https://doi.org/10.2352/ISSN.2470-1173.2020.9.IQSP-214

Comparing common still image quality metrics in recent High Dynamic Range (HDR) and Wide Color Gamut (WCG) representations

Anustup Choudhury, Scott Daly; Dolby Laboratories Inc.; Sunnyvale, CA, USA

Abstract

There are an increasing number of databases describing subjective quality responses for HDR (high dynamic range) imagery with various distortions. The dominant distortions across the databases are those that arise from video compression, which are primarily perceived as achromatic, but there are some chromatic distortions due to 422 and other chromatic sub-sampling. Tone mapping from the source HDR levels to various levels of reduced capability SDR (standard dynamic range) are also included in these databases. While most of these distortions are achromatic, tone-mapping can cause changes in saturation and hue angle when saturated colors are in the upper hull of the of the color space. In addition, there is one database that specifically looked at color distortions in an HDR-WCG (wide color gamut) space. From these databases we can test the improvements to well-known quality metrics if they are applied in the newly developed color perceptual spaces (i.e., representations) specifically designed for HDR and WCG. We present results from testing these subjective quality databases to computed quality using the new color spaces of $J_za_zb_z$ and IC_TC_P, as well as the commonly used SDR color space of CIELAB.

Introduction

High dynamic range (HDR) and wide color gamut (WCG) capability have now become mainstream in consumer TV displays and is making headway into desktop monitors, laptops and mobile device products. In the consumer industry, the term HDR generally means the combination of HDR and WCG, and we will use that shorthand terminology here. HDR systems provide a more complete representation of information that the human visual system can perceive and thus not a simple extrapolation, which makes evaluating content shown on these HDR systems essential. Since subjective evaluations can be time-consuming and expensive, there is a need for objective quality[1] assessment tools.

Various full reference HDR quality metrics such as HDR-VDP-2 (HDR visual difference predictor) [1, 2], DRIM (Dynamic range independent metric) [3], HDR-VQM (HDR video quality measure) [4] have been proposed for image and video quality assessment (IQA/VQA). HDR-VDP2 and HDR-VQM require modeling of both the human visual system (HVS) and the display, whereas DRIM, in addition to HVS modeling, results in three distortion output maps making it more difficult for interpretation. Alternatively, due to lack of HDR objective metrics, LDR/SDR (low/standard dynamic range) metrics were also used to evaluate HDR quality [5, 6]. Examples of full reference SDR metrics that have been used in literature for HDR quality evaluation are MS-SSIM (Multi-scale structural similarity index) [7], IFC [8], VIFp (pixel-based visual information fidelity) [9], FSIM (Feature similarity index) [10], FSITM [11], UQI [12], VIF (visual information fidelity) [9] and so on.

Recent studies [5, 6, 13, 14, 15], have evaluated both HDR and SDR quality metrics for HDR quality assessment. In particular, Hanhart et al. [5] evaluated the performance of 35 objective metrics on a publicly available database [16]. Zerman et al. [6] evaluated the performance of 12 metrics on five different HDR databases. Although the HDR based metrics, HDR-VDP-2 and HDR-VQM outperform existing SDR metrics, modifying some SDR metrics such as MS-SSIM by applying calibrated non-linearities can result in performance close to the HDR based metrics in terms of correlation [5]. Rousselot et al. [15] studied the impact of 12 SDR quality assessment metrics computed in three HDR/WCG color spaces viz., IC_TC_P [17], $J_za_zb_z$ [18] and HDR-*Lab* [19]. Choudhury and Daly [20, 21, 22] used a combination of various HDR and SDR quality metrics in one framework for improved quality assessment and demonstrate state-of-the-art results.

In this work, we propose to use IC_TC_P color representation to compute various IQA metrics. IC_TC_P color representation can be considered to be perceptually uniform for HDR and WCG content thus making the metric computations more accurate. Rousselot et al. [15] has done some preliminary work to evaluate metrics in IC_TC_P color representation amongst other color representations. However, they do not consider the use of this color representation for HDR metrics which we have done in this paper. Furthermore, they use PU [23] transfer function in conjunction with the SDR metrics whereas we use the PQ [24] transfer function. One disadvantage of using PU non-linearity is that it has less sensitivity in the dark regions than PQ partially due to the more limited display capability when PU was developed. Amongst SDR metrics, we also show the results using VIFp, which was not shown in [15]. In addition to that, we show improved performance using our implementation of updated ΔE_{ITP} as compared to the results presented in [15]. Moreover, these improvements can be easily ported over to other techniques that use combinations of quality metrics such as [20, 21, 22] to further improve the state-of-the-art performance.

Methodology

In this section, we describe the various quality metrics that we have used in our evaluation along with the various databases

[1] Some use 'objective quality' to strictly refer to physical measurements. Others, particularly in the perceived quality modeling field, use 'objective' to mean something that can be calculated with a quality model. We use the term in the latter sense in this paper.

that we have considered for our analysis.

Quality Metrics for HDR IQA

Various IQA metrics have been proposed in literature to evaluate human visual quality experience. We also discuss the various transformations that we perform on these metrics. Based on the observations presented in [5, 6, 15], we choose the best performing metrics for evaluation. Amongst HDR metrics, we select HDR-VDP-2 [1, 2] and HDR-VQM [4]. Amongst SDR metrics, we choose MS-SSIM [7], VIFp [9] and FSIM [10] and select CIE ΔE_{00} [25, 26], ΔE_Z [18] and ΔE_{ITP} [27] as candidates for color difference measures.

