# PERFORMANCE EVALUATION OF OBJECTIVE QUALITY METRICS ON HLG-BASED HDR IMAGE CODING

*Yasuko Sugito*
Science and Technology Research Laboratories
NHK
Tokyo, Japan
sugitou.y-gy@nhk.or.jp

*Marcelo Bertalmío*
Dept. of Information and Communication Technologies
Universitat Pompeu Fabra
Barcelona, Spain

**ABSTRACT**

We evaluate the performance of objective quality metrics for high dynamic range (HDR) image coding that uses the transfer function (TF) of the Hybrid Log-Gamma (HLG) method. Previous evaluations of objective metrics for HDR image coding have studied which of them are reliable predictors of perceived quality; however, in those tests, all the non-linear transforms used both for encoding and by the best-performing metrics are essentially very similar and based on visual perception data of detection thresholds for lightness variations. The HLG non-linearity on the other hand is very different, as it is designed for backward compatibility with standard dynamic range (SDR) displays. We test a variety of options for objective metrics, including HLG-based. Our results show that the ranking of metrics for HDR coding changes drastically depending on the TF used for compression.

***Index Terms***— High dynamic range (HDR), Hybrid Log-Gamma (HLG), Transfer function (TF), Image coding, Objective quality metric.

## 1. INTRODUCTION

High dynamic range (HDR) is a technology that supports a wider range of luminance in images than conventional systems, i.e., standard dynamic range (SDR). HDR allows for better detail in dark areas, as well as much brighter highlights. An important technical element in HDR image processing is the transfer function (TF) which defines the transformation between the linear light intensity and non-linear signal value for the purpose of image capture, display, and compression. More complicated TFs are applied to the HDR images to effectively process a wider range of luminance.

For HDR Television (HDR-TV), two different types of TFs are standardized: Hybrid Log-Gamma (HLG) opto-electronic transfer function (OETF) and perceptual quantization (PQ) electro-optical transfer function (EOTF) [1]. Figures 1 and 2 show HDR image coding diagrams using HLG OETF and PQ EOTF, respectively. HLG OETF was designed for backward compatibility with SDR and translates relative scene linear light captured by a camera into a non-linear signal value, whereas PQ EOTF was designed according to a contrast sensitivity function (CSF) and translates a non-linear PQ encoded value into an absolute display light that can be seen by the human eye. Opto-optical TF (OOTF) maps the scene light to the display light and is different for HLG than for PQ. In the diagrams, codec is composed of encoding and decoding processes, and image deterioration can be observed after encoding.

This work has partially received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 761544 (project HDR4EU) and under grant agreement number 780470 (project SAUCE), and by the Spanish government and FEDER Fund grant ref. TIN2015-71537-P (MINECO/FEDER, UE)

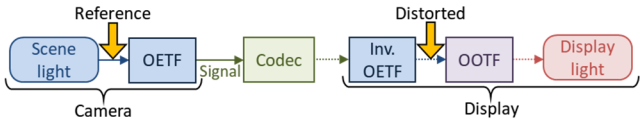


Figure 1. HDR image coding diagram using HLG OETF.

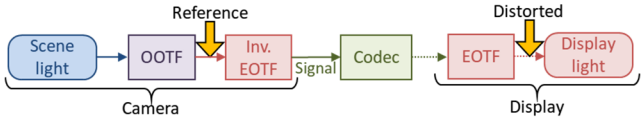


Figure 2. HDR image coding diagram using PQ EOTF.

In image coding, objective quality metrics such as the peak signal-to-noise ratio (PSNR) are used frequently to easily measure the quality of a distorted image relative to an original reference image. Appropriate objective quality metrics accurately emulate human perception and gives results similar to those of a subjective evaluation. Hanhart et al. [2] benchmarked objective quality metrics for HDR image coding using the display adaptive tone mapping method [3] and comparing with subjective evaluation results. They concluded that HDR-VQM [4] and HDR-VDP-2 [5] are the most reliable predictors of perceived quality, and the multi-scale structural similarity index (MS-SSIM) [6] computed in perceptual uniform (PU) space [7] is another lower complexity substitute. However, in those tests, all the non-linear transforms used both for encoding and by the best-performing metrics are essentially very similar and based on visual perception data of detection thresholds for lightness variations in the mappings between signal value and display light, in the same manner as PQ EOTF. Thus, in the PQ image coding case, it is reasonable to use display light as the inputs for the metrics, reference and distorted images in Fig. 2. The HLG non-linearity on the other hand is very different, as it is designed for backward compatibility with SDR displays and not according to CSFs like the PQ EOTF. As Fig. 1 shows, in the case of HLG, scene light might be more suitable for the inputs of metrics.

