Quantitative measurement of contrast, texture, color, and noise for digital photography of high dynamic range scenes

Gabriele Facciolo, Gabriel Pacianotto, Martin Renaudin, Clement Viard, Frédéric Guichard
DxOMark Image Labs, 3 rue Nationale, 92100 Boulogne-Billancourt FRANCE

Abstract

Today, most advanced mobile phone cameras integrate multi-image technologies such as high dynamic range (HDR) imaging. The objective of HDR imaging is to overcome some of the limitations imposed by the sensor physics, which limit the performance of small camera sensors used in mobile phones compared to larger sensors used in digital single-lens reflex (DSLR) cameras. In this context, it becomes more and more important to establish new image quality measurement protocols and test scenes that can differentiate the image quality performance of these devices. In this work, we describe image quality measurements for HDR scenes covering local contrast preservation, texture preservation, color consistency, and noise stability. By monitoring these four attributes in both the bright and dark parts of the image, over different dynamic ranges, we benchmarked four leading smartphone cameras using different technologies and contrasted the results with subjective evaluations.

Introduction

Despite the very tight constraint on form factor, the smartphone camera industry has seen big improvements on image quality in the last few years. To comfortably fit in our pockets, smartphone thickness is limited to a few millimeters. This limits the pixel size and the associated full well capacity, which in turn reduces its dynamic range. The use of multi-image technologies is one of the key contributors to the image quality improvement in the last years. It allows to overcome the limitation of small sensors by combining multiple images taken simultaneously (with multiple sensors, multiple cameras) or sequentially (using bracketed exposures, bursts). Multi-image processing enables many computational photography applications including spatial or temporal noise reduction [9], HDR tone mapping [15, 8, 14], motion blur reduction [10, 11], super-resolution, focus stacking, and depth of field manipulation [25], among others.

The creation of a single image from a sequence of images entails several problems related to the motion in the scene or the camera. We refer to [5, 7, 1] and references therein for a review of methods for evaluating and dealing with these issues. In a previous work [1] we explored these artifacts and provided a first approach to evaluating multi-image algorithms. In this work we focus on the evaluation of the tone mapping of HDR scenes. Since the images must be displayed on screens with limited dynamic range, the tone mapping algorithm becomes a critical part of the system [14]. This process is qualitative in nature as it aims at tricking the observer into thinking that the image shown on a low dynamic range medium has actually a high dynamic range [4, 12]. Nevertheless, as we will see below, the quantitative assessment of some attributes is possible and it fits with our perception of the scene.

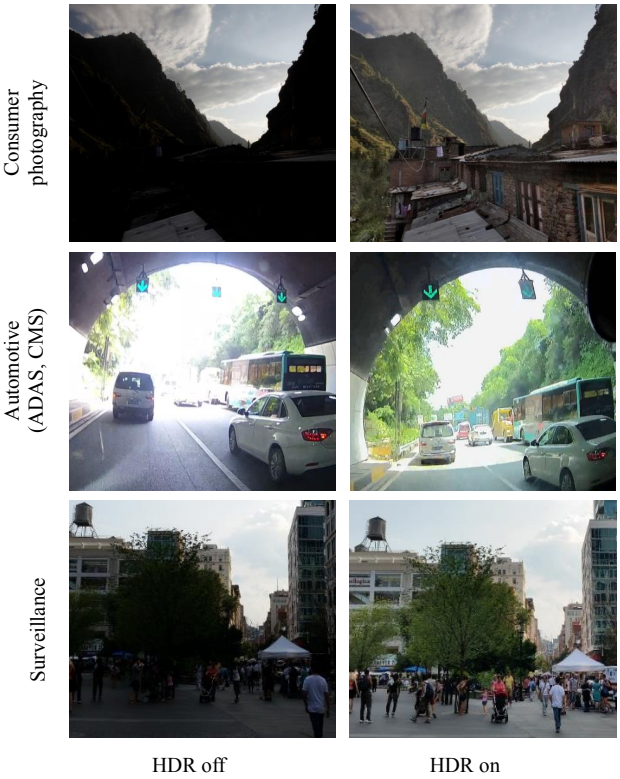


Figure 1. *A typical application of high dynamic range (HDR) imaging for consumer photography is illustrated in the first row. Beyond consumer photography, HDR is also relevant in the context of advanced driver-assistance systems (ADAS) as seen in the second row, and in the context of surveillance as shown in the last row.*

Current image quality measurements are challenged to quantify the performance of these cameras[6, 13, 22] because of the limited dynamic range of the test scenes. Furthermore, the sophistication of algorithms requires more complex test scenes where several image attributes such as local contrast, texture, and color can be measured simultaneously. Beyond consumer photography, image quality in HDR scenes is also very important for other applications such as automotive or surveillance, as illustrated in Figure 1.

The objective of the paper is to present new measurement metrics, test scenes, and a new protocol to assess the quality of digital cameras using HDR technologies. The measurements eval-

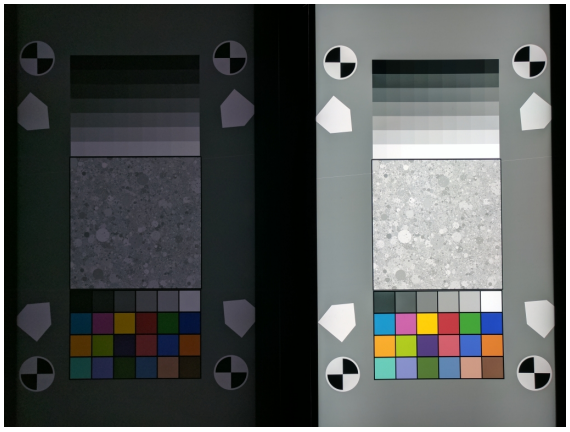


Figure 2. *The proposed setup for measuring the device capacity to capture HDR scenes contains two back-lit targets. Each target contains a color chart, a texture chart, and a grayscale of 63 uniform patches between 0 and 100% of transmittance. The luminance of the light source in the right is always 13000 cd/m^2, while the luminance of the left source varies between 100 and 13000 cd/m^2. The photo is taken with the Device A of our evaluation and the intensity difference corresponds to $\Delta EV = 6$.*

uate local contrast, texture, color consistency, and noise in a laboratory setup where the light intensity as well as color temperature can be adjusted to simulate a wide variety of high dynamic range scenes (Figure 2). The proposed measures are evaluated by benchmarking digital cameras with different HDR technologies, and establishing the correlation between these new laboratory objective measures and human perception of image quality on natural scenes (Figure 3).

