Practical Use Suggests a Re-evaluation of HDR Objective Quality Metrics

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*—Full-reference objective quality metrics for high dynamic range (HDR) images must be validated in terms of their consistency with subjective evaluation results, and previous studies have found a number of metrics that appear to correlate well with the preference of observers. However, those conclusions were based on experiments that do not correspond to the practical use of a current day HDR professional production scenario. In this work, we carry out subjective evaluation experiments for the popular HDR standards of perceptual quantization (PQ) and Hybrid Log-Gamma (HLG), with a state-of-the-art HDR reference monitor used in broadcasting and post-production, and where all the observers participating in the tests are video experts. We find that the ranking of HDR metrics is now substantially different from what was reported earlier, and also that a simple standard dynamic range (SDR) metric can be applied directly to PQ or HLG encoded signals providing excellent results that surpass those of HDR metrics.**

Keywords—High dynamic range (HDR), Perceptual quantization (PQ), Hybrid Log-Gamma (HLG), Objective quality metric, Image coding

I. Introduction

High dynamic range (HDR) imaging helps to express better detail in dark areas as well as much brighter highlights, and is becoming an essential technology for video production. For HDR Television (HDR-TV), two different types of HDR methods are standardized, namely, perceptual quantization (PQ) and Hybrid Log-Gamma (HLG), and they define a non-linear transfer function (TF) between luminance and signal for capturing, displaying, recording, compressing, and transmitting purposes [1].

The PQ electro-optical TF (EOTF) was designed according to Barten's contrast sensitivity function (CSF) [2] and translates a non-linear PQ encoded signal value into an absolute display linear light that comes out of a monitor, whereas the HLG opto-electronic TF (OETF) was designed for backward compatibility with standard dynamic range (SDR) displays and translates relative scene linear light captured by a camera into a non-linear HLG signal value. See Figs. 1 and 2 for coding diagrams illustrating these concepts. The Opto-optical TF (OOTF) maps the scene light to the display light and is different for PQ than for HLG, and the codec is composed of a lossy encoding process followed by decoding, and image deterioration is produced.

In image coding, full-reference objective quality metrics, such as the peak signal-to-noise ratio (PSNR), are frequently used to easily measure the quality of a distorted image, relative to an original reference image. Appropriate objective quality metrics should accurately emulate human perception and give results similar to that of a subjective evaluation. Full-reference objective quality metrics for HDR images have been investigated based on their consistency with subjective evaluation results [3]–[6]. According to these works, HDR video quality metric (VQM) [7], HDR visual difference predictor (VDP) [8], the multi-scale structural similarity index (MS-SSIM) [9] computed in perceptual uniform (PU) space [10], and visual information fidelity (VIF) [11] in PU space were concluded as reliable predictors of perceived quality. But all those metrics take as input the absolute display light, and therefore those studies did not consider the current-day practical use scenario where HDR images come in PQ or HLG form, which can be a camera output or a display input, as shown in Figs. 1 and 2.

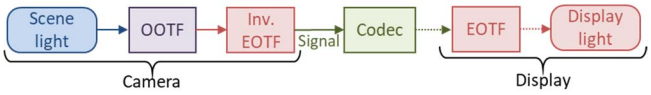

Fig. 1. HDR image coding diagram using PQ method.

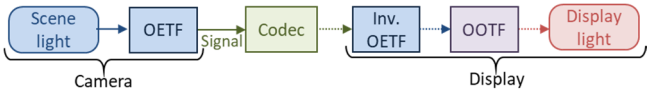

Fig. 2. HDR image coding diagram using HLG method.

In this study, we investigate HDR objective quality metrics in a setting that corresponds to a practical use scenario: this involves not only encoding images in PQ and HLG, but also using a professional HDR reference monitor – employed in broadcasting and post-production – to conduct subjective evaluation experiments where the participants are all video experts. Our main contributions are two. Firstly, we now find that the ranking of HDR metrics is substantially different from what has been reported earlier. And secondly, that a simple SDR metric can be applied directly to PQ or HLG encoded signals providing excellent results that surpass those of HDR metrics.

