SMPTE 2020 ATC Manuscript template

A Benchmark of Objective Quality Metrics for HLG-Based HDR/WCG Image Coding

Yasuko Sugito

NHK, Tokyo, Japan

Trevor Canham

Universitat Pompeu Fabra, Barcelona, Spain

Javier Vazquez-Corral

Universitat Pompeu Fabra, Barcelona, Spain

Marcelo Bertalmío

Universitat Pompeu Fabra, Barcelona, Spain

Written for presentation at the

SMPTE 2020 Annual Technical Conference & Exhibition

Abstract. *In this work, we study the suitability of high dynamic range, wide color gamut (HDR/WCG) objective quality metrics to assess the perceived deterioration of compressed images encoded using the Hybrid Log-Gamma (HLG) method, which is the standard for HDR television.*

Several image quality metrics have been developed to deal specifically with HDR content, although in previous work we showed that the best results (i.e., better matches to the opinion of human expert

observers) are obtained by an HDR metric that consists simply in applying a given standard dynamic range metric, called visual information fidelity (VIF), directly to HLG-encoded images.

However, all these HDR metrics ignore the chroma components for their calculations, e.g., they just consider the luminance channel. For this reason, in the current work, we conduct subjective evaluation experiment in a professional setting using compressed HDR/WCG images encoded with HLG and analyze the ability of the best HDR metric to detect perceivable distortions in the chroma components, as well as the suitability of popular color metrics (including ΔITP_R, which supports parameters for HLG) to correlate with the opinion scores. Our first contribution is to show that there is a need to consider the chroma components in HDR metrics, as there are color distortions that subjects perceive but that the best HDR metric fails to detect. Our second contribution is the surprising result that VIF, which utilizes only the luminance channel, correlates much better with the subjective evaluation scores than the metrics that do consider the color components.

Keywords. *High dynamic range (HDR), wide color gamut (WCG), Hybrid Log-Gamma (HLG), objective quality metric*

Introduction

Full-reference objective quality metrics, such as the peak signal-to-noise ratio (PSNR), are frequently used to assess a degree of deterioration in a compressed image relative to the original image. Good objective metrics are well correlated to subjective assessment results and can be a cost-saving alternative to subjective evaluations.

High dynamic range (HDR) technology is capable of reproducing much brighter highlights and much darker details than what can be achieved with standard dynamic range (SDR) and is becoming essential in video production. An important technical feature of HDR imaging is the transfer function (TF) that determines the conversion from luminance values (which cover a very wide range, e.g. from 0.005 cd/m^2 to 1,000 cd/m^2) to non-linear signal values, which is required in order to capture, display, and compress images effectively. A popular TF for HDR television, standardized in Rec. BT.2100[1] by the International Telecommunications Union – Radiocommunication (ITU-R) and used in HDR broadcasting in the UK, the US, and Japan, is the Hybrid Log-Gamma (HLG) opto-electronic transfer function (OETF). A number of objective metrics specifically dedicated to HDR image coding have been developed, such as the HDR visual difference predictor (HDR-VDP)[2] and the HDR video quality metric (HDR-VQM).[3] In previous works,[4-5] we showed that these metrics are not too accurate for images encoded with HLG, and that the best results (i.e. better matches to the opinion of human expert observers) are obtained by an HDR metric that consists simply in applying a given standard dynamic range metric, called visual information fidelity (VIF),[6] directly to HLG-encoded images.

Wide color gamut (WCG) technology allows for the reproduction of very vivid colors that fall outside the standard color gamut (SCG) of traditional television, Rec. BT.709.[7] It is key to the advancement of realistic image representation, and is commonly associated with HDR given that brighter displays can produce more saturated colors prescribed in Rec. BT.2020.[8] HDR-specific metrics like the ones mentioned above only consider the luminance component, and although some works in the literature have investigated objective quality metrics for HDR/WCG images taking into account the chroma channels,[9-11] to the best of our knowledge, HDR/WCG metrics suited for HLG images have not been introduced.

For this reason, in the present study, we investigate the suitability of existing objective quality metrics for the case of compressed HDR/WCG images encoded with HLG. We conduct subjective evaluation experiments in a professional setting and analyze the ability of the best HDR metric to detect perceivable distortions in the chroma components, as well as the suitability of popular color metrics to correlate with the opinions of observers; this includes the latest color difference metric, ΔITP_R, which supports parameters for HLG and is described in Rec. BT.2124.[12]

Our first contribution is to show that there is a need to consider the chroma components in HDR metrics, as there are color distortions that subjects perceive but that even the best HDR metric fails to detect. Our second contribution is the surprising result that VIF, although it's applied only on the luminance channel, correlates much better with the subjective evaluation scores than the metrics that do consider the color components.

Validation Methodology

Following the methodology of our previous work about HDR objective metrics,[5] here we select a varied set of HLG-encoded HDR/WCG test images, compress them at different levels, perform a subjective evaluation experiment with expert observers to assess the quality of the compressed images, and finally analyze the performance of objective quality metrics in reference to the subjective evaluation results. That is, the same distorted images are assessed both subjectively and objectively. In order to look into the effect of color deteriorations, we evaluated additional synthesized images that have distortion of only either the luma component or the chroma components.

HDR/WCG Test Images

Figures 1 and 2 represent the thumbnails and chromaticity diagrams of twenty HDR/WCG test images, and Table 1 shows specifications of them.

Table 1. Specifications of the HDR/WCG Test Images.

Test image	HDR type	Source
01–06	HLG native	Our experimental content (a diffuse white level is in accordance with the HDR-TV production guideline[13])
07–10	Scene referred	Fairchild's HDR photos[14]
11–14	Display referred	Zurich Athletics 2014 test sequence[15]
15–20	perceptual quantization (PQ)[1] native	HdM-HDR-2014 content[16]

The spatial resolution of the considered images is set to 1,920 × 1,080 pixels by cropping, and then, the images are converted into a nonlinear HLG R'G'B' signal. In each diagram, the largest triangle with the black line borders indicates the WCG of BT.2020,[8] whereas the smallest triangle with the black dotted line borders corresponds to the SCG of BT.709.[7] The middle triangle with the white line borders indicates the color gamut of a 4K HDR/WCG liquid crystal display (LCD) reference monitor (namely, EIZO CG-3145[17]) used in the subjective evaluation experiment. It uses an HLG signal as an input. Each primary color expressed as a Commission International de l'Éclairage (CIE) xy value is derived based on the actual measured value. Moreover, we sample pixel values for every portion of 64 × 64 pixels, corresponding to a 1° field-of-view angle, corresponding to each image. Then, we convert the HLG R'G'B outputs into the xy values and plot the chromaticity values using black-bordered circles. The diagrams represent that 9 out of 20 images have the color gamut that exceeds SCG, and most of them are displayable using the 4K HDR/WCG monitor. Figure 3 illustrates dynamic range and spatial information corresponding to twenty test images. The vertical axis indicates the dynamic range log (L_{max}/L_{min}), where L_{max} and L_{min} are the maximum and minimum luminance after excluding 1% of the brightest and darkest pixels, respectively. The horizontal axis represents the spatial perceptive information obtained from Rec. BT.500.[18] Overall, the resulting graph and diagrams indicate that the test images have a wide coding complexity.