The novelty of this approach is a laboratory setup that allows to create a reproducible high dynamic range scene with the use of two programmable light panels and printed transparent charts. The two light panels allow to measure and trace the gain in contrast and color attained by the multi-imaging technology on scenes with a dynamic range that is increased through predefined stops. Also, the measurements through the proposed setup are independent of the content of the scene. The results of this research will be added to the DxOMark Image Labs testing solution, which includes the hardware setup and software necessary for the measurement: a set of programmable light panels to independently and automatically control light intensity and color temperature; a set of transparent chart with specific test patterns used for the automated qualitative analysis of local contrast, color, texture, and noise; and specific algorithms to compute from the shots the quantitative image quality information for the device under test.

In the next section we describe the proposed objective measures of local contrast, texture, color, and noise. We will remind the rationale behind each measure [1] and describe the laboratory setup conceived so as to evaluate these attributes in HDR images. Then we will evaluate the proposed measures by applying them to four devices and validate the results of objective metrics by correlating them with observations on natural images.

Objective HDR measures

High dynamic range imaging aims at reproducing a greater dynamic range of luminosity than is possible with standard digi-

Figure 3. *Comparison of quality attributes observed in natural and laboratory setups. The two photos correspond to different devices observing the same natural scene under the same conditions. The proposed objective measurements are scene independent and allow to study the rendition by the same devices in a controlled laboratory setting. For instance, the textures shown in the bottom-right (called Dead Leaves pattern) are used in the laboratory to evaluate the texture preservation. Note that in the laboratory shots, textures are reproduced similarly to the textures captured in the natural setting (crops in the bottom-left).*

Figure 4. *An important aspect of HDR rendering is perceptual contrast preservation. The pictures illustrate this as (a) is less contrasted and some colors are lost as compared with (b).*

tal imaging techniques. For our HDR laboratory setup, we use the static scene composed of two diffuse and adjustable light sources (Kino Flo LED 201 DMX devices, DMX for short) as proposed in [1]. This allows to precisely adjust the luminous emittance from 5 to 17000 cd/m^2. In front of the DMX devices we placed two identical transparent prints containing a grayscale, a color, and a texture chart. Our final image contains the two DMX devices as it can be seen in Figure 2. The two DMX devices are then programmed. They begin with the same luminous emittance (13000 cd/m^2) and progressively decrease the left one by reducing one EV each time, until the $\Delta EV = 7$. By stretching the intensities of the two DMX devices we intend to create scenes with increasing dynamic range. For each dynamic range setting we acquire a photograph with the HDR setting and automatic exposure.[1]

The characteristics we want to measure are the preservation of local contrast, texture, color consistency, and noise consistency. Simply scaling the high dynamic range of the scene to fit the dynamic range of the display is not good enough to reproduce the visual appearance of the scene [14]. We want to quantify how the device compresses the HDR scene to fit the display range, while preserving details and local contrast, how colors are altered and how noise is handled.

[1] In most devices exposure can be "forced" so that a point of interest is well exposed (by tapping on it).

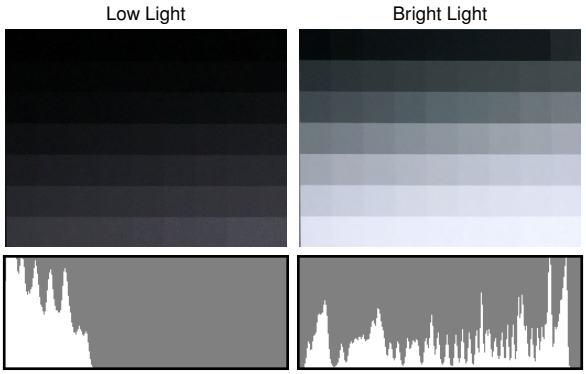


Figure 5. *Local contrast analysis using the entropy. The images show the dark and bright part of the setup (Figure 2) with $\Delta EV = 6$ acquired with device D. The figures correspond to the grayscales, the corresponding normalized histograms, and entropy. Note that a grayscale with many saturated values (left column) have a lower entropy value than an evenly distributed grayscale (right column).*

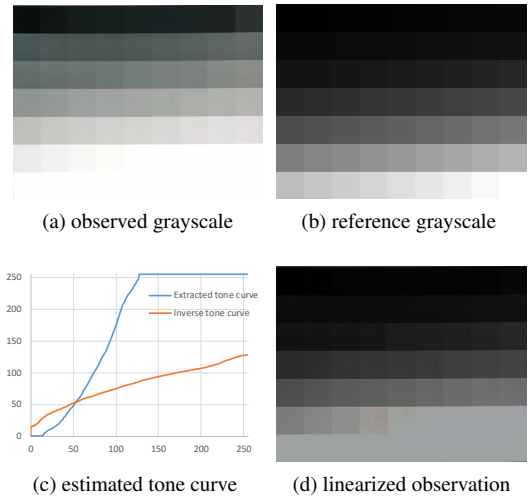
Local contrast preservation. Tone mapping algorithms allow to display an HDR image on a support with limited dynamic range. The local contrast perception is an important part of a good HDR image, as illustrated in Figure 4. The tone mapping algorithm must produce a pleasant rendering of the image while preserving low contrast details [3]. This is usually done by local contrast adaptations, which are inspired on perceptual principles [4] (i.e. humans do not perceive absolute intensities but rather local contrast changes).