II. Validation Method

We evaluated the performance of the objective quality metrics using a method equivalent to that used in existing works [4]–[6]. We assessed the validity of our calculation method using a publicly available database [12] and its results [4].

A. HDR Test Images

Fig. 3 shows the thumbnails of the 20 HDR test images. The spatial resolution of the images was set to 1,920 × 1,080 pixels by cropping, and the images were converted to both PQ and HLG images corresponding to the domains written as "Signal" in Figs. 1 and 2, respectively. Fig. 4 describes the

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)

978-1-5386-8212-8/19/$31.00 ©2019 IEEE

Fig. 3. Thumbnails of the 20 HDR test images.

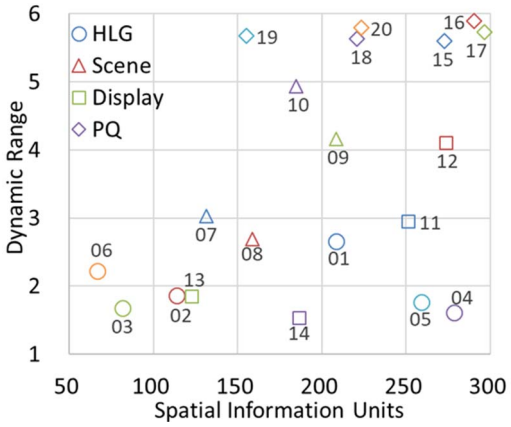

Fig. 4. Characteristics of 20 HDR test images.

characteristics of the 20 test images. The vertical axis indicates the 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 [13]. For the calculation, luminance values of HLG images were used. The graph indicates that the test images have a wide coding complexity.

1) Types of HDR images

To evaluate various cases, we selected the images from four HDR image data sets, with different HDR image formats. Some of them were converted to BT.2020 color space [14] to fit the HLG and PQ methods.

a) HLG native images

Six images in Fig. 3 from 1 to 6 are native HLG images in accordance with the HDR-TV production guideline [15]; that is, 75% of the nominal signal level corresponds to a diffuse white. The original image format is in the HLG-encoded value.

b) Scene-referred images

Four images in Fig. 3, 7 to 10, are Fairchild's HDR photos [16], and the original image format is in the absolute scene light captured by a camera. We converted these images into the relative scene light in the BT.2020 color space using a color space conversion matrix provided in the paper.

c) Display-referred images

Four images in Fig. 3, 11 to 14, are the Zurich Athletics 2014 test sequences [17], and the original image format is in a normalized display light. We regarded the signal value 1 as 4,000 cd/m^2 and the color space as BT.709 [18] because technical information indicated that the images were graded on a SIM2 HDR47ES4MB display in "max" brightness mode. Then, we converted them into absolute display light in the BT.2020 color space.

d) PQ native images

Six images in Fig. 3, 15 to 20, are the color-graded version of the HdM-HDR-2014 content [19], and the original image format is in the PQ-encoded value. The images were graded from 0.005 to 4,000 cd/m^2.

2) Conversion to PQ and HLG images

Fig. 5 and Fig. 6 show the diagrams of the conversions to PQ and HLG images, respectively. The domains denoted by a, b, c, and d in the diagrams correspond to the types of HDR images written in the previous section. PQ EOTF, HLG OETF, HLG OOTF, and their inverse TFs described in BT.2100 [1] were used for the conversion.

a) PQ EOTF

PQ EOTF maps the non-linear normalized PQ signal E' composed of {R', G', B'} in the range [0:1] into absolute display linear light F_D in cd/m^2 composed of {R_D, G_D, B_D} in the range [0:10000] using (1):

$$F_D = 10000 \cdot Y \tag{1}$$

where Y is given by (2):

$$Y = \left(\frac{\max[(E'^{1/m_2} - c_1), 0]}{c_2 - c_3 E'^{1/m_2}} \right)^{1/m_1} \tag{2}$$

and $m_1 = 0.1593017578125$, $m_2 = 78.84375$, $c_1 = 0.8359375$, $c_2 = 18.8515625$, and $c_3 = 18.6875$.

b) HLG OETF

HLG OETF maps the relative scene linear light E composed of {R_S, G_S, B_S} in the range [0:1] into the non-linear HLG signal value E' composed of {R', G', B'} in the range [0:1] using (3):

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

where $a = 0.17883277, b = 1 - 4a,$ and $c = 0.5 - a \cdot \ln(4a)$.