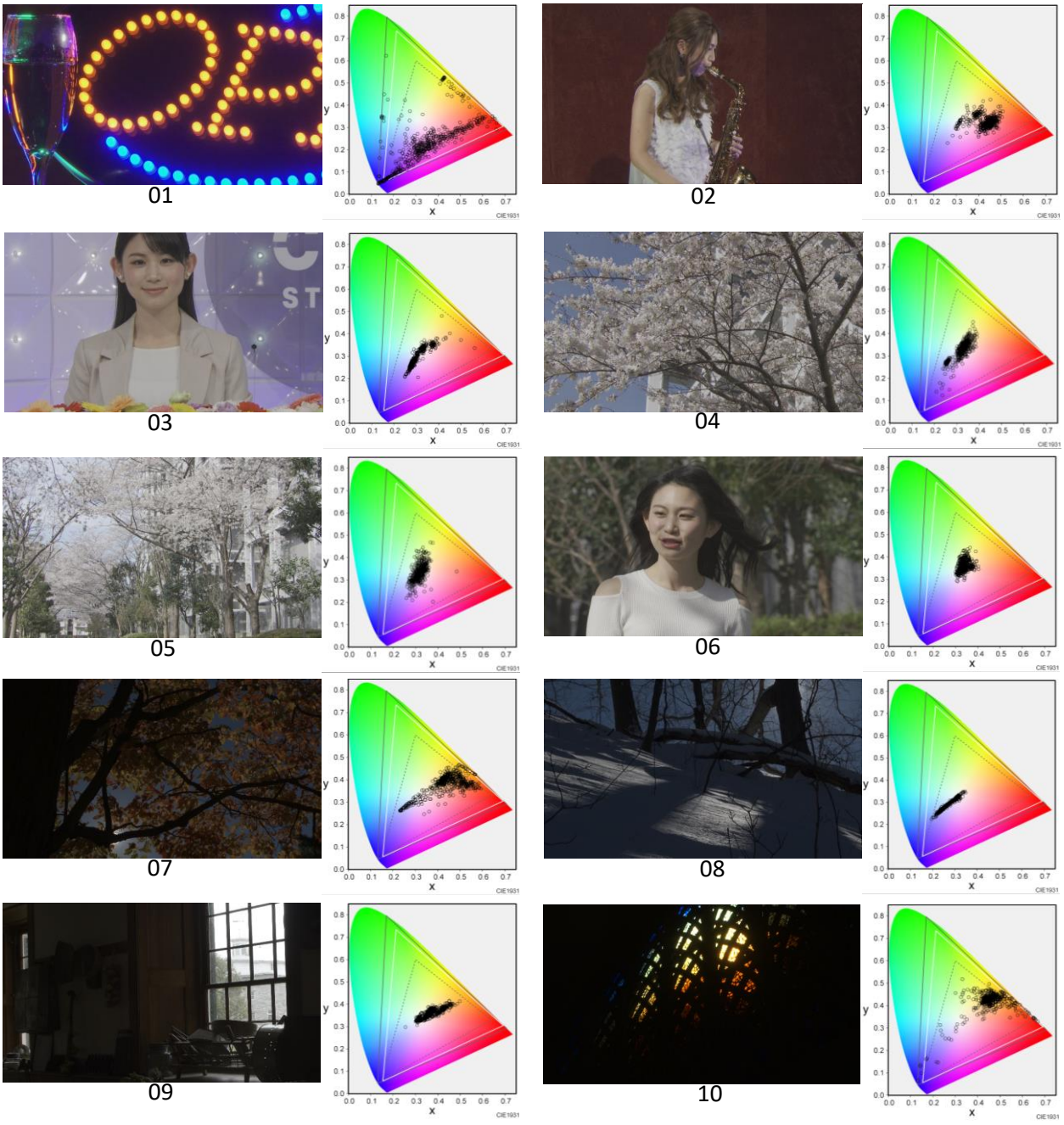


Figure 1. Thumbnails and chromaticity diagrams of the HDR/WCG test images (from 01 to 10).