In this paper, we first encode HDR still images using the HLG method and use an HDR monitor to conduct a subjective evaluation experiment for the compressed images. Then, we validate the performance of the objective quality metrics by calculating the similarity to the subjective evaluation results. Our results show that the ranking of metrics for HDR coding changes drastically depending on the TF used for compression.

## 2. VALIDATION METHOD

We evaluated the performance of objective quality metrics using a method equivalent to that used by Hanhart et al. [2].

### 2.1. HDR image coding experiment

We conducted an HDR image coding experiment using the HLG method to prepare a dataset which consists of various distorted images. For the image coding, high efficiency video coding (HEVC)/H.265 [8] Main 10 Profile was used, and the image format of the encoder input and the decoder output was Y'CbCr 4:2:0 10-bit.

*2.1.1. Pre- and post-processing*

Figures 3 and 4 illustrate the diagrams for the pre- and post-processing steps, respectively. In the pre-processing prior to the encoding, scene light captured by a camera is converted into a non-linear signal value by HLG OETF, defined by the following formula [1].

$$E' = \begin{cases} \sqrt{3E} & 0 \le E \le 1/12 \\ a \cdot \ln(12E - b) + c & 1/12 < E \le 1 \end{cases}$$

where $a = 0.17883277, b = 1 - 4a, c = 0.5 - a \cdot \ln(4a)$. E consists of $R_S G_S B_S$ color components, proportional to scene linear light and normalized to the range of [0:1]. E' is a non-linear signal R'G'B' in the range of [0:1]. The latter processes are conducted to comply with the input format of the encoding. First, the RGB color space is transferred to Y'CbCr, a luma, and two chroma components, complying with ITU-R BT.2020 [9]. Second, the pixel values from 0 to 1 are increased by 1,023 times and rounded to an integer format so they can be treated as10-bit precision values. Finally, the chroma components Cb and Cr are subsampled to a 4:2:0 format. Their image sizes are decreased to half of the original image both horizontally and vertically.

The post-processing phase after the decoding follows the inverse order of the pre-processing stage. After the HLG inverse OETF, the scene light $R_S G_S B_S$ is transferred into the display light $R_D G_D B_D$ in cd/m$^2$ by HLG OOTF defined by the following formulas [1].

$$R_D = \alpha Y_S^{\gamma-1} R_S + \beta \qquad G_D = \alpha Y_S^{\gamma-1} G_S + \beta \qquad B_D = \alpha Y_S^{\gamma-1} B_S + \beta$$

$$Y_S = 0.2627 R_S + 0.6780 G_S + 0.0593 B_S \quad (1)$$

$$\alpha = (L_W - L_B) \qquad \beta = L_B \quad (2)$$

In this paper, we configured the parameters as $L_W = 1{,}000, L_B = 0.005, \gamma = 1.2$, considering the viewing environment of the subjective evaluation experiment.

*2.1.2. HDR test images*

Figure 5 shows thumbnails of the 22 HDR test images. The still images were cropped at 1,920×1,080 pixels. 14 images, from 1 to 14, are Fairchild's HDR photos [10], and we converted them into $R_S G_S B_S$ in BT.2020 color space; the leftmost image format in Fig. 3 indicated as (b). The other 8 images, from 15 to 22, are native HLG images complying with the HDR-TV production guideline [11]: 75% of the nominal signal level corresponds to a diffuse white. The original image format is in the non-linear HLG encoded value, indicated as (c) in Fig. 3.

Figure 6 describes the characteristics of the 22 test images. The vertical axis shows dynamic range: $\log(L_{max}/L_{min})$ where $L_{max}$ and $L_{min}$ are maximum and minimum luminance after excluding 1% of the brightest and darkest pixels, respectively. The horizontal axis shows the spatial perceptive information calculated from ITU-T P.910 [12]. The graph denotes that the test images have wide coding complexity.

*2.1.3. Encoding condition*

We used HEVC Test Model (HM) [13] version 16.17 encoder with all intra Main 10 and fixed QP settings. The target bit-rates were set at 100, 200, 300, and 400 kbits.

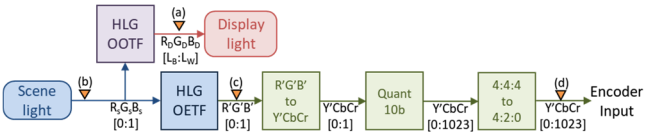


Figure 3. Diagram of pre-processing steps before encoding.


