Towards high dynamic range extensions of HEVC: subjective evaluation of potential 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 the subjective evaluations conducted at EPFL to evaluate the responses to the Call for Evidence (CfE) for High Dynamic Range (HDR) and Wide Color Gamut (WCG) Video Coding issued by Moving Picture Experts Group (MPEG). The CfE on HDR/WCG Video Coding aims to explore whether the coding efficiency and/or the functionality of the current version of HEVC standard can be significantly improved for HDR and WCG content. In total, nine submissions, five for Category 1 and four for Category 3a, were compared to the HEVC Main 10 Profile based Anchor. More particularly, five HDR video contents, compressed at four bit rates by each proponent responding to the CfE, were used in the subjective evaluations. Further, the side-by-side presentation methodology was used for the subjective experiment to discriminate small differences between the Anchor and proponents. Subjective results shows that the proposals provide evidence that the coding efficiency can be improved in a statistically noticeable way over MPEG CfE Anchors in terms of perceived quality within the investigated content. The paper further benchmarks the selected objective metrics based on their correlations with the subjective ratings. It is shown that PSNR-DE1000, HDR-VDP-2, and PSNR-Lx can reliably detect visible differences between the proposed encoding solutions and current HEVC standard.

Keywords: High dynamic range video, subjective evaluation, video coding, video compression.

1. INTRODUCTION

Since the completion of the first edition of the High Efficiency Video Coding (HEVC) standard, several key extensions of its capabilities have been developed to address the needs of an even broader range of applications. Recognizing the rise of High Dynamic Range (HDR) applications and the lack of a corresponding video coding standard, the Moving Picture Experts Group (MPEG) released in February 2015 a Call for Evidence (CfE) for HDR and Wide Colour Gamut (WCG) video coding.[1] The purpose of this CfE 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 HDR and WCG content.

Potential evidence might include among others new video compression algorithms and coding tools, as well as new signal processing techniques, and different colour spaces and transfer functions. The CfE addressed four different categories covering various applications, including backward compatibility with existing Standard Dynamic Range (SDR) content and assuming both, with either normative or non-normative changes to existing HEVC profiles. Note that non-normative changes are categorized as modifications that do not have impact on the decoding process, e.g., color sampling conversion. More particularly, the submission categories are defined as follows:

- Category 1: Single layer solution for HDR
- Category 2: Backward compatible solutions
  - 2a: Backward compatibility with legacy SDR decoders and displays, using an encoding system that has both HDR and SDR inputs

Further author information: (Send correspondence to Philippe Hanhart or Touradj Ebrahimi) E-mail: {firstname.lastname}@epfl.ch

2b: Technology Under Consideration for backward compatibility with legacy SDR decoders and displays, using an encoding system that has only an HDR input
2c: Technology Under Consideration for backward compatibility with legacy SDR displays, but not SDR decoders, using an encoding system that has both HDR and SDR inputs
2d: Technology Under Consideration for backward compatibility with legacy SDR displays, but not SDR decoders, using an encoding system that has only an HDR input

- Category 3:Non-normative changes to the existing HEVC profiles

3a: Main 10 Profile
3b: Scalable Main 10 Profile

Each test condition, i.e. category, is described in more details within the CfE document.[1]

In total, eight companies or aggregations of different companies and one university responded to the CfE and submitted responses to one or more of the different categories. Initially, responses to categories 1, 2b, 3a, and 3b were planned to be tested through formal subjective evaluations. However, based on the large number of responses, it was further agreed that only responses to categories 1 and 3a would be tested in the formal subjective evaluations. Therefore, the results reported in this contribution cover only the five Category 1 submissions and the four Category 3a submissions.

In the context of the CfE preparation for HDR/WCG video coding, HEVC Anchors of the selected content[2] were generated[3] with carefully selected bit rates as test points using official HM software. These Anchors served as reference testing sequences as described in the CfE.[1] In addition, each proponent provided the selected content encoded with a proprietary solution at the same bit rates as an attempt to improve compression efficiency of HEVC Main profiles.

Several improvements relevant to both evaluated categories have been proposed and the most important are briefly listed below. In Category 1, only single layer coding solution is allowed. This setup allows one HDR input to the encoder, which produces a single compressed bit stream associated with that input. A corresponding decoder produces one HDR output to be sent to HDR displays.
The Arris/Dolby/InterDigital response to this category proposes three new technologies aiming at improving color performance as well as general coding efficiency: a perception based color opponent model, IPT-PQ, Cross Plane Chroma Enhancement (CPCE) Filtering, and Adaptive Reshaping and Transfer Function (ARTF). IPT-PQ includes a few modifications to the older IPT color space, which uses a model of the color difference between cones in the Human Visual System, to provide a better fit for HDR and WCG signals.[4] CPCE can further mitigate the distortions caused by chroma downsampling and quantization, and thus improves color performance. ARTF changes the signal characteristics to improve the coding efficiency. In particular, it adaptively re-distributes the codeword based on pixel brightness and re-quantizes the signal among I, P, and T components, which ultimately changes the bit rate allocation for the three components.
FastVDO solution employs the conversion of HDR content into a FastVDO-developed integer color space YFbFr, which was designed to be closely aligned to the YCbCr of BT.709. Consequently a grayscale smoothed luminance signal is generated from the Y component and further used to generate a low bit depth base signal. Both signals are then converted to YFbFr 4:2:0 and coded using the Main 10 HEVC standard.[5]
Philips[6] and Technicolor[7] solutions use a parameter based single layer coding approach and transmit SDR signal with side metadata that enable the reconstruction of the HDR signal.

