# Performance evaluation of objective quality metrics for HDR image compression

Giuseppe Valenzise, Francesca de Simone, Paul Lauga, Frederic Dufaux

# Performance evaluation of objective quality metrics for HDR image compression

Giuseppe Valenzise, Francesca De Simone, Paul Lauga, Frederic Dufaux

Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, Paris, France

## ABSTRACT

Due to the much larger luminance and contrast characteristics of high dynamic range (HDR) images, well-known objective quality metrics, widely used for the assessment of low dynamic range (LDR) content, cannot be directly applied to HDR images in order to predict their perceptual fidelity. To overcome this limitation, advanced fidelity metrics, such as the HDR-VDP, have been proposed to accurately predict visually significant differences. However, their complex calibration may make them difficult to use in practice. A simpler approach consists in computing arithmetic or structural fidelity metrics, such as PSNR and SSIM, on perceptually encoded luminance values but the performance of quality prediction in this case has not been clearly studied. In this paper, we aim at providing a better comprehension of the limits and the potentialities of this approach, by means of a subjective study. We compare the performance of HDR-VDP to that of PSNR and SSIM computed on perceptually encoded luminance values, when considering compressed HDR images. Our results show that these simpler metrics can be effectively employed to assess image fidelity for applications such as HDR image compression.

**Keywords:** High dynamic range, quality assessment, image coding

## 1. INTRODUCTION

High dynamic range (HDR) content has been recently gaining momentum thanks to its ability to reproduce a much wider gamut of luminance and contrast than traditional low dynamic range (LDR) formats. This has motivated research towards novel HDR processing algorithms, including acquisition/generation[1] and compression[2,3] and, consequently, towards methods for assessing the quality of the processed results. In principle, the most accurate way to evaluate image quality is to carry out extensive subjective test campaigns. However, this is often impractical, especially when the number of parameters and testing conditions is large. In addition, the feasibility of subjective testing in the case of HDR content is further reduced by the limited diffusion and the high cost of HDR displays. This calls for the design of automatic and accurate objective quality metrics for HDR content.

In this work, we focus on full-reference quality assessment, where the goal is to assess the *perceptual fidelity* of a processed image with respect to its original (i.e., reference) version. This is the typical scenario, e.g., in image compression, where a picture coded at a certain bitrate is compared to the uncompressed original. In the LDR case, popular metrics, such as the Structural Similarity Index (SSIM),[4] are known to provide good predictions of image quality and even the criticized Peak Signal-to-Noise Ratio (PSNR) produces valid quality measures for a given content and codec type.[5] A key advantage of these metrics is that they can be easily computed through simple pixel operations on LDR images. This is partially due to the fact that LDR pixel values are *gamma-corrected* in the sRGB color space,[6] which not only does compensate for the non-linear luminance response of legacy CRT displays, but also accounts somehow for the lower contrast sensitivity of the human visual system (HVS) at dark luminance levels. In other words, the non linearity of the sRGB color space provides a pixel encoding which is approximately linear with respect to perception.

In the case of HDR, this is no longer the case, since pixel values are proportional to the physical luminance of the scene, while the HVS is sensible to luminance ratios, as expressed by the Weber-Fechner law. In order to take into account luminance masking and other complex aspects of the HVS, some metrics, such as the HDR-VDP,[7,8]

Corresponding author: Giuseppe Valenzise — E-mail: giuseppe.valenzise@telecom-paristech.fr

Additional material available at http://perso.telecom-paristech.fr/~gvalenzi/download.htm

accurately model various stages of visual perception under a broad range of viewing conditions, in such a way to predict and quantify precisely significant visual differences between images. These metrics can provide very good approximations of human perception but require in general a delicate tuning of several parameters in order to be computed, which limits their use in many practical applications. A simpler and more convenient approach is to transform HDR values to *perceptually uniform* quantities and compute arithmetic or structural metrics, such as the PSNR or the SSIM, on them. Typical encodings from HDR to perceptually linear values include the simple logarithm, based on the Weber-Fechner law, or more sophisticated transfer functions such as the PU encoding.[9] These metrics are often used to evaluate HDR image and video compression performance;[3,10] however, it is not clear up to which extent they can provide accurate estimates of the actual visual quality, thus, whether they are a valid alternative to more complex predictors based on HVS modeling.

In this paper, we evaluate the performance of PSNR and SSIM applied to log- or PU-encoded HDR pictures corrupted by one specific type of processing, i.e., image compression. Since PSNR and SSIM are widely used for quality assessment of LDR images, in the following, we will refer to them as LDR metrics. We also analyze the performance of the HDR-VDP algorithm (referred to as HDR-VDP-2 in the original paper of Mantiuk et al.[8]). In terms of image compression, we consider three schemes, which are representative of the state of the art in still image HDR content compression, to build a dataset of compressed images with different levels of distortion. We use this dataset to conduct a subjective experiment and collect subjective mean opinion scores (MOS). Our analysis of the results shows that subjective ratings are well correlated with LDR metrics applied to perceptually linearized HDR values, and thus, that they can be consistently used to evaluate coding performance.