Our measure uses the grayscale part of the charts in Figure 2, which is composed of 63 uniform patches with linearly increasing transmission. Having two grayscales with two different luminance on the same scene allows to measure how a device preserves the local dynamic range of each part of the image. To measure the dynamic range we adopt the metric proposed in [1], which computes the entropy of the normalized histogram $hist_{gs}$ of the grayscale chart

$$Entropy_{gs} = \sum_k hist_{gs}(k) \log \frac{1}{hist_{gs}(k)}. \quad (1)$$

The entropy can be seen as the quantity of information contained in the grayscale chart. A grayscale with many saturated values in the dark or in the bright parts will have an entropy value lower than an evenly distributed grayscale (as illustrated in Figure 5). A grayscale with evenly distributed values will have an entropy equal to the dynamic of the grayscale.The entropy has some clear limitations related to the fact that it does not incorporate spatial information. A dithering grayscale, for instance, can have bad entropy and good visual appearance, and a grayscale with strong halos can have good entropy but bad visual appearance. Nonetheless, in [1] it is shown that the entropy provides a good indicator of the perceived contrast.

In the proposed experimental setup the entropy is measured on each grayscale chart for the different ΔEVs. This will provide information about the contrast trade-offs made by the different tone mapping algorithms.



(a) observed grayscale (b) reference grayscale

(c) estimated tone curve (d) linearized observation

Figure 6. *Tone curve extraction and inversion. Image (a) shows the observed grayscale and (b) the reference one. Matching the patches we estimate the tone curve that maps the reference to the observed one. Then we invert the tone curve avoiding stretching the saturated part (c). The same inverse tone curve is used to linearize the observed Dead Leaves chart. Image (d) illustrates the effect of linearization on the grayscale (a), the non-saturated part should match the reference grayscale.*

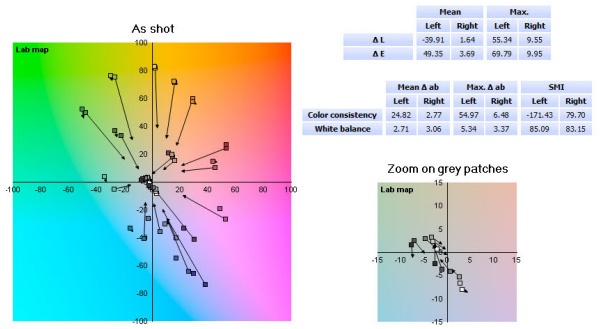
Texture preservation. Preservation of fine details is different from contrast; it is possible to have a locally low contrasted scene with good texture and a locally highly contrasted scene with no texture. The texture preservation measure is designed to evaluate how fine details are preserved after tone mapping and denoising have been applied [16, 17, 18]. The Dead Leaves pattern [16] is used to simulate a texture with natural image properties (see Figure 3), which are hard for post processing to enhance. Let us define the spatial frequency response (SFR) [17, 21] as the measured power spectral density (PSD) of the texture divided by the ideal target power spectral density

$$SFR_{tex}(f) = \sqrt{\frac{PSD_{tex}(f) - PSD_{noise}(f)}{PSD_{ideal}(f)}}, \quad (2)$$

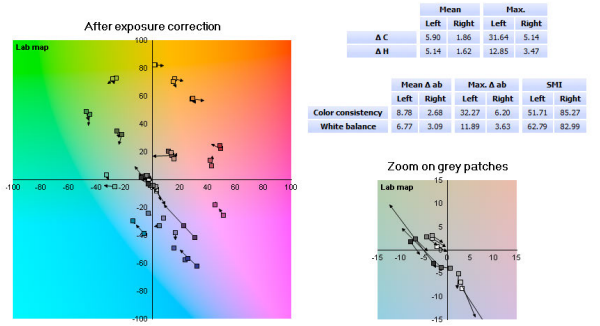
where PSD_{ideal} is the known spectral density of the observed pattern [16], and PSD_{noise} denotes the power spectral density of the noise present in the image, which is measured on uniform patches. Then, the *acutance* metric A is computed. The *acutance* provides a measure of the perceived sharpness of the image and is defined as the weighted average of the texture SFR with the contrast sensitivity function (CSF), which represents the sensitivity of the human visual system to different frequencies

$$A = \int SFR_{tex}(f) CSF(f) df. \quad (3)$$

The acutance gives information about how texture is preserved, however it is contrast dependent. Hence, similarly to [1], a linearization preprocess is applied. The linearization scales the gray levels of the observed image to the levels of the reference chart. Unlike [1] a high resolution tone curve is estimated using the 63 patches of the grayscale (see Figure 2). Then, the inverse tone curve is applied to the Dead Leaves chart. As illustrated in



(a) before exposure correction



(a) after exposure correction

Figure 7. *Color consistency measurement before and after exposure correction. The diagrams illustrate the color difference in the a*,b* plane, for an exposure difference $\Delta EV = 6$ (corresponding to Device A). Without exposure correction the color differences are large because of the nonlinear relations between the luminance and color channels om the CIE L*a*b* color space.*

Figure 6, special care must be taken with the saturation points, in order to avoid singularities in the inversion. In the HDR setup, for each ΔEV the acutance of each chart is computed, which permits to analyze the behavior of the tone mapping algorithm.

It is worth noting that a tone curve would not undo the local adaptation effects of HDR tone mapping. This implies that there is no guarantee that the estimated tone curve is valid on the texture. Nevertheless, the perceptual validation confirms that this setup captures the effects of texture loss.

Color consistency. Color consistency can be described as the ability of a camera to preserve colors at different exposures and at different intensities within the same image (Figure 14). Here, we extend the classic color reproduction evaluation methods [20, 19] to HDR images. Each chart in Figure 2 contains a set of 24 representative colors, inspired by the *Macbeth ColorChecker*.

The classic approach to measuring color consistency consists in capturing charts with calibrated spectral responses under known illumination. However, since the repeatability of the illumination and print properties of the back-lit HDR setup is not as good as that of the ColorChecker, we recommend to measure color consistency with respect to a reference shot of the same chart, which is acquired with $\Delta EV = 0$.