As for the category 3a, only non-normative changes and improvements to HEVC Main 10 profile can be taken into account. Since the HM reference software was used to generate the Anchors, there can be an actual improvement of its capabilities in terms of non-normative coding tricks and optimizations such as, for example, better motion estimation, rate control, adaptive rate allocation, subjectively optimized quantization, enhanced mode decision strategy, etc. Moreover, the HM reference software was developed for SDR content, thus many encoding parameters have been tuned for SDR.
For instance, Qualcomm and Apple provided two approaches for non-normative improvements to HEVC, a non-constant luminance (NCL) approach and a constant luminance (CL) approach. Both approaches use several encoder enhancements to the official HM reference software, whereas identical pre- and post-processing techniques

Table 1: 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	239	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

are used for converting the data in and out from the 4:2:0 format.[8]
BBC introduces additional pre- and post-processing associated with their Hybrid Log-Gamma (HLG) system solution. More particularly, the input HDR video is converted from linear light, half floating point format, to integer value sequence by applying the HLG OETF and subsequent pre-processing is applied to convert the colour space of the source to Y'CbCr with 10-bit depth per component and 4:2:0 chroma format. The Y'CbCr data is then presented as input to a modified HEVC reference implementation, which attaches a Supplementary Enhanced Information (SEI) message to the bitstream containing parameters to allow HDR displays to properly show the decoded content.[9]
Ericsson's contribution uses different ways of calculating Y', Cb, and Cr components to avoid the luminance errors that occur in the Anchor chain. Further, it employs an anti-banding filter in the post-processing before display and modifies the calculation of the rate distortion optimization parameters (lambda) used from the QPs in HM.

This paper presents the details and the results of the subjective quality evaluation performed to benchmark the potential coding technologies submitted in response to the CfE. It also describes the results of correlations between perceived video quality and selected objective metrics. The subjective tests were performed in the form of partial pair comparison, where one video sequence of the pair was always the Anchor as a reference. Overall 48 naïve subjects participated in the subjective experiment, which leads to a total of 24 ratings per video stimuli. The objective metrics are evaluated based on their classification errors.

The remainder of this paper is organized as follows. Section 2 presents details related to subjective evaluation campaign within CfE, such as description of exploited contents, test methodology, and performed statistical analysis together with summary of results. Section 2 describes results of objective measurements and their correlations with perceived video quality. Section 4 concludes the paper.

2. SUBJECTIVE EVALUATION

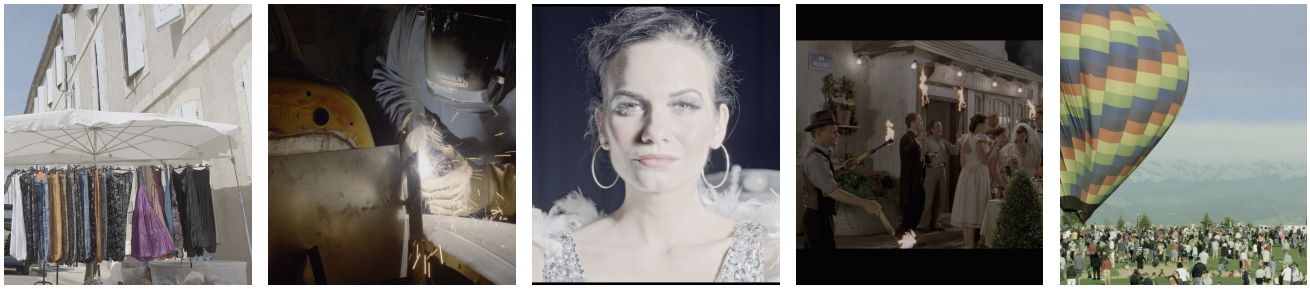
2.1 Dataset

The dataset used for the subjective evaluation tests consists 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. Each video sequence was displayed at 24 fps, which is the native frame rate of the display used in the experiments (see Sec. 2.2), and cut to 240 frames, which corresponds to 10 seconds. Note that the *Market3* sequence was played at a slower frame rate than the original content (50 fps). This solution was evaluated as visually more pleasant than playing every other frame, which created temporal distortions. The coordinates of the cropping window, selected frames, and bit rates are given in Table 1.