HDR-VDP-2 and HDR-VQM were developed for HDR quality assessment. HDR-VDP-2 is a calibrated metric and takes into account models regarding point spread function of the eye, the light-adaptive CSF, and masking within an oriented multi-scale decomposition. While HDR-VDP-2 does not implement spatio-chromatic modeling, it does use colorimetric XYZ based calibration of the input as well as the display primaries to create a calibrated Y signal for analysis by the metric, which is consistent with the achromatic channel of spatial vision. HDR-VQM is a video quality metric computed in PU [23] space and also relies on multi-scale and multi-orientation analysis, as well as simple temporal differences which are pooled. In this setup, we compute HDR-VQM on still images, thus having zero temporal error. Both HDR-VDP-2 and HDR-VQM perform spatial pooling to compute overall quality score. These metrics allow for information about viewing conditions such as viewing distance, ambient and so on.

We considered three metrics that were all designed for SDR content. MS-SSIM is a multi-scale technique that models the quality based on the assumption that the HVS uses structural information from a scene. VIFp analyzes the natural scene statistics and is based on Shannon information. FSIM analyzes high phase congruency, extracting the gradient magnitude to encode contrast information.

Finally, we use three color difference metrics. Traditionally, these are 'pixel-based' operators that compare only corresponding pixels from the reference and distorted images and thus do not consider any spatial process of vision. CIE ΔE_{00} is a color difference measure that includes weighting factors for lightness, chroma and hue along with the ability to handle the relationship between chroma and hue. It was designed for the CIELAB color space, which is limited to SDR. In addition, we analyzed newer metrics derived for HDR applications: ΔE_Z based on the $J_z a_z b_z$ color space [18], and ΔE_{ITP} based on the $IC_T C_P$ [17] color space. ΔE_{ITP} is standardized in ITU-R BT.2124 [27] and utilizes the PQ (ST 2084) transfer function. It differs slightly from the ΔE_{ITP} used in [15] by a rescaling of the B-Y opponent channel. $J_z a_z b_z$ color space also utilizes the PQ (ST 2084) transfer function and the aim is to improve lightness correlation; however the lightness correlation optimization was based on data [19] with a diffuse white of 997 cd/m^2.

Databases

We consider five publicly available databases from different labs comprised of natural images to compare the performance of the different metrics for evaluation. The digital images and subjective scores are made available for independent researchers to do various analysis and metric development. The first database [16] (Database 1) contains 20 reference HDR images. Distorted images were created by compressing the reference images with JPEG XT with various profiles and quality levels. Two different tone mapping operations [28, 29] were used for the base layer. Four different bit rates were chosen using three profiles of JPEG XT. Each image had a resolution of 944 X 1080 pixels (i.e., a crop for a split-screen of a 1920x1080) and were calibrated for a SIM2 HDR monitor.

The second database [6] that we considered is a combination of two different databases [6, 30]. One of them [30] is composed of five original HDR images which were first tone-mapped [31], following which 50 compressed images were obtained using three different coding schemes - JPEG, JPEG2000 and JPEG XT. These images were presented sequentially(one after the other) on a SIM2 HDR47E display and scores were collected from 15 participants. The second database [6] also uses five original HDR images from which 50 compressed images were obtained. They used JPEG and JPEG2000 (with different bit rates) and the SDR images were obtained using two different mapping operations [31, 24]. Overall, the second database (Database 2) has 100 1920 X 1080 images.

The third database [32] (Database 3) contained 10 HDR source images. The distorted images were obtained using a backward compatible scheme where the HDR image is tone mapped using iCAM06 [33] and then this tone mapped image is coded using JPEG codec at seven different bit rates. Finally the compressed image is expanded by an inverse tone mapping operation to the original HDR image. Subjective ratings were given by 27 participants. Two different criteria were used to optimize the quality of the reconstructed HDR resulting in a total of 140 images with a resolution of 1920 X 1080.

One of the limitations of the three databases mentioned above was that these databases did not explore wider color gamut (eg. DCI-P3 or larger). Also, they did not contain any specifically parameterized color artifacts, although some may arise from tone-mapping at the very top and bottom of the color solid (and if any chromatic sub-sampling was used in the compression profiles). In addition, the subjective testing for all three above-mentioned databases were conducted on the same monitor (SIM2 HDR monitor). To introduce more variety in our experimental samples we used two additional databases (Database 4 and 5) that included chromatic distortions as well as images beyond ITU-R BT.709 color gamut (exceeding P3 and nearly up to ITU-R BT.2020 color gamut by use of the Sony BVM-X300 OLED pro monitor).

The fourth database [34] (Database 4) contained eight images distorted using four types of distortion: HEVC compression using four different quantization parameters (QP), HEVC compression without the chroma QP adaptation resulting in chromatic distortions at three different values, three different levels of Gaussian noise and two different types of gamut mismatch (i.e., rendered assuming that ITU-R BT.709 images were interpreted as ITU-R BT.2020 images leading to more saturated colors, and assuming that ITU-R BT.2020 images were interpreted as ITU-R BT.709 images leading to less saturated colors).

The fifth database [15] (Database 5) contained eight images that were compressed with four different QP using three different compression options – Recommended HEVC compression, HEVC compression without chroma QP offset algorithm and HEVC compression with 8 bits quantization for chroma instead of 10 during compression. The images were all using the

ITU-R BT.2020 color gamut.

Please refer to [35] for a more detailed description of the databases.

Metric Computation and Transformation

In order to calculate the quality metrics, the pixel values are first scaled to the range of luminance emitted by the HDR display. We use the technique mentioned in [6] where the HDR pixels are converted to luminance emitted by a hypothetical display such that the there exists a linear response between the minimum and maximum luminance of the display. Since the databases that we considered use two different displays (SIM2 and Sony BVM-X300), we use the parameters belonging to that display for evaluating the respective databases. Valenzise et al. [30] showed that using this linear assumption is equivalent to more sophisticated luminance estimation techniques that require a detailed knowledge of the device. The images were also clipped to the range of the display to mimic the physical clipping performed by the HDR display.

