An extensive performance evaluation of full-reference HDR image quality metrics

Emin Zerman, Giuseppe Valenzise, Frederic Dufaux

Noname manuscript No.
(will be inserted by the editor)

An extensive performance evaluation of full-reference HDR image quality metrics

Emin Zerman · Giuseppe Valenzise · Frederic Dufaux

Received: date / Accepted: date

Abstract High dynamic range (HDR) image and video technology has recently attracted a great deal of attention in the multimedia community, as a mean to produce truly realistic video and further improve the Quality of Experience (QoE) of emerging multimedia services. In this context, measuring the quality of compressed HDR content plays a fundamental role. However, full-reference (FR) HDR visual quality assessment poses new challenges with respect to the conventional low dynamic range case. Quality metrics have to be redesigned or adapted to HDR, and understanding their reliability to predict users' judgments is even more critical due to the still limited availability of HDR displays to perform subjective evaluations. The goal of this paper is to provide a complete and thorough survey of the performance of the most popular HDR FR image quality metrics. To this end, we gather several existing HDR image databases with subjective quality annotations, in addition to a new one created by ourselves. After aligning the scores in these databases, we obtain an extensive set of 690 compressed HDR images, along with their subjective quality. Next, we analyze in depth many FR metrics, including those used in MPEG standardization, using both classical correlation analyses and classification accuracy. We believe that our results could serve as the most complete and comprehensive benchmark of image quality metrics in the field of HDR image compression.

Keywords High dynamic range · quality assessment · image coding · subjective test

E. Zerman
LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013 Paris, France
Tel.: +33-1-45817240
E-mail: emin.zerman@telecom-paristech.fr

G.Valenzise
L2S, CNRS-CentraleSupélec-Université Paris-Sud, Gif-sur-Yvette, France
E-mail: giuseppe.valenzise@l2s.centralesupelec.fr

F.Dufaux
L2S, CNRS-CentraleSupélec-Université Paris-Sud, Gif-sur-Yvette, France
E-mail: frederic.dufaux@l2s.centralesupelec.fr

1 Introduction

High Dynamic Range (HDR) imaging enables to capture, represent and reproduce a wide range of colors and luminous intensities present in everyday life, ranging from bright sunshine to dark shadows [11]. These extended capabilities are expected to significantly improve the Quality of Experience (QoE) of emerging multimedia services with respect to conventional Low Dynamic Range (LDR) technology. Commercial HDR video cameras and displays are becoming available, and parts of the HDR end-to-end delivery chain such as image and video compression are currently matter of standardization activities in MPEG [19,31] and JPEG [53]. In this context, evaluating the visual quality of compressed HDR pictures is of critical importance in order to design and optimize video codecs and processing algorithms.

Evaluating HDR visual quality presents new challenges with respect to conventional LDR quality assessment [45]. The higher peak brightness and contrast offered by HDR increases the visibility of artifacts, and at the same time changes the way viewers focus their attention compared to LDR [41]. Moreover, color distortion assumes a major role in the overall quality judgment, as a result of the increased luminance level [15]. Since these and other factors intervene in a complex

way to determine HDR visual quality, the most accurate approach to assess it is, in general, through subjective test experiments. However, these are expensive to design and implement, require specialized expertize and are time-consuming. Furthermore, in the case of HDR, subjective testing requires specialized devices such as HDR displays, which still have a high cost and a limited diffusion. Therefore, designing and tuning *full-reference* (fidelity) quality metrics for HDR content is very timely, and has motivated research in both the multimedia and computer graphics community in the past few years [3, 35, 42–44].

Two main approaches have been proposed to measure HDR fidelity. On one hand, some metrics require modeling of the human visual system (HVS), such as the HDR-VDP [35] or HDR-VQM [42] metrics for images and videos, respectively. For example, the HDR-VDP metric accurately models the early stages of HVS, including intra-ocular scattering, luminance masking, and achromatic response of the photoreceptors, in order to precisely predict the visibility and strength of per pixel distortion. On the other hand, one can resort to metrics developed in the context of LDR imagery, such as simple arithmetic (PSNR, MSE), structural (SSIM [62] and its multiscale version [61]) and information-theoretic (e.g., VIF [55]) metrics. All these LDR metrics are based on the assumption that pixel values are perceptually linear, i.e., equal increments of pixel values correspond to equivalent changes in the perceived luminance. This is not true in the case of HDR content, where pixel values store *linear* light, i.e., pixels are proportional to the physical luminance of the scene. Instead, human perception has a more complex behavior: it can be approximated by a square-root in low luminance values and is approximately proportional to luminance ratios in higher luminance values, as expressed by the DeVries-Rose and Weber-Fechner laws, respectively [28]. Thus, in order to employ these metrics, the HDR content needs to be perceptually linearized, e.g., using a logarithmic or perceptually uniform (PU) encoding [3].

The capability of both kinds of fidelity metrics to predict viewers' mean opinion scores (MOS) has been assessed in a number of recent subjective studies using compressed HDR pictures [17, 37, 39, 59]. Nevertheless, the results of these studies show sometimes discrepancies in their conclusions about the ability of these metrics to yield consistent and accurate predictions of MOSs. For instance, the correlation values of PU-SSIM, i.e., SSIM metric applied after the PU encoding of [3], differ substantially between the study of Narwaria et al. [43] and that of Valenzise et al. [59]. The difference is basically related to the size and characteristic of the subjective material. In [59], the performance of objective metrics was assessed on a small image database (50 subjectively annotated images), using different coding schemes including JPEG, JPEG 2000 and JPEG-XT. In [43], the authors evaluate metric correlations using a number of subjectively annotated databases, with variegate distortion and, especially, with scores gathered in separated tests (each with their own experimental conditions). Both studies have their advantages and limitations, which renders difficult to extract a simple and clear conclusion about the performance of fidelity metrics. In other cases, such as [17], metrics have been tested on a single type of distortion only (specifically JPEG-XT compression), thus it is desirable to extend those conclusions to more realistic and variegate conditions.

The aim of this paper is to bring more clarity in this field, by providing an extensive, reliable, and consistent benchmark of the most popular HDR image fidelity metrics. To this end, we collected as many as possible publicly available databases of HDR compressed images with subjective scores, in addition to proposing a new one which mixes different codecs and pixel encoding functions. This gives a total of 690 HDR images, which is up to our knowledge the largest set on which HDR metrics have been tested so far. We then align the MOSs of these databases using the iterated nested least square algorithm (INLSA) proposed in [50], in order to obtain a common subjective scale. Based on this data, we analyze the prediction accuracy and the discriminability (i.e., the ability of detecting when two images have different perceived quality) of 25 fidelity metrics, including those currently tested in MPEG standardization.

The main contributions of this paper include:

- the most extensive evaluation (using 690 subjectively annotated HDR images) of HDR full-reference image quality metrics available so far;
- the proposal of a new subjective database with 50 distorted HDR images, combining 3 image codecs and 2 pixel encoding algorithm (SMPTE-2084 Perceptual Quantization [57] and a global tone-mapping operator);
- an evaluation of metric discriminability, that complements the conventional statistical accuracy analysis, based on a novel classification approach.

Assessment of image quality is different from the assessment of video quality, as HVS has different temporal mechanisms. Nevertheless, image quality metrics are often applied to video on a frame-by-frame basis, e.g., PSNR or SSIM. Therefore, the result of this work could be indicative of frame-by-frame objective metrics performance in video as well.

The rest of this paper is organized as follows. Section 2 describes the subjective databases considered within this paper. The alignment procedure is explained in Section 3. In Section 4, existing objective image quality metrics have been compared using both statistical evaluation and a classification approach. Finally, Section 5 concludes the paper.

2 Considered subjective databases

Although there are several publicly available repositories of high-quality HDR pictures [9,10,13,14,48], there is only a small number of subjectively annotated image quality databases. For this study, we selected four publicly available HDR image quality assessment databases, in addition to proposing a new one described in Section 2.5. Each database contains compressed HDR pictures with related subjective scores. The databases differ in size, kind of distortion (codec) and subjective methodology. A brief description of these databases is given in the following, while a summary of their characteristics is reported in Table 1. The interested reader can refer to original publications for further details.

2.1 Database #1 - Narwaria et al. (2013) [39]

In the work of Narwaria et al. [39], a tone mapping based HDR image compression scheme has been proposed and assessed via a subjective test. Subjective scores were collected from 27 observers, using a SIM2 HDR47E S 4K display in a 130 cd/m^2 illuminated room. The participants were asked to rate overall image quality using the Absolute Category Rating with Hidden Reference (ACR-HR) methodology, employing a five-level discrete scale where 1 is bad and 5 is excellent quality. The test material was obtained from 10 pristine HDR pictures, including both indoor and outdoor, natural or computer-generated scenes. The distorted images are generated through a backward compatible scheme [63]: the HDR image is first converted to LDR by using a tone mapping operator (TMO); then, the LDR picture is coded using a legacy image codec; finally, the compressed image is expanded by inverse tone mapping to the original HDR range. The coding scheme in [39] employs iCAM06 [27] as TMO, and JPEG compression at different qualities. In addition, the authors proposed two criteria to optimize the quality of the reconstructed HDR. As a result, a total of 10 contents $\times$ 7 bitrates $\times$ 2 optimization criteria = 140 test images were evaluated. This database is publicly available at `http://ivc.univ-nantes.fr/en/databases/JPEG_HDR_Images/`.

