Effects of display rendering on HDR image quality assessment

Emin Zerman, Giuseppe Valenzise, Francesca de Simone, Francesco Banterle, Frederic Dufaux

Effects of display rendering on HDR image quality assessment

Emin Zerman[a], Giuseppe Valenzise[a], Francesca De Simone[a], Francesco Banterle[b], Frederic Dufaux[a]

[a]Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, Paris, France
[b]Consiglio Nazionale delle Ricerche, Pisa, Italy

ABSTRACT

High dynamic range (HDR) displays use local backlight modulation to produce both high brightness levels and large contrast ratios. Thus, the display rendering algorithm and its parameters may greatly affect HDR visual experience. In this paper, we analyze the impact of display rendering on perceived quality for a specific display (SIM2 HDR47) and for a popular application scenario, i.e., HDR image compression. To this end, we assess whether significant differences exist between subjective quality of compressed images, when these are displayed using either the built-in rendering of the display, or a rendering algorithm developed by ourselves. As a second contribution of this paper, we investigate whether the possibility to estimate the true pixel-wise luminance emitted by the display, offered by our rendering approach, can improve the performance of HDR objective quality metrics that require true pixel-wise luminance as input.

Keywords: High dynamic range, quality assessment, image coding

1. INTRODUCTION

High dynamic range (HDR) display technology has considerably evolved in the last decade,[1] enabling much higher peak luminance and contrast ratios than conventional low dynamic range (LDR) displays can achieve. These novel viewing conditions have called into question traditional image and video quality assessment tools and best practices, and have recently motivated a great deal of research towards assessing HDR visual quality.[2–6]

In our previous work, we have shown that popular metrics commonly used for measuring distortion in compressed LDR images, can effectively be adopted for comparing the visual quality of compressed HDR images, provided that HDR values are scaled to the luminance range of the display and perceptually encoded.[7] Nevertheless, in that study the luminance scaling to fit the limits of the display was a simple linear transfer function. Instead, it is well known that HDR rendering on dual-modulated displays is an inherently local process, where factors as display power constraints, resolution of the back panel, etc., play a key role, making the rendered pixel luminance a highly nonlinear function of the input pixel values.[1,8] The effect of different rendering algorithms has been previously studied, objectively and subjectively, on both simulated[9] and real[10] backlight dimming LCD systems. In their work,[6] Hanhart et al. showed that human observers prefer images displayed at high brightness levels over images visualized at low brightness levels, a result that was previously observed also by Akyuz et al.[11] This brings forth the thought that the quality experienced by humans viewing images on an HDR display may differ due to rendering differences.[10,12] Additionally, quality metrics developed for the assessment of HDR image and video quality[3–5] require as input the per pixel luminance values (expressed in cd/m^2) that an observer in front of the display would see. As a result, different renderings could also have a potential impact on the calculation of objective quality. In spite of this close connection between quality evaluation and HDR visualization, the effect of different rendering on HDR subjective and objective quality assessment has not been sufficiently investigated so far.

The goal of this study is to assess the impact of HDR image rendering on both subjective and objective scores. We first develop a simple, yet effective, HDR image rendering for the SIM2 HDR47 display,[13] and compare it with the proprietary built-in visualization offered by the display. Notice that the SIM2 HDR47 display is widely used with this built-in mode in many subjective studies on HDR image and video compression.[14] The proposed

Corresponding author: Giuseppe Valenzise — E-mail: giuseppe.valenzise@telecom-paristech.fr
Additional material available at http://perso.telecom-paristech.fr/~gvalenzi/download.htm

rendering algorithm has clear differences from the built-in one, e.g., it yields brighter images, with higher local contrast at low luminance levels. Equipped with this new rendering, we conduct a subjective study to judge the quality of compressed HDR images, using the same settings as in our previous work,[7] except that we display images with the new rendering algorithm. We show through a multiple comparison analysis that a different rendering does not affect substantially subjective mean opinion scores (MOS), except for the highest quality levels, where the artifacts of visualization overcome those due to compression. Since, however, typical HDR use-cases entail a high-quality scenario, this suggests that rendering could play an important role in assessing the performance of image processing or compression techniques.

As a second contribution, we consider the effect of display modeling on the computation of objective quality metrics. To this end, we estimate per pixel luminance produced by the display with our rendering algorithm*, and use this as input to quality metrics for both pristine and compressed contents. We compare this with a simple linear model of display response, which scales HDR pixels into the physical bounds of display luminance and clips values that overpass the peak luminance of the device.[7] Surprisingly, our results show that the performance of objective metrics, measured through Spearman rank-order correlation coefficient (SROCC), do not increase significantly by increasing the accuracy of input luminance values, with respect to the simple linear model with clipping. This finding is particularly interesting since a precise estimation of displayed luminance values would require the knowledge of the reproduction device, as well as its characterization. Conversely, we show that a simple linear model, which is almost independent from the display – only peak brightness is needed, but it can be shown that the predictions of the metrics are robust to its changes – can provide results as reliable as if a detailed knowledge of the reproduction display were available.

The rest of the paper is organized as follows. The details of our HDR image rendering algorithm are discussed in Section 2. In section 3, the impact of this developed rendering on the subjective quality is analyzed. Impact of the rendering on the objective quality is discussed in section 4. Section 5 concludes the paper.

