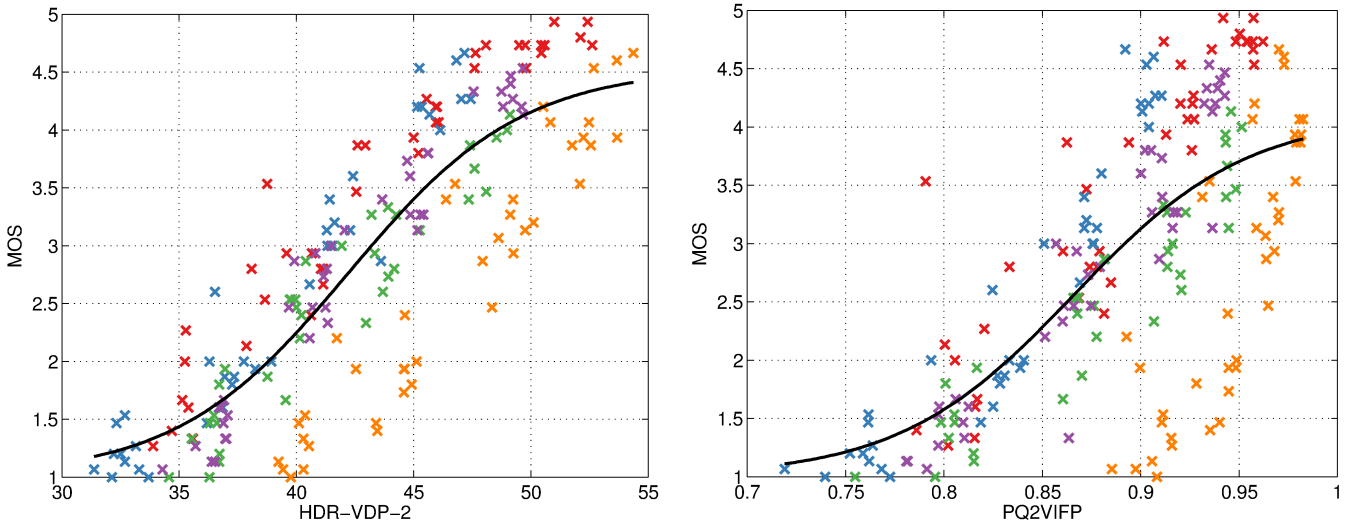


Figure 4: Subjective versus objective results for best performing (assuming A mapping scheme) metrics. The black curve represents the mapping function. For the data points, red, blue, green, purple, and orange color corresponds to *Market3*, *AutoWelding*, *ShowGirl2*, *WarmNight*, and *BalloonFestival* content, respectively.

where $\tilde{M}$ is the predicted MOS, O is the objective metric result, and a, b, c and d are the parameters that define the shape of the logistic mapping function and were determined via the least squares method.

The following properties of the estimation of MOS were considered: linearity, monotonicity, accuracy, and consistency. To this end, four different performance indexes were computed between the ground truth and predicted subjective scores. In particular, the Pearson linear correlation coefficient (PLCC) and Spearman rank order correlation (SROCC) were computed to estimate linearity and monotonicity, respectively. Accuracy and consistency were estimated using the root-mean-square error (RMSE) and outlier ratio (OR), respectively. The OR is the ratio of points for which the error between the predicted and actual MOS values exceeds the 95% confidence interval of MOS values.

To determine whether the difference between two performance index values corresponding to two different metrics is statistically significant, two-sample statistical tests were performed on all four performance indexes. In particular, for the PLCC and SROCC, a Z-test was performed using Fisher z-transformation. For the RMSE, an F-test was performed, whereas a Z-test for the equality of two proportions was performed for the OR. No correction was applied to correct for the multiple comparisons. The statistical tests were performed according to the guidelines of recommendation ITU-T P.1401 [16].

B. Results and Discussions

Figure 4 depicts the scatter plots of subjective versus objective results for the two best performing metrics. As it can be observed, the data points are well concentrated near the mapping curve for HDR-VDP-2, as well as for PQ2VIFP, however they are more scattered for the other metrics, especially in the case of PSNR in linear domain and in PQ-TF domain on the Yu'v' color space, as well as mPSNR, PSNR-DEx, PSNR-Lx, and HDR-VQM, which show higher content dependency. These findings indicate that HDR-VDP-2 and PQ2VIFP have a very high consistency when compared to the other metrics, when all contents are considered.