Color consistency is a single value metric aimed at measuring the capacity of the device to reproduce the same color between two photos, especially between a low dynamic scene and a high dynamic one. To compare colors between images having a different contrast we propose to first apply an exposure correction and then compare the corrected values in the CIE L*a*b* color space. The exposure correction must be done using linear coordinates, this is because (in order to mimic the nonlinear response of the eye) in the in CIE L*a*b* the luminance and chrominance channels are nonlinearly related.

Let us suppose we want to compute the color consistency between two photos, a sample S and a reference R. On each photo, we have a set of uniform color patches. We also know the theoretical color value of those patches expressed in the CIE 1931 XYZ color space. For correcting the exposure we first convert the photos to the CIE 1931 XYZ color space and compute the mean value of each patch (X,Y,Z) on this color space. The exposure correction is done by imposing the luminance of the theoretical patch (X_r,Y_r,Z_r) on the measured patch as:

$$(X'_S,Y'_S,Z'_S) = \frac{Y_r}{Y_S}(X_S,Y_S,Z_S). \tag{4}$$

The impact of the exposure correction is illustrated in Figure 7.

After the exposure correction, we convert the values (X'_R,Y'_R,Z'_R) and (X'_S,Y'_S,Z'_S) to the CIE L*a*b* color space. For each patch we then compute the distance Δab as given by the following formula:

$$\Delta ab = \sqrt{(a^*_S - a^*_R)^2 + (b^*_S - b^*_R)^2}. \tag{5}$$

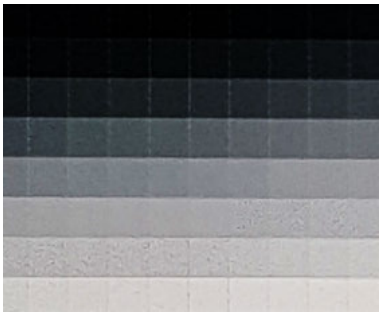
Noise analysis. Noise analysis is particularly interesting in HDR imaging because the multi-image algorithms may end up mixing inconsistent levels of noise in the same image. This can happen when a multi-image fusion algorithm stitches images with incoherent noise as seen in Figure 8(a). It is important to analyze this noise artifact because this incoherence can be interpreted as the presence of texture.

In Figure 8(b) we show the dark part of the sample image, which not only has high levels of noise, but the noise level is also discontinuous. This incoherent noise can be seen between the fifth and sixth lines of the grayscale image, which was the one that originated the curve shown in Figure 8(b). The plot also show the noise levels corresponding to a $\Delta EV = 0$. This differential analysis permits to study the stability of the image quality as the dynamic range is stretched.

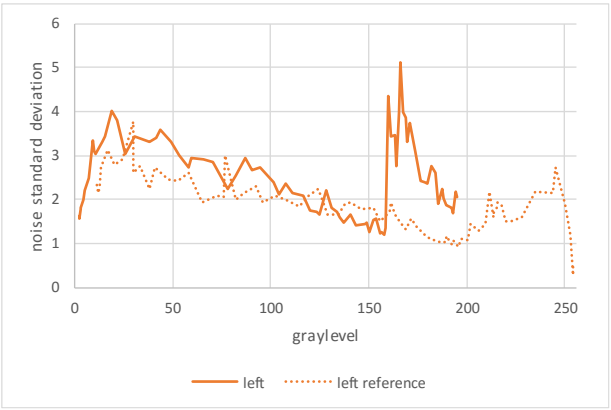
Another important aspect of the noise analysis is the apparent noise level. For that we analyze the evolution of the *visual noise* (defined in ISO 15739) for increasing dynamic ranges. The *visual noise* is a metric that measures noise as perceived by end-user. Prior to computing the visual noise the image is converted to the CIE L*a*b* color space and it is filtered (in the frequency domain) to take into account the sensitivity of the human visual system to different spatial frequencies under the current viewing conditions. Then the visual noise is computed [24] as the base-10 logarithm of the weighted sum of the noise variance estimated on the CIE L*a*b* channels of the filtered image u

$$K \log_{10}[1 + \sigma^2(u_{L*}) + \sigma^2(u_{a*})\sigma^2(u_{b*})]. \tag{6}$$

The noise variances are computed over large uniform areas of the image with known graylevels. We sample seven different



(a) An example of noise artifact due to the HDR stitching.



(b) Noise consistency plot computed on the HDR chart.

Figure 8. *Noise artifact due to the HDR stitching. Notice in (a) the rupture of noise consistency in the 6th and 7th rows. In the plot (b) we can see that the estimated standard deviation of noise in the sample image (which corresponds to the left side of the mire, shown in image (a)) not only has elevated levels of noise, but it also presents discontinuities. The plot also shows the noise level of the reference image, which is acquired with both light panels at the same intensity. This image corresponds to the dark side of the setup with a $\Delta EV = 3$, acquired with the Device B of our evaluation.*

graylevels: the six gray patches present on the ColorChecker, plus the background of the chart. The visual noise for other intensity levels is linearly interpolated from the samples.

Evaluation of four devices

Our final objective is to develop a single metric that quantifies the system performance to simplify comparisons between devices. In this paper we compare the devices using the individual metrics, which will eventually be combined into a single one.

For that purpose, the laboratory setup and the metrics presented above are evaluated by comparing four devices launched between 2014 and 2016. We denote the devices with a letter from A to D, where A is the more recent and D is the oldest one. The interest of comparing these devices is that they permit to observe the evolution of the HDR technology over time. In the next section we will also perform a subjective validation for two of the proposed measures.