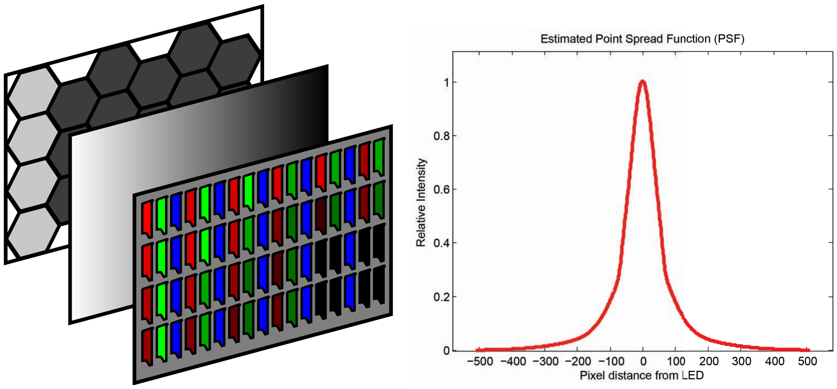
2. HDR IMAGE RENDERING ALGORITHM

The display used in this work is the SIM2 HDR47[13] display, which has a nominal peak luminance of 4250 cd/m^2 and a contrast ratio higher than $4 \cdot 10^6$. This device is a dual-modulated LED-LCD screen, as shown in Fig. 1, where LED and LCD parts can be driven separately. There are 2202 independently controllable LED lights, while the LCD panel consists of 1920 × 1080 pixels (HD resolution).

The SIM2 HDR47 can be used via two different modes: the built-in automatic HDR mode (HDRr), and the DVI Plus mode (DVI+). In HDRr mode, the user supplies the HDR image to a special software, that converts it to Log Luv color space. The Log Luv image is then processed internally by the display, which determines the values of LED's and of LCD pixels *transparently* to the user, i.e., one cannot know the values of single LED's and of LCD pixels obtained in this process. In DVI+, the user can supply the screen with customized LED and LCD values, by formatting standard HD-resolution images so that a small number of pixels are used to signal LED values, while the rest correspond to LCD pixel values. Thus, the task of an HDR rendering algorithm is to determine the values of LED/LCD illumination.

The rendering process on a LED/LCD system, as that displayed in Fig. 1, is essentially a deconvolution problem, i.e., finding the values of the LED's and of the LCD pixels in such a way to minimize the distance from a target input image. In this process, a critical factor is the asymmetry in the resolution of the LED and LCD panels – the number of pixels in the LCD panel is much greater than the number of LED's, and the point spread function (PSF) of the LED diffuser has a size of approximately 1000 × 1000 pixels, which is necessary to avoid discontinuities in the LCD illumination. Another delicate aspect of the display is the LCD *leakage*,[15] due to the non-ideal response of liquid crystals that allow a small percentage of incoming light to pass through them even when they are completely closed (black). Finally, an important aspect is power constraint, i.e., overall brightness should be modulated to account for the maximal power absorption of the display. In practice, this causes some very bright regions of the image to be *clipped*, causing detail loss.[15]

*Unfortunately, a precise estimation of per pixel luminance using SIM2 HDR47 display is not available with the built-in rendering mode.



(a) SIM2 HDR47 Display Parts: LED backlight, light diffuser, and LCD panel

(b) Cross Section of PSF

Figure 1. Layers of an LED/LCD dual-modulated HDR display and Point Spread Function (PSF) of a LED, due to diffuser.

2.1 Details of the proposed rendering

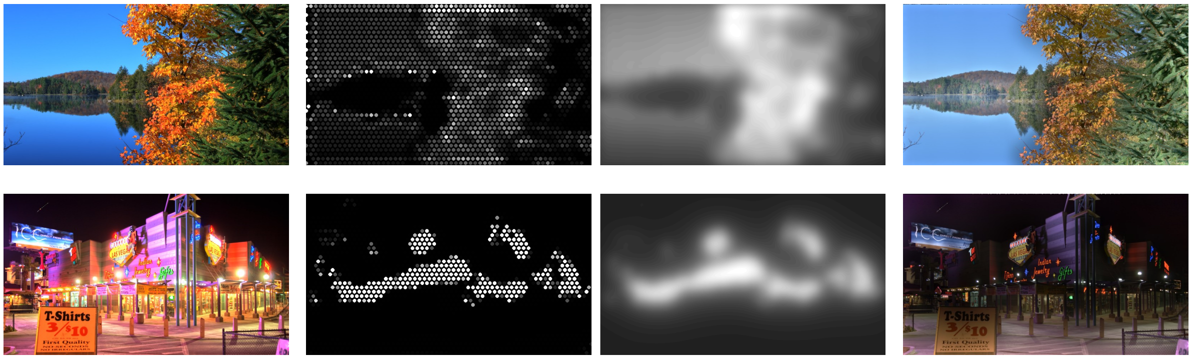
There has been a discrete amount of work[1,8,15,16] dealing with dual-modulation LED displays. The majority of these approaches are based on optimization methods. From a practical point of view, the feasibility of these techniques depends on the size of the problem at hand, i.e., the number of LED and LCD pixels. For instance, Mantel et al.[16] and Burini et al.[8,15] propose a gradient descent optimization for a display with 16 LED segments (8 columns and 2 rows). The backlight is modeled in terms of power used to light the LED's, and the rendering is validated by a subjective test that shows a preference of the observers with respect to alternative simpler rendering approaches. A larger LED setup has been considered by Seetzen at al.,[1] who use a local approximation of the gradient with a single Gauss-Seidel iteration to solve for 760 LED's and 1280×1024 LCD pixels in their prototype HDR display.