HDR images in the databases that we previously mentioned have code value corresponding to linear luminance to account for wider luminance ranges. Since HDR metrics, which are calibrated metrics, require absolute luminance values as input, we use the HDR metrics as-is without any transformation. This is used as baseline for evaluating HDR metrics. The color representations for these include the achromatic perceptual non-linearity derived from luminance (both HDR metrics uses the PU non-linearity). However, these metrics ignore any chromatic distortions. These HDR metrics are advanced, computationally complex, and include CSF effects, masking effects implemented via frequency channels, and spatial pooling. We also compute the HDR metrics after converting the linear luminance to IC_TC_P [17] color representation and then using the I channel (which is the PQ non-linearity applied to the achromatic visual channel). This I channel in the PQ non-linear space is then converted back to linear values and given as input to the HDR metrics. This is denoted by the '_I' suffix (for the HDR metrics).

We use three different variations of computing the SDR metrics -

1. Compute it directly in the linear domain using the physical luminance of the scene.
2. Compute in the PQ domain [24]. We first convert the gamma domain code values to linear luminance for each of the RGB signals and then convert luminance to the PQ domain and denote that using '_PQ' suffix. This is based on the results from [5] that found calibrating the SDR metrics via either the PQ or PU non-linearities always improved their performance compared to applying them directly on the code values of the signal space i.e., the gamma-corrected domain or SDR. The PU non-linearity has less sensitivity in the dark regions than PQ partially due to more limited display capability when PU was developed. Further, such processing focuses the quality on the achromatic channel of human vision, which is known to have the better spatial performance. Any purely color differences (i.e., iso-luminant) are ignored in the SDR analysis due to the models limitations, despite there being some chromatic distortion if 422 and 420 profiles were used, or in the tone-mapping distortions. Then we normalize the RGB color components to [0, 1] range and transform the RGB color space to YC_bC_r color space. The quality score was computed on the luminance (Y) channel since [5] found that using the Y channel alone instead of using the mean of the Y, C_b and C_r color channels resulted in the best performance. We thus consider only the Y channel for the SDR metrics and denote that using '_Y' suffix. This combination of PQ (non-linearity) and Y channel is denoted using '_PQ_Y' suffix.
3. Convert linear RGB pixel values to IC_TC_P color representation and compute using the I channel. This is denoted using '_I' suffix. Please note that the IC_TC_P color representation implicitly applies the PQ non-linearity function.

Algorithm 1 Convert from linear RGB values to IC_TC_P

Input: R, G, B color channels
Output: I, C_T, C_P color channels

procedure I, C_T, C_P

Convert linear R, G, B values (assuming BT. 2100) to linear L, M, S values as

$$L = (1688R + 2146G + 262B)/4096$$

$$M = (683R + 2951G + 462B)/4096$$

$$S = (99R + 309G + 3688B)/4096$$

Convert linear L, M, S to non-linear L', M', S' by applying the PQ transfer function

$$\{L', M', S'\} = EOTF^{-1}(F)$$

where,

$$F = \{L, M, S\}$$

$$EOTF^{-1}(F) = ((c1 + c2Y^{m1})/(1 + c3Y^{m1}))^{m2}$$

and $Y = F/10000$, $m1 = 0.1593$, $m2 = 78.8435$, $c1 = 0.8359 = c3 - c2 + 1$, $c2 = 18.8515$, $c3 = 18.6875$

Convert non-linear L', M', S' to I, C_T, C_P as :

$$I = 0.5L' + 0.5M'$$

$$C_T = (6610L' - 13613M' + 7003S')/4096$$

$$C_P = (17933L' - 17390M' - 543S')/4096$$

end procedure

The IC_TC_P color representation is used by both HDR and SDR metrics and the conversion can be shown using Algorithm 1.

The color difference measures were not computed in the transformed spaces i.e., we do not apply either the PQ or PU non-linearity. CIE ΔE_{00} requires a conversion from RGB to CIELAB color space. In typical use, the white point needs to be input to the CIELAB calculations, which for hard-copy the paper white is commonly used. Perceptually, this makes sense since paper white is usually visible on the border of the print, and the visual system can anchor to it. But this approach has always been a problem for

video where there is no white border surrounding the image. It is particularly a problem for HDR content. Reinhard et al. [36] showed why using an adaptive white point luminance of 1000 cd/m^2 is not ideal and [19] has shown that using a value closer to diffuse white (as opposed to the highlight maximum) produces better results. Choudhury et al. [35] also showed better performance using a value of 100 cd/m^2. We thus use an adapting white point luminance of 100 cd/m^2 (D65 is used as the chromaticity). We compare that with ΔE_{ITP} which is based on the IC_TC_P color representation and requires no white-point assumption. We also compare with ΔE_Z which is based on $J_za_zb_z$ color representation, a different HDR/WCG color representation. The color difference metrics are all pixel computations, that is, no spatial filtering or pooling is modeled. This means the achromatic and chromatic contrast sensitivity functions are not accounted, nor achromatic masking, nor chromatic masking, nor the achromatic-chromatic masking interactions.