Contrast preservation measure. We evaluate the contrast preservation of a device by computing, for different ΔEV, the entropy of the two grayscales in the laboratory setup shown in Figure 2. The results for the four devices considered in the evaluation

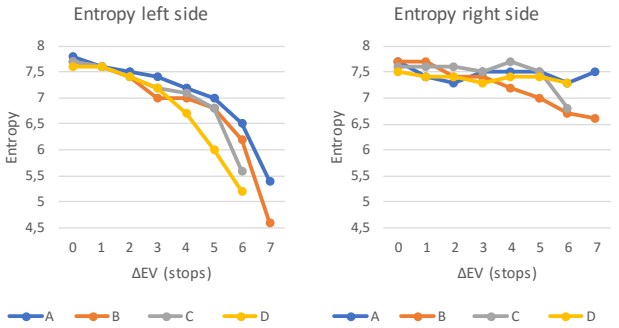


Figure 9. *Contrast preservation measures of four devices in the laboratory setting. The plots show the measured entropy in the dark part (left side) and bright part (right side) of the setup (Figure 2) for increasing ΔEV.*

ΔEV	A	B	C	D
0	7	7	7	7
1	7	7	7	7
2	7	7	7	7
3	7	7	7	7
4	7	7	7	6.85
5	7	6.9	6.9	6.5
6	6.75	6.45	6.2	6.1
7	6.2	5.6	5.45	5.5
SUM ΔEV 4 to 7	26.95	25.95	25.55	24.95

Figure 10. *Aggregated contrast preservation measures of four devices. The table shows the average entropy (from Figure 9 thresholded at 7) for all the ΔEV. We see a strong loss of contrast in the dark part of the setup as ΔEV increases.*

are shown in Figure 9. A high entropy means that the different values of the grayscale are well represented by the device. We note that all the devices tend to preserve the bright part of the scene (right side) and sacrifice the dark part as the ΔEV increases. For all the considered devices these losses correspond to saturation of dark or bright areas.

Taking into account that an entropy above 7 is not perceptually relevant [1] we conclude that, on the bright side of the setup (right) all the tested devices have a similar behavior and we observe that device B has a the tendency of saturating for large ΔEV. From the left side of the setup we see that older devices (from D to A) have a worse contrast preservation, as their entropy curves decline faster for larger ΔEV.

These measures are interesting per-se and could be used to compare against a reference photo taken with $\Delta EV = 0$, or with respect to a reference device. However, to obtain an overall score we must combine the scores on the left and right parts of the setup. We propose to start by thresholding the entropy at 7, then average the thresholded entropies on the two sides to obtain a single score for each ΔEV. Since the entropy is a concave function of the measured dynamic range, averaging the two entropies allows to penalize the case in which just one of the sides is well contrasted, while the other one is poorly contrasted.

The overall score for a device can then be obtained by aggregating the scores for all the considered ΔEV. The aggregated results for the four devices are shown in Figure 10. We can observe that the score improves for more recent devices (from D to

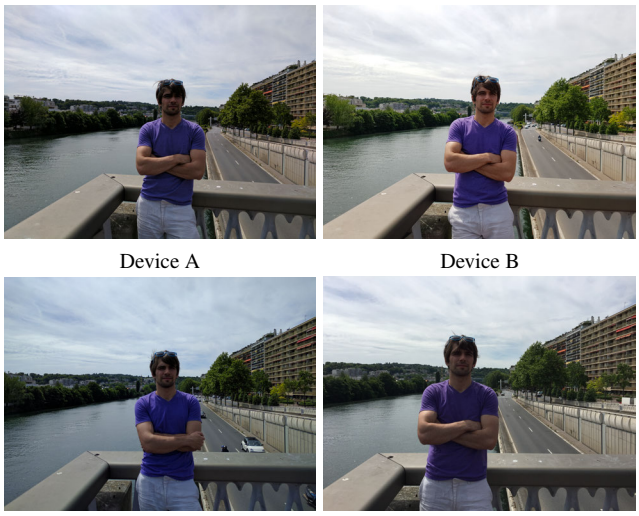


Figure 11. *Comparison of contrast preservation of the four devices. Note that device B seems to be more contrasted than A, despite having a slightly lower score in Figure 10, This is because device B saturates the high and low levels of the image, while device A preserves them, as can be seen in the clouds details.*

A), which is evidence of the improvement of the tone mapping technology over time.

Figure 11 shows an HDR scene captured with the four devices. It is interesting to observe that the image corresponding to device B, despite having a slightly lower score than device A, seems to be more contrasted. This is because device B saturates the high and low levels of the image, while device A preserves them, as can be seen in the cloud details. The perceptual validation conducted in the next section also confirms that the observers indeed prefer device B over device A. It is worth noticing that, this slight saturation of the bright part of the scene for device B could be identified in the laboratory measurement (Figure 9) as the decrease in entropy in the bright part of the setup.

Texture preservation measure. The acutance measurements for all the devices for different ΔEV are summarized in Figure 12. We start by observing that, while a good acutance should be between 0.8 and 1, device B has an acutance larger than 1. This behavior is due to an over-sharpening of the output, which increases the measure but does not produce pleasant results. We shall see in the perceptual validation that indeed, the sharpening is not mistaken as texture by the users.

Devices A and C perform similarly for all the ΔEV, while device D is systematically below them by 0.2 points. We also observe that for large ΔEV all the devices loose texture on the dark part of the setup, which is consistent with the saturation of the dark levels. From these measures we can conclude that devices A and C have the best texture preservation, followed by D and B.

The result of the subjective evaluation presented in the next section confirm the conclusions we reached by analyzing the laboratory measurements of Figure 12.

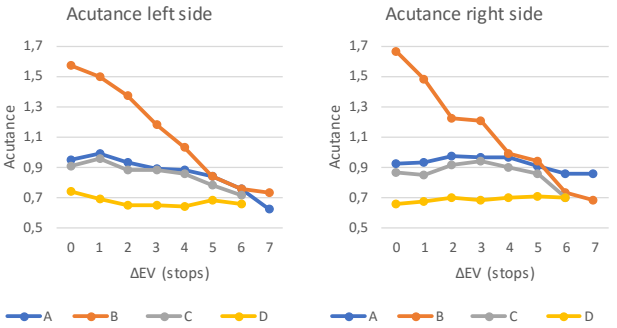


Figure 12. *Texture preservation measures of four devices in the laboratory setting. The plots show the measured acutance in the left and right side of the setup (Figure 2) for increasing ΔEV. We observe that for large ΔEV all the devices loose acutance on the dark part of the setup. A good acutance should be between 0.8 and 1, the acutance above 1 of Device B is due to an over-sharpening of the result, which implies that textures are not well preserved. The best results are obtained by devices A and C, which perform similarly for all the ΔEV on both sides, while device D is systematically below them by 0.2 points.*