In this paper, we find the values of each of the 2202 LED's of the SIM2 HDR47 display by a simple iterative scaling algorithm. The detailed procedure is as follows:

- *Preprocessing.* First, we find the target *display-referred* luminance values from the input HDR image. HDR images are generally *scene-referred*, i.e., they store values proportional to the physical luminance of the scene. However, the luminance range that can be reproduced by an HDR display is clearly inferior to that of the scene. Therefore, the images should be "graded" to the display capabilities manually or by some automatic process, e.g., by using the display-adaptive tonemapping of Mantiuk et al.[17] Here, we assume that the input images have been previously graded to the display, and we just saturate luminance values in excess of the maximum display brightness. We denote the preprocessed image as I.

- *Computation of target LED backlight.* Next, we search for the optimal backlight target luminance map L_{opt} that minimizes the required backlight luminance (to meet the power constraint) *and* maximizes the fidelity to the target pixel values. To do this, we first compute local maxima of the target luminance over 30-pixel radius windows, where 30-pixel is the approximate area corresponding to one LED. To control the effects of LCD leakage (which decrease local contrast), the maximum luminance values allowed for each pixel are found by dividing the target luminance of that pixel by the estimated LCD leakage factor $\epsilon = 0.005$. The LCD leakage factor ϵ is found empirically by measuring LCD leakage in different test patterns, using a Minolta LS-100 luminance meter. Median filtering is applied afterwards in order to avoid any peaks that may be caused by very bright and very small light sources, e.g., stars on a dark sky. This also enables us to meet energetic constraint of the display.

- *Convolution.* Once L_{opt} is computed, we use an iterative procedure to compute the physical LED's values that enable to reproduce it on the display. The LED's are initialized by sampling L_{opt} on the LED grid,



(a) (b) (c) (d)

Figure 2. Examples of rendering for two HDR images. Top row: “MasonLake(1)”; bottom row: “LasVegasStore”. (a) HDR (tone-mapped) images. (b) LED_{opt}. (c) $LED_{opt} * PSF$. (d) LCD images.

which yields an array of LED values $LED^{(0)}$. Given $LED^{(0)}$, the rendered backlight on the display is obtained by convolving the values of the LED’s with the PSF of a single LED: $L^{(0)} = LED^{(0)} * PSF$.

- *Scaling and projection.* A scale map is generated by dividing the target luminance by $L^{(0)}$. By using this scale map, the LED values are scaled as follows:

$$LED^{(1)} = LED^{(0)} \times \left(\frac{L_{opt}}{L^{(0)}} \right). \tag{1}$$

  $LED^{(1)}$ is then clipped to take values in $[0, 1]$, i.e., it is projected onto the set of feasible LED’s values. This and the previous steps are then iterated to obtain $LED^{(1)} \rightarrow L^{(1)} \rightarrow LED^{(2)} \ldots$, until the sum of squared errors $||L_{opt} - L^{(i)}||^2$ falls below a given threshold. The estimate LED_{opt} is checked for the power constraint and further scaled to comply with the maximum power absorption of the display.

- *LCD calculation.* The LCD values of the panel are found by dividing (pixel-wise) each channel of the original image by the result of the previous optimization:

$$LCD_j = \left(\frac{I_j}{LED_{opt} * PSF} \right)^{1/\gamma_j} \tag{2}$$

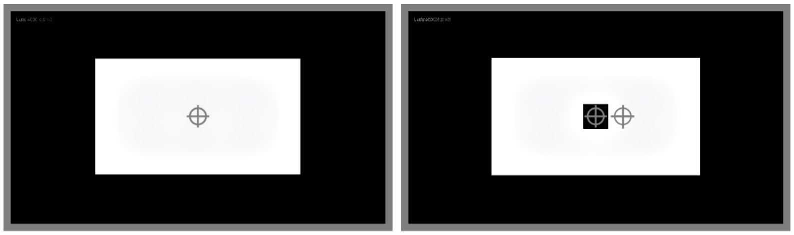
  where $j \in \{R, G, B\}$ is the RGB channel indicator, and γ_j is the gamma correction factor, which has been determined experimentally for each channel.

A Matlab implementation of this algorithm takes an average of 21 seconds (corresponding to about 100 iterations) on an Intel i7-3630QM 2.40 GHz 8 GB RAM PC for rendering a 1920 × 1080 pixels image. Example results of LED backlight LED_{opt} and LCD panel images can be seen in Figure 2.

2.2 Comparison with HDRr mode

We characterize the performance of the rendering algorithm described in Section 2.1 with respect to the built-in HDRr mode in terms of accuracy of brightness rendering and local contrast. Since an evaluation of these two measures on complex content (such as natural images) is per itself a challenging and content-dependent task, we consider here simple stimuli, which also enable a more accurate measurement of displayed luminance using the Minolta LS-100 luminance meter. Specifically, we consider the following test pattern:

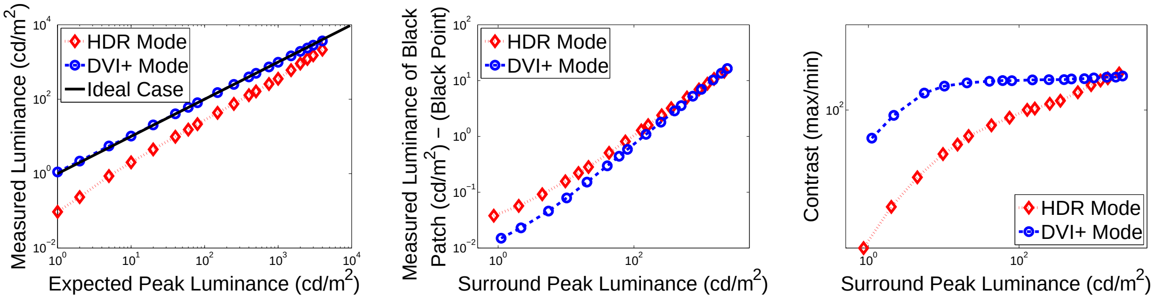
- *Linear brightness response and peak luminance.* We use the pattern of Figure 3(a) to measure the accuracy of produced luminance with respect to the target one. The pattern consists of a white box covering 30% of the display surface, surrounded by a black background. The 30% area is selected as it yields the maximum



(a) Measurement of linear brightness response and peak brightness.