The rest of the paper is organized as follows. We review objective approaches to quality assessment of HDR content in Section 2. The subjective test setup, including the generation of the test material, the test environment and the test methodology, is described in Section 3. We present and discuss the results of our study in Section 4. Finally, Section 5 concludes the paper.

## 2. OBJECTIVE METRICS FOR HDR CONTENT

Automatic quality assessment of low dynamic range pictures has been widely investigated in the past decades and a number of full-reference metrics have been proposed for this purpose, including: metrics that model the HVS (e.g., Sarnoff JND,[11] VDP,[12] Perceptual Distortion Metric[13]); feature-based algorithms;[14] application-specific models (DCTune[15]); structural (SSIM[4] and its multiscale version[16]) and information-theoretic (e.g., VIF[17]) frameworks. For a comprehensive statistical evaluation of these algorithms on LDR content, the interested reader can refer to, e.g., the work of Sheikh et al.[18] At a higher level of abstraction, fidelity metrics can be classified according to whether they include some modeling of the HVS (such as contrast and luminance masking, adaptation mechanisms, etc.), or assume perceptually linearized luminance values. The latter is the case of arithmetic measures such as the mean square error (MSE) and derived metrics, such as PSNR, as well as of structural metrics, such as SSIM, which are largely used in fields such as image/video coding as they offer a good trade-off between simplicity and accuracy.

Metrics based on HVS models are conceived to work in a limited luminance range, i.e., that of standard LCD or CRT displays, but need to be somehow extended to work in the full luminance range of HDR content. In their HDR-VDP[8] metric Mantiuk et al. extended the Visual Difference Predictor of Daly,[12] in order to take into account a number of phenomena that occur in the early stages of the HVS – from intra-ocular light scatter to contrast sensitivity across the full range of visible luminance (scotopic and photopic) and intra/inter-channel contrast masking – which characterize the optical and retinal pathway. The test and references pictures are processed according to this path and the resulting images are decomposed through a multiband filter in such a way to obtain perceptually linearized per-band contrast differences. These quantities are then either mapped to per-pixel probabilities maps of visibility, or they are pooled to produce a single image quality correlate $Q$. The pooling function has been selected and parametrized among several candidates by maximizing Spearman rank-order correlation over a large LDR image dataset (details are found in Section 6.1 of the original HDR-VDP paper[8]). The motivation of this choice is twofold: on one hand, it assures the backward compatibility of the metric to LDR content; on the other hand, it is the only feasible way to optimize the pooling function in the lack of sufficiently large HDR datasets with subjective annotations. Recently, Narwaria et al.[19] computed optimized pooling weights for HDR-VDP over a dataset of HDR compressed images. Their results show that tuning on

HDR data may improve HDR-VDP performance, but the gain is not statistically significant. Thus, in this work, we resort to the default setting in the implementation of Mantiuk et al.*, which we parametrize to account for the viewing conditions described in Section 3.2.

A main disadvantage of HDR-VDP is that it requires a complex calibration of its optical and retinal parameters. A known problem is, e.g., the setting of the peak sensitivity of the photoreceptors – higher values decrease overall sensitivity to contrast. In many practical applications, and especially in the case of coding, it is customary to compute simple arithmetic or structural metrics on *perceptually linearized* HDR values. Perceptual linearization consists in a monotonically increasing mapping of HDR luminance to encoded pixel values. Typical mapping functions include the logarithm, as it expresses Weber-Fechner law on small luminance ranges, or a gamma correction to account for Steven's power law.[20] Aydin et al.[9] observed that the Weber ratio can be assumed to be constant only for luminance values approximatively greater than 500 $cd/m^2$, while for lower luminance levels the detection threshold rises significantly. Thus, they computed a perceptually uniform (PU) encoding under the form of a look-up table, which follows the Weber-Fechner law for luminance larger than 1000 $cd/m^2$, while at the same time it maintains backward compatibility with the sRGB encoding on typical LDR displays brightness ranges. Notice that this mapping requires a rough characterization of the response function of the HDR display in order to transform HDR pixel values into photometric quantities.

Quality assessment for high dynamic range is quite a recent topic, hence there is lack of extensive statistical studies and image datasets to evaluate performance of existing metrics. Perceptual linearization is supported by psycho-visual arguments, but its effectiveness for quality assessment has only been conjectured or just showcased through simple proofs of concepts in the case of PU encoding. Additionally, to the authors' knowledge, the only study on the performance of HDR-VDP on HDR content is the recent work by Narwaria et al.,[19] which considers test material similar to that considered in this paper, i.e., compressed HDR images. The main difference with respect to that study is that, there, the authors compared HDR-VDP with LDR metrics computed over HDR pixel values *without any perceptual linearization*. Therefore, they arrive to the rather expected result that HDR-VDP clearly outperforms LDR metrics and that LDR metrics cannot be used to evaluate HDR content. In this work, we use instead perceptually linearized HDR values, obtained using either logarithm or PU encoding. Under this setting, our results reverse the conclusions found previously and show that, with an appropriate perceptual linearization, well-established metrics that work excellently for LDR image coding can be extended with similar performance to HDR.