Figure 4. Diagram of post-processing steps after decoding.

### 2.2. Subjective evaluation experiment

We conducted a subjective evaluation experiment based on the double stimulus impairment scale (DSIS) method, Variant I [14].

*2.2.1. Experimental setup*

We used a 4K HDR monitor, EIZO CG-3145 prototype, which supports the HLG method and functions as the display in the diagram of Fig. 1. Table I shows the specifications of the monitor. The viewing distance was 1.5 times the picture height (approx. 0.55 m).

Table I. Specifications of the HDR monitor

| | |
|---|---|
| Size | 31.1-inch liquid crystal display monitor (about 0.70 m wide and 0.37 m high) |
| Output format | 4,096×2,160/10-bit |
| Peak luminance | 1,000 cd/m$^2$ |

An original image and the corresponding image to be evaluated were displayed side by side for 10 s. After that, evaluators inputted a five-grade score:

5 imperceptible
4 perceptible, but not annoying
3 slightly annoying
2 annoying
1 very annoying

To present the images and to input and record the scores, we used Psychtoolbox-3 [15] with a 10-bit framebuffer mode.

16 video experts participated in the experiment, and the evaluation was conducted one person at a time. Considering the order effect, the position of the original reference images was given in different orders to the evaluators: 8 of them received the left side and 8 of them received the right. Each evaluator assessed 110 items: four compressed images with different bit-rates and the original image for each test image. The items were displayed in random order.

*2.2.2. Screening the evaluators*

For the screening of the evaluators, we confirmed the individual mean opinion score (MOS) of the original images and the correlation between MOS and the individual score for all the evaluation items. The individual MOS and the correlation were more than 4.2 and 0.88, respectively. Thus, we concluded that there was no outlier.

### 2.3. Objective quality metrics

We calculated 11 types of HDR objective quality metrics, four metrics with display or scene light inputs, two HLG-based metrics, and weighted PSNR (wPSNR), for the luminance component of 88 compressed images, at four different bit-rates (22 test images at the


Figure 5. 22 HDR test images.

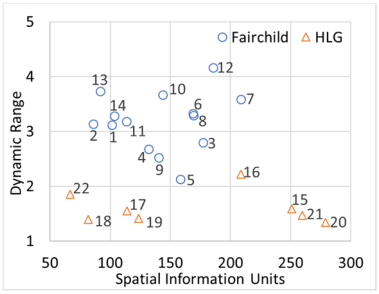


Figure 6. Characteristics of 22 HDR test images.

same bit rate). For the metrics, a tone mapping method such as TF is applied in the first stage.

*2.3.1. HDR-VQM, HDR-VDP-2 and PU_SSIM/MS-SSIM*

We adopted HDR-VQM [4] (VQM), HDR-VDP-2 [5] (VDP2) and MS-SSIM [6] computed in PU space [7] (PU_M) which showed excellent results in Hanhart et al. [2]. In addition, we also used a structural similarity index (SSIM) [16] in PU space (PU_S), which was proposed in [7]. The display configuration on VQM was adjusted to the monitor used for the subjective evaluations.

For the input images, VQM, PU_M, and PU_S apply PU encoding which can transform $10^{-5}$ to $10^{8}$ cd/m$^2$ into perceptually uniform code values and was designed according to a CSF. The red line in Figure 7 shows the relationship between the display light in log and the PU encoded value of the range [$L_B$:$L_W$]. VDP2 employs a CSF based on a simplified version of Barten's CSF [17].

Those metrics were designed to input display light of reference and distorted images in cd/m$^2$. Considering the HLG image coding diagram shown in Fig. 1, we input both the luminance of display and absolute scene light in cd/m$^2$, denoted as $Y_D$ ((a) and (a')) domains in Figs. 3 and 4) and $Y_{AS}$ ((b) and (b') domains in Figs. 3 and 4): $Y_D = 0.2627R_D + 0.6780G_D + 0.0593B_D$ and $Y_{AS} = \alpha Y_S + \beta$, where α, $\beta$ and $Y_S$ are in formulas (1) and (2).