(b) Measurement of black level and local contrast.

Figure 3. Test patterns and measurement spots.



(a) Peak luminance

(b) Black point

(c) Contrast

Figure 4. Comparison of peak brightness, black level luminance and local contrast of HDRr and DVI+, using the test patterns shown in Figure 3.

luminance of the display in HDRr mode. A sequence of test patterns as in Figure 3(a) is generated, for values of brightness levels of the white box ranging from 1 to 4,000 cd/m^2. Figure 4(a) shows the value of luminance in cd/m^2 measured in correspondence of the cross in Figure 3(a), as a function of the target input luminance at the same spot. The black solid line indicates the ideal case of a perfectly linear response, i.e., measured luminance matches exactly the required one. This plot shows that: i) the proposed DVI+ algorithm matches more precisely target luminance; ii) it also achieves a higher peak brightness than HDRr mode.

- *Local contrast.* Local contrast is tested with the pattern in Figure 3(b). This stimulus contains again a white box of 30% of screen area, but in the middle of the white area there is a 64 × 64 pixels square black patch. The small black square width is chosen in order to gauge how LCD leakage affects local contrast in different renderings. We consider several versions of this pattern with different luminance levels of the white region. Figure 4(b) shows the measured luminance of the center black surface versus the measured luminance of the white box. Both measurements spots are shown in Figure 3(b). The plot shows how the black level of the center black square is darker for DVI+ than for HDRr, i.e., the proposed rendering manages to handle better LCD leakage. The effect of this on local contrast, measured as the ratio between the luminance of the white and black patches, is shown in Figure 4(c), which highlights the better local contrast achievable with DVI+.

3. IMPACT ON SUBJECTIVE EVALUATION

In this section we analyze how a different display rendering can affect subjective quality, for the scenario of HDR image compression. Switching from HDRr to DVI+ mode on the SIM2 HDR47 display takes several seconds and requires a manual intervention of the experimenter, thus designing a test presenting the results of both rendering at the same time is not feasible. Therefore, in this paper we design a subjective test with the same test material and conditions as in our previous work,[7] which collected MOS's using HDRr mode, where the only experimental variable that is changed is the rendering mode (DVI+). We summarize here the test environment and methodology. Next, we analyze the differences among the results through analysis of variance and multiple comparisons.

3.1 Test environment and methodology

We use the same test environment and material as for the HDRr dataset of Valenzise et al.,[7] in order to rule out all possible independent variables but the different rendering. That is, the experiment is conducted in a gray surfaced test space that is isolated from all external light sources as it is stated in the BT.500-13 and BT.2022 standards.[18, 19] The amount of ambient light, not directed to the observer, is 20 cd/m^2. The viewers are seated at about 1 meter distance from the display. As test methodology, we employ Double Stimulus Impairment Scale (DSIS),[18] coherently with the HDRr dataset.[7] In this test method, two images, reference image A and distorted image B are shown to subjects in a sequential manner. Before the experiment, a training session has been conducted to familiarize the subjects with the levels of distortion to be expected during the experiment. As for the HDRr dataset, the subjects are asked to rate the distortion appearing in the distorted image B using 5 distinct adjectives ("Very annoying", "Annoying", "Slightly annoying", "Perceptible but not annoying", "Imperceptible"), on a continuous scale between 0 and 100, 0 being "Very annoying" and 100 "Imperceptible".

While conducting the pilot test, it is noticed that the magnitude of distortion in the images is more difficult to judge than for the HDRr case. So, differently from the previous experiment, compressed images are displayed for a duration of 8 seconds instead of the 6 seconds used for the HDRr dataset. There is a total of 50 images in the dataset, spanning several contents and coding conditions as detailed in the original HDRr dataset paper.[7] The experiment is paused during the interactive voting, leaving to the subjects as much time as they wish to complete the task. During the pilot test, it is noted that the average voting time is between 4 and 8 seconds. Hence, the experiment takes approximately 20 minutes.