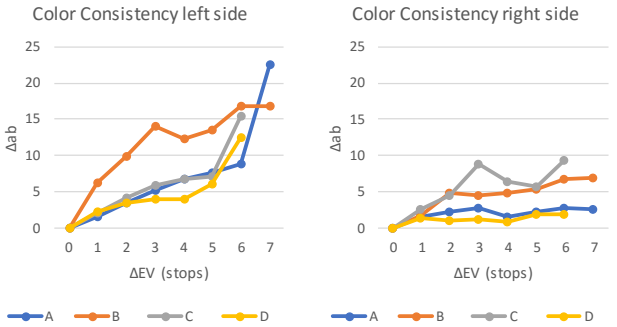
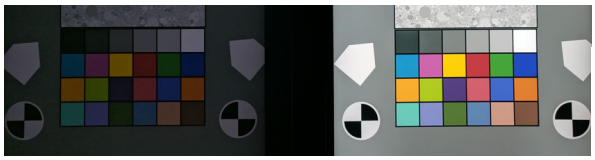


Figure 13. *Color preservation results for the four devices in the laboratory setting. The plots present, for different ΔEVs, the average Δab (Equation 5 averaged over all the ColorChecker patches) computed with respect to a reference image acquired with $\Delta EV = 0$. We see from the graph, that color consistency deviates strongly on the dark of the setup when ΔEV increases, while colors are more consistent on the bright part.*

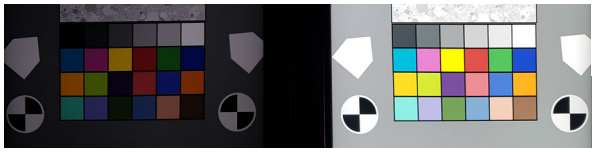
Color consistency measure. For a given device and ΔEV, we propose to measure the color consistency as the average of Δab (Equation 5) computed with respect to a reference image acquired with $\Delta EV = 0$. The average is computed over all the ColorChecker patches. This measure yields two average Δab per shot, one for each side of the setup.

The results of this evaluation are summarized in Figure 13. From the plot corresponding to the bright part of the setup we can see that, as ΔEV increases, devices B and C become less consistent, while devices A and D are better at preserving the colors for all the exposures. However, these differences are not perceptually relevant, as a $\Delta ab < 8$ is barely noticeable.

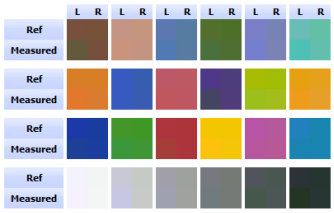
In the dark part of the scene, on the other hand, we observe larger differences. Devices A, C, and D perform similarly up to $\Delta EV = 5$, with color differences below the barely noticeable threshold, for larger ΔEV the errors of all devices rise because of saturation. For Device B however, we observe much higher errors



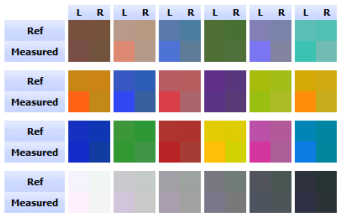
(a) Device A



(a) Device B



(c) Device A: reference vs. measured patch comparison table



(d) Device B: reference vs. measured patch comparison table

Figure 14. *Color consistency evaluation of devices A and B for $\Delta EV = 6$. The images (a) and (b) are crops (for each device) of the laboratory shots with $\Delta EV = 6$. The color consistency plotted in Figure 13 is computed with respect to a reference image taken with $\Delta EV = 0$ (not shown here). The color comparison tables (c) and (d) show (in 2×2 grids) the exposure corrected patches from the left (L) and right (R) side for the Reference and Measured images. Note that while both images (a) and (b) have the same exposure, the colors reproduced by device B are less consistent as seen in the table (d) and in the image (b), particularly the orange and yellow patches.*

even for small ΔEV. To illustrate the impact of a large Δab we show in Figure 14 the ColorChecker for the shots of devices A and B with $\Delta EV = 6$. Note that while both images have the same exposure, the colors reproduced by device B are less consistent as seen in the corresponding comparison table and in the image (particularly visible in the orange and yellow patches).

In conclusion the best color consistency across ΔEV is attained by devices A and D, followed closely by device C, and then device B.

Noise analysis. Figure 15 shows (for the four devices) the evolution of the visual noise computed for a value L*=50 (CIE L*a*b*) with an increasing ΔEV. This measure is proportional to the perceived noise, a visual noise below 1 is not visible in bright light conditions (above 300lux), and below 3 is not visible in low light conditions. The two plots correspond to each side of the setup (low light and bright light). The low light conditions (below 300lux) are only attained on left side for ΔEV 6 and 7.

On the bright side of the setup all the devices remain within a visual noise of 2, with devices B and D strictly below 1. On the dark part of the setup, for all the devices except B, we see a strong increase of visual noise as ΔEV increases. Device B maintains a low visual noise, at the expense of the textures, by applying a stronger denosing.

In Figure 14(a,b) we can compare the images corresponding to $\Delta EV = 6$, for the devices A and B. We can easily see that device A (with a visual noise of 5) is indeed noisier than the image produced by device B (which is strongly denoised). A visual noise below 6 is not necessarily bad, and may even be a design choice. Visual noise levels above 6, on the other hand, are more disturbing. In conclusion, except for device B (which applies a strong denoisng), device A has the lowest visual noise followed by devices C and D.