Table II reports the linearity, monotonicity, accuracy, and consistency indexes, as defined in Sec. III-A. The fitting, as defined in Sec. III-A, was applied in two different ways

A) on all contents at once and
B) on each content separately.

In the latter case, the performance indexes were computed separately on each content and then averaged across contents.

Results of the first case show that HDR-VDP-2 has the best correlation with human perception of visual quality (with PLCC and SROCC values above 0.86), followed by VIFP computed in the PQ-TF domain. However, the statistical tests demonstrate that HDR-VDP-2 is statistically significantly better than PQ2VIFP on the PLCC, SROCC, and RMSE indexes, whereas there is not sufficient evidence to show statistical differences on the OR index. Nevertheless, the OR of HDR-VDP-2 is statistically significantly lower than for the other metrics. All other metrics have poor correlation with perceived quality and a large prediction error.

The second case is related to codec optimization scenario, where it is more important to know that an increase (decrease) in the metric value computed on a specific content will correspond to an increase (decrease) in visual quality, rather than to be able to relate a metric score of any content to an absolute quality level. Results show that most metrics achieve a relatively high correlation with perceived quality as most correlation coefficients are above 0.8. As the mapping is applied on each content separately, metrics that showed strong content dependency previously achieve significantly

Table II: Linearity, monotonicity, accuracy, and consistency indexes for the different metrics. Left column for each index reports the results considering all contents at once, right column for each index reports the results as an average over all contents.

Metric	PLCC		SROCC		RMSE		OR	
mapping	A	B	A	B	A	B	A	B
HDR-VDP-2	0.863	0.962	0.866	0.954	0.591	0.300	0.485	0.224
PQ2VIFP	0.726	0.921	0.700	0.899	0.806	0.421	0.546	0.341
PQ2MS-SSIM	0.640	0.937	0.594	0.921	0.900	0.379	0.633	0.332
tPSNR-Yyuv	0.633	0.912	0.604	0.883	0.906	0.433	0.663	0.322
tPSNR-Yxyz	0.633	0.915	0.609	0.889	0.907	0.422	0.679	0.342
tPSNR-Yyupvp	0.633	0.915	0.609	0.889	0.907	0.422	0.679	0.342
avLumaPSNR	0.631	0.924	0.603	0.900	0.909	0.403	0.653	0.337
tPSNR-G	0.629	0.914	0.596	0.886	0.910	0.422	0.643	0.322
tPSNR-X	0.604	0.917	0.610	0.897	0.933	0.419	0.725	0.342
tPSNR-XYZ	0.602	0.909	0.595	0.884	0.935	0.431	0.735	0.332
tPSNR-RGB	0.577	0.892	0.573	0.868	0.957	0.451	0.730	0.331
tPSNR-R	0.572	0.898	0.563	0.879	0.961	0.450	0.684	0.371
tPSNR-YUV	0.572	0.894	0.567	0.867	0.961	0.444	0.704	0.336
tPSNR-Z	0.558	0.893	0.519	0.877	0.972	0.470	0.674	0.362
tPSNR-B	0.540	0.872	0.506	0.848	0.985	0.500	0.658	0.372
PQ2SSIM	0.508	0.934	0.414	0.921	1.009	0.382	0.709	0.316
PSNR-L100	0.508	0.958	0.335	0.952	1.009	0.310	0.668	0.219
PSNR-DE1000	0.485	0.896	0.290	0.882	1.024	0.491	0.628	0.452
avColorPSNR	0.468	0.841	0.460	0.832	1.035	0.596	0.725	0.524
tPSNR-V	0.467	0.755	0.458	0.753	1.036	0.700	0.725	0.559
tPSNR-U	0.465	0.807	0.386	0.784	1.036	0.610	0.745	0.483
HDR-VQM	0.462	0.944	0.375	0.930	1.039	0.358	0.709	0.286
PSNR-L1000	0.447	0.961	0.258	0.953	1.048	0.301	0.668	0.219
avLumaErr	0.440	0.932	0.428	0.908	1.052	0.387	0.719	0.312
mPSNR	0.433	0.840	0.444	0.816	1.056	0.576	0.735	0.462
PSNR-DE100	0.405	0.890	0.263	0.878	1.071	0.498	0.658	0.457
avColorErr	0.400	0.842	0.415	0.835	1.073	0.594	0.730	0.524
tPSNR-up	0.389	0.605	0.373	0.575	1.079	0.821	0.750	0.597
tPSNR-vp	0.369	0.672	0.329	0.630	1.088	0.746	0.750	0.542
tPSNR-Yupvp	0.356	0.644	0.357	0.686	1.094	0.682	0.765	0.544
PSNR-B	0.302	0.951	0.131	0.946	1.116	0.340	0.684	0.288
PSNR-G	0.297	0.956	0.182	0.942	1.118	0.315	0.689	0.226
PSNR-R	0.288	0.957	0.205	0.951	1.121	0.309	0.699	0.229

better performance in this case. The top performing metrics with PLCC and SROCC values above 0.9, RMSE below 0.4, and OR below 0.3 are HDR-VDP-2, PSNR-Lx, PSNR computed in the linear domain, and HDR-VQM. It should be noted that because of the relatively low number of data points per content, no statistical tests were performed in this case.