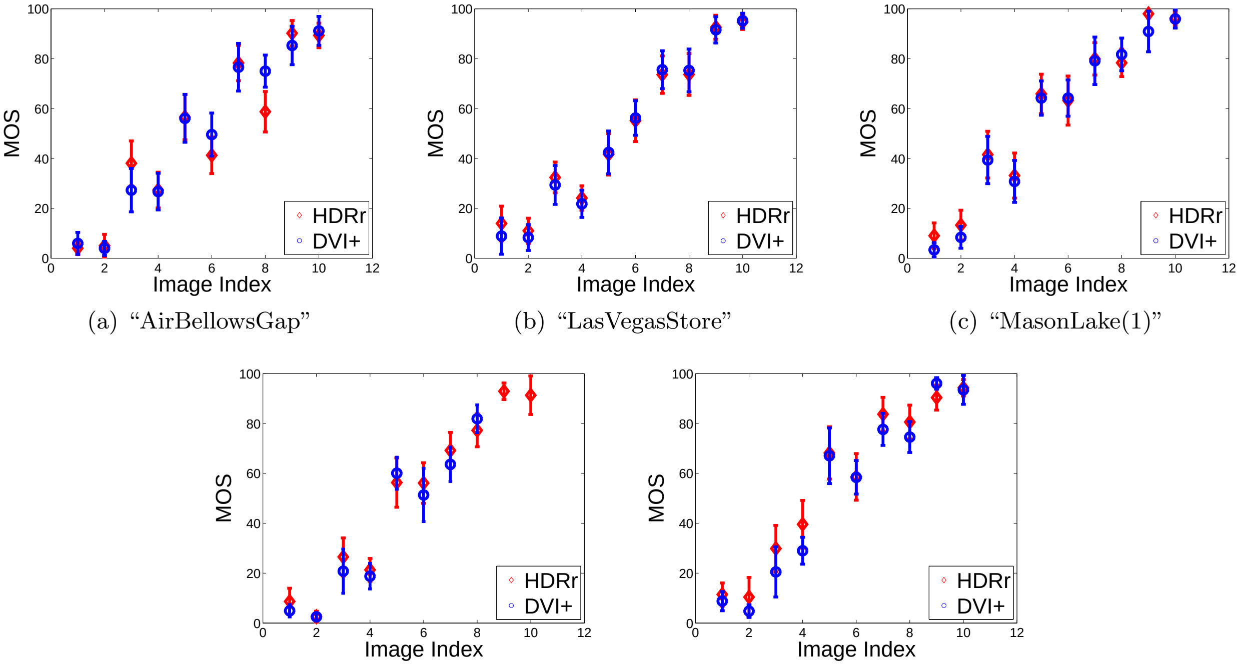
3.2 Experiment Results

Sixteen people (fourteen men and two women) participated in the subjective experiment with the developed rendering. The subjects were aged between 23 and 39, and the average age was 27.75. All the subjects reported normal or corrected-to-normal vision. Two of the subjects were found to be outliers with the standard detection procedure.[18] The mean opinion score (MOS) and confidence interval (CI) for each of the 50 tested images are calculated after outlier removal, assuming that scores follow a *t-Student* distribution.

The resulting MOS for each content are shown in Figure 5. After concluding the tests, we noticed that two samples of "Perceptible" level of the "RedwoodSunset" content were erroneously repeated twice in place of the corresponding "Imperceptible" level. Hence, we excluded them from this comparison. The results of DVI+ are compared with the HDRr MOS's published with the associated dataset.[7] These plots show a substantial level of agreement between the scores obtained with the two renderings (overall, the MOS's collected using HDRr and DVI+ have a linear correlation of 0.99), with some differences for some specific contents such as "AirBellowsGap" and "UpheavalDome". A qualitative analysis shows that the distortion in "UpheavalDome" becomes more visible, due to an increased brightness of the rendering, while for "AirBellowsGap" the opposite happens, i.e., details and blocking artifacts become invisible around the sun region, which is clipped in our DVI+ rendering since its brightness is much higher than for HDRr mode. Examples of the latter phenomena are illustrated in Figure 6.

More details about the differences produced by the two renderings are obtained by performing a one-way analysis of variance, followed by multiple comparison analysis on HDRr and DVI+ MOS's separately. Changes in the results of multiple comparison may reveal significant differences in the relative perceived quality levels of the stimuli with the two renderings. The results of multiple comparison analysis are reported in Figure 7, where the two binary matrices have been obtained by comparing all the pairs of MOS's in each dataset, and applying the Tukey's honestly significant difference criterion. A black entry in the matrix indicates that no statistical evidence that the corresponding pair of MOS values are significantly different has been found. In both Figure 7(a) and (b), we can observe that stimuli are grouped around five clusters, which correspond approximately to the five adjectives of the quality scale (reported for convenience in the figure).

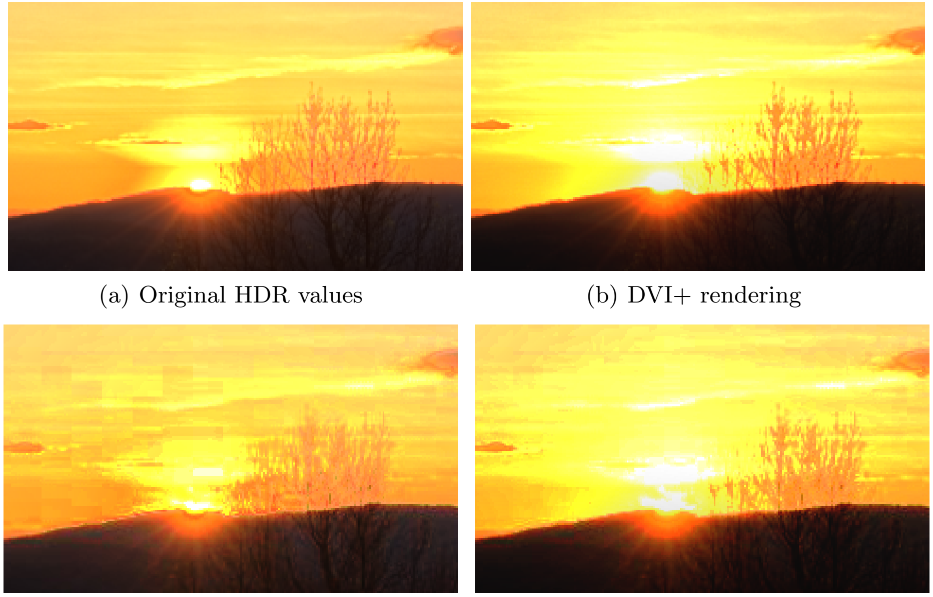
A qualitative evaluation of Figure 7 suggests that the clustering of stimuli MOS's does not change significantly with the two rendering modes. In the highest quality levels, i.e., "Perceptible" and "Imperceptible", though, the results are more intertwined. Considering only these two adjectives (i.e., 190 pairs), there are only 26 pairs