Experimental Results and Discussion

In this section, we test the performance of various metrics and the proposed transformations on the five databases mentioned in the previous section. We evaluate the performance of the metrics by comparing the subjective scores with the scores predicted from the different metrics using a standardized method [37] used by the video quality experts group (VQEG). In that standardized approach, a monotonic logistic function is used to fit the objective prediction to the subjective scores as follows:

$$f = \alpha + \frac{\beta}{1 + e^{-\gamma.(x-\delta)}}, \tag{1}$$

where f is the fitted objective score, x is the predicted score using different techniques and $\alpha, \beta, \gamma, \delta$ are the parameters that define the shape of the logistic fitting function.The fit is computed by minimizing the least squares error between the subjective and the fitted objective scores. This mapping function is used to mimic the fact that high-level cognitive processes are required to map the lower-level perceptions to a score. The rationale is that the various metrics can model low-level perception, but that high-level cognitive processes are required to arrive at a score. As a simplified model of this internal mapping step, the logistic function with variable parameters is currently being used as a surrogate until better understanding is achieved. Please note that the subjective scores for each database have been made available by the respective authors.

We use the following four standard evaluation procedures and criteria [37] to measure the performance – Pearson Linear Correlation Coefficient (PLCC) and Root Mean Square Error (RMSE) for measuring prediction accuracy, Spearman Rank-Order Correlation Coefficient (SROCC) for prediction monotonicity and Outlier Ratio (OR) is used to determine prediction consistency. Lower values of RMSE and OR, and higher values of PLCC and SROCC indicates better performance.

We compare the performance of the different metrics and report the results in Tables 1, 2, 3, 4 and 5 for databases 1, 2, 3, 4 and 5 respectively. As mentioned in the previous section, we consider two different variations of HDR metrics. The two variations for HDR-VDP-2 are shown as HDR-VDP-2 (Computed using linear luminance values) and HDR-VDP-2_I (Suffix '_I' shows that it is computed using I channel from IC_TC_P color representation). In the latter case, only the I channel is given as input. Likewise, for HDR-VQM. The three different variations of SDR metrics (in particular MS-SSIM) are shown as MS-SSIM (Computed using linear luminance values), MS-SSIM_PQ_Y (Suffix '_PQ_Y' shows that it is computed using Y channel in '_PQ' non-linear domain) and MS-SSIM_I (Computed using I channel from IC_TC_P color representation). Likewise, for VIFp and FSIM. The three different color difference metrics are shown as ΔE_{00}, ΔE_Z and ΔE_{ITP}. For each database, the best performing HDR metric is highlighted in bold; best SDR metric is italicized and best color difference metric is underlined. For some metrics we show results from previous publications. Those are denoted by 'From [Reference#]'. For instance, in Table 1, we present results of MS-SSIM using linear values from [5].

We can also see that the overall performance of all the metrics (including HDR, SDR and color difference) can be improved using IC_TC_P color representation. For instance on database 5, the PLCC of HDR-VQM increases from 0.771 to 0.832; PLCC of MS-SSIM increases from 0.472 in linear to 0.847 in PQ using Y channel to 0.872 using I from IC_TC_P color representation and PLCC of ΔE_{ITP} is 0.693 compared to the PLCC of ΔE_{00}, which is 0.398.

In general, we can see that HDR metrics are better in performance than SDR metrics, which in turn are better than color difference metrics. The exception being databases 4 and 5, where SDR metrics in perceptual domain are slightly better than HDR metrics. The overall trend is not surprising because the HDR models are more advanced and closely mimic the human visual system by taking into account point spread function of the eye, masking effects and contain additional information such as having orientation channels. In case of HDR metrics, their calibration and HVS front-end non-linearities evenly distribute the perception of distortions across the image's full tone-scale resulting in improved performance. While applying such front-end non-linearities to the SDR metrics does improve their performance, on average it doesn't outperform HDR metrics.

Database 1 seems less selective and most metrics already have very high correlation and low error on that database. We observe that the overall performance of all the metrics is poorer on Database 4 compared to other databases. This might be due to the fact that Database 4 has a wide variety of artifacts. Some distortions such as gamut mismatch might be clearly visible but not associated with loss of quality for some viewers.

Amongst the HDR metrics, HDR-VDP-2 outperforms HDR-VQM on all databases except database 1. The performance is similar on database 3 (they are tied on database 3 for PLCC). Both metrics are the most computationally expensive, with HDR-VDP-2 being about 4x that of HDR-VQM. HDR-VQM is intended as a temporal video metric, while HDR-VDP-2 was intended to be used with still images. That aspect may help explain the behavior of HDR-VQM, whereas the high computational cost of HDR-VDP-2 helps its performance, and the video capability (while being applied to still images) of HDR-VQM may cause it to perform negatively with the other databases.

Using the I channel from IC_TC_P color representation consistently improves the results for both HDR metrics across each of the five databases. Applying HDR metrics using absolute linear values can be considered to be state-of-the-art and we further improve its performance using IC_TC_P color representation. While

Table 1: Performance comparison on Database 1 [16]

Method	PLCC	SROCC	RMSE	OR
HDR METRICS				
HDR-VDP-2 From [15]	0.951	0.951	0.48	0.370
HDR-VQM From [15]	0.961	0.957	0.428	0.392
HDR-VDP-2_I	0.960	0.956	0.349	0.341
HDR-VQM_I	0.962	0.957	0.341	0.379
SDR METRICS				
MS-SSIM From [5]	0.854	0.877	0.652	0.758
VIFp From [5]	0.827	0.834	0.705	0.666
FSIM From [5]	0.893	0.916	0.564	0.683
MS-SSIM_PQ_Y From [20, 21, 22]	0.932	0.926	0.448	0.491
VIFp_PQ_Y From [20, 21, 22]	0.925	0.922	0.473	0.516
FSIM_PQ_Y From [20, 21, 22]	0.917	0.916	0.494	0.583
MS-SSIM_I	0.942	0.936	0.418	0.450
VIFp_I	0.926	0.922	0.473	0.483
FSIM_I	0.933	0.938	0.450	0.562
COLOR DIFFERENCE METRICS				
ΔE_{00}	0.794	0.790	0.764	0.645
ΔE_Z	0.667	0.671	0.938	0.737
$\underline{\Delta E_{ITP}}$	0.836	0.837	0.687	0.637