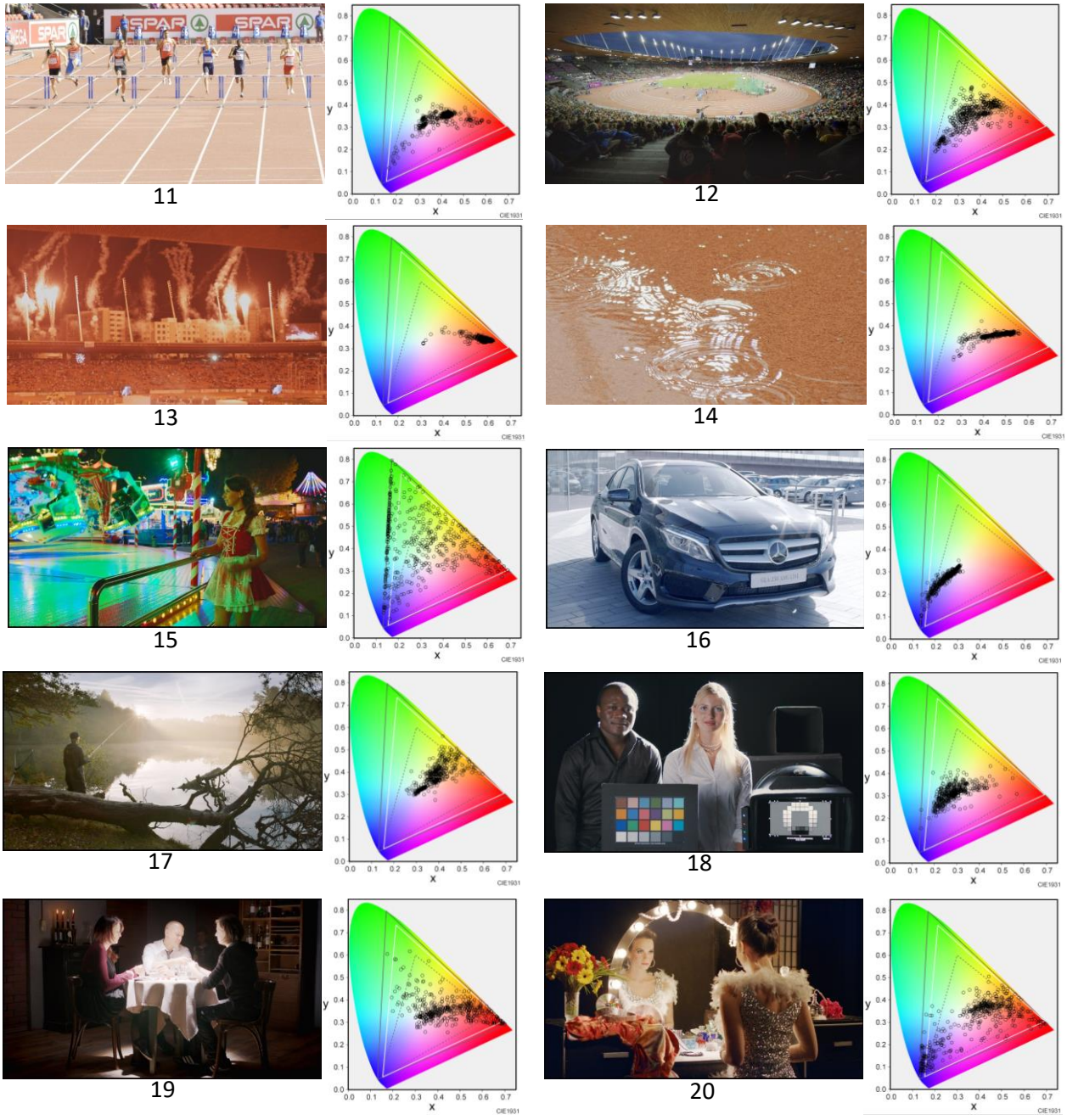


Figure 2. Thumbnails and chromaticity diagrams of the HDR/WCG test images (from 11 to 20).

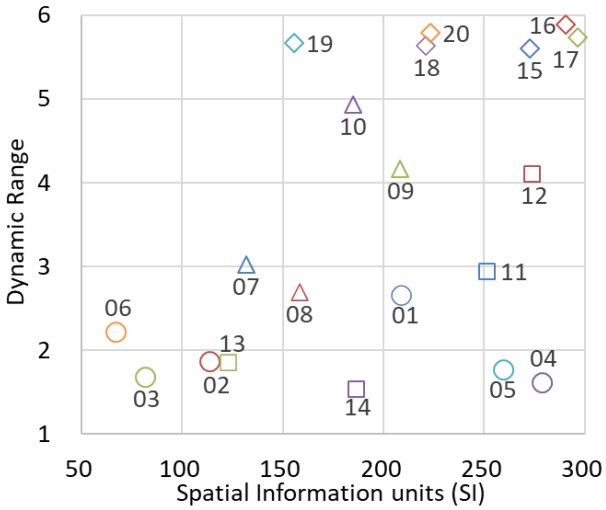


Figure 3. Dynamic range and spatial information corresponding to 20 HDR/WCG test images.

HDR/WCG Image Coding and Generation of Synthesized Test Images

To prepare various distorted images, we compress twenty input HLG images using high efficiency video coding (HEVC)/H.265[19] and versatile video coding (VVC). Figure 4 illustrates an HLG image coding diagram and the process of synthesizing test images. The image coding procedures conform with the HEVC common test conditions concerning HDR/WCG images.[20]

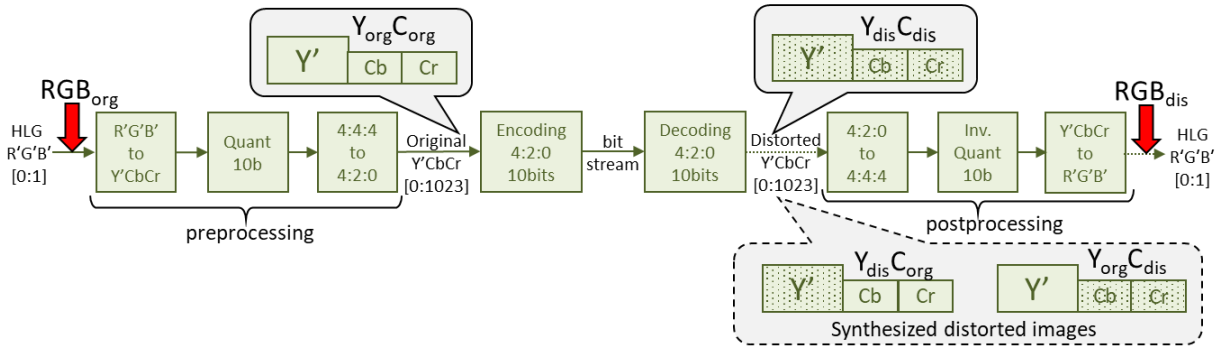


Figure 4. HLG image coding diagram and synthesizing test images.

The image format of the encoder input and decoder output is Y'CbCr 4:2:0 10-bit. During the preprocessing step of the encoder, an HLG encoded signal in BT.2100 R'G'B' 4:4:4 is transferred to Y'CbCr 4:2:0 10 bits, which has one luma and two subsampled chroma components. As a result of subsampling, the image sizes of the Cb and Cr components are decreased by half with respect to the original image both horizontally and vertically. We denote this original Y'CbCr image as $Y_{org}C_{org}$.