Fig. 5. Conversion of HDR image into PQ image.

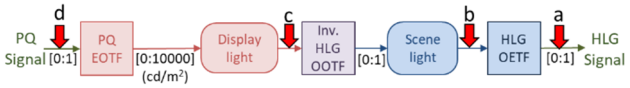

Fig. 6. Conversion of HDR image into HLG image.

c) HLG OOTF

HLG OOTF maps the relative scene linear light E composed of $\{R_S, G_S, B_S\}$ in the range [0:1] into the absolute display linear light F_D in cd/m^2 composed of $\{R_D, G_D, B_D\}$ in the range $[0:L_W]$ using the (4) – (7):

$$R_D = \alpha Y_S^{\gamma-1} R_S \quad (4)$$

$$G_D = \alpha Y_S^{\gamma-1} G_S \quad (5)$$

$$B_D = \alpha Y_S^{\gamma-1} B_S \quad (6)$$

$$Y_S = 0.2627 R_S + 0.6780 G_S + 0.0593 B_S \quad (7)$$

where $\alpha = L_W$ is the nominal peak luminance of a display and γ is the system gamma determined by L_W: $\gamma = 1.2 + 0.42 \log_{10}(L_W/1000)$. In this study, we configured the parameters as $L_W = 1{,}000$ for the HLG and scene-referred images and as $L_W = 4{,}000$ for the PQ and display-referred images.

B. HDR Image Coding Experiment

We performed an HDR image coding experiment for PQ and HLG images to prepare various distorted images. For the image coding, the high efficiency video coding (HEVC)/H.265 [20] Main 10 Profile, which is employed for HDR broadcasting [21], and versatile video coding (VVC), which is a new video coding method considered to be the subsequent standard of HEVC, were used.

1) Pre- and post-processing

The image format of the encoder input and decoder output was Y'CbCr 4:2:0 10-bit. For the pre-processing of the encoder, the PQ- or HLG-encoded signal in the BT.2020 R'G'B' was transferred to Y'CbCr, which has one luma and two chroma components. Then, the pixel values from 0 to 1 were amplified to 64–940 (for luma component Y') or 64–960 (for chroma components Cb and Cr) in an integer format to treat them as 10-bit precision values. Finally, the chroma components Cb and Cr are subsampled to a 4:2:0 format. Their image sizes were decreased to half of the original image both horizontally and vertically. The post-processing phase after decoding follows the inverse order of the pre-processing stage. For the color space conversion, the 10-bit quantization, 4:2:0 chroma subsampling, and the inverse processing, we used the procedures in accordance with the HEVC common test conditions for the HDR images [22].

2) Encoding condition

a) HEVC

We utilized the HEVC Test Model (HM) [23] version 16.19 encoder with all intra Main 10 and fixed QP settings. A total of 20 PQ and HLG images were encoded at the target bit rates of 100, 200, 300, and 400 kbits considering bit rates of intra pictures on broadcasting. For the PQ images, configurations to change the quantization parameter (QP) calculation methods of luma and chroma components were set based on the common test conditions [22]. The settings include the scale factors of chroma components depending on capture and representation color space. To set the factors, we treated the capture color space of the test images from 1 to 6, from 7 to 14, and from 15 to 20 as BT.2020 [14], BT.709 [18], and DCI-P3 [24], respectively. A total of 80 types (20 image × 4 bit rate) of HEVC-encoded images were generated for each PQ and HLG images.

a) VVC

We used the VVC Test Model (VTM) [25] version 3.0 encoder with all intra-conditions at 10-bit precision and fixed QP settings. Ten PQ and HLG images (test images 1, 3, 5, 8, 10, 12, 14, 15, 17, and 19) were encoded at the same target bit rates as those of HEVC: 100, 200, 300, and 400 kbits. For PQ images, we configured the same parameters for the luma and chroma QP calculation methods as that of HEVC based on the VVC common test condition for HDR images [26]. A total of 40 types (10 image × 4-bit rate) of VVC-encoded images were generated for each PQ and HLG images.