## 3. SUBJECTIVE TEST SETUP

### 3.1 Test material

#### 3.1.1 Selection of original content

We analyzed several HDR images from the HDR photographic survey dataset,[21] as potential test material to be included in our experiment. The resolution of the pictures was downscaled to meet our display's resolution, equal to $1920 \times 1080$ pixels. We focused on high quality images where typical HDR acquisition artifacts such as ghosting are not present. In order to select material with sufficiently diverse characteristics, we compute the following three features for each image:

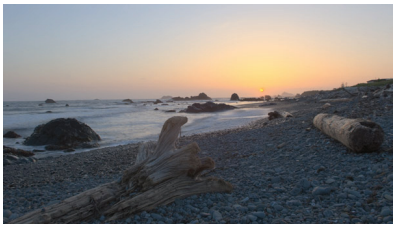
- The *key* $k \in [0, 1]$ of the picture,[22] which gives a measure of the overall brightness of the scene and is defined as:
$$k = \frac{\log L_{\text{avg}} - \log L_{\min}}{\log L_{\max} - \log L_{\min}}, \tag{1}$$
where the average luminance is computed as $\log L_{\text{avg}} = \sum_{ij} \log(L(i,j) + \delta)/N$, with $N$ being the number of pixels in the image, $L(i,j)$ the luminance of pixel $(i,j)$, and $\delta$ is a small offset to avoid the singularity occurring for black pixels. $L_{\min}$ and $L_{\max}$ are the minimum and maximum relative luminance values of the image, computed after excluding 1% of brightest and darkest pixels in order to make the method robust against outliers.

*Available at `http://sourceforge.net/projects/hdrvdp/` (version 2.1.3).

(a) "AirBellowsGap"


(b) "LasVegasStore"

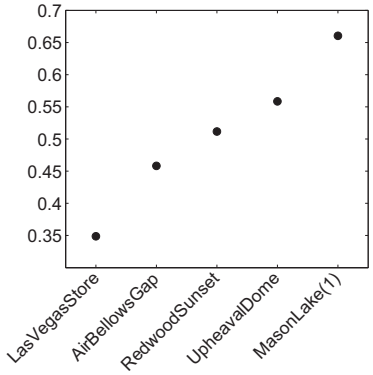
(c) "MasonLake(1)"



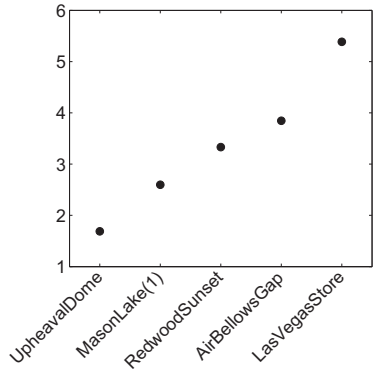
(d) "RedwoodSunset"

(e) "UpheavalDome"

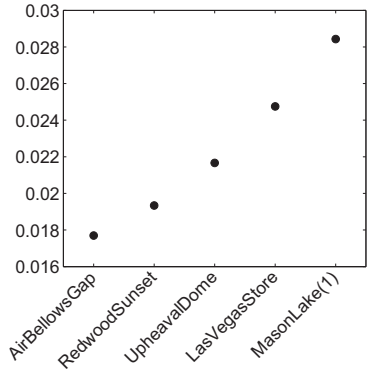
Figure 1. HDR images used for the test (tone mapped version).




(a) Key




(b) Dynamic Range




(c) Spatial Information

Figure 2. Characteristics of the selected HDR test images (contents are ordered for increasing value of each feature).

- The image *dynamic range* $DR = L_{\max}/L_{\min}$, with $L_{\min}$ and $L_{\max}$ computed as above.
- The *spatial perceptual information* $SI$,[23] which describes image spatial complexity and is related to coding complexity. For an LDR image, spatial information is defined as the standard deviation of the output of a Sobel operator applied to the image. The LDR image in our case is obtained using Reinhard's photographic tone reproduction operator.[24]

Based on the semantic interest of each content and on the diversity of the considered characteristics, we selected the five images shown in Fig. 1. Fig. 2 reports the content characteristics for the selected material. Two additional images, shown in Fig. 3, were used for training the subjects.