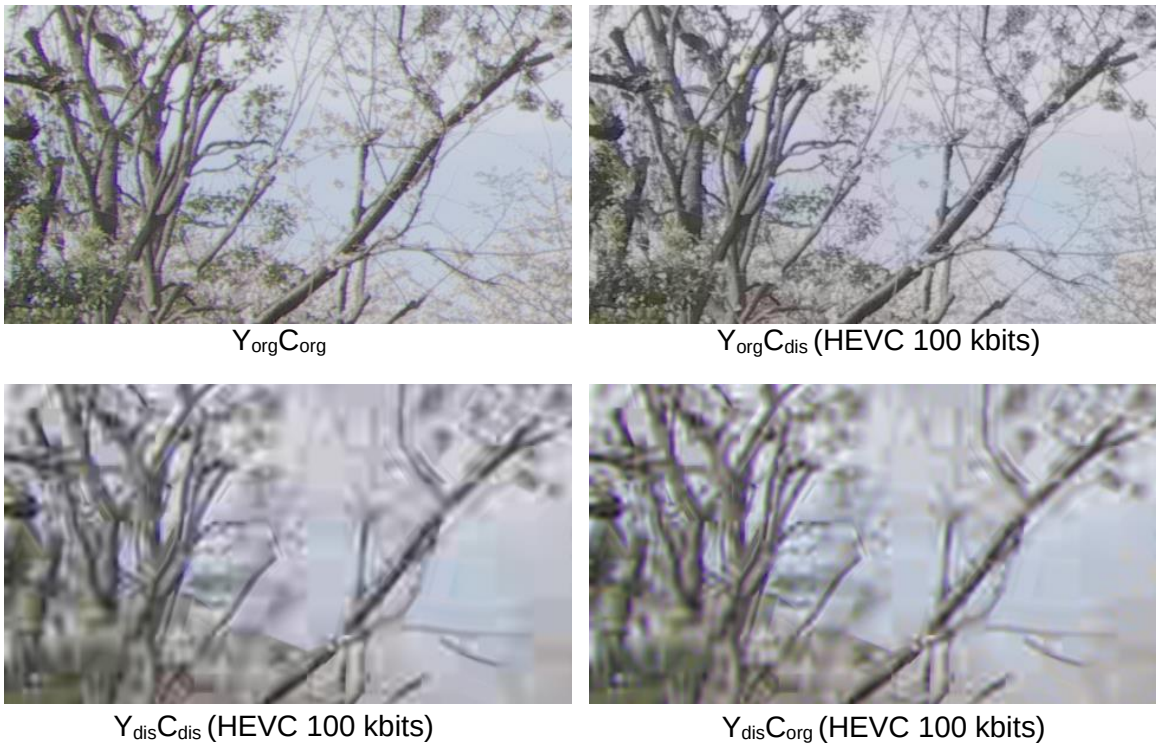
Table 2 describes the encoding conditions for HEVC and VVC. After encoding, image deterioration can be observed in both luma and chroma components. Therefore, we denote the

compressed Y'CbCr images as $Y_{dis}C_{dis}$. After decoding, the postprocessing phase follows the inverse order of the preprocessing stage.

Table 2. Encoding Conditions for HEVC and VVC.

	HEVC	VVC
Original images	20 (from 01 to 20)	10 (01, 03, 05, 08, 10, 12, 14, 15, 17, and 19)
Encoder	HEVC Test Model (HM)[21] ver. 16.19	VVC Test Model (VTM)[22] ver. 3.0
Configurations	Intra only, 4:2:0, 10-bit precision	
Target bitrates	100, 200, 300, and 400 kbits with the fixed quantization parameter (QP) setting	
Encoded images	80	40

We also generate the synthesized images based on $Y_{org}C_{org}$ and $Y_{dis}C_{dis}$. One is $Y_{dis}C_{org}$, composed of the compressed Y' and uncompressed Cb and Cr components, and another is $Y_{org}C_{dis}$, comprised of the opposite components, as can be seen in Figure 4. Figure 5 illustrates an example of test images, the left upper part of 05.



$Y_{org}C_{org}$ $Y_{org}C_{dis}$ (HEVC 100 kbits)

$Y_{dis}C_{dis}$ (HEVC 100 kbits) $Y_{dis}C_{org}$ (HEVC 100 kbits)

Figure 5. Example of test images (cropping by 480 × 270 from 05).

For example, if an objective quality metric relies on only achromatic component, the calculation results for $Y_{dis}C_{dis}$ and $Y_{dis}C_{org}$ with the same luma component Y_{dis} should be approximately the same. The purpose is to confirm whether there is a significant difference in subjective evaluation results in such cases so as to prove the necessity of incorporating chroma components into the

calculations of the objective quality metrics. Moreover, the subjective results can be utilized to investigate the performance of metrics.

Subjective Evaluation Experiment

In the present study, we conduct a subjective evaluation experiment referring to Rec. BT.500[18] and BT.2100.[1]

Table 3 represents the experimental conditions used in the performed subjective assessments.

Table 3. Experimental Conditions for Subjective Assessments.

	$Y_{dis}C_{dis}$	$Y_{dis}C_{org}$, $Y_{org}C_{dis}$, and $Y_{org}C_{org}$
Monitor	31.1-in. HDR/WCG LCD monitor (approximately 0.70 m wide × 0.37 m high) 4,096 × 2,160/10-bit/1,000 cd/m^2	
Viewing condition	Compliant with TABLE 3 of BT.2100[1]	
Viewing distance	1.5 picture height (approximately 0.55 m)	
Surround luminance	5 cd/m^2	
Presentation method	SDSCE method[18]	
Grading method	DSIS method[18]	
Date of experiment	December 2018	March 2020
Observers	16 video experts	15 video experts
Distorted images	120 $Y_{dis}C_{dis}$ generated from 20 original images (from 01 to 20)	60 $Y_{dis}C_{org,}$ 60 $Y_{org}C_{dis}$, and 9 $Y_{org}C_{org}$ generated from 9 original images (01, 05, 10, 11, 14, 15, 16, 19, and 20)

We utilize a 31.1-in. 4K HDR/WCG LCD monitor supporting the HLG method.[17] The peak luminance denoted as L_W in TABLE 5 of BT.2100[1] is 1,000 cd/m^2, which is a common value for the HLG method. Here, the viewing distance is set to 1.5 times the picture height.

The presentation method is based on the simultaneous double stimulus for the continuous evaluation (SDSCE) method.[18] As shown in Figure 6, an original image (RGB_{org} in Figure 4) and the corresponding distorted one (RGB_{dis} in Figure 4) to be evaluated are displayed side by side on a mid-gray background (approximately 50 cd/m^2) during 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. Then, each observer evaluates the deterioration level of a test image relatively to the reference image using the five-grade impairment scale corresponding to the double stimulus impairment scale (DSIS) method (5 – imperceptible; 4 – perceptible, but not annoying; 3 – slightly annoying; 2 – annoying; and 1 – very annoying).[18]

The experiment is conducted in two batches, and a total of 16 and 15 experts who are familiar with HDR videos from the viewpoint of research purposes have participated in each experiment, respectively. As a reference image used in the experiment is RGB_{org} represented in Figure 4, we also test $Y_{org}C_{org}$ images that are not compressed by the encoder but are distorted by subsampling, in order to compare with $Y_{org}C_{dis}$.

Figure 6. Image presentation method.

Performance Evaluation of the Objective Quality Metrics

We select six objective quality metrics that can be used to calculate the results using both achromatic and chromatic components, including five types of color difference metrics. In addition, we calculate the best HDR objective metric – VIF in an HLG signal – for reference.

ΔE_{00}

CIE DE2000 (ΔE_{00}) is a color difference metric calculated according to equation 1.[23]

$$\Delta E_{00}(L_1^*, a_1^*, b_1^*; L_2^*, a_2^*, b_2^*) = \sqrt{\left(\frac{\Delta L'}{k_L S_L}\right)^2 + \left(\frac{\Delta C'}{k_C S_C}\right)^2 + \left(\frac{\Delta H'}{k_H S_H}\right)^2 + R_T \left(\frac{\Delta C'}{k_C S_C}\right)\left(\frac{\Delta H'}{k_H S_H}\right)} \qquad (1)$$

This metric utilizes the CIE L*a*b* color space that is designed in such a way that the same amount of numerical changes in these values corresponds to roughly the same amount of perceptual changes. Although this metric is not intended for HDR/WCG images, it achieves good performance in the case of the HDR/WCG image database.[9] We calculate ΔE_{00} for each pixel between an original image and a distorted image and then average the values across all pixels in an image. Other color difference metrics are calculated in the similar manner.