C. Subjective Evaluation Experiment

We performed a subjective evaluation experiment referring to ITU-R BT.500 [27].

1) Experimental setup

We used a 4K HDR reference monitor (i.e., EIZO CG-3145 [28] prototype), which supports both the PQ and HLG methods and functions as "Display" in the diagrams of Figs. 1 and 2, respectively. TABLE I presents the monitor specifications. The monitor can display an all-white background at 1,000 cd/ m^2 and control the brightness by pixel. These features are similar to the latest 31-in. 4K professional master monitor (i.e., Sony BVM-HX310 [29]).

The viewing environment was set according to TABLE 3 of BT.2100 [1], which establishes a reference viewing environment for the critical viewing of HDR program material or completed programs to provide repeatable results from one facility to another when viewing the same material. TABLE II illustrates the viewing environment of the experiment. In this experiment, we focused on the luminance reproducibility: e.g., a signal value corresponding to 800 cd/m^2 should be displayed at 800 cd/m^2. The monitor displayed several PQ images (11–20) after clipping at the peak luminance, 1,000 cd/m^2.

The presentation method was based on the simultaneous double stimulus for continuous evaluation (SDSCE) method [27]. As shown in Fig. 7, an original image and the corresponding image to be evaluated were displayed side by side on a mid-gray background (approximately 50 cd/m^2) for 10 s. Considering the order effect, the position of the original reference images was given in different orders to the subjects, of which half of them received the left side and the other half received the right side. Thereafter, the subjects inputted a five-grade score based on the double stimulus impairment scale (DSIS) method [27], (5, imperceptible; 4, perceptible, but not annoying; 3, slightly annoying; 2, annoying; and 1, very annoying). To present the images and input and record the

TABLE I. SPECIFICATIONS OF THE HDR MONITOR

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

TABLE II. VIEWING ENVIRONMENT OF THE EXPERIMENT

Surround and periphery	Neutral grey at D65
Luminance of surround	5 cd/ m^2
Luminance of periphery	≤ 5 cd/ m^2
Ambient lighting	Avoid light falling on the screen
Viewing distance	1.5 picture height (approx. 0.55 m)
Peak luminance of display	1,000 cd/ m^2
Minimum luminance of display (black level)	0.005 cd/ m^2

Fig. 7. Image presentation method.

scores, we used Psychtoolbox-3 [30] with a 10-bit frame buffer mode.

A total of 16 video experts who often see HDR contents participated in the experiment, which is composed of two 30-min sessions. Between the sessions, they took a break of at least 30 min. The evaluation was conducted one person at a time. Prior to the start of the experiment, verbal instructions to show the evaluation method were provided to the subjects, and then they tested the training samples to become familiar with the operation. For these samples, 3 HDR images out of the 20 test images were used. The first and second sessions were for HLG and PQ images, respectively (this information was not provided to the subjects prior to the experiment), and 130 items, including 10 different original images, were assessed per session. These items were randomly displayed.

2) Subjective evaluation results

For the screening of the subjects, we confirmed the individual mean opinion score (MOS) of the original images and the correlation between the MOS and individual score for all the evaluation items. The individual MOS was between 4.55 and 5.00, whereas the correlation was between 0.86 and 0.94. Thus, no outlier was present.

Fig. 8 shows the MOS distribution of the encoded images. The distribution spread evenly aside from the values of 4.5–5.

D. Objective Quality Metrics

We calculated ten types of objective quality metrics for each PQ and HLG images.

1) Metrics for the display light

We adopted HDR-VQM [7], HDR-VDP-2 [8], MS-SSIM [9] computed in PU space [10] (PU_MS-SSIM), and VIF [11] computed in PU space (PU_VIF), which obtained excellent results in previous works [3]–[6]. For these HDR metrics, a tone mapping method such as TF is applied in the first stage. PU, which is used for HDR-VQM, PU_MS-SSIM, and PU_VIF, is a TF that transforms the luminance of 10^{-5} to 10^{8} cd/m^2 into perceptually uniform code values and was