### 3.1.2 Production of test material

We produced the test material by compressing the selected images using different codecs and coding conditions. Due to the huge bulk of available LDR images, the most promising HDR image coding techniques are those that offer backward compatibility with legacy LDR pictures. These schemes are based on a scalable approach,[25] where an LDR base layer is obtained by tone mapping the original HDR and is then coded using available LDR codecs such as JPEG or JPEG 2000. The tone mapping function is inverted at the decoder to reconstruct an approximation of the original HDR. Additionally, an enhancement layer that stores the differences (or ratios)

(a) "DevilsBathtub" (b) "PaulBunyan"

Figure 3. Training images (tone mapped version).

between the original and the inverse tone mapped images can be also transmitted as header information. In addition to the usual settings to optimize in the LDR case (e.g., quantization parameters, transform size, etc.), the choice of the tone mapping operator (TMO) is critical and can lead to different coding performance.[26] Instead of using a tone mapping designed for rendering on a LDR display, we implemented the minimum-MSE TMO proposed by Mai et al.,[3] which is the global TMO that minimizes the reconstruction error after tone mapping and inverse tone mapping.

Thus, we consider the following three coding schemes:

- JPEG with minimum-MSE TMO (applied to each color channel) and no enhancement layer. We coded each content with a JPEG quality factor $QF$ ranging from 20 to 100, with a step of 5, producing a total of 17 rate points $\times$ 5 contents = 85 images.
- JPEG 2000 with minimum-MSE TMO (applied to each color channel) and no enhancement layer. We sampled 15 target bitrates in the range 0.06 bpp up to 1.75 bpp, giving a total of 75 images.
- JPEG XT,[2] which is the new standardization initiative (ISO/IEC 18477) of JPEG for backward compatible encoding of HDR images. JPEG XT produces a LDR bitstream compatible with the JPEG standard. There are several proposals so far for coding the enhancement layer. In the reference implementation that we adopted[†], the TMO is a content dependent linear map, followed by a gamma adaption with exponent 2.2 to compensate for the sRGB gamma. Encoding of residuals is performed in a lossy manner in the spatial domain. The base and enhancement layer quality is controlled by two quality factors, which take values on $[0, 100]$ and that we varied as follows: $QF_b \in [40, 70, 90, 100]$ and $QF_e \in [50, 75, 80, 90, 95]$, respectively. This yields 100 coded images.

We screened all the 260 images, produced with the coding conditions described above, and we selected a subset of them in such a way to respect the following requirements: i) all the levels of the MOS scale (described in Section 3.3) should be equally represented; ii) all codecs and contents should be equally present; and iii) the length of the actual test should be reasonable, i.e., it should not be longer than 20 minutes without pauses. Distortions with the JPEG and JPEG 2000 codecs, when seen on the HDR display, are similar to analogous distortions in LDR pictures. As for the JPEG XT codec, its distortion has characteristics similar to JPEG: specifically, the noise has the same typical blocking structure; however, as $QF_e$ increases, JPEG XT images have less ringing artifacts than JPEG ones. Finally, we observed that, for some contents, even with the highest considered bitrates, none of the used lossy coding schemes was able to produce imperceptible distortions (i.e., the highest level of the considered MOS scale) on the HDR display. This confirmed the findings of Aydin et al.[9] that distortions are much more perceptible on brighter screens. In those cases, we used the original (uncompressed) content as test image. These samples were excluded from the performance analysis of the objective metrics in order to avoid any bias due to the choice of an arbitrary maximum value for the PSNR. As a result of the screening phase, we retained a set of 50 images to use for the test (details about the exact coding parameters of the test dataset, as well as coded images, are available as supplementary material on the reference author's website).

[†]JPEG document `wg1n6639` in the JPEG document repository, version 0.8 (February 2014).

### 3.2 Test environment

The HDR images were displayed on a SIM2 HDR47 display,[27] which has HD1080 resolution with a declared contrast ratio higher than $4 \cdot 10^6$. Using a light probe, we verified the linear response of the monitor and we measured a peak luminance of approximately 4250 $cd/m^2$ when 60% of the screen surface is white.

We set up a test space with mid gray non-reflective background, isolated from external sources of lights, as recommended in the BT.500-13 and BT.2022 standards.[28,29] Differently from the conclusions reported by Rempel et al.,[30] we assessed during a pilot test that viewing sessions longer than a few minutes in a completely dark environment might cause visual fatigue. Therefore, we placed two lamps at 6500K color temperature behind the HDR screen to ensure ambient illumination while avoiding the presence of any direct light source (apart from the HDR display) in the field of view of the user. The ambient light measured in front of the screen, when this is off, is of approximatively 20 $cd/m^2$. Viewers participated individually to test sessions, sitting at a distance of approximately 1 meter, which corresponds to an angular resolution of about 40 pixels per degree.

### 3.3 Test methodology

The subjective quality evaluation has been performed following the Double Stimulus Impairment Scale (DSIS) methodology.[28] Particularly, pairs of images, i.e. stimuli A and B, were sequentially presented to the user. The user was told about the presence of the original (reference) image, always stimulus A in the pair, having the best expected quality. She/he was asked to rate the level of annoyance of the visual defects that she/he may observe in stimulus B using a continuous quality scale ranging from 0 to 100, associated to 5 distinct adjectives ("Very annoying", "Annoying", "Slightly annoying", "Perceptible", "Imperceptible"). Each test session was run with one user, sitting centered in front of the display, and the ratings were collected using paper scoring sheets.