ΔE_S

S-CIELAB (ΔE_S) has been developed to simulate spatial blurring by a human visual system (HVS).[24] To realize this, a spatial Gaussian filter is applied to input images before calculating the color difference ΔE in the CIE L*a*b* color space. Similarly as ΔE_{00}, ΔE_S achieves a good result for HDR/WCG images[10] apart from the HDR/WCG applications.

ΔE_{ITP}

ΔE_{ITP} described in BT.2124[12] has been introduced specifically for HDR/WCG images. It relies on the display referenced PQ IC_TC_P color space defined in TABLE 7 of BT.2100,[1] as shown in equation 2:

$$\Delta E_{ITP} = 720 \times \sqrt{(I_1 - I_2)^2 + (T_1 - T_2)^2 + (P_1 - P_2)^2} \qquad (2)$$

where $I = I$, $T = 0.5 \times C_T$, and $P = C_P$ in IC_TC_P. Figure 7 illustrates the calculation process of ITP based on an HLG signal.

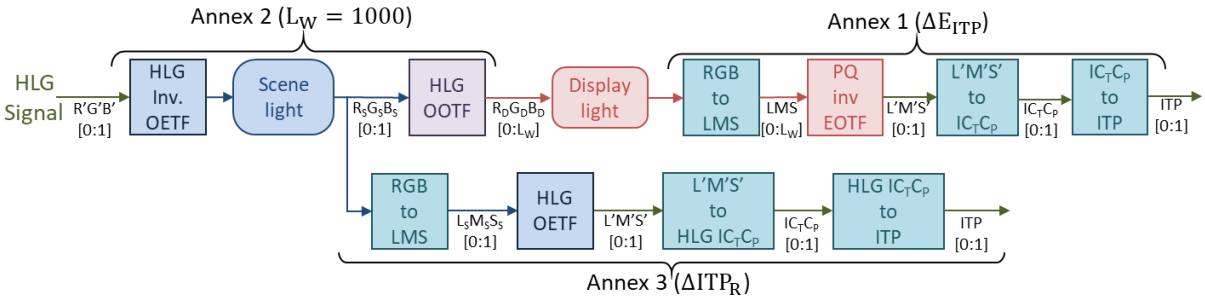


Figure 7. Calculation process of ITP for ΔE_{ITP} and ΔITP_R based on an HLG signal.

In calculation, an estimate of the absolute display light in cd/m^2, as seen by human eyes, is required. To this end, we apply HLG opto-optical TF (OOTF) used to map the relative scene light to the display light and set the peak luminance L_W to 1,000 to adapt the monitor used in subjective assessment.

ΔITP_R

ΔITP_R described in BT.2124[12] is an extension of ΔE_{ITP} and can be directly applied to the scene-referred relative signals, such as those considered in the HLG method. The derivation process is represented in the lower part of Figure 7. To perform the conversion from HLG IC_TC_P to ITP, specific parameters are defined as follows: $I = I$; $T = 0.5 \times 1.823698 \times C_T$; $P = 1.887755 \times C_P$.

ΔE_Z

ΔE_Z is a color difference metric for HDR/WCG images.[25] The metric is calculated from the $J_Za_Zb_Z$ perceptually uniform color space, which was studied based on PQ IC_TC_P color space and has more uniformity than that of IC_TC_P. In common with IC_TC_P, PQ inverse EOTF is included in the conversion from display light to $J_Za_Zb_Z$.

$FSIM_C$

The feature similarity index (FSIM) has been developed based on the observation that HVS considers an image mainly according to its low-level features: specifically, a phase congruency (PC) and an image gradient magnitude.[26] $FSIM_C$ incorporates the chromatic information into the calculation procedure and is defined using equation 3:

$$FSIM_C = \frac{\sum_{x\in\Omega} S_L(x)\cdot[S_C(x)]^{\lambda}\cdot PC_m(x)}{\sum_{x\in\Omega} PC_m(x)} \quad (3)$$

Here, S_L and S_C denote the luminance and chrominance similarity measures, respectively. In the paper, the weight of the chrominance components λ was experimentally determined to 0.03.

VIF

In our previous work, VIF, which was derived from a statistical model for natural scenes, a model for image distortion, and an HVS model in an information-theoretical setting,[6] demonstrated excellent performance compared with the other considered HDR metrics using only an achromatic component.[5] We have inputted a 10-bit luminance signal of HLG R'G'B', after applying equations 4 and 5 as follows:

$$Y = 0.262700 \times R + 0.677998 \times G + 0.059302 \times B \quad (4)$$

$$Y_{10b} = round(1023 \times Y) \quad (5)$$

Performance Evaluation

We evaluate the performance of the considered objective quality metrics in the same manner as in the previous related works.[4-5, 9-11] To investigate the similarity between the objective quality metric and the results of a subjective evaluation, we conduct the curve fitting of the following logistic function based on the least square method, as shown in equation 6:

$$\hat{y} = a + \frac{b}{1+\exp\big(-c(x-d)\big)} \quad (6)$$

where x and $\hat{y}$ denote the result of the objective metric and the predicted mean opinion score (MOS), respectively. The true MOS y, which is obtained from a subjective evaluation, 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 is 249 in total, as can be seen in Table 3. We assess the performance in terms of the Pearson linear correlation coefficient (PLCC), the Spearman rank order correlation coefficient (SROCC), and the root-mean-square error (RMSE), concerning the correlation between y_i and $\hat{y}_i$.

Experimental Results

In the present research, we aimed to prove the necessity of including chroma components into HDR/WCG metrics. Then, we assessed the performance of the considered objective metrics, defined according to the rules provided in the previous section.

The need for image quality metrics to consider the chromatic components

First, we aimed to confirm the following hypothesis: VIF could not be applied to distinguish deterioration in chroma components. During subjective evaluation, observers assigned a score for an RGB distorted image that was, for example, generated based on $Y_{dis}C_{dis}$, relatively to RGB_{org}. Hereinafter, we denote the VIF result for this combination of images as VIF (org, $Y_{dis}C_{dis}$).