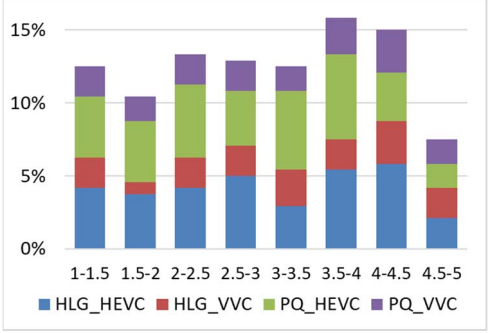

Fig. 8. MOS distribution of the encoded images.

designed according to a CSF. HDR-VDP-2 employs a CSF based on a simplified version of Barten's CSF [2]. These four metrics were designed to input the display light of the reference and distorted images in cd/m^2. We calculated the display light from the PQ and HLG images in the same manner as that shown in Figs. 5 and 6. By considering the display used for the subjective experiment, we clipped the inputs of the metrics from 0.005 to 1,000 cd/m^2.

2) Metrics for the PQ- and HLG-encoded signals

In practical applications, the preferred inputs of the metrics are the PQ- or HLG-encoded signals, which can be a camera output and a display input. Then, we applied MS-SSIM and VIF to the PQ and HLG signals, which are equivalent to the MS-SSIM and VIF after inverse PQ EOTF or HLG OETF, instead of PU TF. We also employed the MS-SSIM and VIF after transforming the PQ/HLG signal into the HLG/PQ signal in the same manner as that shown in Figs. 5 and 6. For example, we calculated VIF in HLG for the PQ signal in addition to VIF in PQ.

To calculate combined results of both PQ and HLG images, we applied three types of MS-SSIM and VIF: MS-SSIM/VIF in PQ, in HLG, and in each domain. For instance, in "VIF in PQ/HLG," VIF was calculated after HLG/PQ images were transformed into the PQ/HLG signal. On the other hand, in "VIF in each domain," VIF was calculated with the original PQ and HLG signals without any conversion.

E. Curve Fitting

To investigate the similarity between the objective quality metric and the results of the subjective evaluation, we conducted curve fitting of the following logistic function [4] using the least square method, shown in (8):

$$\hat{y} = a + \frac{b}{1+\exp(-c(x-d))} \tag{8}$$

where x and $\hat{y}$ are the result of the objective metric and predicted MOS, respectively. The true MOS y corresponding to x exists. The variables a, b, c, and d are selected to minimize $\sum_{all\ evaluation\ items\ i}(y_i - \hat{y}_i)^2$. The number of items was 120 for the PQ or HLG image and 240 for the combination of PQ and HLG images.

III. Results

TABLES III, IV, and V present the Pearson linear correlation coefficient (PLCC), root-mean-square error (RMSE), and Spearman rank order correlation coefficient (SROCC) for PQ, HLG, and the combination of PQ and HLG images, respectively. These values were derived from a set of the true MOS y and predicted MOS $\hat{y}$, as presented in the previous section. Each yardstick on the right side of the values marks the minimum distance of a statistically significant difference at a significance level of 0.05. (e.g., the PLCC results in TABLE III indicate that VIF in HLG is significantly better than HDR-VDP-2, MS-SSIM in HLG, and HDR-VQM; HDR-VQM is significantly worse than other metrics.)

Fig. 9 illustrates the fitted curve, the calculation results of an objective metric in the horizontal axis, and the corresponding true MOSs in the vertical axis for the combination of PQ and HLG images corresponding to the results of TABLE V. The shapes and colors of the markers correspond to the images shown in Fig. 4. The edge types of

TABLE III. RESULTS FOR PQ IMAGES

Metric	PLCC	RMSE	Metric	SROCC
VIF in HLG	**0.9568**	**0.3157**	VIF in HLG	**0.9550**
PU_VIF	0.9444	0.3570	PU_VIF	0.9429
VIF in PQ	0.9405	0.3689	VIF in PQ	0.9377
MS-SSIM in PQ	0.9305	0.3978	MS-SSIM in PQ	0.9276
PU_MS-SSIM	0.9303	0.3982	PU_MS-SSIM	0.9265
HDR-VDP-2	0.9204	0.4245	MS-SSIM in HLG	0.9237
MS-SSIM in HLG	0.9188	0.4285	HDR-VDP-2	0.9198
HDR-VQM	0.8176	0.6251	HDR-VQM	0.8073