*2.3.2. HLG_SSIM/MS-SSIM*

PU_S and _M consist of SSIM and MS-SSIM after the PU encoding, respectively. The inputs of SSIM and MS-SSIM are comparable to the signal value after TF, (c) and (c') domains in Figs. 3 and 4. Then, we used HLG OETF instead of PU. Since the HLG OETF output is normalized from 0 to 1, we multiplied that by 481.8884, which is the absolute difference of the maximum and minimum signal value of PU encoding in Fig. 7. The scaled HLG curve is shown in the green line in Fig. 7. As Figs. 3 and 4 show, the inputs of HLG_SSIM (HLG_S) and MS-SSIM (HLG_M) are in the (b) and (b') domains.

*2.3.3. wPSNR*

wPSNR is an HDR objective metric used for the international standardization meeting of versatile video coding (VVC), a new video coding scheme, and is defined by the following formulas [18]:

$$\text{wPSNR} = 10 * \log\frac{X^2}{wMSE}$$

$$\text{wMSE} = \sum_{all\ pixels\ i} w_i\big(luma(x_{orig,i})\big) * \big(x_{orig,i} - x_{dec,i}\big)^2$$

where X, w, and $x_{orig}$ and $x_{dec}$ are the maximum pixel value, weight, and original and encoded pixel values, respectively. Figure 8 shows the mapping from a 10-bit luma value to the weight.

In the meeting, wPSNR for HLG sequences is computed after converting the sequences to the Y'CbCr 4:2:0 10-bit PQ format [18]. However, we applied the metric to the Y'CbCr 4:2:0 10-bit HLG format corresponding to the (d) and (d') domains in Figs. 3 and 4.

### 2.4. Curve fitting

To investigate the similarity between an objective quality metric and the results of the subjective evaluation, we conducted curve fitting of the following formula [2] using the least square method.

$$\hat{y} = a + \frac{b}{1 + \exp(-c(x - d))}$$

In the formula, $x$ and $\hat{y}$ are a result of an objective metric and the predicted MOS, respectively. There is the true MOS y corresponding to $x$, and the variables a, b, c, and d are chosen to minimize $\sum_{all\ evaluation\ items\ i}(y_i - \hat{y}_i)^2$.

## 3. RESULTS

Table II shows Pearson linear correlation coefficient (PLCC), root mean square error (RMSE), and Spearman rank order correlation coefficient (SROCC) for each metric. These values were derived from a set of the true MOS y and the predicted MOS $\hat{y}$, as shown in the previous section.

Figure 9 shows a fitted curve and the results of an objective

Table II. Similarity results

| PLCC | | RMSE | | SROCC | |
|---|---|---|---|---|---|
| HLG_M | 0.9276 | HLG_M | 0.4463 | HLG_M | 0.9146 |
| $Y_D$_PU_M | 0.9175 | $Y_D$_PU_M | 0.4751 | $Y_D$_PU_M | 0.9036 |
| $Y_D$_VDP2 | 0.9163 | $Y_D$_VDP2 | 0.4783 | $Y_D$_VDP2 | 0.9026 |
| wPSNR | 0.9126 | wPSNR | 0.4883 | $Y_D$_PU_S | 0.9010 |
| $Y_D$_PU_S | 0.8959 | $Y_D$_PU_S | 0.5307 | wPSNR | 0.8971 |
| HLG_S | 0.8734 | HLG_S | 0.5817 | HLG_S | 0.8743 |
| $Y_{AS}$_PU_S | 0.8613 | $Y_{AS}$_PU_S | 0.6068 | $Y_{AS}$_PU_S | 0.8475 |
| $Y_{AS}$_PU_M | 0.8599 | $Y_{AS}$_PU_M | 0.6097 | $Y_{AS}$_VDP2 | 0.8421 |
| $Y_{AS}$_VDP2 | 0.8460 | $Y_{AS}$_VDP2 | 0.6368 | $Y_D$_VQM | 0.8374 |
| $Y_D$_VQM | 0.8066 | $Y_D$_VQM | 0.7060 | $Y_{AS}$_PU_M | 0.7971 |
| $Y_{AS}$_VQM | 0.7028 | $Y_{AS}$_VQM | 0.8497 | $Y_{AS}$_VQM | 0.7236 |

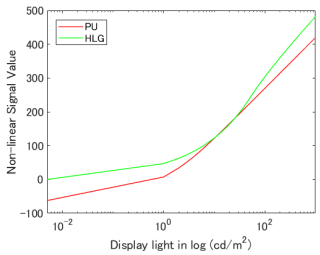


Figure 7. PU and HLG encodings and display light.