The data was stored in uncompressed 16 bit TIFF files, in 12 bit non-linearly quantized (using Dolby PQ EOTF) RGB signal representation, using the SDI data range (code values from 16 up to 4076) and BT.2020 RGB color space. The side-by-side video sequences were generated using the HDRMontage tool from the HDRTools package.[10]

2.2 Test environment

The experiments were conducted at EPFL's Multimedia Signal Processing Group (MMSPG) test laboratory, which fulfills the recommendations for subjective evaluation of visual data issued by ITU-R.[11] The test room is



(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.

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 test results by avoiding unintended influence of external factors. In the experiments, the luminance of the background behind the monitor was about 20 cd/m^2. The ambient illumination did not directly reflect off of the display.

To display the test stimuli, a full HD (1920 × 1080 pixels) 42" Dolby Research HDR RGB backlight dual modulation display (aka Pulsar) was used. The monitor has the following specifications: full DCI P3 color gamut, 4000 cd/m^2 peak luminance, low black level (0.005 cd/m^2), 12 bits/color input with accurate and reliable reproduction of color and luminance. In every session, three subjects assessed the displayed test video content simultaneously. They were seated in one row perpendicular to the center of the monitor, at a distance of about 3.2 times the picture height, as suggested in recommendation ITU-R BT.2022.[12]

2.3 Test methodology

Two video sequences were presented simultaneously in side-by-side fashion. Since only one full HD 1920 × 1080 HDR monitor was available, each video was cropped to 950 × 1080 pixels with 20 pixels of black border separating the two sequences. One of the two video sequences was always the Anchor, with a randomized position on the screen (either on the left or on the right). The other video sequence was the Proponent to be evaluated, at the same (targeted) bit rate as the Anchor.

Subjects were asked to judge which video sequence in a pair ('left' or 'right') has the best overall quality, considering fidelity of details in textured areas and color rendition. The option 'same' was also included to avoid random preference selections.

2.4 Test planning

Before the experiments, a consent form was handed to subjects for signature and oral instructions were provided to explain the evaluation task. 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 the test session to highlight the areas where distortions can be visible. Eleven training samples were manually selected by expert viewers. First, two samples, one of high quality and one of low quality, without any difference between left and right, were selected from the *AutoWelding* sequence. The purpose of these two examples was that subjects could get familiar with HDR content, as this content has both dark and bright luminance levels and fast luminance temporal changes, and see the extreme levels of quality observed in the test material. Then, one sample from *AutoWelding* with large visible difference was presented to illustrate the main differences that can be observed between the left and right video sequences, i.e., loss of texture/details and color artifacts. Finally, for each of the remaining contents, two samples were presented (one example with large difference and one example with small differences) in the following order: *Market3*, *BalloonFestival*, *ShowGirl2*, and *WarmNight*. 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 6 test sessions. Each test session was composed of 30-31 basic test cells, corresponding to approximately 14 minutes each. To reduce contextual effects, the stimuli orders of display were randomized, whereas the same content was never shown consecutively. The test material was randomly distributed over the six test sessions.

Each subject took part to exactly three sessions. Three dummy pairs, whose scores were not included in the results, were included at the beginning of the first session to stabilize the subjects' ratings. Between the sessions, the subjects took a 14-minute break.

A total of 48 naïve subjects (16 females and 32 males) took part in the experiments, leading to a total of 24 ratings per test sample. Subjects were between 18 and 49 years old with an average and median of 25.3 and 24 years of age, respectively. All subjects were screened for correct visual acuity and color vision using Snellen and Ishihara charts, respectively.

2.5 Statistical analysis

No outlier detection was performed on the raw scores, since there is no international recommendation or a commonly used outlier detection technique for paired comparison results.

For each test condition, i.e., combination of content, algorithm, and bit rate, the winning frequency of the Anchor, w_{Ai}, winning frequency of the Proponent, w_{Pi}, and tie frequency, t_i, are computed from the obtained subjective ratings. Note that $w_{Ai}+w_{Pi}+t_i = N$, where N is the number of subjects. To compute the preference probability of selecting the proponent version over the Anchor, p_P, ties are considered as being half way between the two preference options:

$$p_P = \frac{w_{Pi}}{N} + \frac{t_i}{2N}$$

.

To determine whether the visual quality difference between the Proponent and the Anchor is statistically significant, a statistical hypothesis test was performed. As ties are split equally between the two preference options, the data roughly follows a Bernoulli process $B(N,p)$, where N is the number of subjects and p is the probability of success in a Bernoulli trial and was set to 0.5, considering that, a priori, the Anchor and Proponent have the same chance of success. Figure 2 shows the Cumulative Distribution Function (CDF) for Binomial distribution with $N = 24$ and $p = 0.5$. The CDF is used to determine the critical region for the statistical test.