TABLE IV. RESULTS FOR HLG IMAGES

Metric	PLCC	RMSE	Metric	SROCC
PU_MS-SSIM	**0.9344**	**0.3911**	VIF in HLG	**0.9341**
VIF in HLG	0.9336	0.3934	PU_MS-SSIM	0.9241
MS-SSIM in PQ	0.9252	0.4166	MS-SSIM in HLG	0.9191
PU_VIF	0.9200	0.4303	MS-SSIM in PQ	0.9167
MS-SSIM in HLG	0.9180	0.4354	PU_VIF	0.9157
HDR-VDP-2	0.9137	0.4461	HDR-VDP-2	0.9100
VIF in PQ	0.9089	0.4578	VIF in PQ	0.9047
HDR-VQM	0.7706	0.6996	HDR-VQM	0.7650

TABLE V. COMBINED RESULTS FOR BOTH PQ AND HLG IMAGES

Metric	PLCC	Metric	RMSE	Metric	SROCC
VIF in HLG	**0.9442**	VIF in HLG	**0.3599**	VIF in HLG	**0.9433**
VIF in each domain	0.9347	VIF in each domain	0.3886	VIF in each domain	0.9323
PU_VIF	0.9313	PU_VIF	0.3980	PU_VIF	0.9288
PU_MS-SSIM	0.9305	PU_MS-SSIM	0.4002	PU_MS-SSIM	0.9242
MS-SSIM in PQ	0.9263	MS-SSIM in PQ	0.4118	MS-SSIM in HLG	0.9209
VIF in PQ	0.9238	MS-SSIM in each domain	0.4222	VIF in PQ	0.9213
MS-SSIM in each domain	0.9224	VIF in PQ	0.4256	MS-SSIM in PQ	0.9206
MS-SSIM in HLG	0.9177	MS-SSIM in HLG	0.4342	MS-SSIM in each domain	0.9217
HDR-VDP-2	0.9154	HDR-VDP-2	0.4400	HDR-VDP-2	0.9150
HDR-VQM	0.7929	HDR-VQM	0.6659	HDR-VQM	0.7866

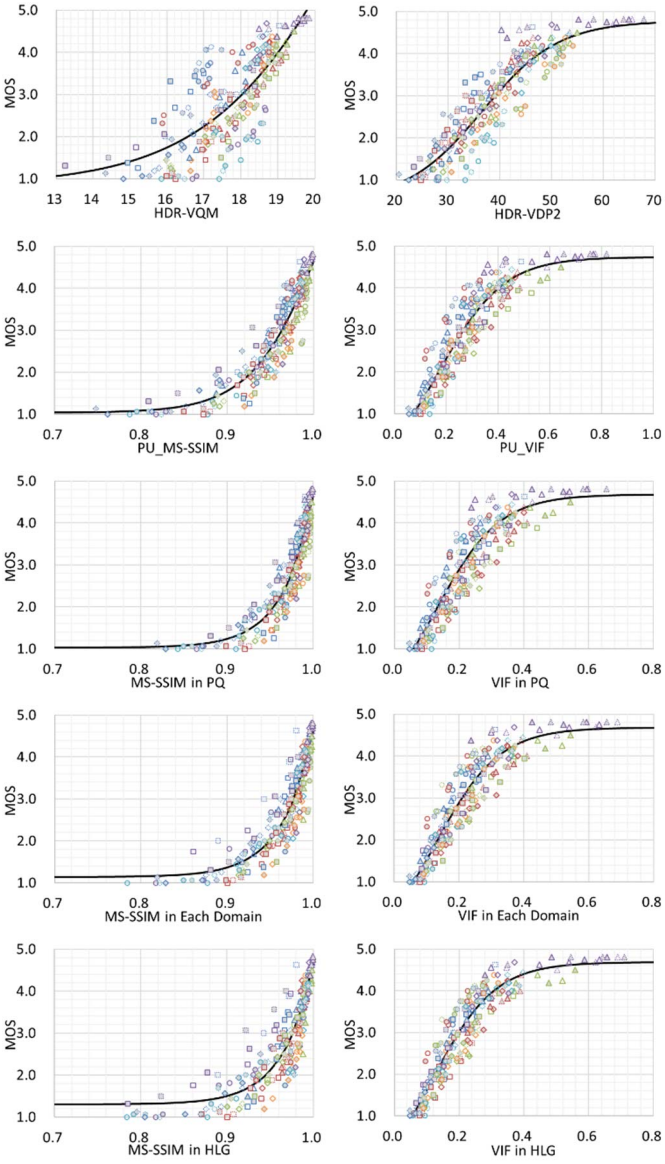

Fig. 9. Fitted curve, objective metric, and MOS for combined results.