(a) “AirBellowsGap” (b) “LasVegasStore” (c) “MasonLake(1)”

(d) “RedwoodSunset” (e) “UpheavalDome”

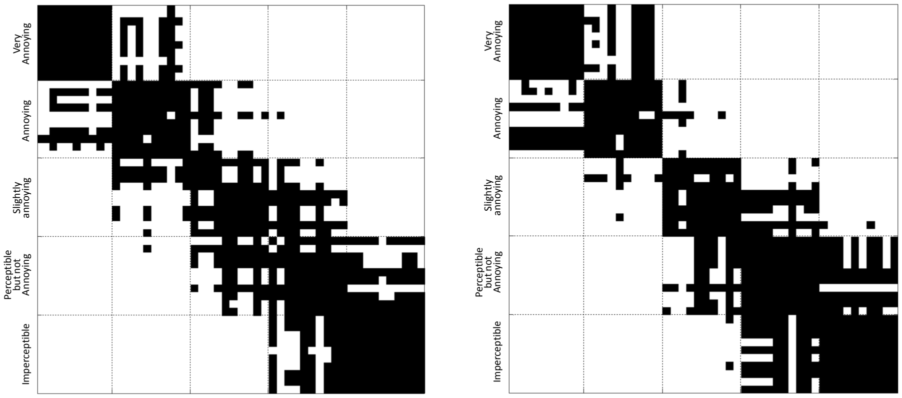
Figure 5. Changes in the Mean Opinion Score by different renderings for the tested contents.



(a) Original HDR values (b) DVI+ rendering

(c) Stimulus no.8, HDR values (d) Stimulus no.8, DVI+ rendering

Figure 6. A detail of the “AirBellowsGap” content showing clipping effects on small and very bright regions. Stimulus number 8 corresponds to JPEG compression with a quality factor of 90. (a) and (c) show the original and compressed HDR values as stored in the HDR file. (b) and (d) are the output of DVI+ rendering. Here the clipping artifacts overcome compression artifacts, i.e., the latter become invisible and thus the MOS of this stimulus is significantly higher with DVI+ rendering. Images are tone-mapped for visualization purposes.



(a) HDRr mode (b) DVI+ mode

Figure 7. Multiple comparison results for MOS of subjective experiments with different renderings. Each of the 50 rows/columns in each matrix corresponds to a pair of MOS's. For convenience, stimuli are grouped according to their adjectives, as found in the test material selection procedure.[7]

of stimuli whose quality appears to be significantly different with DVI+. For HDRr, this number grows up to 41. As overall the proportion of significantly different pairs of stimuli is the same in HDRr and DVI+ mode, this suggests that with DVI+ subtle details become less visible at higher quality levels, i.e., displaying artifacts overcome compression artifacts. Conversely, the higher brightness and local contrast offered by DVI+ rendering make distortion differences more visible at lower quality levels, with respect to HDRr mode.

4. IMPACT ON OBJECTIVE EVALUATION

The techniques for measuring HDR image quality can be broadly divided into two classes. On one hand, metrics such as the HDR-VDP[3] accurately model visual perception in such a way to predict and quantify significant visual differences between images. On the other hand, many quality metrics commonly used in the case of LDR imaging directly assume that input values are *perceptually linear* in order to compute meaningful operations on pixels. The perceptual linearization is implicitly done for the case of LDR images by the gamma encoding of sRGB. In the case of HDR signals, a typical mapping function is the perceptually uniform (PU) encoding.[12] Both HDR-VDP and PU-metrics (metrics computed on PU-encoded values) require as input photometric values of the displayed images. Generally speaking, these values can be estimated by the display rendering algorithm. In practice, using the HDRr mode of SIM2 HDR47 display, the displayed luminance is not known. Therefore, in our previous work,[7] displayed luminance values have been estimated assuming a simple linear response of the display, with saturation at the maximum display luminance, i.e., $L_{out} = \max(L_{in}, L_{max})$, where L_{in} is the target luminance to display [†], and $L_{max} = 4500\,cd/m^2$. However, the results of Section 2.2 suggest that this linear model might be quite inaccurate to describe HDRr rendering, and this could hinder the computation of objective metrics.

An advantage of the DVI+ rendering algorithm described in Section 2.1 is that it can accurately estimate per pixel displayed luminance, that can be fed as input to HDR quality metrics. In this section we compare the performance of several objective metrics when their input is provided by either a simple linear model of the display, or by a sophisticated estimate obtained through the knowledge of rendering algorithm. Specifically, we compute the predictions of six quality metrics computed on either L_{out} values (denoted as "Linear" in the following) or on our DVI+ estimate, and correlate them with the MOS scores obtained from the subjective experiment discussed in Section 3. Considered full-reference metrics include the peak signal to noise ratio (PSNR), the structural similarity index (SSIM)[20] and its multi-scale version,[21] the information fidelity criterion (IFC),[22] the visual information fidelity (VIF),[23] and the HDR-VDP 2.2.[3] The source code for the objective quality metrics

[†] The values L_{in} may depend in fact from the format of input HDR images, e.g., conventionally .hdr files store pixel values v_i such that $L_{in} = 179 \cdot v_i$, while for .exr $L_{in} = v_i$.