To determine whether the proponent provides statistically significant results, a one-tailed binomial test was performed at 5% significance level with the following hypotheses:

H0: Proponent is equal or worse than Anchor
H1: Proponent is better than Anchor

In this case, the critical region for the preference probability over Anchor, p_P, is $\left[\frac{16}{24}, 1\right]$, as the CDF for 16 or more successful trials is above 95% (see Figure 1, $B(16, 24, 0.5) = 0.9680$). Therefore, if there are 16 or more votes in favor of the Proponent, the null hypothesis can be rejected.

Similarly, to determine whether the Proponent provides statistically significantly lower visual quality than the Anchor, a one-tailed binomial test was performed at 5% significance level:

H0: Proponent is equal or better than Anchor
H1: Proponent is worse than Anchor

In this case, the critical region for the preference probability over Anchor, p_P, is $\left[0, \frac{7.5}{24}\right]$, as the CDF for 7.5 or less successful trials is below 5% (see Fig. 2, $B(8, 24, 0.5) = 0.0758$). Note that the Binomial distribution is not defined for non-integer values, and that extension is usually obtained using the floor function. Therefore, if there are 7.5 or less votes in favor of the proponent, the null hypothesis can be rejected.

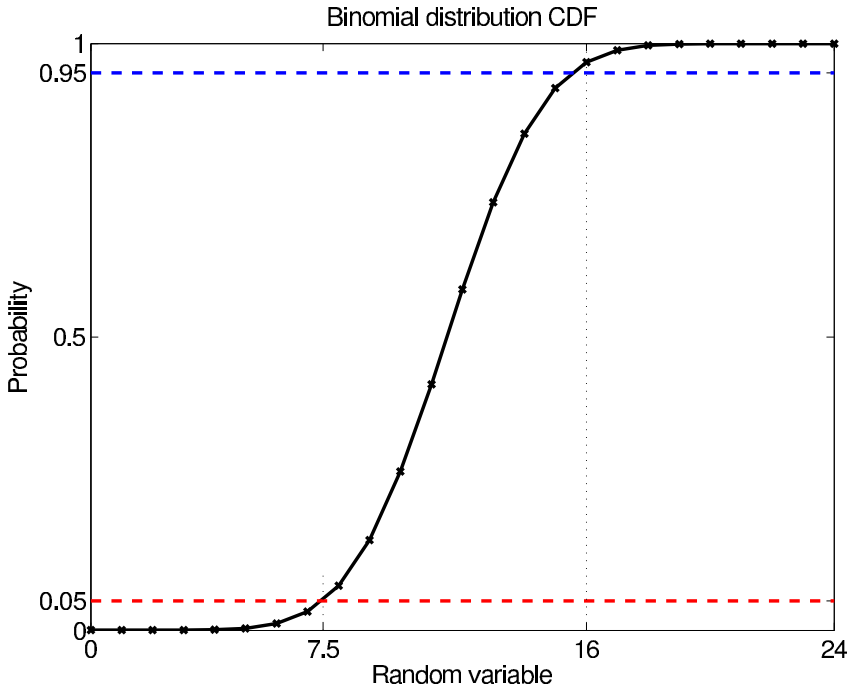


Figure 2: Cumulative distribution function for Binomial distribution with $N = 24$ and $p = 0.5$.

2.6 Results

Figure 3 reports the preference probability of selecting the Proponent version over the Anchor for each content separately. Category 1 submissions (P11, P12, P13, P14, and P22) are plotted with plain lines, while Category 3a submissions (P31, P32, P33, and P34) are plotted with dashed lines. Values on or above the horizontal upper dashed line provide statistically significant visual quality superior to the Anchor, while values on or below the horizontal lower dashed line provide statistically significant inferior visual quality when compared to the Anchor.

As it can be observed, there is evidence that potential coding technologies can do better than the Anchor in a statistically significant way, especially for contents *Market3* and *BalloonFestival*. For instance, on content *ShowGirl2*, Proponent P22 provides statistically significant superior visual quality when compared to the Anchor at rates R1 to R3. Improvements can also be observed for Proponents P11 and P12. Regarding content *WarmNight*, Proponents P32 and P22 outperform the Anchor for rates R2 to R4. Proponents P31 and P11 also show gains for specific rate points. Finally, for content *AutoWelding*, Proponent P32 provides gain for rates R2 to R4, while Proponent P12 is at the limit for the rate R1.

In general, Proponent P32 seems to perform better on dark contents than on bright contents. Regarding P14, wrong colors were observed throughout the test material, probably due to a wrong color transformation, as well as occasional green noise in the table scene on content *WarmNight*. Regarding the selection of contents, bright scenes are better to perceive color artifacts, especially in whitish parts, and loss of details and high frequencies, especially in textured areas. Sequences such as *ShowGirl2* and *Market3* are good for testing HDR compression. On the other hand, sequences with a wide dynamic range and strong luminance temporal changes, such as *AutoWelding* although good for demonstrating HDR, may not be necessarily best to assess HDR compression performance. Dark scenes are important too, as HDR is not only about high brightness, but it might be hard to see the improvements in these sequences, especially if the previous test sequence was bright, due to the adaptation time of the human eye.