the markers correspond to the types of the video coding methods, namely, HEVC (solid line) and VVC (dotted line). Meanwhile, the marker fill colors correspond to the HDR methods, namely, PQ (filled in gray) and HLG (filled in white).

IV. CONSIDERATIONS

Generally, the VIF calculated in the HLG signal exhibited excellent results. Considering that excellent results are obtained in both PQ and HLG images, an accurate comparison between the different types of HDR methods and this metric is possible. In TABLE V, the VIF calculated in each PQ and HLG domain without any conversion showed the second-best results. In a practical application, this type of metric is very useful.

Although HDR-VQM showed excellent results in the previous studies [4], [6], the validation in this study was significantly worse than the other metrics. Also, in some cases, HDR-VDP-2 was significantly worse than VIF in HLG. One reason for these may be the differences in the subjective evaluation conditions. In the previous works, a SIM 2 47-in. display was used for the subjective evaluations [4]–[6], and the subjects were non-experts. Meanwhile, for this evaluation an HDR reference monitor was used, and the subjects were video experts who were familiar with HDR images. In these types of studies, subjective evaluation results are treated as the ground truth. Therefore, accurately performing a subjective evaluation experiment to meet requirements and use cases is considerably important. For example, when an HDR reference monitor which can display a PQ signal up to 4,000 cd/m^2 is released, we should conduct subjective evaluation experiments again using the monitor.

Similar to the technique used in the previous study [4], PU_MS-SSIM exhibited good results. The MS-SSIM is calculated by weighting the results of the SSIM [31] for multi-scale (down-converted) five-layered images. We confirmed the multi-scaling effect. TABLE VI shows the results, where M indicates the level of the multi-scaling; that is, M = 1 is the original image size, M = 2 is half size of the original image (M = 1) in both horizontal and vertical directions, and M = 3, M = 4, and M = 5 are half sizes of

TABLE VI. RESULTS OF PU_SSIM FOR MULTI-SCALE IMAGES

	PQ			HLG		
	PLCC	***SROCC***	***RMSE***	***PLCC***	***SROCC***	***RMSE***
PU_SSIM (M = 1)	0.8941	0.8817	0.4863	0.8766	0.8658	0.5282
PU_SSIM (M = 2)	**0.9341**	**0.9294**	**0.3875**	**0.9389**	**0.9304**	**0.3780**
PU_SSIM (M = 3)	0.9243	0.9222	0.4142	0.9278	0.9186	0.4096
PU_SSIM (M = 4)	0.9057	0.9053	0.4601	0.8920	0.8809	0.4963
PU_SSIM (M = 5)	0.8743	0.8765	0.5270	0.8271	0.8120	0.6171

images in M = 2, M = 3, and M = 4, respectively. The results showed that M = 2 is significantly better than M = 1 and that M = 3 and M = 4 are better than M = 1. The corresponding results of the MS-SSIM also exhibited a similar tendency. From this, down-converting input images may be effective to emulate human perception, achieving a global image structure. The VIF, which showed excellent results in this experiment, also employs a multi-scaling type, that is, a steerable pyramid.

V. CONCLUSIONS

We have carried out an evaluation of 10 types of full-reference objective quality metrics for HDR images in a professional setting, practical use scenario. Our results show that the ranking of HDR metrics is now quite different from what was reported in previous studies. And also, that the SDR metric VIF applied directly on the PQ or HLG signal provides an excellent result and outperforms the best HDR metrics, suggesting that VIF can be very useful for HDR in practice (as it does not require performing signal conversions).