To validate that VIF (org, Y_XC_{org}) ≡ VIF (org, Y_XC_{dis}) for x=org and dis, we checked that VIF (org, Y_XC_{org}) ≈ VIF (org, Y_XC_{dis}) and VIF (Y_XC_{org}, Y_XC_{dis}) ≈ 1 for each condition. Figures 8 and 9 represent the VIF results for the Y_{dis} and Y_{org} images, respectively. In these graphs, the horizontal axis denotes a type of a distorted image: HEVC or VVC and bitrates. For example, H300 means HEVC encoding at 300 kbits. A red bar corresponding to the left vertical axis illustrates the absolute difference between VIF (org, Y_XC_{org}) and VIF (org, Y_XC_{dis}), and this should be 0. Then, blue x marks corresponding to the right vertical axis depict the results of VIF (Y_XC_{org}, Y_XC_{dis}), which are expected to be 1, the maximum value of VIF.

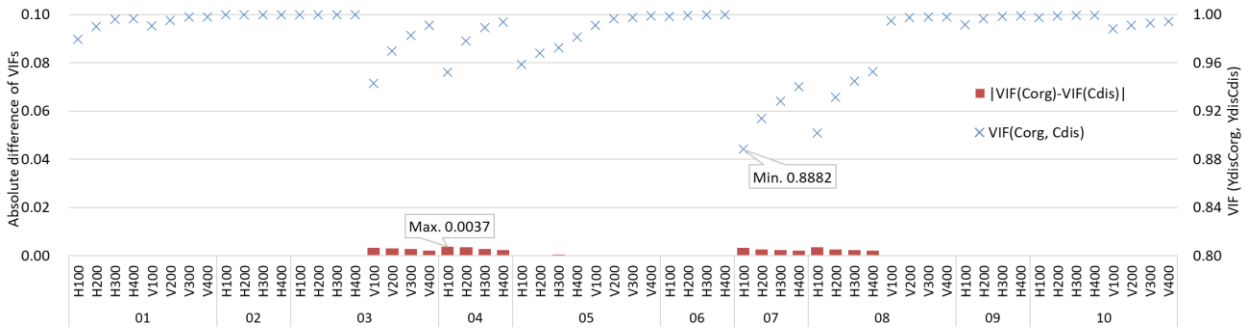


Figure 8. VIF results for $Y_{dis}C_{org}$ and $Y_{dis}C_{dis}$.

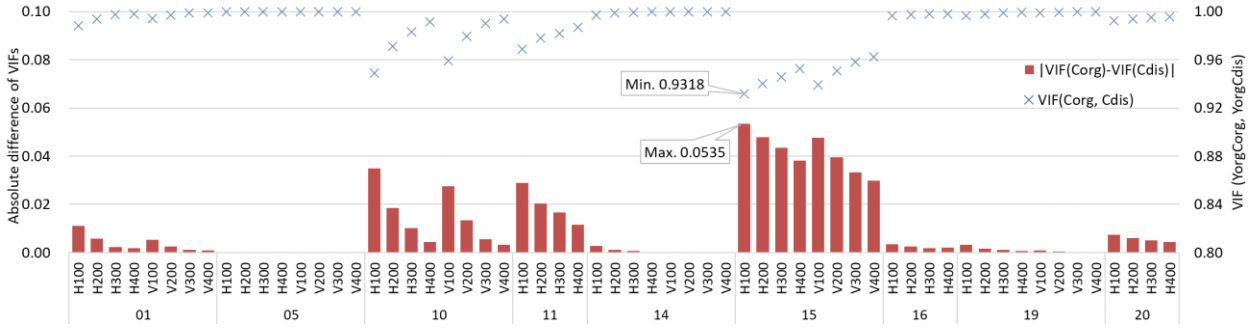


Figure 9. VIF results for $Y_{org}C_{org}$ and $Y_{org}C_{dis}$.

Thereafter, we conducted the Welch's t-test at the 5% significance level based on individual subjective test scores. Table 4 represents the number of conditions that demonstrate a significant difference.

Table 4. Number of Conditions that Exhibit Significant Difference.

Test image	$Y_{org}C_{org} > Y_{org}C_{dis}$	$Y_{dis}C_{org} < Y_{dis}C_{dis}$
01	1/8	-
05	8/8	-
10	-	-
11	4/4	-
14	5/8	2/8
15	4/8	-
16	3/4	3/4
19	3/8	-
20	4/4	-

For instance, 5/8 indicates that 5 out of 8 conditions exhibit a significant difference between the MOS values corresponding to Y_xC_{org} and Y_xC_{dis}. In that case, the differences are always as follows: $Y_{org}C_{org} > Y_{org}C_{dis}$ for the Y_{org} images, and $Y_{dis}C_{org} < Y_{dis}C_{dis}$ for the Y_{dis} images.

Objective Quality Metrics used in HDR/WCG Image Coding

Figure 10 and 11 illustrate the relationship between the considered objective metrics (the horizontal axis) and MOS (the vertical axis). The shapes and colors of markers correspond to image numbers, as shown in Figure 3. It can be seen that the markers filled using similar colors, gray, white, and white with dotted borders are indicating $Y_{dis}C_{dis}$, $Y_{dis}C_{org}$, $Y_{org}C_{dis}$, and $Y_{org}C_{org}$, respectively. The fitting curve is marked in each graph in black. Table 5 represents PLCC, SROCC, and RMSE for each metric.

Table 5. Correlations and Errors of the Considered Objective Metrics.

	PLCC	SROCC	RMSE
VIF	0.9202	0.8424	0.4414
$FSIM_C$	0.8618	0.8047	0.5718
ΔITP_R	0.6615	0.6620	0.8455
ΔE_S	0.6443	0.6549	0.8623
ΔE_{ITP}	0.6406	0.6282	0.8657
ΔE_Z	0.6327	0.6327	0.8731
ΔE_{00}	0.6058	0.6020	0.8970

Consideration

The need for image quality metrics to consider the chromatic components

Analyzing the VIF calculation results represented in Figures 8 and 9, overall, it could be concluded that the lower the bitrate, the larger the difference with respect to the ideal value. Further verification is required, but this might be caused by the upsampling processes corresponding to 4:2:0 and 4:4:4. However, the difference was negligible as the VIF of 0.88 could be deemed sufficiently close to 1, considering the relationship of the VIF values and MOS discussed in our previous work. Moreover, although |VIF (org, $Y_{org}C_{org}$) - VIF (org, $Y_{org}C_{dis}$)| tended to be larger than |VIF (org, $Y_{dis}C_{org}$) - VIF (org, $Y_{dis}C_{dis}$)|, this observation could be disregard, as VIF (org, $Y_{org}C_x$) ≈ 1, meaning that VIF regards that the distorted images are intrinsically difficult to find the difference from the original ones. Therefore, the results indicated that the VIF calculated based on the luminance component could not accurately detect distortion in chroma components.