Table 1 Number of observers, subjective methodology, number of stimuli, compression type and tone mappings employed in the HDR image quality databases used in this paper. TMOs legend: *AS*: Ashikmin, *RG*: Reinhard Global, *RL*: Reinhard Local, *DR*: Durand, *Log*: Logarithmic, *MT*: Mantiuk.

No	Obs.	Meth.	Stim.	Compr.	TMO
#1 [39]	27	ACR-HR	140	JPEG[1]	iCAM [27]
#2 [40]	29	ACR-HR	210	JPEG 2000[1]	AS [2] RG [52] RL [52] DR [12] Log
#3 [25]	24	DSIS	240	JPEG-XT	RG [52] MT [33]
#4 [59]	15	DSIS	50	JPEG[1] JPEG 2000[1] JPEG-XT	Mai [32]
#5	15	DSIS	50	JPEG[1] JPEG 2000[1]	Mai [32] PQ [36, 57]

The analysis in [39] shows that Mean Squared Error (MSE) and Structural Similarity Index Measure (SSIM) perform well in estimating human predictions and ordering distorted images when each content is assessed separately. However, these results do not apply when different contents are considered at the same time. HDR-VDP-2 was found to be the best performing (in terms of linear correlation with MOSs) metric, but not statistically different from the metric proposed in [38].

2.2 Database #2 - Narwaria et al. (2014) [40]

Narwaria et al. [40] evaluate subjectively the impact of using different TMOs in HDR image compression. The test material includes 6 original scenes, both indoor and outdoor, from which a total of 210 test images were created using JPEG 2000 image compression algorithm after the application of several TMOs, including Ashikmin [2], both local and global versions of Reinhard [52], Durand [12], and logarithmic TMO. The experiment setup was the same as in Narwaria et al. (2013) Database #1 described above. The subjective test is conducted with 29 observers using ACR-HR methodology.

Results show that the choice of TMO greatly affects the quality scores. It is also found that local TMOs, with the exception of Durand's, generally yield better results than global TMOs as they tend to preserve more

[1] The distorted images are generated through a scalable coding scheme [63]: the HDR image is converted to LDR using a TMO; then, the LDR picture is encoded & decoded by a legacy codec; finally, the image is converted back to HDR range.

details. No evaluation of objective quality metrics is reported in the original paper [40].

2.3 Database #3 - Korshunov et al. (2015) [25]

In the study of Korshunov et al. [25], an HDR image quality database, publicly available at `http://mmspg.epfl.ch/jpegxt-hdr`, has been created using backward-compatible JPEG-XT standard [53] with different profiles and quality levels. For this database, 240 test images have been produced, using either Reinhard [52] or Mantiuk [33] TMO for the base layer, 4 bit rates for each original image and 3 profiles of JPEG-XT. The test room was illuminated with a 20 lux lamp, and a SIM2 HDR display was used. At any time, 3 observers took the test simultaneously. The subjective scores were collected from 24 participants, using Double Stimulus Impairment Scale (DSIS) Variant I methodology, i.e., images were displayed side-by-side, one of the images was the reference and the other the distorted one.

This subjective databases has been used in the work of Artusi et al. [1]. In this work, an objective evaluation of JPEG-XT compressed HDR images has been carried out. The results show that LDR metrics such as PSNR, SSIM, and multi-scale SSIM (MSSIM) give high correlation scores when they are used with the PU encoding of [3], while the overall best correlated quality metric is HDR-VDP-2.

2.4 Database #4 - Valenzise et al. (2014) [59]

Valenzise et al. [59] were the first to collect subjective data with the specific goal to analyze the performance of HDR image fidelity metrics. Their database is composed of 50 compressed HDR images, obtained from 5 original scenes in the Fairchild HDR image survey [14]. Three different coding schemes have been used to produce the test material, i.e., JPEG, JPEG 2000 and JPEG-XT. In the first two cases, the HDR image is first tone mapped to LDR using the minimum-MSE TMO proposed by Mai et al. [32]. The images were displayed on a SIM2 HDR47E S 4K display, with an ambient luminance of 20 cd/m^2. Subjective scores were collected using DSIS methodology, i.e., pairs of images (original and distorted) were presented to the viewers, who had to evaluate the level of annoyance of distortion in the second image on a continuous quality scale ranging from 0 to 100, where 0 corresponds to very annoying artifacts and 100 to imperceptible artifacts. Fifteen observers rated the images. The database is available at `http://perso.telecom-paristech.fr/~gvalenzi/download.htm`.

The results of this study showed that LDR fidelity metrics could accurately predict image quality, provided that the display response is somehow taken into account (in particular, its peak brightness), and that a perceptually uniform (PU) encoding [3] is applied to HDR pixel values to make them linear with respect to perception.

2.5 Database #5 - New subjective database

In addition to the databases described above, we construct a new subjective HDR image database of 50 images, as an extension to our previous work [59]. The new database features 5 original contents, selected in such a way to be representative of different image features, including the dynamic range, image key and spatial information. The five contents are shown in Figure 1. The images *"Balloon"*, *"FireEater2"*, and *"Market3"* are chosen among the frames of the MPEG HDR sequences proposed by Technicolor [29]. *"Showgirl"* is taken from Stuttgart HDR Video Database [16]. *"Typewriter"* is from HDR photographic survey dataset [14]. All images have either 1920×1080 pixels spatial resolution, or are zero-padded to have the same resolution.

Similarly to [59], the test images are obtained by using a backward compatible HDR coding scheme [63], using JPEG and JPEG 2000 (with different bitrates) as LDR codecs. We did not include JPEG-XT in this experiment, since some of the contents we selected (e.g., *"Showgirl"* and *"Typewriter"*) were already part of the Database #3. In order to convert HDR to LDR, we use two options: *i)* the TMO of Mai et al. [32]; and *ii)* the electro-optical transfer function SMPTE ST 2084 [36, 57], commonly known as Perceptual Quantization (PQ). The latter is a fixed, content-independent transfer function which has been designed in such a way that the increments between codewords have minimum visibility, according to Barten's contrast sensitivity function [5]. We choose this transfer function as an alternative to tone mapping, as it has been proposed as the anchor scheme in current MPEG HDR standardization activities [31]. Both PQ and Mai et al.'s TMO are applied per color channel.

The test environment and methodology are carefully controlled to be the same as in Database #4 (Valenzise et al. (2014)) [59]. The DSIS methodology is employed, where the reference image is shown for 6 seconds, followed by 2 seconds of mid-gray screen and 8 seconds of degraded image. The asymmetry in timing between distorted and reference image is determined in a pilot test, taking into account the fact that the reference image is shown several times, while the degraded image is different at each round and requires a longer evaluation interval. After both the original and distorted image

(a) "Balloon" (b) "FireEater2" (c) "Market3" (d) "Showgirl" (e) "Typewriter"

Fig. 1 Original contents for the new proposed image database described in Section 2.5, rendered using the TMO in [34].

are displayed, the observer takes all the time she/he needed to rate the level of annoyance on the same continuous scale as in [59]. The sequence of tested images is randomized to avoid context effects [8]. Moreover, too bright ("Market3") and too dark ("FireEater2") stimuli are not placed one after another in order to avoid any masking caused by sudden brightness change. In addition to randomization, stabilizing images (one from each content and featuring each quality level) are shown in the beginning of the experiment to stabilize viewers' votes (which are discarded for those images).

In addition to the contents reported in Figure 1, a small subset of the stimuli of Database #4 was included in the test. This enabled to align the two databases, #4 and #5, in order for the corresponding MOS values to be on the same scale [51]. Thus, in the following we will refer to the union of these two databases as Database #4 & 5.

A panel of 15 people (3 women, 12 men; average age of 26.8 years), mainly Ph.D. students naive to HDR technology and image compression, participated to the test. Subjects reported normal or corrected-to-normal vision. The outlier detection and removal procedure described in BT.500-13 [21] resulted in no detected outlier. Then, mean opinion scores and their confidence interval (CI) were computed assuming data follows a *t-Student* distribution[2].

3 Alignment of Database MOSs

During the training phase, the subjects are generally instructed to use the whole range of grades (or distortions) in the scale while evaluating. However, the quality of the test material for different experiments may not be the same when they are compared to each other. The viewers may not share the same understanding and expectations of image or video quality. Hence, the MOS values generally do not show the absolute quality of the stimuli. In Fig. 2(a), we observe the MOS distribution for non-aligned databases as a function of the HDR-VQM metric. Due to the characteristics of the experiments and test material, a similar level of impairment in the subjective scale may correspond to very different values of the objective metrics. Therefore, in order to use in a consistent way the MOS values of different subjective databases, these need to be mapped onto a common quality scale.