IV. CONCLUSION

In this paper, the detailed results of the subjective evaluation using the DSIS method to assess the responses to the CfE for HDR and WCG video coding were reported. The results show that a number of proposals submitted as response to CfE can noticeably improve state of the art standard HDR/WCG video coding technology that was used to generate the CfE anchors. Subjective results advise that a sequential presentation instead of a simultaneous presentation could be considered to reduce the color differences induced by viewing angle dependency of Sim2 monitor. The subjective results were used to evaluate and benchmark the performance of several objective metrics for HDR video quality assessment. Results show that HDR-VDP-2 and PQ2VIFP are good generic predictors of visual quality as they show less content dependency than the other metrics.

REFERENCES

[1] ISO/IEC JTC1/SC29/WG11 MPEG, "Call for Evidence (CfE) for HDR and WCG Video Coding," Doc. N15083, Geneva, Switzerland, Feb. 2015.

[2] P. Hanhart, M. Řeřábek, and T. Ebrahimi, "Towards high dynamic range extensions of HEVC: subjective evaluation of potential coding technologies," in *Proceedings of SPIE 9599*, ser. Applications of Digital Image Processing XXXVIII, Aug. 2015.

[3] M. Azimi, A. Banitalebi-Dehkordi, Y. Dong, M. T. Pourazad, and P. Nasiopoulos, "Evaluating the Performance of Existing Full-Reference Quality Metrics on High Dynamic Range (HDR) Video Content," in *International Conference on Multimedia Signal Processing (MMSP)*, Nov. 2014.

[4] M. Řeřábek, P. Hanhart, P. Korshunov, and T. Ebrahimi, "Subjective and objective evaluation of hdr video compression," in *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Feb. 2015.

[5] A. M. Tourapis and D. Singer, "HDRTools: Software updates," ISO/IEC JTC1/SC29/WG11 MPEG, Doc. M35471, Geneva, Switzerland, Feb. 2015.

[6] ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Jan. 2012.

[7] ITU-R BT.2022, "General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays," International Telecommunication Union, Aug. 2012.

[8] F. De Simone, L. Goldmann, J.-S. Lee, and T. Ebrahimi, "Towards high efficiency video coding: Subjective evaluation of potential coding technologies," *Journal of Visual Communication and Image Representation*, vol. 22, no. 8, pp. 734–748, 2011.

[9] G. W. Snedecor and W. G. Cochran, *Statistical Methods*. Iowa State University Press, 1989.

[10] R. Brondijk, R. Goris, and R. van der Vleuten, "An Objective Metric for HDR," ISO/IEC JTC1/SC29/WG11 MPEG, Doc. M36464, Warsaw, Poland, June 2015.

[11] R. van der Vleuten, R. Brondijk, and R. Goris, "Objective Testing Proposal," ISO/IEC JTC1/SC29/WG11 MPEG, Doc. M35866, Geneva, Switzerland, Feb. 2015.

[12] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: A Calibrated Visual Metric for Visibility and Quality Predictions in All Luminance Conditions," *ACM Transactions on Graphics*, vol. 30, no. 4, July 2011.

[13] M. Narwaria, M. Perreira Da Silva, and P. Le Callet, "HDR-VQM: An objective quality measure for high dynamic range video," *Signal Processing: Image Communication*, vol. 35, pp. 46–60, 2015.

[14] S. Miller, M. Nezamabadi, and S. Daly, "Perceptual Signal Coding for More Efficient Usage of Bit Codes," in *SMPTE Conferences*, vol. 2012, no. 10, 2012, pp. 1–9.

[15] J. Munkberg, P. Clarberg, J. Hasselgren, and T. Akenine-Möller, "High dynamic range texture compression for graphics hardware," *Transactions on Graphics (TOG)*, vol. 25, no. 3, pp. 698–706, 2006.

[16] ITU-T P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," International Telecommunication Union, July 2012.