However, the subjective assessment results, as shown in Table 4, indicated that distortion in chroma components could be distinguished perceptually. Considering that the difference was mainly in Y_{org}, the degradation level on color components could be easily detected if a luma component was not significantly distorted. Therefore, we demonstrated the necessity of incorporating chroma components into the objective quality metrics.

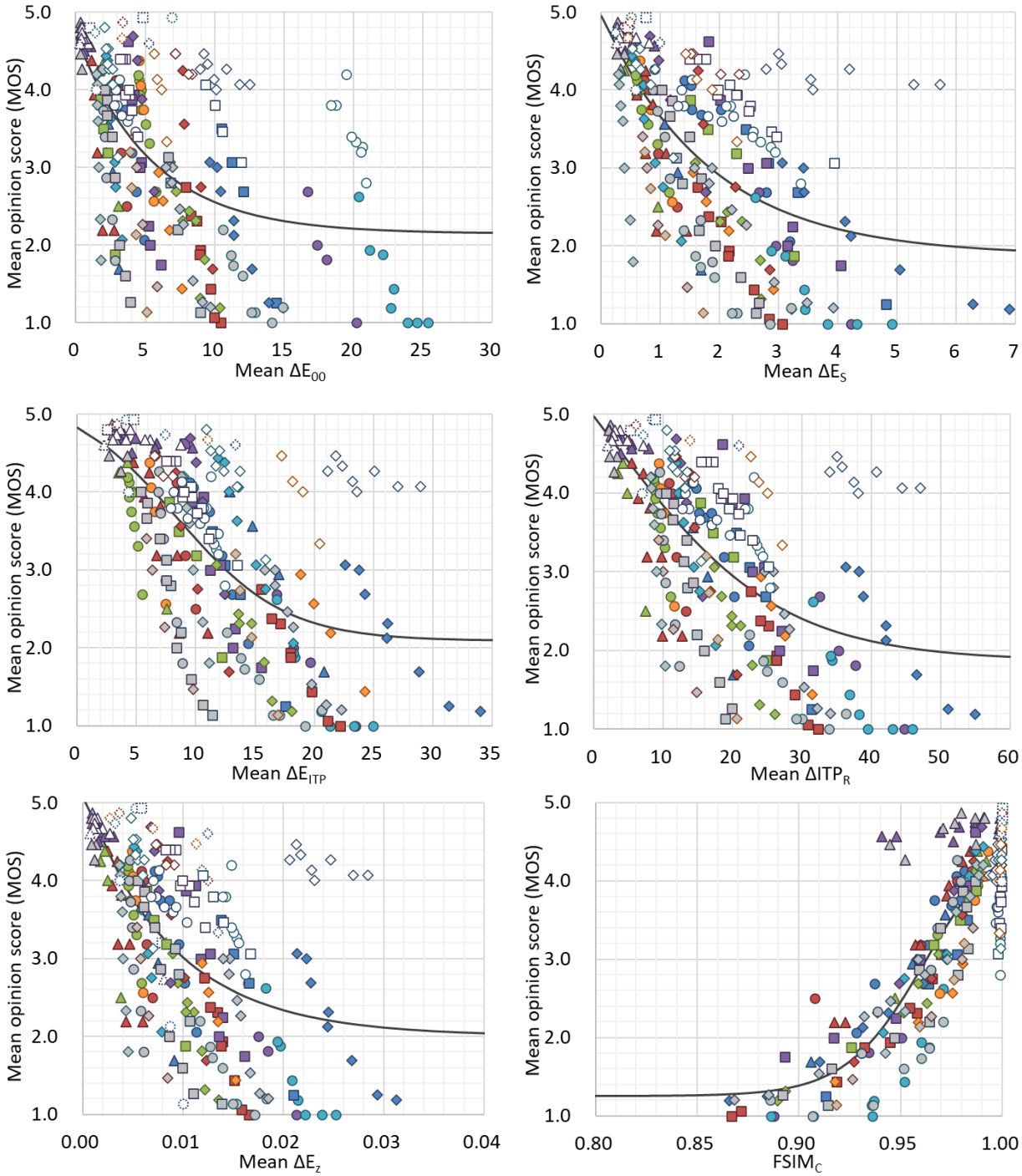


Figure 10. Relationship between the objective metrics and MOS (metrics for luminance and chroma components).

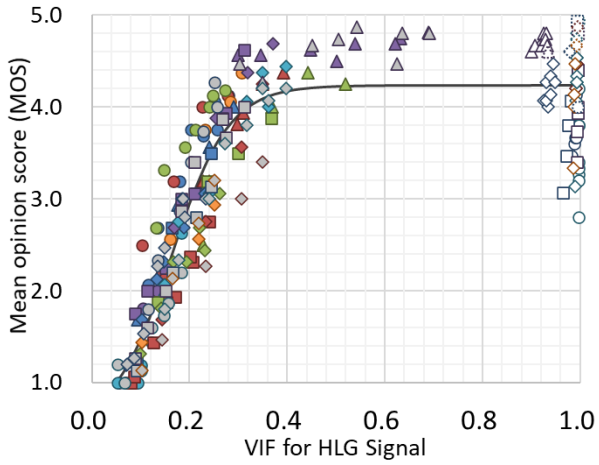


Figure 11. Relationship between the objective metrics and MOS (VIF, a metric for an achromatic component).

Objective Quality Metrics for HDR/WCG Image Coding

As shown in Table 5, VIF achieved much better performance compared with the other metrics that considered color components. Although ΔITP_R was intended to facilitate the HLG method, we found that it demonstrated poor performance in HLG image coding. In addition, the tendency in the color difference metrics varied corresponding to each test image.

For example, ΔE_{ITP}, ΔITP_R, and ΔE_Z emphasize the distortion of the test image 15 (according to blue diamonds on the right side of graphs represented in Figure 10). Considering the fact that 15 consists of multiple vivid green pixels, this could be due to the characteristics of the IC_TC_P color space that differed from perception, specifically, in the low luminance green range.[27] Though ΔE_Z utilizes the $J_Za_Zb_Z$ color space, which improved the color uniformity of IC_TC_P in the cyan and green colors,[25] the performance against such colors seems to be intrinsically quite similar to that of IC_TC_P. Table 6 describes the performance of ΔE_{ITP}, ΔITP_R, and ΔE_Z without singularities. It could be seen that they improved considerably compared with the original results represented in Table 5.

Table 6. Correlations and Errors of ΔE_{ITP}, ΔITP_R, and ΔE_Z without 15.

	PLCC	SROCC	RMSE
ΔITP_R	0.7486	0.7307	0.7417
ΔE_{ITP}	0.7238	0.6945	0.7718
ΔE_Z	0.7162	0.7042	0.7807

Similarly, ΔE_{00} highlighted the deterioration in the test images 04 and 05 (represented by the purple and cyan circles on the right side of the graph represented in Figure 10). These images tended to exhibit large distortion in the luminance component.