Table 2: Performance comparison on Database 2 [6, 30]

Method	PLCC	SROCC	RMSE	OR
HDR METRICS				
HDR-VDP-2 From [15]	0.938	0.928	10.12	0.44
HDR-VQM From [15]	0.930	0.917	10.72	0.53
HDR-VDP-2_I	0.944	0.932	9.96	0.45
HDR-VQM_I	0.933	0.918	10.66	0.53
SDR METRICS				
MS-SSIM	0.636	0.660	22.91	0.79
VIFp	0.721	0.754	21.04	0.6
FSIM	0.796	0.801	17.97	0.74
MS-SSIM_PQ_Y From [20, 21, 22]	0.879	0.873	14.20	0.56
VIFp_PQ_Y From [20, 21, 22]	0.924	0.914	11.31	0.62
FSIM_PQ_Y From [20, 21, 22]	0.889	0.885	13.56	0.56
MS-SSIM_I	0.907	0.899	12.51	0.57
VIFp_I	0.919	0.910	11.69	0.59
FSIM_I	0.904	0.900	12.65	0.52
COLOR DIFFERENCE METRICS				
ΔE_{00}	0.613	0.599	23.59	0.75
ΔE_Z	0.538	0.514	25.16	0.78
$\underline{\Delta E_{ITP}}$	0.714	0.729	20.81	0.74

the HDR metrics still have their PU font-end non-linearities, as opposed to PQ, inputting the I signal from IC_TC_P instead of the Y signal does slightly reshape the effective spectral aspects of the achromatic signal. We tried another variation of evaluating the HDR metrics where we give as input the PQ non-linear values. However, this results in worse performance since both these HDR metrics implicitly apply PU non-linearity to the inputs (not shown). Thus the front-end non-linearity would be applied twice. Also the pooling done at the end make such assumptions and does not directly translate to the PQ non-linearity.

Amongst the different variations of SDR metrics, SDR metrics computed in the linear domain consistently has the worst results. This finding is consistent with that of [5, 6] who showed those results in databases 1 and 2. We observe similar trends in databases 3, 4 and 5. This is expected because SDR metrics are designed for gamma encoded images with small range of luminance values (typically in the range of 0.1 - 100 cd/m^2) whereas the HDR images have a much larger dynamic range. Using PQ transformation improves the prediction of SDR metrics. Amongst the two variations that use PQ non-linearity we find that using IC_TC_P color representation (denoted using '_I' suffix) is better than using YC_bC_r (denoted using '_PQ_Y'). IC_TC_P implicitly applies PQ transformation as seen in Algorithm 1. Similar to HDR metrics, using IC_TC_P color representation results in best performance although most of the performance improvement comes from using PQ non-linearity essentially being the difference between the spectral details of I and Y. There were a few cases when using I from IC_TC_P resulted in worse performance than using PQ in Y viz., for database 3 where PLCC of VIFp drops from 0.924 for '_PQ_Y' to 0.919 for '_I'. However, the drop in performance is not much. We also find that MS-SSIM is the best among SDR metrics on databases 1, 3 and 4; VIFp works the best on database 2 and FSIM works the best on database 5. A surprising result is that FSIM computed in IC_TC_P color representation outperforms the more sophisticated HDR metrics on database 5.

Amongst the color difference metrics, we observe that overall ΔE_{ITP} outperforms ΔE_{00} and ΔE_Z on four out of the five databases. Rousselot et al. [15] also find similar trends in performance with regards to color difference metrics (they found a precursor to ΔE_{ITP} outperforming ΔE_{00} and ΔE_Z). That version didn't achieve as high correlation values (at least with regards to ΔE_{ITP}) as us. For instance, they report a PLCC of 0.8065 on Database 1 compared to our PLCC of 0.836 using DEITP. It is unclear if the slight difference in the metric caused the difference, or other possible unstated assumptions in their calculations. The improved performance of ΔE_{ITP} over ΔE_{00} is expected because the achromatic non-linearity of ΔE_{ITP} (encoding I using PQ) is known to better match the human visual threshold for the HDR luminance range [38]. In particular, the L* achromatic non-linearity of CIELAB is known to fail for dark regions (luminance levels less than 1 cd/m^2).

One surprising result was that ΔE_Z, which has been designed for HDR/WCG applications frequently performed poorly. It has been found previously that the offset for blue linearity correction in the $J_za_zb_z$ space could cause mismatches. Additionally, the

Table 3: Performance comparison on Database 3 [32]

Method	PLCC	SROCC	RMSE	OR
HDR METRICS				
HDR-VDP-2 From [15]	0.898	0.891	0.563	0.586
HDR-VQM From [15]	0.894	0.887	0.532	0.514
HDR-VDP-2_I	0.905	0.898	0.426	0.642
HDR-VQM_I	0.905	0.900	0.436	0.457
SDR METRICS				
MS-SSIM	0.527	0.521	0.854	0.828
VIFp	0.543	0.491	0.844	0.814
FSIM	0.523	0.507	0.857	0.778
MS-SSIM_PQ_Y	0.853	0.841	0.523	0.657
VIFp_PQ_Y	0.692	0.641	0.725	0.757
FSIM_PQ_Y	0.840	0.827	0.544	0.635
MS-SSIM_I	0.855	0.844	0.520	0.650
VIFp_I	0.679	0.623	0.738	0.750
FSIM_I	0.841	0.827	0.543	0.642
COLOR DIFFERENCE METRICS				
$\underline{\Delta E_{00}}$	0.718	0.697	0.700	0.764
ΔE_Z	0.649	0.633	0.765	0.771
ΔE_{ITP}	0.639	0.632	0.773	0.785