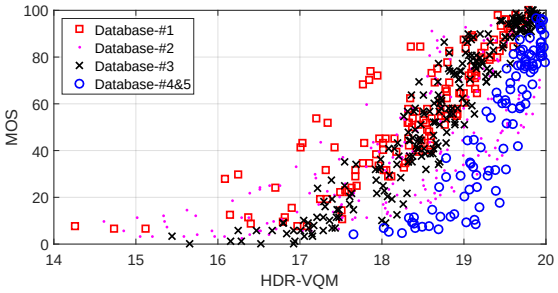
In order to align the MOS values of all five HDR image databases, we use the *iterated nested least square* algorithm (INLSA) proposed in [50][3]. This algorithm requires objective parameters for the alignment, under the assumption that those are sufficiently well correlated and linear with respect to MOS. Therefore, we selected the five most linear and most correlated objective quality metrics: HDR-VDP-2.2, HDR-VQM, PU-IFC, PU-UQI, and PU-VIF (the calculation of PU-metrics will be explained in detail in Sec. 4.1). The INLSA algorithm first normalizes MOS scores from each source in the $[0, 1]$ interval, and then aligns them by solving two least square problems: first, the MOS values are corrected by an affine transformation in order to span the same subjective scale; second, the MOS values are aligned to the corresponding objective values by finding the optimal (in least-square sense) combination of weights such that the corrected MOSs can be predicted as a linear combination of objective parameters. These two steps, prediction and correction, are repeated iteratively till some convergence criterion is met. Details about the algorithm can be found in [50].

The scatter plots of MOS values and HDR-VQM metric values after alignment can be seen in Fig. 2.(b). It can be observed that data points having similar HDR-VQM values have similar MOS values after INLSA alignment. After the alignment, all the MOS values have been mapped onto a common subjective scale, and they can be used in the evaluation of the objective quality metrics.

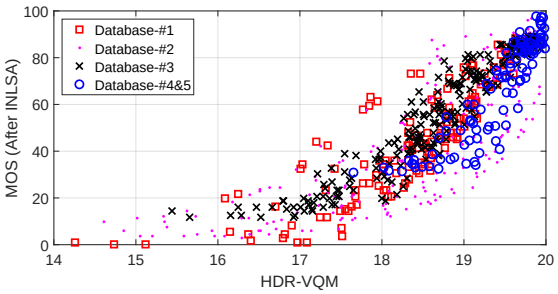
From Fig. 2(b) and initial observations of the test images, we notice that images in Database #2 [40] have very different characteristics compared to others, and MOS values are much more scattered than other databases after the alignment. This is mainly due to the

[2] These scores, together with the test images, are available upon request to the authors and will be made public if this manuscript is accepted for publication.

[3] INLSA implementation on Matlab has been downloaded from `http://www.its.bldrdoc.gov/resources/video-quality-research/guides-and-tutorials/insla-code.aspx`



(a) Before INLSA (RMSE = 44.04)



(b) After INLSA (RMSE = 41.05)

Fig. 2 Plots of MOS vs HDR-VQM scores before and after INLSA alignment. The INLSA algorithm scales MOS values so that images which have similar objective scores also have similar MOS values. In order to compare the scatter plot quantitatively, the root mean squared error (RMSE) of the data is reported for each case.

characteristics of this database, i.e., the stimuli were mainly obtained by changing the tone mapping algorithm used in the compression, including many TMOs which are definitely not adapted to be used in coding as they produce strong color artifacts in the reconstructed HDR image, and that are therefore not used in any practical coding scheme. Also, different kinds of distortion are present simultaneously, such as color banding, saturation etc. In some cases, it is noticed that false contours have been generated, and some color channels were saturated. Initial inspection of both test images and objective metric results indicate that the considered metrics do not capture the effect of color on quality as humans do.

As viewers were rating very different distortions with respect to the other databases, which instead contain similar kinds of visual impairments, Database #2 is very challenging for all the quality metrics we considered in this work. Therefore, in order to provide a complete overview of the performance of HDR fidelity metrics, in the following we report results both with and without including Database #2 in the evaluations.

4 Analysis of Objective Quality Metrics

After the alignment of MOS values of the databases, we obtain an image data set consisting of 690 (or 480 images if Database #2 is excluded) images compressed using JPEG, JPEG-XT, and JPEG 2000. In this section, we provide a thorough analysis of the performance of several HDR image fidelity metrics, both from the point of view of prediction accuracy and of their ability to tell whether two images are actually perceived as being of different quality.

4.1 Objective Quality Metrics under Consideration

We include in our evaluation a number of commonly used full-reference image quality metrics, including the mean square error (MSE), peak signal to noise ratio (PSNR), structural similarity index (SSIM) [62], multi-scale SSIM (MSSIM) [61], information fidelity criterion (IFC) [56], universal quality index (UQI) [60], VIF [55], and pixel based VIF. In addition to those metrics, we consider HDR-VDP-2.2 [43], HDR-VQM [42], additional full-reference metrics recently proposed for HDR video such as mPSNR, tPSNR, CIE ΔE 2000 [58], and spatial extension of CIE ΔE 2000 [65] which is computed with S-CIELAB model.

In order to calculate quality metrics, we first scale pixel values to the range of luminance emitted by the HDR displays used in each subjective experiments. This is especially important for those metrics such as HDR-VDP 2.2 which rely on physical luminance. In order to compute these values, we convert HDR pixels into luminance emitted by a hypothetical HDR display, assuming it has a linear response between the minimum and maximum luminance of the display. As the same display (i.e. SIM2 HDR47E S 4K) has been used in all the experiments, we have selected the same parameters for all experiments, i.e., 0.03 cd/m^2 and 4250 cd/m^2 for minimum an maximum luminance, respectively. Although the emitted luminance on HDR displays depends on many factors and is not exactly a linear function of input pixel values, we found in our previous work that, it is adequately close to linear [64] and from a practical point of view, this simple linear assumption is equivalent to more sophisticated luminance estimation techniques which require a detailed knowledge of the reproduction device [59].

The objective quality metrics under consideration can be grouped as following:

– **HDR-specific metrics:** HDR-VDP-2.2 and HDR-VQM are recent fidelity metrics developed for HDR image and video, respectively. They model several

phenomena that characterize the perception of HDR content, and thus requires some knowledge of viewing conditions (such as distance from the display, ambient luminance, etc.). The mPSNR is PSNR applied on an exposure bracket extracted from the HDR image, and then averaged across exposures.

- **Color difference metrics:** we use CIE ΔE 2000 (denoted as CIE ΔE_{00}), which entails a color space conversion in order to get perceptually uniform color differences [30], and its spatial extension [65] (denoted as CIE ΔE_{00}^{S}). More sophisticated color appearance models have not been considered in this study, as their use in quality assessment has been marginal so far. However they are an interesting aspect to investigate in future work.
- **LDR metrics applied after a transfer function:** LDR metrics such as MSE, PSNR, VIF, SSIM, MSSIM, IFC, and UQI. To compute these LDR metrics we use:
  - Physical luminance of the scene directly, denoted as *Photometric-*,
  - Perceptually uniform [3] encoded pixel values, denoted as *PU-*,
  - Logarithmic coded pixel values, denoted as *Log-*, or
  - Perceptually quantized [36, 57] pixel values. For this case, only tPSNR-YUV has been considered as in [58].

When possible, we use the publicly available implementation of these metrics, i.e., HDR-VDP-2.2.1 available at `http://sourceforge.net/projects/hdrvdp/files/hdrvdp/`, HDR-VQM available at `http://www.sourceforge.net/projects/hdrvdp/files/hdrvdp/`, HDRtools version 0.4 [58] developed within MPEG, the MeTriX MuX library for Matlab, available at `http://foulard.ece.cornell.edu/gaubatz/metrix_mux/`.

4.2 Statistical Analysis

The performance of the aforementioned fidelity metrics has been evaluated in terms of *prediction accuracy*, *prediction monotonicity*, and *prediction consistency* [8]. For prediction accuracy, Pearson correlation coefficient (PCC), and root mean square error (RMSE) are computed. Spearman rank-order correlation coefficient (SROCC) is used to find the prediction monotonicity, and outlier ratio (OR) is calculated to determine the prediction consistency. These performance metrics have been computed after a non-linear regression performed on objective quality metric results using a logistic function, as described in the final report of VQEG FR Phase I [54]. This logistic function is given in Eqn. 1:

$$Y_i = \beta_2 + \frac{\beta_1 - \beta_2}{1 + e^{-(\frac{X_i - \beta_3}{|\beta_4|})}}, \tag{1}$$

where X_i is the objective score for the i^{th} distorted image, and Y_i is the mapped objective score. It tries to minimize the least-square error between the MOS values and the objective results. This fitting has been done using the `nlinfit` function of Matlab to find optimal β parameters for each objective quality metric. After fitting, the performance scores have been computed using the mapped objective results, Y_i, and MOS values.

The results of these performance indexes (SROCC, PCC, RMSE, and OR) have been computed for each database separately, as well as considering all the data together. The results are reported in Tables 2-5. The aligned data scores have been denoted as **"Combined"**, and **"Except Database #2"** for the data aligned excluding Database #2 as explained in Section 3.

These results show that the performance of many fidelity metrics may significantly vary from one database to another, due to the different characteristics of the test material and of the subjective evaluation procedure. In particular, Database #2 is the most challenging for all the considered metrics, due to its more complex distortion features, as discussed in Section 3. Despite the variations across databases, we can observe a consistent behavior for some metrics. Photometric-MSE is the worst correlated one, for all databases. This is expected as mean square error is computed on photometric values, without any consideration of visual perception phenomena. On the other hand, HDR-VQM, HDR-VDP-2.2 Q, and PU-MSSIM are the best performing metrics, with the exception of Database #2.