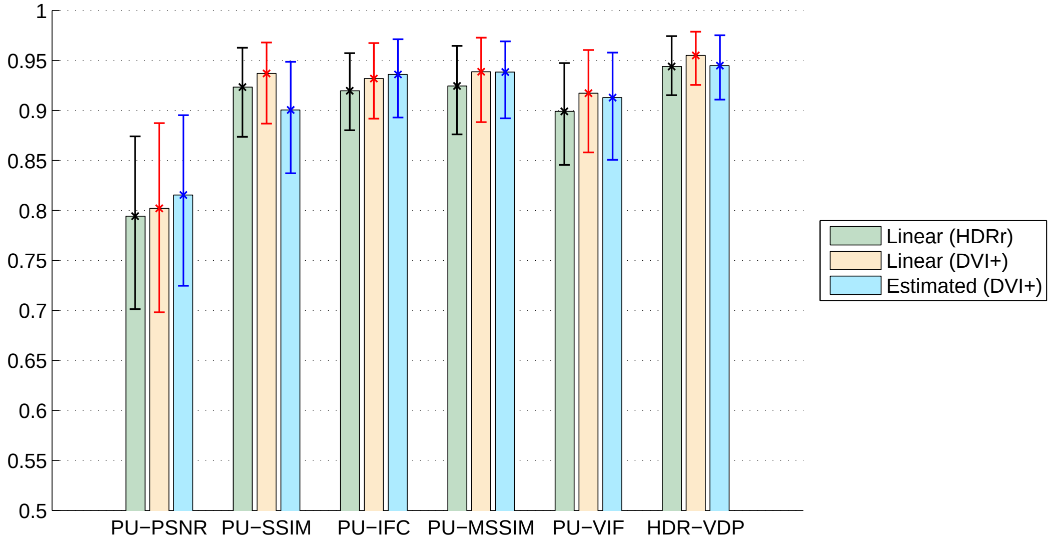


Figure 8. SROCC with 95% confidence intervals for three scenarios: i) displayed luminance computed with Linear model and MOS's collected with HDRr mode;[7] ii) displayed luminance computed with Linear model and MOS's collected with our DVI+; iii) displayed luminance estimated by DVI+ rendering and MOS' collected with our DVI+.

is taken from `http://sourceforge.net/projects/hdrvdp/files/hdrvdp/` for HDR-VDP-2.2, and from `http://foulard.ece.cornell.edu/gaubatz/metrix_mux/` for other objective metrics. All the metrics except HDR-VDP are computed on PU encoded values.[12]

Due to the limited size of the dataset, we evaluate the performance of metric predictions using a non-parametric index such as the Spearman rank-order correlation coefficient (SROCC), which measures the degree of monotonicity of MOS estimates. In addition to SROCC values, we also compute confidence intervals of the correlation coefficients using bootstrap (`bootci` Matlab function with bias-corrected accelerated percentile method, 2000 bootstrap repetitions). Figure 8 reports the SROCC values with their 95% confidence intervals for the linear model and our DVI+ estimate. We also report for comparison the results of our previous experiment,[7] i.e., the SROCC between metrics computed using the linear model and MOS's obtained with HDRr rendering. As Figure 8 illustrates, the three sets of correlations are very close to each other, and there is no clear gain in using more accurate luminance as input to HDR metrics. To confirm this observation, we tested the significance of the difference of SROCC's for each metric, using the method for comparing dependent [‡] correlation coefficients proposed by Zou.[24] This method constructs a confidence interval for the difference of the correlation coefficients. If this interval contains zero the null hypothesis that the two correlations are equal must be retained. Based on this test, we found the two following results: i) the linear model to compute displayed luminance gives statistically indistinguishable performance for two different renderings (HDRr and DVI+, respectively); ii) an accurate knowledge of displayed luminance (with DVI+ rendering) does not significantly increase the performance of objective metrics with respect to the linear model – in fact, for the case of PU-SSIM the SROCC coefficients decreases significantly (although PU-SSIM takes values very close to one, which makes difficult to understand the discriminability of this metric in practice for HDR).

This result is quite surprising, as it contradicts somehow the assumptions of many HDR quality metrics, which compute fidelity using displayed physical luminance as input. A possible explanation for this phenomenon is that, despite the differences between HDRr and DVI+ rendering, the reproduced outputs are highly correlated, as the MOS analysis of Section 3 shows. Furthermore, the saturation in the linear model reduces the effect of outliers in the scene-referred HDR, and contributes significantly to improve its performance. On the other hand,

[‡] The dependency of the two correlations is apparent, due to the fact that they are computed on the same dataset of images. In addition, the correlations of Linear (DVI+) and Estimated (DVI+) are overlapping, since they are computed against the same MOS values.

the DVI+ rendering has a globally linear behavior for the majority of rendered pixels – clipped regions are limited to highlights such as the sun in Figure 6, but the saturation in the linear model actually produces a very similar result. This justifies the effectiveness of the linear model for the DVI+ rendering. Finally, there is one important caveat to take into account. The results presented here are valid for a very specific, although popular, processing task, i.e., HDR image compression, and it is known that for simple additive distortion even simple arithmetic metrics such as the PSNR perform quite well.[25] The generalization of this to more complex types of distortions is left to future work.