Regarding the test methodology, the side-by-side presentation was beneficial to discriminate small differences between the Anchor and Proponents. Repetition also helps improving discrimination power. However, if an absolute ranking of the Proponents is required, the Anchor should be replaced by the source uncompressed reference and an absolute impairment scale should be used, as in a regular DSIS test.

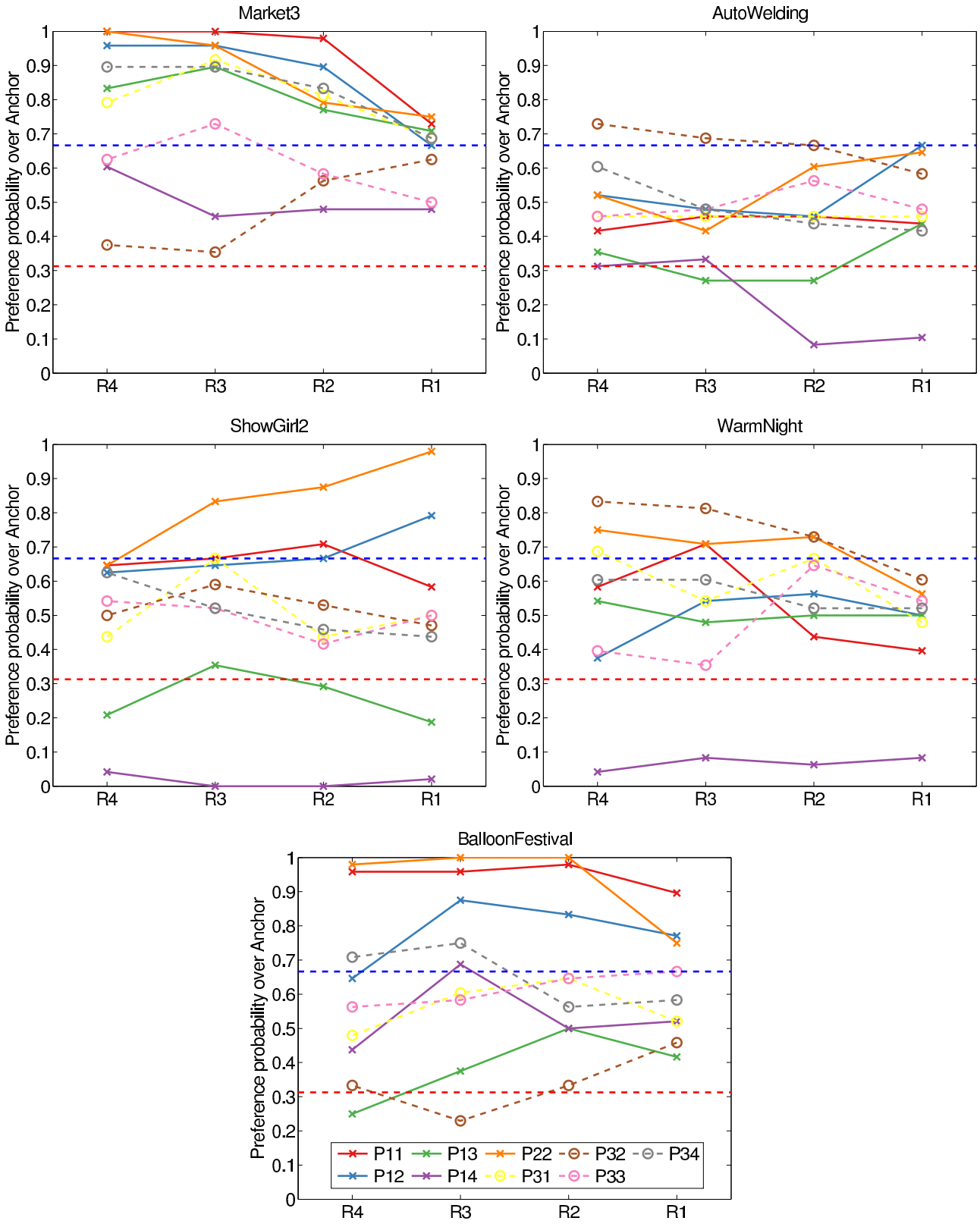


Figure 3: Preference probability of selecting the Proponent version over the Anchor.