When we analyze objective metrics for each transfer function, we observe that Photometric-IFC is the best correlated and Photometric-MSE is the worst in the linear domain; Log-SSIM is the best correlated and Log-VIF is the worst in the logarithmic domain. Among the objective metric results in PU domain, PU-MSSIM and PU-SSIM display high correlation coefficients, while PU-MSE is the again the worst performer. Comparing the three transfer functions, PU is the most effective, as PU-MSSIM and PU-SSIM achieve performance very close to HDR-VDP-2.2 Q and HDR-VQM. In general, metrics which are based on MSE and PSNR (PU-MSE, Log-MSE, PU-PSNR, mPSNR, etc.) yield worse results compared to other metrics. Instead, more advanced LDR metrics such as IFC, UQI, SSIM, and MSSIM yield much better results. We also notice that mPSNR, tPSNR-YUV, and CIE ΔE 2000, which have been recently used in MPEG standardization activities, perform rather poorly in comparison to the others.

Table 2 Pearson Correlation Coefficient (PCC) Results for Each Database and for Aligned Data

Metric	Database #1	Database #2	Database #3	Database #4&5	Combined	Except Database #2
Photometric-MSE	0.4051	0.1444	0.7080	0.5095	0.3651	0.6987
Photometric-PSNR	0.4409	0.2564	0.7132	0.5594	0.5166	0.6506
Photometric-SSIM	0.5016	0.3583	0.8655	0.6708	0.6441	0.7462
Photometric-IFC	0.7781	0.8234	0.9183	0.8195	0.8344	0.7680
Photometric-UQI	0.7718	0.8208	0.8846	0.7876	0.8312	0.7667
Photometric-VIF	0.7603	0.5076	0.8666	0.6144	0.6264	0.8452
PU-MSE	0.4824	0.3309	0.8559	0.8024	0.6273	0.7710
PU-PSNR	0.5297	0.3269	0.8606	0.8009	0.6271	0.7761
PU-SSIM	0.8661	0.7049	0.9532	0.9201	0.8441	0.9016
PU-IFC	0.7910	0.8422	0.9201	0.8566	0.8569	0.8024
PU-MSSIM	0.8847	0.7236	0.9564	0.9038	0.8570	0.9210
PU-UQI	0.7823	**0.8507**	0.8768	0.7777	0.8367	0.7637
PU-VIF	0.7845	0.7583	0.9349	0.9181	0.8574	0.8655
Log-MSE	0.6114	0.5314	0.8856	0.8820	0.6844	0.7872
Log-PSNR	0.6456	0.5624	0.8870	0.8819	0.7001	0.7923
Log-SSIM	0.8965	0.8035	0.9235	0.8255	0.8418	0.8401
Log-IFC	0.7919	0.8366	0.9167	0.8551	0.8530	0.8034
Log-UQI	0.7837	0.8268	0.8786	0.7830	0.8285	0.7592
Log-VIF	0.5079	0.6202	0.8354	0.7065	0.6049	0.6889
HDR-VDP-2.2 Q	**0.8989**	0.5482	0.9531	**0.9408**	0.7590	0.9261
HDR-VQM	0.8949	0.7932	**0.9612**	0.9332	**0.8807**	**0.9419**
mPSNR	0.6545	0.6564	0.8593	0.8587	0.7434	0.7959
tPSNR-YUV	0.5784	0.4524	0.8319	0.7789	0.6580	0.7718
$CIE\ \Delta E_{00}$	0.6088	0.2553	0.7889	0.6082	0.4979	0.7752
$CIE\ \Delta E_{00}^{S}$	0.6167	0.3331	0.8793	0.7322	0.5783	0.7929

Table 3 Spearman Rank-Ordered Correlation Coefficient (SROCC) Results for Each Database and for Aligned Data

Metric	Database #1	Database #2	Database #3	Database #4&5	Combined	Except Database #2
Photometric-MSE	0.3881	0.1235	0.7227	0.5711	0.3417	0.7174
Photometric-PSNR	0.4018	0.2783	0.7183	0.5737	0.4991	0.6520
Photometric-SSIM	0.4953	0.3063	0.8792	0.6770	0.6357	0.7610
Photometric-IFC	0.7684	0.8254	0.9179	0.8109	0.8354	0.7708
Photometric-UQI	0.7495	0.8299	0.8686	0.8017	0.8310	0.7650
Photometric-VIF	0.7482	0.4915	0.8723	0.4864	0.6010	0.8376
PU-MSE	0.4791	0.2959	0.8617	0.8065	0.6108	0.7750
PU-PSNR	0.4791	0.2959	0.8617	0.8065	0.6108	0.7750
PU-SSIM	0.8553	0.7234	0.9503	0.9121	0.8525	0.9080
PU-IFC	0.7786	0.8433	0.9165	0.8489	0.8573	0.8044
PU-MSSIM	0.8711	0.7363	0.9517	0.8969	0.8570	0.9198
PU-UQI	0.7612	**0.8608**	0.8569	0.7932	0.8358	0.7606
PU-VIF	0.7634	0.7662	0.9306	0.9083	0.8560	0.8627
Log-MSE	0.5943	0.5843	0.8892	0.8719	0.6730	0.7917
Log-PSNR	0.5943	0.5843	0.8892	0.8710	0.6802	0.7917
Log-SSIM	**0.8935**	0.7869	0.9268	0.8179	0.8448	0.8424
Log-IFC	0.7782	0.8420	0.9140	0.8482	0.8529	0.8049
Log-UQI	0.7622	0.8232	0.8592	0.7960	0.8285	0.7563
Log-VIF	0.4884	0.5908	0.8385	0.6653	0.6346	0.6885
HDR-VDP-2.2 Q	0.8911	0.5727	0.9503	**0.9298**	0.7634	0.9357
HDR-VQM	0.8874	0.8126	**0.9572**	0.9193	**0.8779**	**0.9416**
mPSNR	0.6133	0.6496	0.8648	0.8521	0.7381	0.7970
tPSNR-YUV	0.5324	0.4342	0.8374	0.7901	0.6497	0.7722
$CIE\ \Delta E_{00}$	0.5883	0.2551	0.7824	0.5951	0.4837	0.7761
$CIE\ \Delta E_{00}^{S}$	0.5979	0.3096	0.8779	0.7430	0.5816	0.7955

Table 4 Root Mean Squared Error (RMSE) Results for Each Database and for Aligned Data (Please note that, in order to have comparable results, RMSE values were calculated after all MOS values are scaled to the range of [0,100].)

Metric	Database #1	Database #2	Database #3	Database #4&5	Combined	Except Database #2
Photometric-MSE	23.526	27.459	22.163	25.684	24.204	17.910
Photometric-PSNR	23.096	26.791	22.000	24.742	22.262	19.012
Photometric-SSIM	22.261	25.907	15.719	22.138	19.888	16.665
Photometric-IFC	16.164	15.748	12.426	17.105	14.328	16.032
Photometric-UQI	16.364	15.850	14.635	18.392	14.455	16.071
Photometric-VIF	16.715	23.909	15.659	23.551	20.267	13.378
PU-MSE	22.540	26.187	16.232	17.814	20.247	15.942
PU-PSNR	21.826	26.225	15.984	17.874	20.251	15.787
PU-SSIM	12.861	19.683	9.489	11.688	13.939	10.831
PU-IFC	15.744	14.963	12.295	15.403	13.401	14.939
PU-MSSIM	11.995	19.153	9.165	12.775	13.396	9.754
PU-UQI	16.030	**14.586**	15.093	18.765	14.238	16.162
PU-VIF	15.956	18.089	11.142	11.828	13.381	12.539
Log-MSE	20.362	23.508	14.574	14.067	18.956	15.437
Log-PSNR	19.651	22.945	14.494	14.071	18.566	15.275
Log-SSIM	11.400	16.520	12.038	16.847	14.033	13.578
Log-IFC	15.713	15.201	12.540	15.477	13.571	14.905
Log-UQI	15.984	15.611	14.988	18.567	14.560	16.295
Log-VIF	22.167	21.769	17.249	21.126	20.704	18.146
HDR-VDP-2.2 Q	**11.276**	23.209	9.496	**10.120**	16.926	9.447
HDR-VQM	11.481	16.900	**8.657**	10.725	**12.313**	**8.410**
mPSNR	19.455	20.934	16.053	15.298	17.390	15.158
tPSNR-YUV	20.992	24.748	17.418	18.721	19.577	15.918
$CIE\ \Delta E_{00}$	20.414	26.830	19.285	23.694	22.548	15.813
$CIE\ \Delta E_{00}^{S}$	20.256	26.165	14.949	20.330	21.211	15.254