We are currently extending our work in several directions, including the development of a novel HDR metric based on vision models and extending the work to video.

ACKNOWLEDGMENT

The authors would like to thank Enago (www.enago.jp) for the English language review.

REFERENCES

[1] Recommendation ITU-R BT.2100-2, "Image parameter values for high dynamic range television for use in production and international programme exchange," 2018.

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

[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," International Conference on Multimedia Signal Processing (ICMSP), Venice, Italy, 2014.

[4] 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, vol. 2015, no. 1, 2015, pp.1-18.

[5] T. Vigier, L. Krasula, A. Milliat, M. Pereira da Silva, and P. Le Callet, "Performance and robustness of HDR objective quality metrics in the context of recent compression scenarios," Digital Media Industry and Academic Forum 2016, Santorini, Greece, 2016, pp.59-64.

[6] E. Zerman, G. Valenzise, and F. Dufaux, "An extensive performance evaluation of full-reference HDR image quality metrics," Quality and User Experience, vol. 2, no. 1, 2017, p. 5.

[7] 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.

[8] 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.

[9] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multi-scale Structural Similarity for Image Quality Assessment," The 37th Asilomar Conf. on Signals, Systems, and Computers, 2003, Pacific Grove, CA, USA, 2003, pp.1398-1402.

[10] 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, 2008.

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

[12] P. Korshunov, P. Hanhart, T. Richter, A. Artusi, R. Mantiuk, and T. Ebrahimi, "Subjective quality assessment database of HDR images compressed with JPEG XT," 7th International Workshop on Quality of Multimedia Experience (QoMEX), Costa Navarino, Messinia, Greece, 2015.

[13] Recommendation ITU-T P.910, "Subjective video quality assessment methods for multimedia applications," 2008.

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

[15] Report ITU-R BT.2408-1, "Operational practices in HDR television production," 2018.

[16] M.D. Fairchild, "The HDR photographic survey," 15th Color and Imaging Conference (CIC 2007), Albuquerque, NM, USA, vol. 2007, no. 1, pp.233-238, 2007.

[17] "EBU Technology & Innovation – Zurich athletics," Internet: https://tech.ebu.ch/testsequences/zurich_athletics [Jan. 19, 2019]

[18] Recommendation ITU-R BT.709-6, "Parameter values for the HDTV standards for production and international programme exchange," 2015.

[19] J. Froehlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling, and H. Brendel, "Creating cinematic wide gamut HDR video for the evaluation of tone mapping operators and HDR-displays," Proc. SPIE 9023, Digital Photography X, 90230X, 2014.

[20] 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," 2018.

[21] ARIB STD-B32 Version 3.11-E1 (Fascicle 1), "Video Coding, Audio Coding, and Multiplexing Specifications for Digital Broadcasting," 2018.

[22] E. François, J. Sole, J. Ström, and P. Yin "Commmon Test Conditions for HDR/WCG video coding experiments," JCTVC-Z1020, 2017.

[23] "High Efficiency Video Coding (HEVC) | JCT-VC," Internet: http://hevc.hhi.fraunhofer.de/ [Jan. 19, 2019]

[24] RP 431-2:2011, "- SMPTE Recommended Practice - D-Cinema Quality — Reference Projector and Environment," 2011.

[25] "Versatile Video Coding (VVC) | JVET," Internet: https://jvet.hhi.fraunhofer.de/ [Jan. 19, 2019]

[26] A. Segall, E. François, S. Iwamura, and D. Rusanovskyy, "JVET common test conditions and evaluiation procedures for HDR/WCG video," JVET-L1011, 2018.

[27] Recommendation ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," 2012.

[28] "HDR Reference Monitor ColorEdge Prominence CG3145 – EIZO," Internet: https://www.eizo.com/products/coloredge/cg3145/ [Jan. 19, 2019]

[29] "BVM-HX310," Internet: https://pro.sony/en_GB/products/broadcastpromonitors/bvm-hx310 [Jan. 19, 2019]

[30] M. Kleiner, D.H. Brainard, D. Pelli, A. Ingling, R. Murray, and C. Broussard, "What's new in Psychtoolbox-3," Perception, vol. 36, no. 14, 2007, pp.1-16.

[31] 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, 2004.