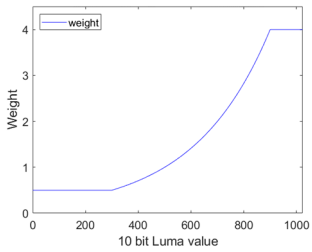


Figure 8. Weight curve of wPSNR for 10-bit images.

metric in the horizontal axis and the corresponding true MOSs in the vertical axis for all 11 metrics. Circle and triangle markers show the results of Fairchild's and HLG test images, respectively, and each test image is plotted in a different color.

## 4. CONSIDERATIONS

Overall, in Table II, HLG_M, $Y_D$_PU_M, $Y_D$_VDP2, wPSNR, and $Y_D$_PU_S show few differences and higher similarity than other metrics in all the PLCC, RMSE, and SROCC. Thus, the five metrics can be referred to as reliable metrics in this validation of HDR image coding with the HLG method.

Regarding VQM, VDP2, PU_S, and PU_M, results of the display light are always better than those of scene light. Therefore, the inputs of the above metrics should be display light, according to the original uses of the metrics.

VDP2 and VQM are the best metrics in Hanhart et al. [2]. In this validation, $Y_D$_VDP2 shows the third best result in all the three types of similarity; however, $Y_D$_VQM shows the lowest reliability except for $Y_{AS}$_ metrics. Considering the good performance of $Y_D$_PU_M and _S, VQM's processing after PU encoding might not be suitable for HLG image coding.

HLG_M shows the best result among all the 11 metrics. This suggests that a metric with a tone mapping method which is equivalent to the TF used for compression can be a reliable metric. HLG_M is always much better than HLG_S as with PU_M and _S. Since there is room to consider more combinations of TFs and post processes, changing the process after HLG OETF may improve performance.

Although wPSNR must be applied after the conversion from HLG to PQ systems, it also works well in the HLG system alone. The reason for this might be related to the weight curve shown in Fig. 8. The sigmoid curve is similar to the response of photoreceptors to light intensity in the human eye [19]. Thus, this curve could emulate human perception, and the metric might show reliability even with a simple SDR metric, like PSNR.

In view of the practical applications, VDP2 has a problem with the processing time. Moreover, the reference input in the display light is outside of the HLG image coding process as shown in (a) in Fig. 3. Therefore, wPSNR as well as HLG_M should be great alternatives because their complexity is much lower than that of VDP2 and their domains of the inputs are along the HLG image coding process.

The results of this performance evaluation show that the ranking of objective metrics for HDR image coding according to their similarity to subjective evaluation depends on the TFs used for compression. Careful assessment is required for comparing different types of TFs in HDR image coding.

## 5. CONCLUSIONS

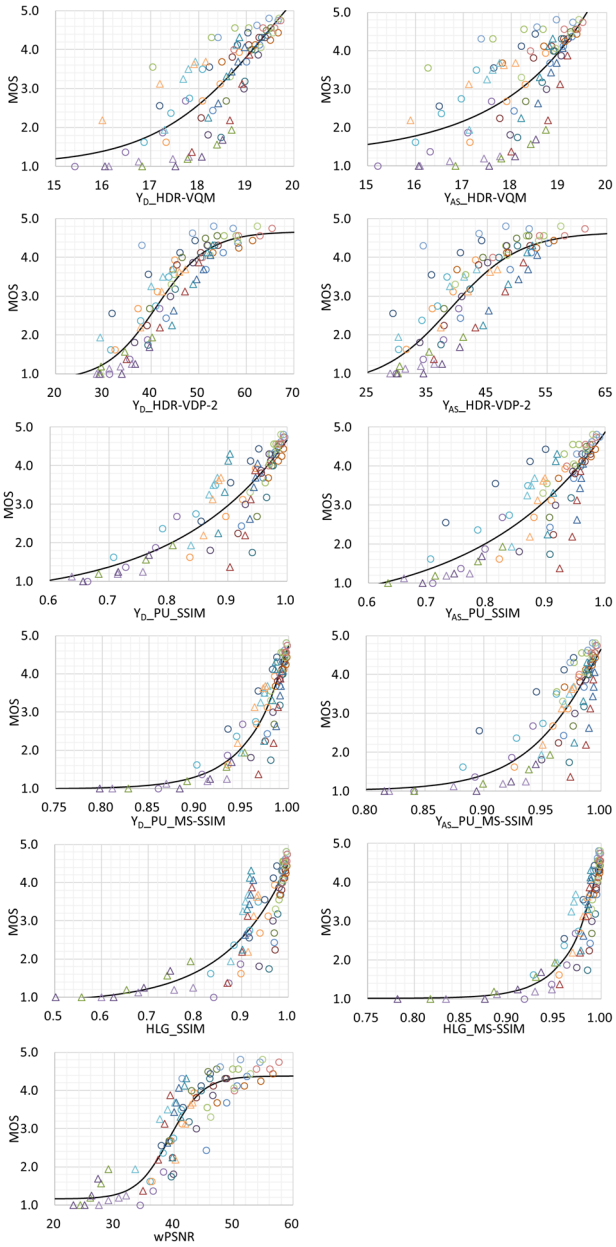


Figure 9. Fitted curve, objective metric and MOS.