Table 5 Outlier Ratio (OR) Results for Each Database and for Aligned Data

Metric	Database #1	Database #2	Database #3	Database #4&5	Combined	Except Database #2
Photometric-MSE	0.750	0.933	0.787	0.830	0.838	0.744
Photometric-PSNR	0.771	0.905	0.767	0.820	0.810	0.729
Photometric-SSIM	0.821	0.938	0.679	0.780	0.790	0.681
Photometric-IFC	0.750	0.871	0.546	0.610	0.658	0.637
Photometric-UQI	0.707	0.871	0.558	0.640	0.664	0.629
Photometric-VIF	0.679	0.948	0.617	0.800	0.796	0.596
PU-MSE	0.857	0.933	0.633	0.680	0.768	0.635
PU-PSNR	0.779	0.919	0.579	0.660	0.774	0.640
PU-SSIM	0.714	0.948	0.404	0.560	0.645	0.456
PU-IFC	0.750	0.886	0.500	0.610	0.619	0.629
PU-MSSIM	0.607	0.933	0.388	0.570	0.625	**0.446**
PU-UQI	0.664	0.848	0.583	0.680	0.648	0.615
PU-VIF	0.700	0.943	0.450	0.520	0.632	0.629
Log-MSE	0.843	0.924	0.592	0.570	0.694	0.646
Log-PSNR	0.786	0.919	0.588	0.580	0.745	0.667
Log-SSIM	0.643	0.876	0.525	0.570	0.681	0.560
Log-IFC	0.750	**0.833**	0.529	0.610	0.636	0.627
Log-UQI	0.671	0.843	0.579	0.630	0.652	0.627
Log-VIF	0.807	0.924	0.654	0.730	0.864	0.694
HDR-VDP-2.2 Q	0.586	0.938	**0.342**	**0.490**	0.733	0.475
HDR-VQM	**0.514**	0.890	0.392	0.530	**0.607**	0.448
mPSNR	0.771	0.895	0.667	0.610	0.720	0.642
tPSNR-YUV	0.800	0.952	0.625	0.670	0.774	0.656
$CIE\ \Delta E_{00}$	0.743	0.924	0.675	0.760	0.833	0.669
$CIE\ \Delta E_{00}^{S}$	0.793	0.933	0.613	0.710	0.813	0.669

We also evaluate the significance of the difference between the considered performance indexes, as proposed in ITU-T Recommendation P.1401 [23]. The results are provided in Fig. 3 and Fig. 4 for "Combined" and "Except Database #2" cases respectively. The bars indicate statistical equivalence between the quality metrics. We observe that the performance of HDR-VQM in the combined database is significantly different from all others while PU-MSSIM, PU-VIF, and some other metrics have essentially equivalent performance across the combined databases. Although HDR-VDP-2.2 has a lower performance on combined dataset compared to its performance on individual databases, it is among the three most correlated metrics with HDR-VQM and PU-MSSIM on the case excluding Database #2. Interestingly, the HDR-VQM metric, which has been designed to predict *video* fidelity, gives excellent results also in the case of static images, and is indeed more accurate on Database #2 than HDR-VDP-2.2. Furthermore, we notice that all metrics except CIE ΔE_{00} and CIE ΔE_{00}^{S} consider only luminance values. Although CIE ΔE_{00} and CIE ΔE_{00}^{S} have been found to be among the most relevant color difference metrics among others in a recent study [47], they have lower correlation scores when compared to luminance-only metrics. In fact, this result is not in disagreement with [47], which did not consider compression artifacts in the experiments, as the impact of those on image quality was deemed to be much stronger than color differences. Thus, our analysis confirms that luminance artifacts such as blocking, etc., play a dominant role in the formation of quality judgments, also in the case of HDR.

4.3 Discriminability Analysis

MOS values are *estimated* from a sample of human observers, i.e., they represent expected values of random variables (the perceived annoyance or quality). Therefore, MOS are as well random variables which are known with some uncertainty, which is typically represented by their confidence intervals [21]. As a result, different MOS values could correspond to the same underlying distribution of subjective scores and two images with different MOS might indeed have the same visual quality in practice (with confidence level). The performance scores considered in Section 4.2 assume instead that MOS values are deterministically known, and that the goal of fidelity metrics is to predict them as precisely as possible, without taking into account whether two different subjective scores do actually correspond to different quality. Therefore, in the following we consider another evaluation approach, which aims at assessing if an objective fidelity metric is able to discriminate whether two images have significantly different subjective quality.

The intrinsic variability of MOS scores is not a completely new problem, and several approaches have been proposed in the literature to take this into account while evaluating objective metrics. Brill et al. [6] introduced the concept of *resolving power* of an objective metric, which indicates the minimum difference in the output of a quality prediction algorithm such that at least $p\%$ of viewers (where generally $p = 95\%$) would observe a difference of quality between two images. This approach has also been standardized in ITU Recommendation J.149 [22], and used in subsequent work [4, 18, 46, 49]. Nevertheless, this technique has a number of disadvantages. Resolving power is computed after transforming MOS to a common scale, which requires applying a fitting function; however, the fitting problem could be ill-posed in some circumstances, yielding incorrect results. Also, the resolving power in the common scale corresponds to a variable metric resolution in the original scale, which makes it difficult to interpret. Moreover, it is not always possible to fix the level of significance p to be the same for different metrics, as there could be cases when the percentage of observers seeing a difference between image qualities is lower than p for any metric difference values. Finally, the results of this approach are generally evaluated in a qualitative manner, e.g., by considering how the number of correct decisions, false rankings, false differentiations, etc., vary as a function of objective metric differences [6, 18]; conversely, a compact, quantitative measure is desirable in order to fairly compare different metrics. Another approach to this problem has been recently proposed by Krasula et al. [26]. In their paper, Krasula et al. find the accuracy of an objective image or video quality metric by transforming the problem into a classification problem. For this purpose, they find z-score of subjective scores and the difference of objective scores for each pair of stimuli, and then find the accuracy of the metric by calculating classification rates.

Due to the factors above limiting the effectiveness of resolving power, in this work we propose an alternative approach in the *original* scale of the metric similar to what has been presented in Krasula et al. [26], which enables to evaluate its discrimination power while avoiding the shortcomings discussed above. Despite the similarities, the implementation and the data processing steps of their work and the proposed algorithm are not the same. Therefore, we give the details of the proposed algorithm below in order to clarify differences.

The basic idea of the proposed method is to convert the classical *regression* problem of accurately predicting MOS values, into a *binary classification* (detection)

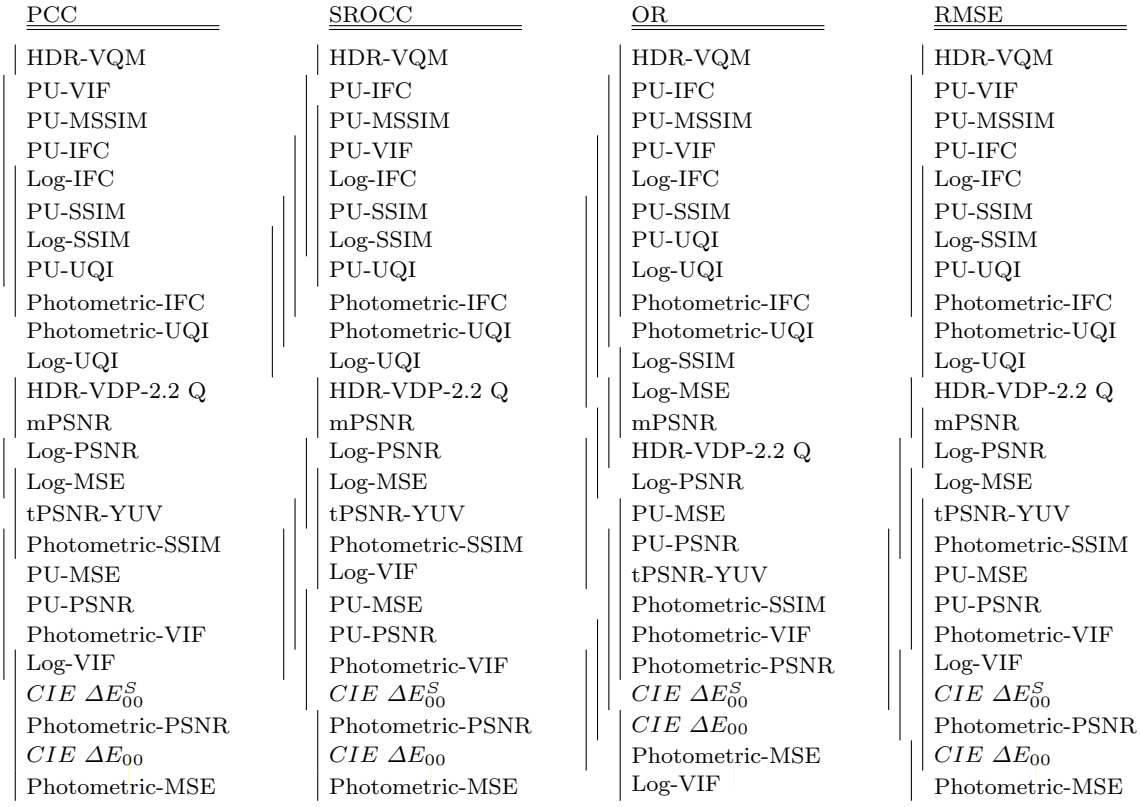


Fig. 3 Statistical analysis results for correlation indices for combined data according to ITU-T Recommendation P.1401 [23]. The bars signify statistical equivalence between the quality metrics if they have the same bar aligned with two quality metrics; e.g., there is a statistically significant difference between HDR-VQM and all the other metrics considered in terms of PCC, SROCC, and RMSE.

problem [24]. We denote by $S(I)$ and $O(I)$ the subjective (MOS) and objective quality of stimulus I, respectively, for a certain objective quality metric. Given two stimuli I_i, I_j, we model the detection problem as one of choosing between the two hypotheses $\mathcal{H}_0$, i.e., there is no significant difference between the visual quality of I_i and I_j, and $\mathcal{H}_1$, i.e., I_i and I_j have significantly different visual quality. Formally:

$$\begin{aligned} &\mathcal{H}_0 : S(I_i) \cong S(I_j); \\ &\mathcal{H}_1 : S(I_j) \ncong S(I_j), \end{aligned} \tag{2}$$

where we use $\cong$ (resp. $\ncong$) to indicate that the means of two populations of subjective scores (i.e., two MOS values) are the same (resp. different). Given a dataset of subjective scores, it is possible to apply a pairwise statistical test (e.g., a two-way *t-test* or *z-test*) to determine whether two MOSs are the same, at a given significance level. In our work, we employ a one-way analysis of variance (ANOVA), with Tukey's honestly significant difference criterion to account for the multiple comparison bias [20], as it is also stated as the ideal way to find statistical significance in [26]. Figure 5(a) shows the results of ANOVA on our combined database, thresholded at a confidence level of 95% (i.e., 5% significance). For convenience of visualization, MOS values have been sorted in ascending order before applying ANOVA. White entries represent MOS pairs which are statistically indistinguishable.