Each image was shown for 6 seconds, after being introduced by a gray screen showing the kind of stimulus (A or B) and the image pair number. Before the visualization of the next pair of images, the user was shown a 5 seconds long gray screen with a "Vote" message to enter the quality rate for the test stimulus. In practice, to allow a detailed exploration of the high resolution content, each user was left free to pause the interface and take as much time as needed in order to visually inspect each image and rate the quality of the test image.

The pairs of stimuli were presented in random order, different for each viewer, with the constraint that no consecutive pairs concerning the same content will occur.

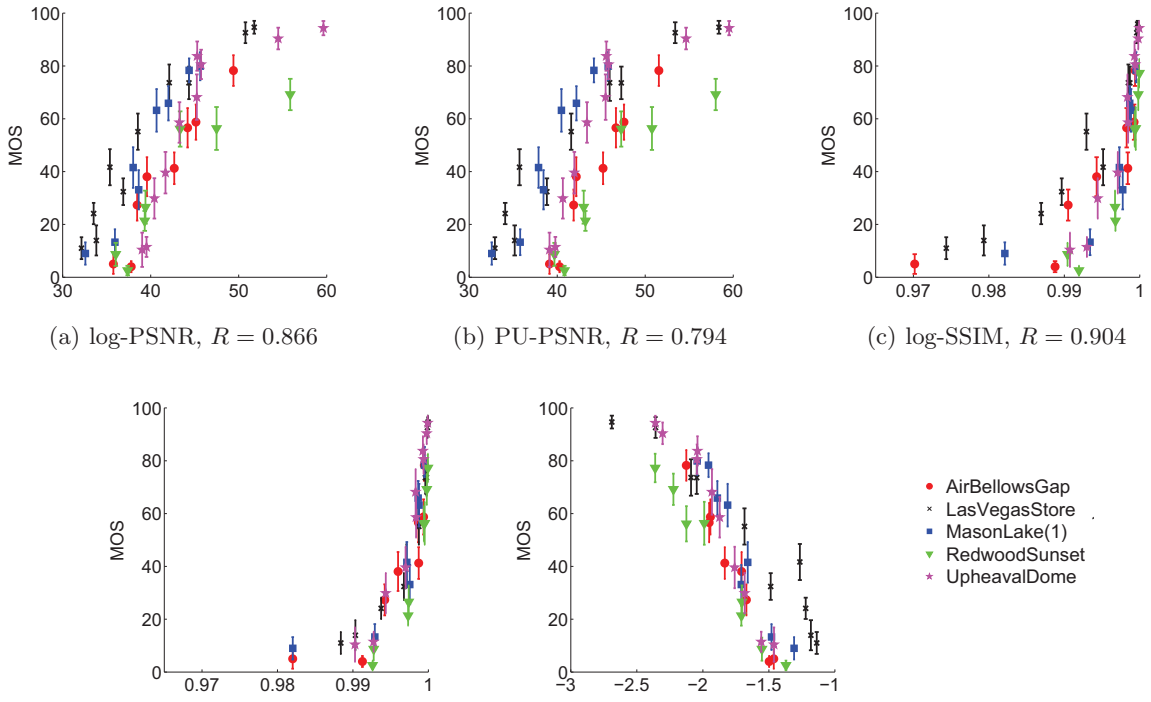
## 4. RESULTS AND ANALYSIS

Fifteen observers (four women, eleven men, average age 30.8 years old) took part in our subjective test. The viewers reported correct color vision and visual acuity and wore corrective glasses when needed. The subjective data has been processed by first detecting outliers, following the standard procedure described in[28] for the DSIS method. No outliers have been detected. The mean opinion score (MOS) and the 95% confidence interval (CI) have then been computed, assuming that the scores are following a $t$-Student distribution.

The metrics considered for evaluation are the HDR-VDP and the classic PSNR and SSIM measures. The PSNR and SSIM have been computed on either a logarithmic mapping (log-PSNR and log-SSIM) or a PU encoding[‡] (PU-PSNR and PU-SSIM) of HDR values. Particularly, since pixel values in the considered HDR images were display-referred, before applying the logarithmic mapping or PU encoding, these have been converted to actual luminance values by multiplying them by the luminance efficacy at equal energy white, i.e., by the constant factor 179. Then, the response of the HDR display, which is approximately linear within its black level (0.03 $cd/m^2$) and maximum luminance (4250 $cd/m^2$) and saturates over this value, has been simulated. The obtained luminance values have been used as input for the HDR-VDP metric, used with the default settings (the only option being specified is the type of display, i.e., "lcd-led").