3. OBJECTIVE QUALITY METRICS

This session describes the results of correlations between perceived video quality and objective measurements. Based on recent works on HDR quality assessment,[13–16] we selected the following objective metrics:

- Metrics computed in linear domain
  - PSNR_DEx: PSNR of mean of absolute value of deltaE2000 metric, derived with x as reference luminance value
  - 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
  - HDR-VDP-2[17]
  - HDR-VQM[16]
- Metrics computed in PQ-TF domain[18]
  - tPSNR-x: PSNR computed on x component
  - PQ2SSIM
  - PQ2MS-SSIM
  - PQ2VIFP: VIF pixel based version
- Metrics computed using multi-exposure[19]
  - mPSNR

SSIM, MS-SSIM, and VIFP 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 $[0.005, 4000]$ 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 HDRTool[10] (v0.9). For contents *ShowGirl2* and *WarmNight*, the top and bottom black borders were discarded when computing the metrics.

3.1 Classification errors

Since we don't have the mean opinion scores (MOS) for each individual test sequence as in a typical subjective evaluation, e.g., by using a Double Stimulus Impairment Scale (DSIS) methodology, we cannot compute the correlation between subjective and objective scores. Instead, we have to use another approach to evaluate the performance of the objective metrics, e.g., by computing the classification errors as defined in recommendation ITU-T J.149.[20] A classification error is performed when the objective metric and subjective test lead to different conclusions on a pair of video sequences, A and B, for example. Three types of error can happen:

a) *False Tie*, the least offensive error, which occurs when the subjective test says that A and B are different while the objective scores say that they are identical,
b) *False Differentiation*, which occurs when the subjective test says that A and B are identical while the objective scores say that they are different,
c) *False Ranking*, the most offensive error, which occurs when the subjective test says that A is better than B while the objective scores say the opposite.

The analysis was performed by computing the objective metric on the Anchor and Proponent video sequences and checking whether the results of the comparison based on objective measurements matches that of the subjective evaluations. The percentage of *Correct Decision*, *False Tie*, *False Differentiation*, and *False Ranking* were recorded from all Anchor versus Proponent pairs, for each content and bit rate, as a function of the difference in the metric values, ΔOM.

As ΔOM increases, more pairs of data points are considered as equivalent by the objective metric. This reduces the occurrences of *False Differentiations* and *False Rankings*, but increases the occurrence of *False Ties*. As ΔOM tends towards 0, the occurrence of *False Tie* will tend towards 0 and the occurrence of *False*

*MeTriX MuX v1.1: `http://foulard.ece.cornell.edu/gaubatz/metrix_mux/`
†HDR-VDP-2 v2.1.1: `http://hdrvdp.sourceforge.net/`
‡HDR-VQM v1: `http://sites.google.com/site/narwariam/hdr-vqm/`

Differentiation will tend towards the proportion of pairs of video sequences that were declared equivalent by the subjective test.

The relative frequencies are plot as a function of ΔOM. Ideally, the occurrence of *Correct Decision* should be maximized and the occurrence of *False Ranking* should be minimized when the ΔOM tends towards 0. The occurrences of *False Differentiations* and *False Rankings* should decrease as fast as possible as ΔOM increases. Based on this, different graphs corresponding to different metrics can be compared to determine the best metric for the application under analysis.

3.2 Results

Figure 4 reports the classification errors for each metric separately. Even though the results are reported in the native scale of the metric instead of a common scale, it is still possible to compare the classification errors of the different metrics by looking at the relative ΔOM ratio (ΔOM divided by the maximum value of ΔOM) rather than the absolute ΔOM.

Subjective results reported in Sec. 2.6 showed that there were many cases where the Proponent version was providing similar quality when compared to the Anchor. More precisely, in 55% of the cases, no statistically significant difference was observed between Proponent and Anchor, while the difference was statistically significant in 45% of the cases. These values determine the plateau for the *Correct Decision* and *False Tie* frequencies, i.e., if the threshold on ΔOM is set to infinite, all pairs of video sequences are considered as equal for the objective metric, which will lead to a *Correct Decision* frequency of 55%, as 55% of the pairs were evaluated as not statistically different in the subjective evaluations. Similarly, the plateau for the *False Tie* frequency is 45%.

On Fig. 4, dashed lines indicate ΔOM that maximizes the *Correct Decision* frequency. As it can be observed, the maximum of *Correct Decision* is between 0.55 and 0.71. In particular, for HDR-VQM and mPSNR, the highest *Correct Decision* frequency corresponds to the plateau, i.e., the metric cannot distinguish quality. The results for HDR-VQM are quite surprising, as this is the only metric designed to assess quality of HDR video sequences and it was reported to have a relatively low outlier ratio.[16] The reason might be due to the data used by Narwaria *et al.* to train and validate the metric in their experiments on video quality. In particular, they used seven computer generated contents and only three real scenes, while it is known that computer generated content has very different noise characteristics. Additionally, they used their own backward-compatible HDR compression scheme to generate distortions, which might be very different from that of the algorithms considered in the CfE evaluations.

The PSNR metric provides similar results on the different components considered in this study. In all cases, the highest *Correct Decision* frequency is about 60%, which means that it cannot reliably detect visible differences. Additionally, the *False Ranking* frequency decay is very slow, i.e., the probability of making the wrong decision remains, even for large relative ΔOM ratio. It is known that PSNR is not good at handling different types of artifacts,[21] which explains the relatively low performance when comparing compression algorithms based on different schemes.