In order to decide between $\mathcal{H}_0$ and $\mathcal{H}_1$, similar to Krasula et al. [26], we consider the simple test statistic $\Delta_{ij}^O = |O(I_i) - O(I_j)|$, i.e., we look at the difference between the objective scores for the two stimuli and compare it with a threshold τ, that is:

$$\text{Decide: } \begin{cases} \mathcal{H}_0 & \text{if } \Delta_{ij}^O \leq \tau \\ \mathcal{H}_1 & \text{otherwise.} \end{cases} \tag{3}$$

For a given value of τ, we can then label the set of stimuli as being equivalent or not, as shown in Figure 5(b). The performance of the detector in (3) depends on the choice of τ. We call *true positive rate (TPR)* the ratio of images with different MOSs correctly classified as being of different quality, and *false positive rate (FPR)* the ratio of images with equal MOSs incorrectly classified as being of the different quality. By varying the value of τ, we can trace a Receiver Operating Characteristic (ROC) curve, which represents the TPR at

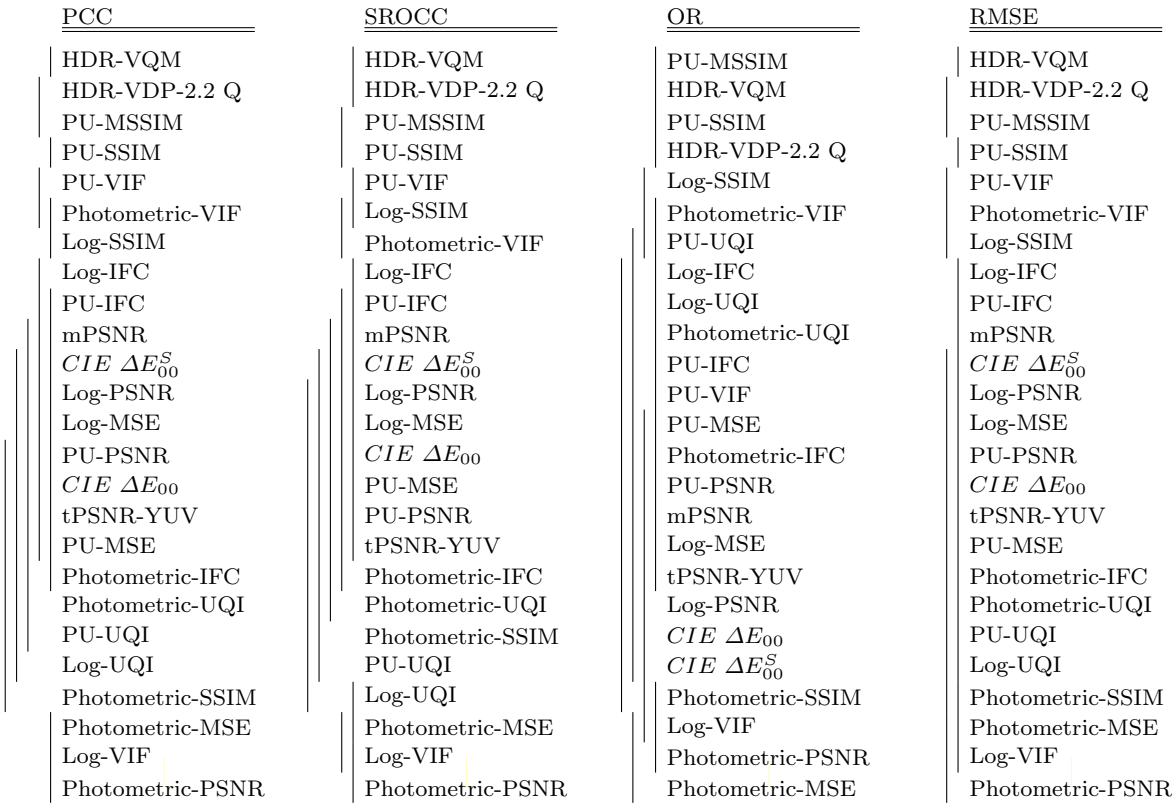
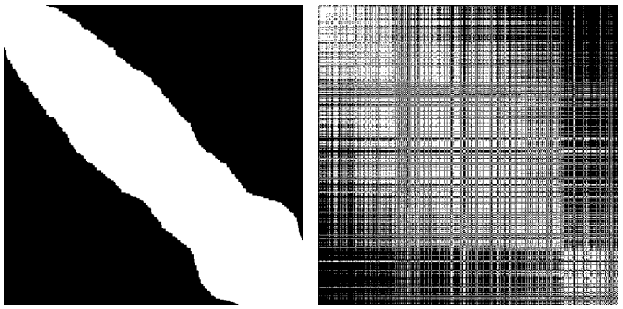


Fig. 4 Statistical analysis results for correlation indices for combined data excluding Database #2 according to ITU-T Recommendation P.1401 [23]. The bars signify statistical equivalence between the quality metrics if they have the same bar aligned with two quality metrics; e.g., HDR-VDP-2.2 Q, HDR-VQM, PU-SSIM and PU-MSSIM are statistically equivalent to each other in terms of OR.



(a) MOS equivalence matrix at 95% confidence level

(b) HDR-VDP-2.2 Q estimated equivalence matrix (τ fixed for maximum accuracy)

Fig. 5 Equivalence maps for the (sorted) combined database. White entries correspond to $S(I_i) \cong S(I_j)$, black to $S(I_i) \ncong S(I_j)$.

a given value of FPR [24]. The area under the ROC curve (AUC) is higher when the overlap between the marginal distributions of Δ_{ij}^O under each hypothesis, that is, $p(\Delta_{ij}^O; \mathcal{H}_0)$ and $p(\Delta_{ij}^O; \mathcal{H}_1)$, is smaller. Therefore, the AUC is a measure of the *discrimination power* of an objective quality metric.

Table 6 reports the AUC values for the combined case and the combination without Database-#2. In addition to the area under the ROC curve, we also compute the balanced classification accuracy, which is an extension of the conventional accuracy measure to unbalanced datasets, i.e., where the number of positive and negative samples is different [7]:

$$Acc = \frac{2 \times TP}{TP + FN} + \frac{2 \times TN}{TN + FP}. \tag{4}$$

In Table 6 we report the maximum classification accuracy, $Acc^* = \max_\tau Acc$, which characterizes the global detection performance, as well as the value of the detector threshold at FPR = 5%, that is,

$$\tau_{.05} = \min\{\tau : p(\Delta_{ij}^O > \tau; \mathcal{H}_0) \leq 0.05\}, \tag{5}$$

which indicates the minimum value of τ in order to keep below 5% the probability of incorrectly classifying two stimuli as being of different quality. This latter measure provides somehow the *resolution* of an objective metric (with a 5% tolerance) in the original metric scale.

These results in Table 6 are complemented with the percentage of correct decisions (CD) in [6], which is to be compared with Acc^*. Furthermore, we present

Table 6 Results of discriminability analysis: area under the ROC curve (AUC), threshold τ at 5% false positive rate, maximum classification accuracy. We report for comparison the fraction of Correct Decisions (CD) at 95% confidence level as proposed in [6]. For CD, '–' indicates that the 95% confidence level cannot be achieved.