Figures 4 and 5 show each MOS, with its CI, versus the value of the metric. The two figures depict the same scatter plots, but in the first one the results for each content are highlighted, while in the second the dependency on the codec is shown. We report also the corresponding Spearman rank order correlation coefficient $R$, in modulus, computed on the entire set of MOS points. The use of a non parametric correlation coefficient avoids

[‡]A publicly available implementation can be found at `http://resources.mpi-inf.mpg.de/hdr/fulldr_extension/`.




(a) log-PSNR, $R = 0.866$ (b) PU-PSNR, $R = 0.794$ (c) log-SSIM, $R = 0.904$

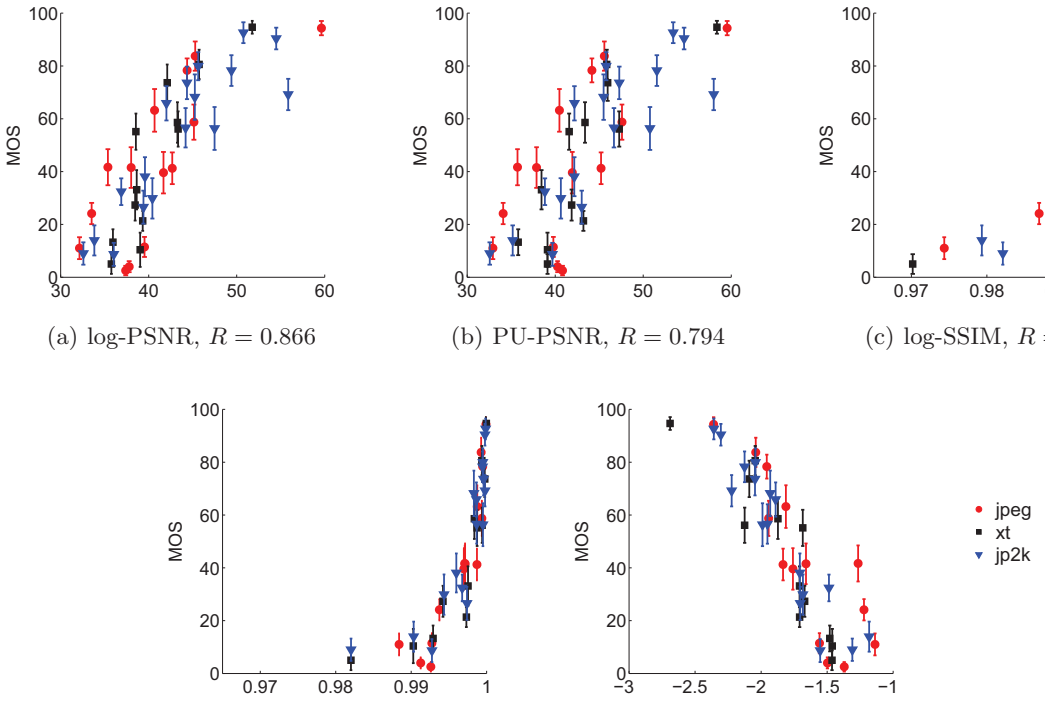
(d) PU-SSIM, $R = 0.923$ (e) HDR-VDP Q, $R = 0.889$

Figure 4. Scatter plots of MOS vs objective metric values, highlighing content dependency. The modulus of the Spearman rank order correlation coefficient $R$ computed on the entire set of test images is shown in the caption of each subfigure. The legend identifies each content.

any need for non linear fitting in order to linearize the values of the objective metrics, which may be questionable due to the relatively small size of our subjective groundtruth dataset.

The scatter plots clearly show a low level of dispersion, which indicates overall good prediction performance of all the considered metrics. More specifically, the correlation coefficient demonstrate that PU-SSIM provides an excellent prediction in terms of ranking the subjective quality of the considered HDR images. The values of the modulus of the correlation coefficient for each metric, per content and per codec, are also reported in Table 1. As it can be seen, overall the best performing metric is the PU-SSIM, followed by log-SSIM and HDR-VDP. The results obtained with two additional metrics based on HDR-VDP, i.e. $P_{\text{avg}}$ and $P_{50}$, have also been reported in the table, since these measures have been used in existing studies in the literature to evaluate coding or inverse tone mapping.[1,31] $P_{\text{avg}}$ is the average probability of detection, computed as the mean of the probability map output by HDR-VDP. $P_{50}$ is the fraction of pixels in the distorted image having more than 50% probability of being detected. While on some contents these metrics could provide good subjective MOS predictions, they are not consistent across several contents and codecs, and thus in general they cannot be used for supra-threshold quality assessment.