Regarding SSIM, MS-SSIM, and VIFP computed in the PQ domain, results show that these metrics achieve similar results to PSNR in terms of *Correct Decision*. They have a faster decay for the *False Ranking* frequency, but a slower for the *False Differentiation* frequency. Surprisingly, MS-SSIM shows slightly lower performance than SSIM in terms of *Correct Decision*, while the multiscale approach usually improves performance for SDR content.

PSNR-DE1000 shows the highest *Correct Decision* frequency with a peak at about 0.71, but the *False Ranking* and *False Differentiation* frequencies are not null at the peak. PSNR-L100, PSNR-L1000, and HDR-VDP-2 seem to be better alternatives, as they have a faster decay for the *False Ranking* frequency and reach similar *Correct Decision* frequency for a *False Ranking* frequency of 0. In particular, PSNR-L100 and PSNR-L1000 show slightly less *False Differentiation* compared to HDR-VDP-2. Considering that HDR-VDP-2 has a very high complexity and requires a lot of processing time when compared to the other metrics, PSNR-Lx seems to be a good alternative.

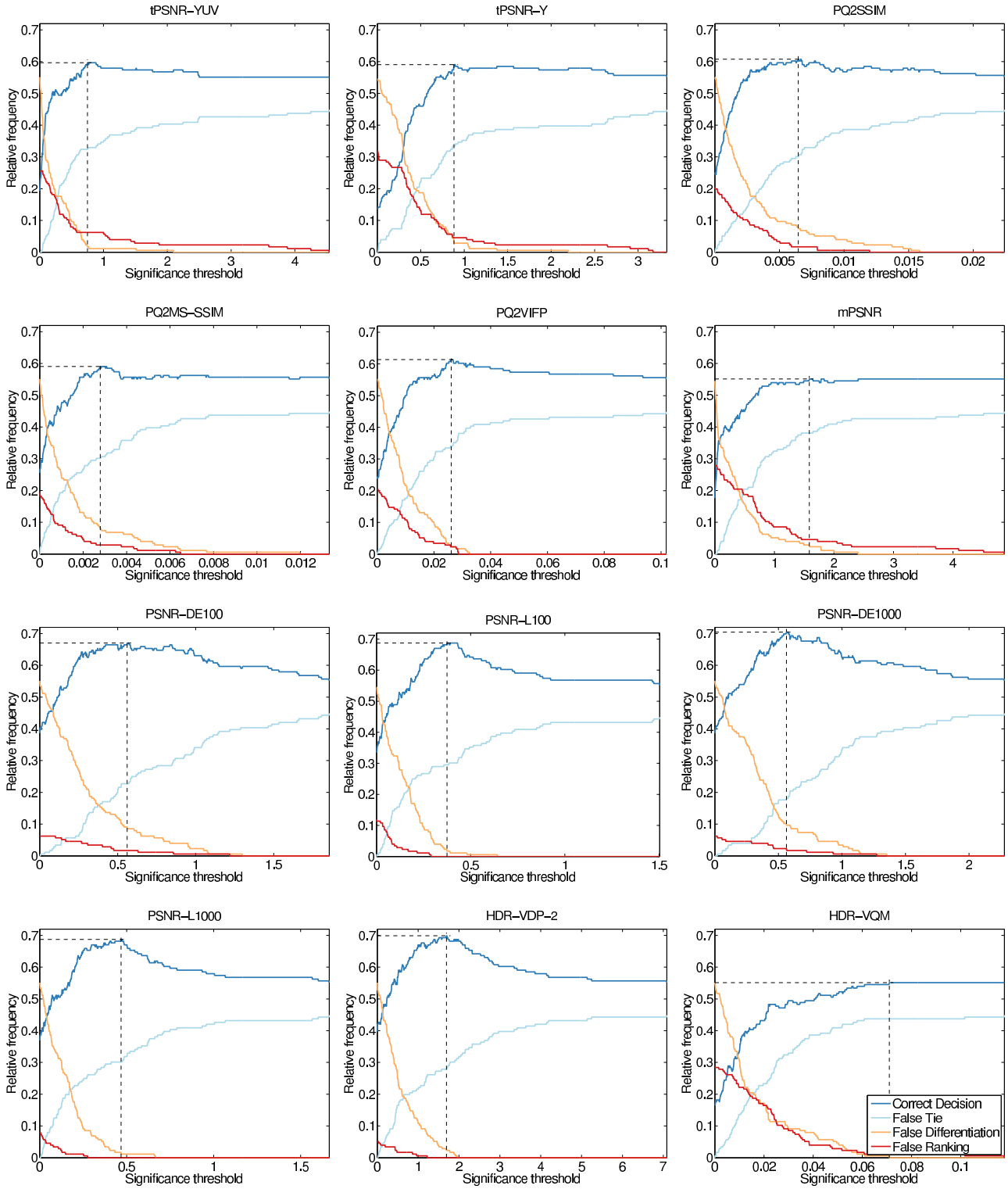


Figure 4: Frequencies of classification error.