	Combined				**Except Database #2**			
Metric	AUC	$\tau_{.05}$	Acc*	CD [6]	AUC	$\tau_{.05}$	Acc*	CD [6]
Photometric-MSE	0.532	34894.476	0.530	–	0.644	34894.476	0.614	0.317
Photometric-PSNR	0.576	24.798	0.556	–	0.633	18.135	0.596	0.249
Photometric-SSIM	0.609	0.070	0.590	–	0.677	0.057	0.633	0.306
Photometric-IFC	0.716	5.784	0.666	0.398	0.675	7.554	0.629	0.340
Photometric-UQI	0.765	0.333	0.707	0.380	0.730	0.381	0.678	0.296
Photometric-VIF	0.605	0.730	0.585	0.204	0.717	0.730	0.654	0.446
PU-MSE	0.596	431.687	0.580	–	0.677	431.687	0.645	0.379
PU-PSNR	0.625	20.047	0.593	–	0.715	15.350	0.661	0.380
PU-SSIM	0.721	0.057	0.663	0.399	0.804	0.035	0.725	0.512
PU-IFC	0.729	6.081	0.676	0.451	0.694	7.880	0.643	0.386
PU-MSSIM	0.737	0.092	0.680	0.434	0.838	0.054	0.758	0.598
PU-UQI	0.770	0.312	0.711	0.391	0.730	0.408	0.678	0.286
PU-VIF	0.782	0.419	0.719	0.463	0.802	0.455	0.735	0.493
Log-MSE	0.600	0.522	0.587	0.253	0.687	0.036	0.653	0.393
Log-PSNR	0.668	21.195	0.624	0.256	0.729	15.251	0.668	0.395
Log-SSIM	0.717	0.130	0.664	0.394	0.762	0.068	0.696	0.407
Log-IFC	0.725	6.074	0.673	0.443	0.694	7.840	0.642	0.382
Log-UQI	0.769	0.359	0.711	0.368	0.728	0.408	0.676	0.272
Log-VIF	0.634	0.311	0.593	0.217	0.666	0.210	0.635	0.282
HDR-VDP-2.2 Q	0.689	24.084	0.630	0.300	0.850	18.441	0.759	0.622
HDR-VQM	**0.791**	1.723	**0.727**	**0.487**	**0.893**	1.320	**0.816**	**0.684**
mPSNR	0.690	13.840	0.648	0.278	0.727	13.840	0.671	0.381
tPSNR-YUV	0.636	16.452	0.603	0.178	0.708	14.396	0.658	0.367
$CIE\ \Delta E_{00}$	0.580	7.608	0.559	0.168	0.721	6.657	0.669	0.332
$CIE\ \Delta E^{S}_{00}$	0.602	7.677	0.575	0.187	0.723	6.718	0.668	0.349

the results of statistical significance evaluation of the reported AUC values according to the guidelines presented in Krasula et al. [26]. The results of this statistical significance evaluation are presented in Fig. 6. The results show that HDR-VQM is the best performing metric, and PU-VIF and PU-MSSIM perform better than most of the considered metrics. Although its performance is reduced in the combined case, HDR-VDP-2.2 Q also is statistically better than other metrics in the case excluding Database #2.

We notice that, in general, the values of CD are much lower than Acc^*. This is due to the fact that the method in [6] not only aims at distinguishing whether two images have the same quality, but also to determine which is the one with better quality. Thus the classification task is more difficult, as there are three classes – equivalent, better or worse – to label. Indeed, we observe a certain coherence between our approach and [6], and with the statistical analysis in Section 4.2: the best performing metrics are HDR-VQM and those based on PU transfer function such as PU-MSSIM, PU-VIF, and PU-SSIM. Nevertheless, our analysis provides a better insight on the discrimination power of fidelity metrics compared to [6], and gives practical guidelines on which should be the minimal differences between the objective scores of two images in order to claim that those have different visual quality. Finally, the fact that, even for the best performing metrics in terms of correlation with MOSs, maximum accuracy saturates at 0.8 suggests that there is still space for improving existing HDR objective quality measures, as far as discriminability (and not only prediction accuracy) is included in the evaluation of performance.

5 Conclusion

In this paper, we conduct an extensive evaluation of full-reference HDR image quality metrics. For this purpose, we collect four different publicly available HDR image databases for compression distortion and a newly created one. In order to have consistent MOS values across all databases, we align subjective scores using the INLSA algorithm. After the alignment, a total of 690 compressed HDR images have been evaluated using several full-reference HDR image quality assessment metrics. The performance of these fidelity metrics has

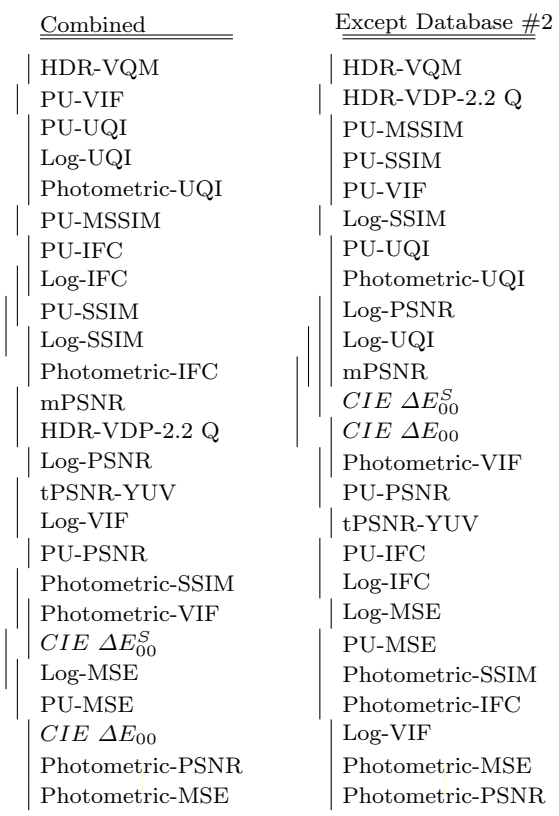


Fig. 6 Statistical analysis results for the discriminability analysis, according to the procedure described in Krasula et al. [26]. The bars signify statistical equivalence between the quality metrics if they have the same bar aligned with two quality metrics. It can be said that among PU-UQI, Log-UQI, and Photometric-UQI, there is not any statistically significant difference. Whereas, there is a statistically significant difference between HDR-VQM and all the other metrics considered.

been assessed from two different perspectives: on one hand, by looking at the quality estimation as a regression problem, using conventional statistical accuracy and monotonicity measures [8]; on the other hand, by focusing on the ability of objective metrics to discriminate whether two stimuli have the same perceived quality.

Our analysis shows that recent metrics designed for HDR content, such as HDR-VQM and to some extent HDR-VDP-2.2, provide accurate predictions of MOSs, at least for compression-like distortion. We also confirm the findings in previous work [17, 59] that legacy LDR image quality metrics have good prediction and discrimination performance, provided that a proper transformation such as PU encoding is done beforehand. This somehow suggests that the quality assessment problem for HDR image compression is similar to the case of LDR, if HDR pixels are properly preprocessed. Yet, the absolute performance figures of these metrics show that, when databases with heterogeneous characteristics are merged (database #2 in our experiments), none of the tested metrics provides highly reliable predictions. All but two of the considered metrics are computed on the luminance channel only. Interestingly, the non color-blind metrics, CIE ΔE_{00} and CIE ΔE^S_{00}, displays poor performance in our evaluation, similar to other MSE-based metrics. While other studies report different results in terms of correlation with MOSs [19], we believe that a partial explanation for these results is that in the case of coding artifacts, the structural distortion (blocking, blur) in the luminance channel dominates the color differences, captured by CIE ΔE_{00} and CIE ΔE^S_{00} . The important aspect of color fidelity metrics for HDR content, however, is still little understood and is part of our current research.

Finally, the alternative evaluation methodology proposed in this work, based on the discriminability of a metric, provides a complementary perspective on the performance of objective quality metrics. It recognizes the stochastic nature of MOSs, which are samples from a population and hence are known with some uncertainty. Therefore, we consider the quality estimation task as one of detecting when images have significantly different quality. The relevance of this alternative point of view is demonstrated by the amount of efforts to go beyond classical statistical measures such as correlation in the last decade, from the seminal work of Brill et al. [6] to the very recent work of Krasula et al. [26], developed in parallel to our study. These analyses show that, even for metrics which can accurately predict MOS values, the rate of incorrect classifications is still quite high (20% or more). This suggests that novel and more performing object quality metrics could be designed, provided that new criteria such as discriminability are taken into account alongside the correlation indices used to find statistical accuracy.

Acknowledgements The work presented in this paper is supported by BPI France and Région Ile-de-France, in the framework of the FUI18 PLEIN PHARE project.

References

1. Artusi A, Mantiuk RK, Richter T, Hanhart P, Korshunov P, Agostinelli M, Ten A, Ebrahimi T (2015) Overview and evaluation of the JPEG XT HDR image compression standard. Journal of Real-Time Image Processing pp 1–16
2. Ashikhmin M (2002) A tone mapping algorithm for high contrast images. In: Proceedings of the 13th Eurographics workshop on Rendering, Eurographics Association, pp 145–156
3. Aydın TO, Mantiuk R, Seidel HP (2008) Extending quality metrics to full luminance range images. In: Electronic Imaging 2008, International Society for Optics and Photonics