Table 4: Performance comparison on Database 4 [34]

Method	PLCC	SROCC	RMSE	OR
HDR METRICS				
HDR-VDP-2 From [15]	0.871	0.868	12.55	0.457
HDR-VQM From [15]	0.867	0.833	14.72	0.561
HDR-VDP-2_I	0.874	0.834	11.46	0.322
HDR-VQM_I	0.853	0.807	12.31	0.354
SDR METRICS				
MS-SSIM	0.675	0.692	17.43	0.489
VIFp	0.832	0.811	13.08	0.333
FSIM	0.76	0.783	15.36	0.427
MS-SSIM_PQ_Y	0.867	0.827	11.77	0.270
VIFp_PQ_Y	0.798	0.777	14.29	0.354
FSIM_PQ_Y	0.814	0.793	13.72	0.322
MS-SSIM_I	0.879	0.849	11.24	0.270
VIFp_I	0.826	0.792	13.32	0.343
FSIM_I	0.820	0.796	13.50	0.322
COLOR DIFFERENCE METRICS				
ΔE_{00}	0.273	0.219	22.73	0.604
ΔE_Z	0.304	0.296	22.51	0.666
$\underline{\Delta E_{ITP}}$	0.390	0.320	21.75	0.656

final equations of the $J_z a_z b_z$ color space was optimized using a dataset with a diffuse white at 997 cd/m^2, which might be applicable for outdoor prints, but is not ideal for video displays.

Database 3 was a challenge for ΔE_{ITP}. Database 3 was unique in that it was not conducted in a dark surround; the ambient surround was 130 cd/m^2. Since ΔE_{ITP} is designed for most critical conditions (when a display has the largest physical dynamic range with no black level elevation due to screen reflections), it tended to over-predict the differences in this higher ambient case (i.e., it under-predicts the quality). Another unique factor for Database 3 is that they did not use a full-reference methodology. Rather, ACR (absolute category rating) methodology is used without using a labeled reference. This method uses a single stimulus. Without knowing what the image should look like, which is important to aspects of creative intent, there is the possibility for more variability in individual interpretation. This is a particular issue for color, as often incorrect colors may be plausible, but would be deviations from artistic intent and thus affect elements of the narrative (e.g., symbolism, emotion).

Our findings are summarized in Table 6.

Conclusion & Future Work

In this paper we evaluate several image quality metrics to assess the quality of HDR content. Specifically, we consider various HDR, SDR and color difference metrics. We use five different databases containing a wide variety of distortions for evaluation. We find that overall using HDR metrics result in the best performance followed by SDR metrics and finally by color difference metrics. We used a few variations of these metrics and found that the performance of all the metrics (including HDR, SDR and color difference) can be improved using IC_TC_P color representation. For HDR metrics which traditionally uses linear luminance values, using the I channel from IC_TC_P results in improved performance. For SDR metrics, using PQ transfer function significantly improves the results when compared to using in linear domain and using IC_TC_P color representation further improves its performance. Amongst color difference metrics, ΔE_{ITP} has better performance than both ΔE_{00} and ΔE_Z.

Both state-of-the-art HDR metrics (HDR-VDP-2 and HDR-VQM) are modeled in PU domain. We would like to make these metrics work in PQ domain because the PU non-linearity has less sensitivity in the dark regions than PQ. Also, more advanced spatio-chromatic modeling could further improve performance of the color difference metrics. Further improvements in analysis by re-coding the the HDR metrics to actually replace the achromatic PU non-linearity with a color IC_TC_P front-end would allow for a better understanding of the gains possible from use of the IC_TC_P color representation.

References

[1] Mantiuk, R., Kim, K. J., Rempel, A. G., and Heidrich, W., "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Trans. Graph.* **30**, 40:1–40:14 (July 2011).

[2] Narwaria, M., Mantiuk, R., Silva, M. P. D., and Callet, P. L., "HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images," *Journal of Electronic Imaging* **24**, 24 – 24 – 3 (2015).

[3] Aydin, T., Mantiuk, R., Myszkowski, K., and Seidel, H. P., "Dynamic range independent image quality assessment," *ACM Trans. Graph.* **27**, 69:1–69:10 (Aug. 2008).

[4] Narwaria, M., Silva, M. P. D., and Callet, P. L., "HDR-VQM: An Objective Quality Measure for High Dynamic Range Video," *Signal Processing: Image Communica-*

Table 5: Performance comparison on Database 5 [15]

Method	PLCC	SROCC	RMSE	OR
HDR METRICS				
HDR-VDP-2 From [15]	0.860	0.867	11.3	0.354
HDR-VQM From [15]	0.771	0.773	14.11	0.531
HDR-VDP-2_I	0.871	0.874	10.99	0.302
HDR-VQM_I	0.832	0.833	12.30	0.427
SDR METRICS				
MS-SSIM	0.472	0.413	19.56	0.666
VIFp	0.648	0.641	16.90	0.645
FSIM	0.661	0.648	16.65	0.541
MS-SSIM_PQ_Y	0.847	0.841	11.78	0.406
VIFp_PQ_Y	0.822	0.825	12.70	0.479
FSIM_PQ_Y	0.890	0.895	10.10	0.343
MS-SSIM_I	0.872	0.871	10.86	0.395
VIFp_I	0.833	0.830	12.35	0.427
FSIM_I	0.893	0.896	9.99	0.364
COLOR DIFFERENCE METRICS				
ΔE_{00}	0.398	0.311	20.35	0.593
ΔE_Z	0.329	0.288	20.95	0.708
$\underline{\Delta E_{ITP}}$	0.693	0.698	16.01	0.583

tion **35**, 46–60 (July 2015).