In terms of content- and distortion-dependency our results confirm widely known observations concerning the scope of validity of most objective metrics.[5] On one hand, the codec-dependent results show that all metrics suffer to some extent from content-dependency in their prediction capability. On the other hand, the content-dependent results clearly indicate that even perfect (ranking) prediction is reachable for some contents (i.e. results of PU-SSIM for content "LasVegasStore" and "RedwoodSunset"). Of course these results must be interpreted by taking into account the limited set of distortions which are characterizing our test database. In this sense, we believe that an extension of our subjective database will be useful to provide more test samples of the same content and confirm these preliminary results when a wider set of distortions is considered. Finally, it is interesting to notice that the range of PSNR values obtained in this work is significantly higher than that commonly encountered in the case of LDR image compression, as reported, e.g., by De Simone et al.[32]




(a) log-PSNR, $R = 0.866$ (b) PU-PSNR, $R = 0.794$ (c) log-SSIM, $R = 0.904$

(d) PU-SSIM, $R = 0.923$ (e) HDR-VDP Q, $R = 0.889$

Figure 5. Scatter plots of MOS vs objective metric values, highlighing codec dependency. The modulus of the Spearman rank order correlation coefficient $R$ computed on the entire set of test images is shown in the caption of each subfigure. The legend identifies each codec (jpeg = JPEG, xt = JPEG XT, jp2k = JPEG 2000).

## 5. CONCLUSIONS

Quality assessment of high dynamic range content poses new challenges with respect to the effectiveness of well-established fidelity metrics, which have being used for several years in numerous low dyamic range processing tasks. The scene-referred nature of HDR images entails that those metrics, such as the popular PSNR or SSIM, cannot be used on relative luminance values, and therefore, new metrics based on an accurate prediction of the human visual system such as the HDR-VDP have been proposed. At the same time, it has been conjectured but not systematically proved that, under some appropriate perceptual linearization of HDR values, metrics used for the LDR can be extended to higher dyamic range.

In this paper we have carried out a subjective study to gain a better insight of the question. We have focused on backward-compatible HDR image compression, which is currently an active topic as demonstrated by the JPEG XT standardization initiative. Our analyses on existing arithmetic and structural metrics, computed using different perceptual linearization encodings, show that these quality measures can perform as good as or

Table 1. Spearman correlation coefficients (modulus) calculated for each content and codec (jpeg = JPEG, xt = JPEG XT, jp2k = JPEG 2000), with maximum correlation values for each column highlighted in bold.

| | overall | AirBel. | LasVeg. | Mason. | Redw. | Uphea. | jpeg | xt | jp2k |
|---|---|---|---|---|---|---|---|---|---|
| PU-PSNR | 0.794 | **0.976** | 0.963 | **0.976** | 0.952 | **0.987** | 0.591 | 0.797 | 0.835 |
| PU-SSIM | **0.923** | 0.952 | **1** | 0.952 | **1** | 0.975 | **0.942** | **0.944** | 0.887 |
| log-PSNR | 0.866 | **0.976** | 0.963 | **0.976** | 0.976 | **0.987** | 0.753 | 0.832 | 0.887 |
| log-SSIM | 0.904 | 0.952 | 0.975 | 0.928 | 0.952 | 0.975 | 0.907 | 0.881 | 0.872 |
| HDR-VDP $Q$ | 0.889 | 0.952 | 0.987 | **0.976** | 0.952 | **0.987** | 0.802 | 0.909 | **0.924** |
| HDR-VDP $P_{\text{avg}}$ | 0.445 | 0.857 | 0.975 | 0.833 | 0.952 | 0.612 | 0.103 | 0.496 | 0.463 |
| HDR-VDP $P_{50}$ | 0.444 | 0.833 | 0.963 | 0.833 | 0.976 | 0.612 | 0.134 | 0.503 | 0.463 |

even better than the complex HDR-VDP. At the same time, they do not require delicate calibration procedures.

We point out that the results in this paper are valid for the specific case of HDR backward-compatible compression, which produces very similar distortion to the LDR case. An open question is whether the same conclusions may be drawn on other forms of HDR-specific distortion, e.g., inverse tone mapping. More complex metrics such as the HDR-VDP, instead, are expected to adapt to new test conditions and artifacts, thanks to their accurate model of HVS. Also, it has been observed that imperceptible distortion on LDR displays become perceptible on HDR bright displays. This implies that the range of legacy metrics is different for LDR or HDR content, thus further study on the extension of these quality measures to HDR is necessary.

## REFERENCES


[1] Banterle, F., Ledda, P., Debattista, K., and Chalmers, A., "Inverse tone mapping," in [*Proceedings of the 4th International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia*], *GRAPHITE '06*, 349–356, ACM, New York, NY, USA (2006).

[2] Richter, T., "On the standardization of the JPEG XT image compression," in [*Picture Coding Symposium (PCS)*], 37–40 (Dec 2013).

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

[4] Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E., "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004).

[5] Huynh-Thu, Q. and Ghanbari, M., "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters* **44**, 800–801 (June 2008).

[6] Anderson, M., Motta, R., Chandrasekar, S., and Stokes, M., "Proposal for a standard default color space for the internetsrgb," in [*Color and Imaging Conference*], **1996**(1), 238–245, Society for Imaging Science and Technology (1996).