4. Barkowsky M (2009) Subjective and objective video quality measurement in low-bitrate multimedia scenarios. Citeseer
5. Barten PG (1999) Contrast sensitivity of the human eye and its effects on image quality, vol 72. SPIE press
6. Brill MH, Lubin J, Costa P, Wolf S, Pearson J (2004) Accuracy and cross-calibration of video quality metrics: new methods from ATIS/T1A1. Signal Processing: Image Communication 19(2):101–107
7. Brodersen KH, Ong CS, Stephan KE, Buhmann JM (2010) The balanced accuracy and its posterior distribution. In: Pattern recognition (ICPR), 2010 20th international conference on, IEEE, pp 3121–3124
8. De Simone F (2012) Selected contributions on multimedia quality evaluation. PhD thesis
9. Debevec PE, Malik J (2008) Recovering high dynamic range radiance maps from photographs. In: ACM SIGGRAPH 2008 classes, ACM, p 31
10. Drago F, Mantiuk R (2004) MPI HDR image gallery. `http://resources.mpi-inf.mpg.de/hdr/gallery.html`, accessed: 2015-11-15
11. Dufaux F, Le Callet P, Mantiuk R, Mrak M (2016) High Dynamic Range Video: From Acquisition, to Display and Applications. Academic Press
12. Durand F, Dorsey J (2002) Fast bilateral filtering for the display of high-dynamic-range images. In: ACM transactions on graphics (TOG), ACM, vol 21, pp 257–266
13. EMPA MT (2013) Empa HDR image database. `http://empamedia.ethz.ch/hdrdatabase/index.php`, accessed: 2015-11-15
14. Fairchild MD (2007) The HDR photographic survey. In: Color and Imaging Conference, Society for Imaging Science and Technology, vol 2007, pp 233–238
15. Fairchild MD (2013) Color appearance models. John Wiley & Sons
16. Froehlich J, Grandinetti S, Eberhardt B, Walter S, Schilling A, Brendel H (2014) Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays
17. Hanhart P, Bernardo MV, Pereira M, Pinheiro AM, Ebrahimi T (2015a) Benchmarking of objective quality metrics for HDR image quality assessment. EURASIP Journal on Image and Video Processing 2015(1):1–18
18. Hanhart P, Řeřábek M, Ebrahimi T (2015b) Towards high dynamic range extensions of HEVC: subjective evaluation of potential coding technologies. In: SPIE Optical Engineering+ Applications, International Society for Optics and Photonics
19. Hanhart P, Rerabek M, Ebrahimi T (2016) Subjective and objective evaluation of HDR video coding technologies. In: 8th International Conference on Quality of Multimedia Experience (QoMEX)
20. Hogg RV, Ledolter J (1987) Engineering statistics. Macmillan Pub Co
21. ITU-R (2012) Methodology for the subjective assessment of the quality of television pictures. ITU-R Recommendation BT. 500-13
22. ITU-T (2004) Method for specifying accuracy and cross-calibration of video quality metrics (VQM). ITU-T Recommendation J.149
23. ITU-T (2012) Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. ITU-T Recommendation P.1401
24. Kay SM (1998) Fundamentals of statistical signal processing: Detection theory, vol. 2
25. Korsunov P, Hanhart P, Richter T, Artusi A, Mantiuk R, Ebrahimi T (2015) Subjective quality assessment database of HDR images compressed with JPEG XT. In: 7th International Workshop on Quality of Multimedia Experience (QoMEX)
26. Krasula L, Fliegel K, Le Callet P, Klíma M (2016) On the accuracy of objective image and video quality models: New methodology for performance evaluation. In: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), IEEE, pp 1–6
27. Kuang J, Johnson GM, Fairchild MD (2007) iCAM06: A refined image appearance model for HDR image rendering. Journal of Visual Communication and Image Representation 18(5):406–414
28. Kundu MK, Pal SK (1986) Thresholding for edge detection using human psychovisual phenomena. Pattern Recognition Letters 4(6):433–441
29. Lasserre S, LeLéannec F, Francois E (2013) Description of HDR sequences proposed by technicolor. ISO/IEC JTC1/SC29/WG11 JCTVC-P0228, IEEE, San Jose, USA
30. Luo MR, Cui G, Rigg B (2001) The development of the CIE 2000 colour-difference formula: CIEDE2000. Color Research & Application 26(5):340–350
31. Luthra A, Francois E, Husak W (2015) Call for Evidence (CfE) for HDR and WCG Video Coding. ISO/IEC JTC1/SC29/WG11 MPEG2014 N 15083
32. Mai Z, Mansour H, Mantiuk R, Nasiopoulos P, Ward W R Heidrich (2011) Optimizing a tone curve for backward-compatible high dynamic range image and video compression. IEEE Trans on Image Processing 20(6):1558–1571
33. Mantiuk R, Myszkowski K, Seidel HP (2006) A perceptual framework for contrast processing of high dynamic range images. ACM Transactions on Applied Perception (TAP) 3(3):286–308
34. Mantiuk R, Daly S, Kerofsky L (2008) Display adaptive tone mapping. In: ACM Transactions on Graphics (TOG), ACM, vol 27, p 68
35. Mantiuk R, Kim K, Rempel A, Heidrich W (2011) HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. In: ACM Trans. on Graphics, ACM, vol 30, p 40
36. Miller S, Nezamabadi M, Daly S (2012) Perceptual signal coding for more efficient usage of bit codes. In: SMPTE Conferences, Society of Motion Picture and Television Engineers, vol 2012, pp 1–9
37. Narwaria M, Da Silva MP, Le Callet P, Pépion R (2012a) Effect of tone mapping operators on visual attention deployment. In: SPIE Optical Engineering+ Applications, International Society for Optics and Photonics
38. Narwaria M, Lin W, McLoughlin IV, Emmanuel S, Chia LT (2012b) Fourier transform-based scalable image quality measure. IEEE Transactions on Image Processing 21(8):3364–3377
39. Narwaria M, Da Silva MP, Le Callet P, Pepion R (2013) Tone mapping-based high-dynamic-range image compression: study of optimization criterion and perceptual quality. Optical Engineering 52(10):102,008–102,008
40. Narwaria M, Da Silva MP, Le Callet P, Pépion R (2014a) Impact of tone mapping in high dynamic range image compression. In: VPQM, pp pp–1
41. Narwaria M, Da Silva MP, Le Callet P, Pepion R (2014b) Tone mapping based HDR compression: Does it affect visual experience? Signal Processing: Image Communication 29(2):257–273

42. Narwaria M, Da Silva MP, Le Callet P (2015a) HDR-VQM: An objective quality measure for high dynamic range video. Signal Processing: Image Communication 35:46–60
43. Narwaria M, Mantiuk RK, Da Silva MP, Le Callet P (2015b) HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images. Journal of Electronic Imaging 24(1):010,501–010,501
44. Narwaria M, Le Callet P, Valenzise G, De Simone F, Dufaux F, Mantiuk R (2016a) HDR image and video quality prediction. High Dynamic Range Video: From Acquisition, to Display and Applications
45. Narwaria M, da Silva MP, Le Callet P, Valenzise G, De Simone F, Dufaux F (2016b) Quality of experience and HDR: concepts and how to measure it. High Dynamic Range Video: From Acquisition, to Display and Applications
46. Nuutinen M, Virtanen T, Häkkinen J (2016) Performance measure of image and video quality assessment algorithms: subjective root-mean-square error. Journal of Electronic Imaging 25(2):023,012–023,012
47. Ortiz-Jaramillo B, Kumcu A, Philips W (2016) Evaluating color difference measures in images. In: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), IEEE, pp 1–6
48. pfstools (2015) pfstools HDR image gallery. `http://pfstools.sourceforge.net/hdr_gallery.html`, accessed: 2015-11-15
49. Pinson M, Wolf S (2008) Techniques for evaluating objective video quality models using overlapping subjective data sets. US Department of Commerce, National Telecommunications and Information Administration
50. Pinson MH, Wolf S (2003) An objective method for combining multiple subjective data sets. In: Visual Communications and Image Processing 2003, International Society for Optics and Photonics, pp 583–592
51. Pitrey Y, Engelke U, Barkowsky M, Pépion R, Le Callet P (2011) Aligning subjective tests using a low cost common set. In: Euro ITV
52. Reinhard E, Stark M, Shirley P, Ferwerda J (2002) Photographic tone reproduction for digital images. In: ACM Transactions on Graphics, ACM, vol 21, pp 267–276
53. Richter T (2013) On the standardization of the JPEG XT image compression. In: Picture Coding Symposium (PCS), 2013, IEEE, pp 37–40
54. Rohaly AM, Libert J, Corriveau P, Webster A, et al (2000) Final report from the video quality experts group on the validation of objective models of video quality assessment. ITU-T Standards Contribution COM pp 9–80
55. Sheikh HR, Bovik AC (2006) Image information and visual quality. Image Processing, IEEE Transactions on 15(2):430–444
56. Sheikh HR, Bovik AC, De Veciana G (2005) An information fidelity criterion for image quality assessment using natural scene statistics. Image Processing, IEEE Transactions on 14(12):2117–2128
57. SMPTE (2014) High dynamic range electro-optical transfer function of mastering reference displays. SMPTE ST 2084
58. Tourapis AM, Singer D (2015) HDRTools: Software updates. In: ISO/IEC JTC1/SC29/WG11 MPEG2015/M35471, IEEE, Ed., Geneva, Switzerland
59. Valenzise G, De Simone F, Lauga P, Dufaux F (2014) Performance evaluation of objective quality metrics for HDR image compression. In: SPIE Optical Engineering+ Applications, International Society for Optics and Photonics
60. Wang Z, Bovik AC (2002) A universal image quality index. Signal Processing Letters, IEEE 9(3):81–84
61. Wang Z, Simoncelli EP, Bovik AC (2003) Multiscale structural similarity for image quality assessment. In: Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on, IEEE, vol 2, pp 1398–1402
62. Wang Z, Bovik A, Sheikh H, Simoncelli E (2004) Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13(4):600–612
63. Ward G, Simmons M (2006) JPEG-HDR: A backwards-compatible, high dynamic range extension to JPEG. In: ACM SIGGRAPH 2006 Courses, ACM, New York, NY, USA, SIGGRAPH '06
64. Zerman E, Valenzise G, Dufaux F (2016) A dual modulation algorithm for accurate reproduction of high dynamic range video. In: Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), 2016 IEEE 12th, IEEE, pp 1–5
65. Zhang X, Wandell BA (1997) A spatial extension of CIELAB for digital color-image reproduction. Journal of the Society for Information Display 5(1):61–63