Even though $FSIM_C$ achieved the second best performance results, this could be attributed to the performance of the luminance component. This was because the contribution rate of chroma

components λ was set to 0.03 only in equation 3. In fact, as can be seen in Figure 10, the $FSIM_C$ values for the Y_{org} images are almost 1 similarly as for VIF in Figure 11.

Considering the success of the video multimethod assessment fusion (VMAF),[28] the objective quality metrics for videos, aggregating the metrics with different trends using machine learning techniques can be considered as a possible solution for a problem of identifying objective metrics suitable for HLG image coding. As the similar approaches for the HDR[29] and HDR/WCG[11] images have been already examined, we will continue to extend this study in this way.

Conclusion

In the present study, we investigated the suitability of objective quality metrics for compressed HDR/WCG images encoded with HLG. Our main findings are three. Firstly, that the chroma components must be taken into account by objective metrics, because HDR metrics that just consider the luminance channel fail to detect color distortions that observers can notice. Secondly, that the results of metrics that do consider color are nonetheless worse than the results of VIF, despite the fact that VIF is only applied on the luminance. And thirdly, that each color difference metric behaves differently depending on the type of test image.

Based on these results, as future work, we will study the aggregation of metrics into a single one as a solution to the open problem of finding an objective metric for HLG-encoded HDR/WCG content.

References

1. International Telecommunications Union – Radiocommunication (ITU-R), Recommendation BT.2100-2, “Image parameter values for high dynamic range television for use in production and international programme exchange,” 07/2018.
2. M. Narwaria, R. Mantiuk, M. Perreira da Silva, and P. Le Callet. "HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images." *Journal of Electronic Imaging*, vol. 24, no. 1, 010501, Jan. 2015.
3. M. Narwaria, M. Perreira da Silva, and P. Le Callet, “HDR-VQM: An Objective Quality Measure for High Dynamic Range Video,” *Signal Processing: Image Communication*, vol. 35, pp. 46-60, Jul. 2015.
4. Y. Sugito, and M. Bertalmío, "PERFORMANCE EVALUATION OF OBJECTIVE QUALITY METRICS ON HLG-BASED HDR IMAGE CODING," *in Proc. 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 96-100, 2018.
5. Y. Sugito, and M. Bertalmío, "Practical Use Suggests a Re-evaluation of HDR Objective Quality Metrics," *in Proc. 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1-6, 2019.
6. H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430-444, Feb. 2006.

7. International Telecommunications Union – Radiocommunication (ITU-R), Recommendation BT.709-6, “Parameter values for the HDTV standards for production and international programme exchange,” 06/2015.

8. International Telecommunications Union – Radiocommunication (ITU-R), Recommendation BT.2020-2, “Parameter values for ultra-high definition television systems for production and international programme exchange,” 10/2015.

9. A. Choudhury, J. Pytlarz, and S. Daly, "HDR and WCG Image Quality Assessment Using Color Difference Metrics," *in Proc. SMPTE Annual Technical Conference (ATC) 2019*, Los Angeles, California, pp. 1-22, 2019.

10. M. Rousselot, O. Le Meur, R. Cozot, and X. Ducloux, “Quality assessment of HDR/WCG images using HDR uniform color spaces,” *Journal of Imaging*, vol. 5, no. 1, pp. 1-40, Jan. 2019.

11. M. Rousselot, X. Ducloux, O. L. Meur and R. Cozot, "Quality Metric Aggregation for HDR/WCG Images," *in Proc. 2019 IEEE International Conference on Image Processing (ICIP)*, pp. 3786-3790, 2019.

12. International Telecommunications Union – Radiocommunication (ITU-R), Recommendation BT.2124-0, “Objective metric for the assessment of the potential visibility of colour differences in television,” 01/2019.

13. International Telecommunications Union – Radiocommunication (ITU-R), Report BT.2408-3, “Guidance for operational practices in HDR television production,” 07/2019.

14. M.D. Fairchild, “The HDR photographic survey,” *in Proc. 15th Color and Imaging Conference (CIC 2007)*, pp. 233-238, 2007.

15. “EBU Technology & Innovation – Zurich athletics,” Sep. 8, 2020. [Online], Available: https://tech.ebu.ch/testsequences/zurich_athletics

16. 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,” *in Proc. SPIE 9023, Digital Photography X*, 90230X, 2014.

17. “HDR Reference Monitor ColorEdge Prominence CG3145 – EIZO,” Jul. 8, 2020. [Online], Available: https://www.eizo.com/products/coloredge/cg3145/

18. International Telecommunications Union – Radiocommunication (ITU-R), Recommendation BT.500-14, “Methodologies for the subjective assessment of the quality of television images,” 10/2019.

19. International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) 23008-2:2017, “High Efficiency Coding and Media Delivery in Heterogeneous Environments—Part 2: High Efficiency Video Coding.” International Telecommunications Union – Telecommunication (ITU-T), Recommendation H.265, “High Efficiency Video Coding,” Feb. 2018.

20. E. François, J. Sole, J. Ström, and P. Yin “Common Test Conditions for HDR/WCG video coding experiments,” JCTVC-Z1020, 2017.

21. “High Efficiency Video Coding (HEVC) | JCT-VC,” Jul. 8, 2020. [Online], Available: http://hevc.hhi.fraunhofer.de/

22. “Versatile Video Coding (VVC) | JVET,” Jul. 8, 2020. [Online], Available: https://jvet.hhi.fraunhofer.de/

23. G. Sharma, W. Wu, and E. N. Dalal, “The CIEDE2000 color difference formula: Implementation notes, supplementary test data, and mathematical observations,” *Color Research & Application*, vol. 30, no. 1, pp. 21-30, Feb. 2005.

24. X. Zhang, and B. A. Wandell, "A spatial extension of CIELAB for digital color image reproduction." *Journal of The Society for Information Display*, vol. 5, pp. 61-63, 1997.

25. M. Safdar, G. Cui, Y.J. Kim, and M.R. Luo, "Perceptually uniform color space for image signals including high dynamic range and wide gamut," *Opt. Express*, vol. 25, no.13, pp. 15131-15151, Jun. 2017.

26. L. Zhang, L. Zhang, X. Mou and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378-2386, Aug. 2011.

27. A. Pytlarz, and E. G. Pieri, “How close is close enough?” *in Proc. International Broadcasting Convention (IBC) 2017 Conference*, pp.1-9, 2017.

28. “Toward A Practical Perceptual Video Quality Metric,” Jul. 8, 2020. [Online] Available: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652

29. A. Choudhury, and S. Daly, "Combining Quality Metrics for Improved HDR Image Quality Assessment," *in Proc. 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 179-184, 2019.