[7] Mantiuk, R., Daly, S., Myszkowski, K., and Seidel, H.-P., "Predicting visible differences in high dynamic range images - model and its calibration," in [*Human Vision and Electronic Imaging X, IS&T/SPIE's 17th Annual Symposium on Electronic Imaging (2005)*], Rogowitz, B. E., Pappas, T. N., and Daly, S. J., eds., **5666**, 204–214 (2005).

[8] Mantiuk, R., Kim, K., Rempel, A., and Heidrich, W., "HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions," in [*ACM Trans. on Graphics*], **30**(4), 40, ACM (2011).

[9] Aydın, T. O., Mantiuk, R., and Seidel, H.-P., "Extending quality metrics to full luminance range images," in [*Electronic Imaging 2008*], 68060B–68060B, International Society for Optics and Photonics (2008).

[10] Koz, A. and Dufaux, F., "Methods for improving the tone mapping for backward compatible high dynamic range image and video coding," *Signal Processing: Image Communication* **29** (Feb. 2013).

[11] Lubin, J., "A visual discrimination model for imaging system design and evaluation," *Vision models for target detection and recognition* **2**, 245–357 (1995).

[12] Daly, S., "The visible differences predictor: an algorithm for the assessment of image fidelity," in [*Digital images and human vision*], 179–206, MIT Press (1993).

[13] Winkler, S., [*Digital video quality: vision models and metrics*], John Wiley & Sons (2005).

[14] Miyahara, M., Kotani, K., and Algazi, V., "Objective picture quality scale (PQS) for image coding," *IEEE Transactions on Communications* **46**(9), 1215–1226 (1998).

[15] Watson, A. B., "Dctune: A technique for visual optimization of dct quantization matrices for individual images," in [*Sid International Symposium Digest of Technical Papers*], **24**, 946–946, SOCIETY FOR INFORMATION DISPLAY (1993).

[16] Wang, Z., Simoncelli, E., and Bovik, A., "Multiscale structural similarity for image quality assessment," in [*Proc. 37th Asilomar Conference on Signals, Systems and Computers*], **2**, 1398–1402 (2004).

[17] Sheikh, H. and Bovik, A., "Image information and visual quality," *IEEE Transactions on Image Processing* **15**(2), 430–444 (2006).

[18] Sheikh, H., Sabir, M., and Bovik, A., "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing* **15**, 3440–3451 (Nov 2006).

[19] Narwaria, M., Da Silva, M., Le Callet, P., and Pepion, R., "On improving the pooling in HDR-VDP-2 towards better HDR perceptual quality assessment," in [*IS&T/SPIE Electronic Imaging*], 90140N–90140N, International Society for Optics and Photonics (2014).

[20] Stevens, S., "On the psychophysical law.," *Psychological review* **64**(3), 153 (1957).

[21] Fairchild, M. D., "The HDR photographic survey," in [*Color and Imaging Conference*], **2007**(1), 233–238, Society for Imaging Science and Technology (2007).

[22] Akyüz, A. O. and Reinhard, E., "Color appearance in high-dynamic-range imaging," *J. Electronic Imaging* **15**(3), 033001 (2006).

[23] ITU-T, "Subjective Video Quality Assessment Methods for Multimedia Applications." ITU-T Recommendation P.910 (Apr 2008).

[24] Reinhard, E., Stark, M., Shirley, P., and Ferwerda, J., "Photographic tone reproduction for digital images," in [*ACM Transactions on Graphics*], **21**(3), 267–276, ACM (2002).

[25] Ward, G. and Simmons, M., "JPEG-HDR: A backwards-compatible, high dynamic range extension to JPEG," in [*ACM SIGGRAPH 2006 Courses*], *SIGGRAPH '06*, ACM, New York, NY, USA (2006).

[26] Narwaria, M., 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**(10), 102008–102008 (2013).

[27] SIM2, "http://www.sim2.com/hdr/," (June 2014).

[28] ITU-R, "Methodology for the Subjective Assessment of the Quality of Television Pictures." ITU-R Recommendation BT. 500-13 (Jan 2012).

[29] ITU-R, "General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays." ITU-R Recommendation BT. 2022 (Aug 2012).

[30] Rempel, A. G., Heidrich, W., Li, H., and Mantiuk, R., "Video viewing preferences for hdr displays under varying ambient illumination," in [*Proceedings of the 6th Symposium on Applied Perception in Graphics and Visualization*], *APGV '09*, 45–52, ACM, New York, NY, USA (2009).

[31] Zhang, Y., Naccari, M., Agrafiotis, D., Mrak, M., and Bull, D., "High dynamic range video compression by intensity dependent spatial quantization in HEVC," in [*Picture Coding Symposium*], 353–356 (2013).

[32] De Simone, F., Goldmann, L., Lee, J., and Ebrahimi, T., "Performance analysis of VP8 image and video compression based on subjective evaluations," in [*SPIE Optics and Photonics, Applications of Digital Image Processing XXXIV*], International Society for Optics and Photonics (2011).