4. CONCLUSION

In this paper, the detailed results of the subjective and objective evaluations conducted at EPFL to assess the responses to the Call for Evidence (CfE) for High Dynamic Range (HDR) and Wide Color Gamut (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. In terms of objective measures, PSNR-DE1000, HDR-VDP-2, and PSNR-Lx can reliably differentiate between the proposed encoding solutions and HEVC Main 10 profiles used as Anchors.

ACKNOWLEDGMENTS

This work has been conducted in the framework of the FP7 EC EUROSTAR funded Project - Transcoders Of the Future TeleVision (TOFuTV) and QoE-Net Initial Training Network (H2020-MSCA-ITN-2014).

REFERENCES

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

[2] ISO/IEC JTC1/SC29/WG11 MPEG, "Selected test content and timeline for HDR single layer anchors generation for 111th MPEG meeting." Doc. m35480, Geneva, Switzerland (February 2015).

[3] ISO/IEC JTC1/SC29/WG11 MPEG, "Report on the anchors generation for HDR /WCG video coding." Doc. M35852, Geneva, Switzerland (February 2015).

[4] ISO/IEC JTC1/SC29/WG11 MPEG, "Response to Call for Evidence for HDR and WCG Video Coding: Arris, Dolby and InterDigital." Doc. m36264, Warsaw, Poland (July 2014).

[5] ISO/IEC JTC1/SC29/WG11 MPEG, "An Efficient Dual-Stream Approach for HDR Video Coding (Cat. 1)." Doc. m36251, Warsaw, Poland (July 2014).

[6] ISO/IEC JTC1/SC29/WG11 MPEG, "Philips response to CfE for HDR and WCG." Doc. m36251, Warsaw, Poland@MISCISO/IECJTC1/SC29/WG11MPEG2014Efficient, author = ISO/IEC JTC1/SC29/WG11 MPEG, title = An Efficient Dual-Stream Approach for HDR Video Coding (Cat. 1), howpublished = Doc. m36251, Warsaw, Poland, month = July, year = 2014, date-added = 2015-07-13 13:02:31 +0000, date-modified = 2015-07-13 13:02:31 +0000, owner = rerabek, timestamp = 2014.11.03 (July 2014).

[7] ISO/IEC JTC1/SC29/WG11 MPEG, "Technicolor's response to CfE for HDR and WCG (category 1) - Single layer HDR video coding with SDR backward compatibility." Doc. m36251, Warsaw, Poland (July 2014).

[8] ISO/IEC JTC1/SC29/WG11 MPEG, "Single layer non-normative (category 3a) NCL and CL responses to the Call for Evidence on HDR/WCG." Doc. m36256, Warsaw, Poland (June 2015).

[9] ISO/IEC JTC1/SC29/WG11 MPEG, "BBC's response to CfE for HDR Video Coding (Category 3a)." Doc. m36256, Warsaw, Poland (June 2015).

[10] ISO/IEC JTC1/SC29/WG11 MPEG, "HDRTools: Software updates." Doc. M35471, Geneva, Switzerland (February 2015).

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

[12] 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 (August 2012).

[13] Hanhart, P., Bernardo, M., Korshunov, P., Pereira, M., Pinheiro, A., and Ebrahimi, T., "HDR image compression: a new challenge for objective quality metrics," in [*Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*], (September 2014).

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

[15] Rerabek, M., Hanhart, P., Korshunov, P., and Ebrahimi, T., "Subjective and objective evaluation of hdr video compression," in [*9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*], (February 2015).

[16] Narwaria, M., Silva, M. P. D., and Callet, P. L., "Hdr-vqm: An objective quality measure for high dynamic range video," *Signal Processing: Image Communication* **35**(0), 46–60 (2015).

[17] Mantiuk, R., Kim, K. J., Rempel, A. G., and Heidrich, W., "Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," in [*ACM SIGGRAPH 2011 Papers*], *SIGGRAPH '11*, 40:1–40:14, ACM, New York, NY, USA (2011).

[18] Miller, S., Nezamabadi, M., and Daly, S., "Perceptual Signal Coding for More Efficient Usage of Bit Codes," in [*SMPTE Conferences*], **2012**(10), 1–9 (2012).

[19] Munkberg, J., Clarberg, P., Hasselgren, J., and Akenine-Möller, T., "High dynamic range texture compression for graphics hardware," *Transactions on Graphics (TOG)* **25**(3), 698–706 (2006).

[20] ITU-T J.149, "Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM)." ITU (Mar. 2004).

[21] Huynh-Thu, Q. and Ghanbari, M., "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters* **44**, 800–801 (June 2008).