[5] Hanhart, P., Bernardo, M., Pereira, M., Pinheiro, A. M. G., and Ebrahimi, T., "Benchmarking of objective quality metrics for HDR image quality assessment," *EURASIP Journal on Image and Video Processing* **2015**, 39 (Dec 2015).

[6] Zerman, E., Valenzise, G., and Dufaux, F., "An extensive performance evaluation of full-reference HDR image quality metrics," *Quality and User Experience* **2**, 5 (Apr 2017).

[7] Wang, Z., Simoncelli, E. P., and Bovik, A. C., "Multiscale structural similarity for image quality assessment," in [*37th Asilomar Conference on Signals, Systems and Computers*], **2**, 1398–1402, IEEE (Nov. 2003).

[8] Sheikh, H. R., Bovik, A. C., and de Veciana, G., "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing* **14**, 2117–2128 (Dec 2005).

[9] Sheikh, H. R. and Bovik, A. C., "Image information and visual quality," *IEEE TIP* **15**, 430–444 (Feb 2006).

[10] Zhang, L., Zhang, L., Mou, X., and Zhang, D., "FSIM: A feature similarity index for image quality assessment," *IEEE TIP* **20**, 2378–2386 (Aug 2011).

[11] Nafchi, H. Z., Shahkolaei, A., Moghaddam, R. F., and Cheriet, M., "FSITM: A feature similarity index for tone-mapped images," *IEEE Signal Processing Letters* **22**, 1026–1029 (Aug 2015).

[12] Wang, Z. and Bovik, A. C., "A universal image quality index," *IEEE Signal Processing Letters* **9**, 81–84 (March 2002).

[13] Azimi, M., Banitalebi-Dehkordi, A., Dong., Y., Pourazad, M., and Nasiopoulos, P., "Evaluating the performance of existing full-reference quality metrics on high dynamic range (HDR) video content," in [*International Conference on Multimedia Signal Processing*], (November 2014).

[14] Hanhart, P., Rerabek, M., and Ebrahimi, T., "Subjective and objective evaluation of hdr video coding technologies," in [*QoMEX*], 1–6 (June 2016).

[15] Rousselot, M., Le Meur, O., Cozot, R., and Ducloux, X., "Quality assessment of hdr/wcg images using hdr uniform color spaces," *Journal of Imaging* **5**(1) (2019).

[16] Korshunov, P., Hanhart, P., Richter, T., Artusi, A., Mantiuk, R., and Ebrahimi, T., "Subjective quality assessment database of HDR images compressed with jpeg xt," in [*QoMEX*], 1–6 (May 2015).

[17] "BT2100: Image parameter values for high dynamic range television for use in production and international programme exchange," *International Telecommunication Union* (July 2016).

[18] Safdar, M., Cui, G., Kim, Y. J., and Luo, M. R., "Perceptually uniform color space for image signals including high dynamic range and wide gamut," *Opt. Express* **25**, 15131–15151 (Jun 2017).

[19] Fairchild, M. D. and Chen, P.-H., "Brightness, lightness, and specifying color in high-dynamic-range scenes and images," in [*Image Quality and System Performance VIII*], Farnand, S. P. and Gaykema, F., eds., **7867**, 233 – 246, International Society for Optics and Photonics, SPIE (2011).

[20] Choudhury, A. and Daly, S., "HDR image quality assessment using machine-learning based combination of quality metrics," in [*2018 IEEE Global Conference on Signal and Information Processing*], 91–95 (2018).

[21] Choudhury, A. and Daly, S., "Combining quality metrics using machine learning for improved and robust HDR image quality assessment,," in [*IS&T Electronic Imaging, Image Quality and System Performance XVI*], (January 2019).

[22] Choudhury, A. and Daly, S., "Combining quality metrics for improved HDR image quality assessment," in [*2nd IEEE Conference on Multimedia Information Processing and Retrieval, MIPR 2019, San Jose, CA, USA, March 28-30, 2019*], 179–184 (2019).

[23] T. Aydin, R. Mantiuk, H. S., "Extending quality metrics to full luminance range images," (2008).

[24] Miller, S., Nezamabadi, M., and Daly, S., "Perceptual signal coding for more efficient usage of bit codes," in [*The 2012 Annual Technical Conference Exhibition*], 1–9 (Oct 2012).

[25] Luo, M. R., Cui, G., and Rigg, B., "The development of the CIE 2000 colour-difference formula: CIEDE2000," *Color Research & Application* **26**(5), 340–350 (2001).

[26] ISO/CIE 11664-6:2014(E), "Colorimetry - Part 6: CIEDE2000 Colour-Difference Formula," Standard (2014).

[27] ITU-R BT. 2124, "Objective metric for the assessment of the potential visibility of colour differences in television," Standard (Jan 2019).

[28] Mantiuk, R., Myszkowski, K., and Seidel, H.-P., "A perceptual framework for contrast processing of high dynamic range images," *ACM Trans. Appl. Percept.* **3**, 286–308 (July 2006).

[29] Reinhard, E., Stark, M., Shirley, P., and Ferwerda, J., "Photographic tone reproduction for digital images," *ACM Trans. Graph.* **21**, 267–276 (July 2002).