5. CONCLUSIONS

In this paper, we have analyzed the impact of a different display rendering on both subjective and objective quality assessment of compressed HDR images. For this purpose, we have developed a simple iterative rendering algorithm for the widely used SIM2 HDR47 display, which yields higher brightness and contrast than the built-in HDRr visualization tool. In addition, our DVI+ rendering can estimate accurately the true pixel-wise luminance of displayed images.

Using this rendering, we have conducted a subjective study to complement our previously proposed HDRr dataset. Test conditions are kept as similar as possible to single out the differences in mean opinion scores due to the only varying factor, i.e., visualization. Our results show that, overall, MOS's are not dramatically impacted by the employed rendering, although in some cases small and localized compression artifacts might become invisible due to rendering artifacts. At the same time, distortion may become more visible in darker or uniform regions, due to increased brightness.

From the point of view of objective quality metrics, our experiments do not bring enough evidence to support the hypothesis that giving accurate estimates of displayed luminance in input to HDR image quality metrics does bring significant advantages or changes over using a simple linear model of the display response. This result has important practical implications, since it suggests that HDR quality estimation can be performed with only a rough knowledge of the characteristics of the reproduction device.

Finally, these results are valid for the assessment of image compression. A much more interesting and growing scenario is that of HDR video quality assessment, where temporal masking plays a key role, and where also DVI+ rendering is much more challenging due to aspects such as flickering. This is matter of our current and future work.

REFERENCES

[1] Seetzen, H., Heidrich, W., Stuerzlinger, W., Ward, G., Whitehead, L., Trentacoste, M., Ghosh, A., and Vorozcovs, A., "High dynamic range display systems," in [*ACM SIGGRAPH 2004 Papers*], *SIGGRAPH '04*, 760–768, ACM, New York, NY, USA (2004).

[2] 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).

[3] Narwaria, M., Mantiuk, R. K., Da Silva, M. P., and Le Callet, P., "Hdr-vdp-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images," *Journal of Electronic Imaging* **24**(1), 010501–010501 (2015).

[4] Narwaria, M., Da Silva, M. P., and Le Callet, P., "Hdr-vqm: An objective quality measure for high dynamic range video," *Signal Processing: Image Communication* **35**, 46–60 (2015).

[5] Aydin, T. O., Mantiuk, R., Myszkowski, K., and Seidel, H.-P., "Dynamic range independent image quality assessment," in [*ACM Transactions on Graphics (TOG)*], **27**(3), 69, ACM (2008).

[6] Hanhart, P., Korshunov, P., Ebrahimi, T., Thomas, Y., and Hoffmann, H., "Subjective quality evaluation of high dynamic range video and display for future tv," (2014).

[7] Valenzise, G., De Simone, F., Lauga, P., and Dufaux, F., "Performance evaluation of objective quality metrics for hdr image compression," in [*SPIE Optical Engineering+ Applications*], 92170C–92170C, International Society for Optics and Photonics (2014).

[8] Burini, N., Mantel, C., Nadernejad, E., Korhonen, J., Forchhammer, S., and Pedersen, J., "Block-based gradient descent for local backlight dimming and flicker reduction," *Display Technology, Journal of* **10**, 71–79 (Jan 2014).

[9] Korhonen, J., Mantel, C., Burini, N., and Forchhammer, S., "Modeling the color image and video quality on liquid crystal displays with backlight dimming," in [*Visual Communications and Image Processing (VCIP), 2013*], 1–6, IEEE (2013).

[10] Mantel, C., Burini, N., Korhonen, J., Nadernejad, E., and Forchhammer, S., "Quality assessment of images displayed on lcd screen with local backlight dimming.," in [*QoMEX*], 48–49 (2013).

[11] Akyüz, A. O., Fleming, R., Riecke, B. E., Reinhard, E., and Bülthoff, H. H., "Do hdr displays support ldr content?: A psychophysical evaluation," in [*ACM SIGGRAPH 2007 Papers*], *SIGGRAPH '07*, ACM, New York, NY, USA (2007).

[12] 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).

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

[14] Luthra, A., Francois, E., and W., H., "Call for evidence (CfE) for HDR and WCG video coding," in [*ISO/IEC JTC1/SC29/WG11 MPEG2014/N15083*], (Feb. 2015).

[15] Burini, N., Nadernejad, E., Korhonen, J., Forchhammer, S., and Wu, X., "Modeling power-constrained optimal backlight dimmingfor color displays," *J. Display Technol.* **9**, 656–665 (Aug 2013).

[16] Mantel, C., Burini, N., Nadernejad, E., Korhonen, J., Forchhammer, S., and Pedersen, J., "Controlling power consumption for displays with backlight dimming," *Display Technology, Journal of* **9**, 933–941 (Dec 2013).

[17] Mantiuk, R., Daly, S., and Kerofsky, L., "Display adaptive tone mapping," in [*ACM Transactions on Graphics (TOG)*], **27**(3), 68, ACM (2008).

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

[19] 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).

[20] 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).

[21] 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).

[22] Sheikh, H. R., Bovik, A. C., and De Veciana, G., "An information fidelity criterion for image quality assessment using natural scene statistics," *Image Processing, IEEE Transactions on* **14**(12), 2117–2128 (2005).

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

[24] Zou, G. Y., "Toward using confidence intervals to compare correlations.," *Psychological methods* **12**, 399–413 (Dec. 2007).

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