We validated 11 objective metrics for HLG-based HDR image coding by calculating the similarity between each metric and the subjective evaluation results. Our experiments show that HLG_MS-SSIM, PU_MS-SSIM, HDR-VDP-2, wPSNR, and PU_SSIM are reliable metrics; however, HDR-VQM shows the lowest performance, in contrast with the very good results it's reported to obtain with CSF-based HDR coding.

We conclude that objective metrics should be mindfully selected based on the tone mapping method used for HDR image coding. Considering the practical applications, HLG_MS-SSIM and wPSNR might be better metrics for HDR image coding with the HLG method.

For future work, we will continue to study the validation with different TFs and objective metrics. In addition, we will consider the performance on color components as well as on luminance.

## 6. REFERENCES


[1] Recommendation ITU-R BT.2100-1, “Image parameter values for high dynamic range television for use in production and international programme exchange,” Jun. 2017.

[2] P. Hanhart, M.V. Bernado, M. Pereira, A.M.G. Pinheiro, and T. Ebrahimi, “Benchmarking of objective quality metrics for HDR image quality assessment,” EURASIP Journal on Image and Video Processing, 2015(1), pp.1-18, Dec. 2015.

[3] R. Mantiuk, S. Daly, and L. Kerofsky, “Display Adaptive Tone Mapping,” Proceedings of ACM SIGGRAPH 2008, ACM, Art.68.1-10, Aug. 2008.

[4] 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, Elsevier, 2015, 35, pp.46-60, Jul. 2015.

[5] 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 Trans. Graph., 2011, vol.30, pp.40:1-40:14, Jul. 2011.

[6] Z. Wang, E.P. Simoncelli, and A.C. Bovik, “Multi-scale Structural Similarity for Image Quality Assessment,” IEEE Asilomar Conf. on Signals Systems, and Computers, pp.1398-1402, Nov. 2003.

[7] T.O. Aydm, R. Mantiuk, H-P. Seidel, “Extending quality metrics to full luminance range images,” Proc. SPIE 6806, Human Vision and Electronic Imaging XIII, 68060B, Feb. 2008.

[8] ISO/IEC 23008-2:2017, “High efficiency coding and media delivery in heterogeneous environments - Part 2: High Efficiency Video Coding,” Oct. 2017. | Recommendation ITU-T H.265 (2018), “High Efficiency Video Coding,” Feb. 2018.

[9] Recommendation ITU-R BT.2020-2, “Parameter values for ultra-high definition television systems for production and international programme exchange,” Oct. 2015.

[10] M.D. Fairchild, “The HDR photographic survey,” Color and Imaging Conference, vol. 2007, no. 1, pp.233-238, Jan. 2007.

[11] Report ITU-R BT.2408-1, “Operational practices in HDR television production,” Apr. 2018.

[12] Recommendation ITU-T P.910, “Subjective video quality assessment methods for multimedia applications,” Apr. 2008.

[13] http://hevc.hhi.fraunhofer.de/

[14] Recommendation ITU-R BT.500-13, “Methodology for the subjective assessment of the quality of television pictures,” Jan. 2012.

[15] M. Kleiner, D.H. Brainard, D. Pelli, A. Ingling, R. Murray, and C. Broussard, “What's new in Psychtoolbox-3,” Perception. 36. pp.1-16, Jan. 2007.

[16] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, Apr. 2004.

[17] P.G.J. Barten, “Contrast sensitivity of the human eye and its effects on image quality,” Eindhoven, Technische Universiteit Eindhoven, 1999.

[18] A. Segall, E. François, and D. Rusanovskyy, “JVET common test conditions and evaluation procedures for HDR/WCG video,” JVET-J1011, Apr. 2018.

[19] C.W. Oyster, “The Human Eye: Structure and Function,” Sinauer Associates, Sunderland, MA, 1999.