[30] Valenzise, G., Simone, F. D., Lauga, P., and Dufaux, F.,

Table 6: Summary of results. The overall best metric is bold and italicized

Database	Key characteristics of database	Best HDR Metric	Best SDR Metric	Best Color Difference Metric
Database 1 [16]	JPEG-XT coding, chroma sub-sample SIM2 monitor (BT. R. 709) Dim ambient	***HDR-VQM in ITP***	MS-SSIM in ITP	ΔE_{ITP}
Database 2 [6, 30]	JPEG, JPEG-XT, JPEG2000 Backward Compatible SIM2 monitor (BT. R. 709) Mid ambient	***HDR-VDP-2 in ITP***	VIFp in PQ_Y	ΔE_{ITP}
Database 3 [32]	JPEG Backwards Compatible HDR-SDR Tone-mapping SIM2 monitor (BT. R. 709) Bright ambient No Reference	***HDR-VQM in ITP***	MS-SSIM in ITP	ΔE_{00}
Database 4 [34]	HEVC (incl. chroma) Gaussian noise Gamut mapping BVM-X300 monitor (P3) Mid ambient	HDR-VDP-2 in ITP	***MS-SSIM in ITP***	ΔE_{ITP}
Database 5 [15]	HEVC (incl. chroma) WCG + HDR BVM-X300 monitor (P3) Mid ambient	HDR-VDP-2 in ITP	***FSIM in ITP***	ΔE_{ITP}
Comments	Full Reference (unless specified) Dim ambient = home theater Mid ambient = living room night Bright ambient = daytime office			

"Performance evaluation of objective quality metrics for hdr image compression," in [*SPIE optical engineering + applications, International Society for Optics and Photonics*], (2014).

[31] Mai, Z., Mansour, H., Mantiuk, R., Nasiopoulos, P., Ward, R., and Heidrich, W., "Optimizing a tone curve for backward-compatible high dynamic range image and video compression," *IEEE Transactions on Image Processing* **20**, 1558–1571 (June 2011).

[32] Narwaria, M., Perreira Da Silva, M., Le Callet, P., and Pepion, R., "Tone mapping-based high-dynamic-range image compression: Study of optimization criterion and perceptual quality," *Optical Engineering* **52**, 102008 (10 2013).

[33] Kuang, J., Johnson, G. M., and Fairchild, M. D., "icam06: A refined image appearance model for hdr image rendering," *Journal of Visual Communication and Image Representation* **18**(5), 406 – 414 (2007). Special issue on High Dynamic Range Imaging.

[34] Rousselot, M., Auffret, E., Ducloux, X., Le Meur, O., and Cozot, R., "Impacts of viewing conditions on hdr-vdp2," in [*EUSIPCO*], 1442–1446 (Sept 2018).

[35] Choudhury, A., Pytlarz, J., and Daly, S., "HDR and WCG image quality assessment using color difference metrics," in [*SMPTE 2019 Annual Technical Conference and Exhibition*], (Oct 2019).

[36] Reinhard, E., Stauder, J., and Kerdranvat, M., "An assessment of reference levels in hdr content," *SMPTE Motion Imaging Journal* **128**, 20–27 (April 2019).

[37] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," (2003).

[38] Hoffman, D. M., Johnson, P. V., Kim, J. S., Vargas, A. D., and Banks, M. S., "240hz oled technology properties that can enable improved image quality," *Journal of the Society for Information Display* **22**(7), 346–356 (2014).

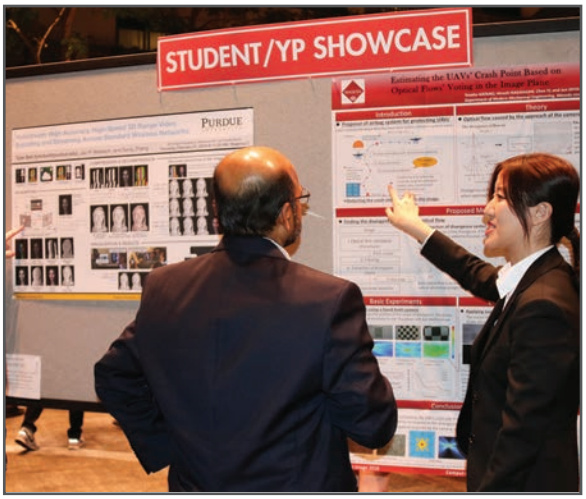
JOIN US AT THE NEXT EI!

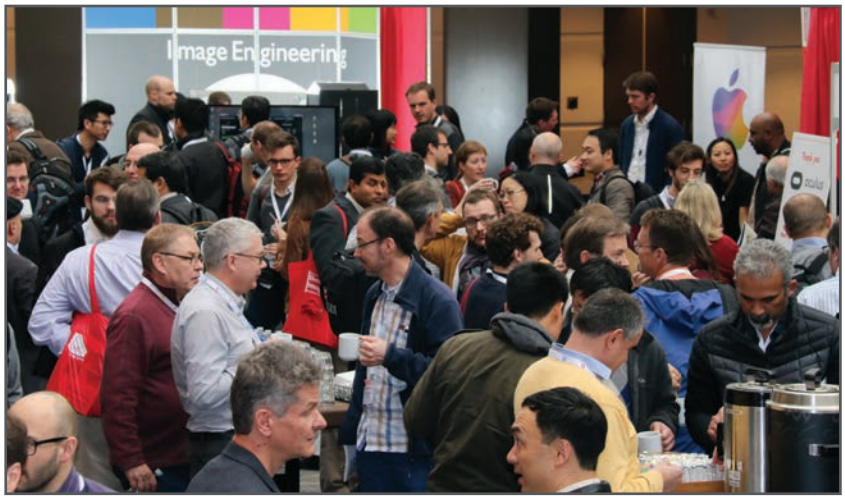
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org