Perceptual validation of texture and contrast measures

To validate the results of the texture and contrast measures we conducted a subjective evaluation. For our evaluation, six natural HDR scenes were shot with the four devices in auto exposure

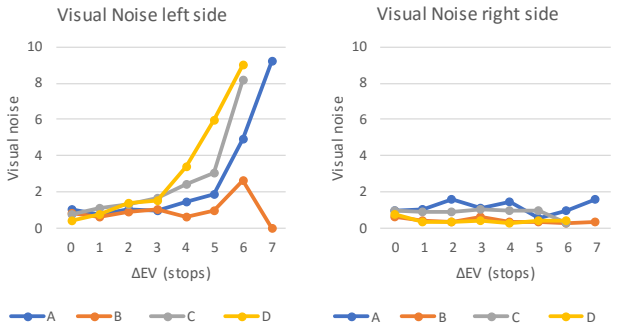


Figure 15. *Visual Noise for a value L*=50 (CIE L*a*b*) for an increasing ΔEV for the four devices in the evaluation. This measure is proportional to the perceived noise, a visual noise below 1 is not visible in bright light conditions (above 300lux), and below 3 is not visible in low light conditions. The two plots correspond to each side of the setup (dark and bright). For all devices, except B, we see a strong increase of visual noise on the dark part as ΔEV increases. Device B which applies a strong denoising to the results.*

Figure 16. *The six scenes used in the subjective evaluation of the contrast preservation, and from which the crops for evaluating the texture preservation (Figure 17) are extracted.*

(a) Device A (b) Device B

Figure 17. *Subjective evaluation of HDR texture preservation. The subfigures show the six crops of textured parts of the images in Figure 16, for two devices.*

mode. These scenes (shown in Figure 16) were acquired on a cloudy day and had a dynamic range of around 7 to 8 stops. We define the dynamic range of a scene as the exposure difference between a picture well-exposed on the brightest part of the scene and a picture well-exposed on the darkest part. We measure this by bracketing the scene with a DSLR, increasing the exposure by 1 stop in each image.

Fifteen subjects participated in the subjective evaluation. The evaluation used a two-alternative forced choice method (described below) which presents the observers with two images and asks to rank them. For the evaluation of the contrast preservation measure we present the subjects with the entire images (Figure 16) and ask the observer to choose the image with better contrast. For the evaluation of the texture preservation measure we present pairs of crops containing preselected textured parts of the scenes (shown in Figure 17) and ask the observer to choose the image in which the texture is best preserved.

Two-alternative forced choice evaluation. For the subjective evaluation of the texture and contrast measures we used a forced-choice method. In [2] the authors compared different perceptual quality assessment methods and their potential in ranking computer graphics algorithms. The forced-choice pairwise comparison method was found to be the most accurate from the tested methods.

In forced choice, the observers are shown a pair of images (of the same scene) corresponding to different devices and asked to indicate an image that better preserves texture (or contrast). Observers are always forced to choose one image, even if they see no difference between them (hence the forced-choice name). There is no time limit or minimum time to make the choice. The ranking is then given by n_S, the number of times one algorithm is preferred to others assuming that all pairs are compared. The ranking score is normalized $\hat{p} = n_S/n$ by dividing with the number of tests containing the algorithm n. So that $\hat{p}$ can be associated to a probability of choosing a given algorithm.

By modeling the forced-choice as a binomial distribution we can compute confidence interval of the ranking score $\hat{p}$ using the formula

$$\hat{p} \pm z\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})}, \tag{7}$$

where z is the target quantile. This formula is justified by the central limit theorem. However, the central limit theorem applies poorly to this distribution when the sample size is less than 30, or when the proportion $\hat{p}$ is close to 0 or 1. For this reason we adopt the Wilson interval [23]

$$\frac{1}{1+\frac{1}{n}z^2}\left[\hat{p}+\frac{1}{2n}z^2 \pm \sqrt{\frac{1}{n}\hat{p}(1-\hat{p})+\frac{1}{4n^2}z^2}\right], \tag{8}$$

which has good properties even for a small number of trials and/or an extreme probability.

Results and analysis. Figure 18 summarizes the results of subjective evaluation for texture preservation. Devices A and C are identified by the subjects as the best performing. This is coherent with the results of the acutance measurements seen in Figure 12, where devices A and C have very similar scores. Moreover, as mentioned above, the observers penalized the over-sharpening introduced by device B placing it slightly below device C.

Figure 19 presents the results of the subjective evaluation of contrast preservation. The subjective evaluation ranks devices B and A as the best performing and then devices C and D. This is coherent (except for the inversion B,A) with the ranking based on the laboratory measurement, shown in Figure 10, which ranks the devices as: A, B, C, and D.

Let us concentrate on the inversion between the objective measurement and the subjective evaluation results for the devices A and B. Closer inspection of this case reveals that the verdict of the contrast measure is correct. Device A better preserves the dynamic range, while the device B tends to saturate the brights and dark areas of the image. However, this saturation is associated to more contrast by the human observers, hence the higher perceptual score. This is a nuanced point that highlights a limitation of the proposed entropy-based measure. Addressing this issue would require a more accurate modeling of human visual system to capture this preference for slightly saturated images.

Conclusion

In this paper we presented a novel laboratory setup that creates a high dynamic reproducible scene with the use of two light

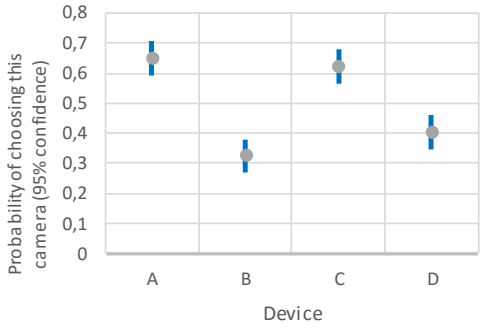


Figure 18. *Subjective evaluation of HDR texture preservation for the four devices. The plot presents the results of the forced choice evaluation of texture preservation. The values represent the probability of an observer to choose the result of a device over the others. We see that, according the human observers, textures are better preserved by devices A and C.*

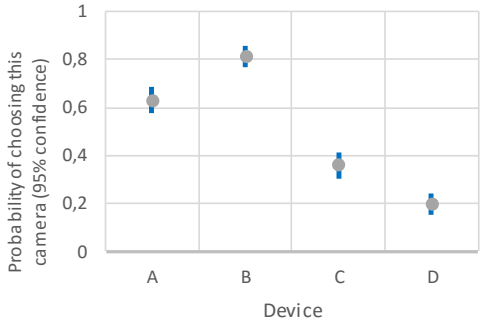


Figure 19. *Subjective evaluation of HDR exposure preservation for the four devices. The plot presents the results of the forced choice evaluation of contrast preservation. The values represent the probability of an observer to choose the result of a device over the others. We see that, according the human observers, contrast is better preserved by devices B and A.*

panels and printed transparent charts. The use of the two programmable light panels allows to measure and trace the gain in contrast, texture, and color from the HDR technology for scenes with a dynamic range getting higher through predefined stops. Improved image quality measures [1] are also proposed, allowing the automated analysis of the test scenes. In addition, the measures obtained with the proposed laboratory setup are independent of the content of the scene. Validation of the measures along with a benchmark of different devices was also presented, highlighting the key findings of the proposed HDR measures.

References

[1] Renaudin, M., Vlachomitrou, A. C., Facciolo, G., Hauser, W., Sommelet, C., Viard, C., and Guichard, F. (2017). Towards a quantitative evaluation of multi-imaging systems. Electronic Imaging, 2017(12), 130-140. 1, 2, 3, 5, 9

[2] Mantiuk, R. K., Tomaszewska, A., and Mantiuk, R. (2012). Comparison of four subjective methods for image quality assessment. In Computer Graphics Forum 31(8) 2478-2491. 8

[3] Mertens, T., Kautz, J., and Van Reeth, F. (2009). Exposure Fusion: A Simple and Practical Alternative to High Dynamic Range Photography. Computer Graphics Forum, 28(1), 161–171. 3

[4] Land, E. H., and McCann, J. J. (1971). Lightness and Retinex Theory. Journal of the Optical Society of America, 61(1), 1–11. 1, 3

[5] Srikantha, A., and Sidibé, D. (2012). Ghost detection and removal for high dynamic range images: Recent advances. Signal Processing: Image Communication, 27(6), 650-662. 1

[6] Chen, Y., and Blum, R. S. (2009). A new automated quality assessment algorithm for image fusion. Image and Vision Computing, 27(10), 1421-1432. 1

[7] Tursun, O. T., Akyüz, A. O., Erdem, A., and Erdem, E. (2015). The state of the art in HDR deghosting: A survey and evaluation. In Computer Graphics Forum (Vol. 34, No. 2, pp. 683-707). 1

[8] Hasinoff, S. W., Durand, F., and Freeman, W. T. (2010). Noise-optimal capture for high dynamic range photography. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on (pp. 553-560). IEEE. 1

[9] Buades, A., Lou, Y., Morel, J. M., and Tang, Z. (2010). Multi image noise estimation and denoising. 1

[10] Hee Park, S., and Levoy, M. (2014). Gyro-based multi-image deconvolution for removing handshake blur. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3366-3373). 1

[11] Delbracio, M., and Sapiro, G. (2015). Burst deblurring: Removing camera shake through fourier burst accumulation. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2385-2393). IEEE. 1

[12] Eagleman, D. M. (2001). Visual illusions and neurobiology. Nature Reviews Neuroscience, 2(12), 920-926. 1

[13] Eilertsen, G., Unger, J., Wanat, R., and Mantiuk, R. (2013). Survey and evaluation of tone mapping operators for HDR video. In ACM SIGGRAPH 2013. 1

[14] Reinhard, E., Heidrich, W., Debevec, P., Pattanaik, S., Ward, G., and Myszkowski, K. (2010). High dynamic range imaging: acquisition, display, and image-based lighting. Morgan Kaufmann. 1, 2

[15] Debevec, P. E., and Malik, J. (2008). Recovering high dynamic range radiance maps from photographs. In ACM SIGGRAPH 2008 classes (p. 31). ACM. 1

[16] Cao, F., Guichard, F., and Hornung, H. (2010). Dead leaves model for measuring texture quality on a digital camera. In Digital Photography (p. 75370). 3

[17] McElvain, J., Campbell, S. P., Miller, J., and Jin, E. W. (2010). Texture-based measurement of spatial frequency response using the dead leaves target: extensions, and application to real camera systems. In IST/SPIE Electronic Imaging (pp. 75370D-75370D). International Society for Optics and Photonics. 3

[18] Kirk, L., Herzer, P., Artmann, U., and Kunz, D. (2014). Description of texture loss using the dead leaves target: current issues and a new intrinsic approach. In IST/SPIE Electronic Imaging (pp. 90230C-90230C). International Society for Optics and Photonics. 3

[19] Cao, F., Guichard, F., and Hornung, H. (2008). Sensor spectral sensitivities, noise measurements, and color sensitivity. In Electronic Imaging 2008 (pp. 68170T-68170T). International Society for Optics and Photonics. 4

[20] Schanda, J. (Ed.). (2007). Colorimetry: understanding the CIE system. John Wiley and Sons. 4

[21] Artmann, U. (2015). Image quality assessment using the dead leaves target: experience with the latest approach and further investigations. In SPIE/IST Electronic Imaging (pp. 94040J-94040J). International Society for Optics and Photonics. 3

[22] Ledda, P., Chalmers, A., Troscianko, T., and Seetzen, H. (2005,

July). Evaluation of tone mapping operators using a high dynamic range display. In ACM Transactions on Graphics (TOG) (Vol. 24, No. 3, pp. 640-648). ACM. 1

[23] Wilson, E. B. Probable Inference, the Law of Succession, and Statistical Inference. J. Am. Stat. Assoc. 22, 209–212 (1927). 8

[24] Kleinmann, J. and Wueller, D. (2007). Investigation of two methods to quantify noise in digital images based on the perception of the human eye. In SPIE/IST Electronic Imaging. International Society for Optics and Photonics. 4

[25] Hauser, W., Neveu, B., Jourdain, J.-B., Viard, C., and Guichard, F. (2018). Image quality benchmark of computational bokeh. Electronic Imaging 2018. 1