Towards a quantitative evaluation of multi-imaging systems

Martin Renaudin, Anna-Cecilia Vlachomitrou, Gabriele Facciolo, Wolf Hauser,Clement Sommelet, Clement Viard, Frédéric Guichard
DxOMark Image Labs, 3 rue Nationale, 92100 Boulogne-Billancourt FRANCE

Abstract

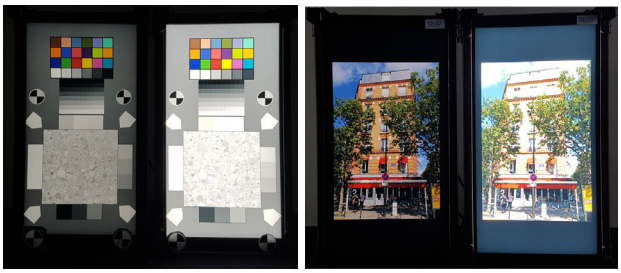
Nowadays many cameras embed multi-imaging (MI) technology without always giving the option to the user to explicitly activate or deactivate it. MI means that they capture multiple images, combine them and give a single final image, letting sometimes this procedure being completely transparent to the user. One of the reasons why this technology has become very popular is that natural scenes may have a dynamic range that is larger than the dynamic range of a camera sensor. So as to produce an image without under- or over-exposed areas, several input images are captured and later merged into a single high dynamic range (HDR) result. There is an obvious need for evaluating this new technology. In order to do so, we will present laboratory setups conceived so as to exhibit the characteristics and artifacts that are peculiar to MI, and will propose metrics so as to progress toward an objective quantitative evaluation of those systems.

On the first part of this paper we will focus on HDR and more precisely on contrast, texture and color aspects. Secondly, we will focus on artifacts that are directly related to moving objects or moving camera during a multi-exposure acquisition. We will propose an approach to measure ghosting artifacts without accessing individual source images as input, as most of the MI devices most often do not provide them. Thirdly, we will expose an open question arising from MI technology about how the different smartphone makers define the exposure time of the single reconstructed image and will describe our work around a time-measurement solution. The last part of our study concerns the analysis of the degree of correlation between the objective results computed using the proposed laboratory setup and subjective results on real natural scenes captured using HDR ON and OFF modes of a given device.

Introduction

Multi-image (MI) computational photography applications have received lots of attention in recent years, mostly driven by the smartphone market. MI technology involves combining multiple shots of a scene into a single image. These shots can be taken simultaneously (multi-sensors, multi-cameras) or sequentially. There exist many applications of MI: spatial or temporal noise reduction [11], high dynamic range (HDR) [17, 9, 10, 16], motion blur reduction [12, 13], super-resolution, focus stacking, depth of field manipulation, high frame rate, among others.

The creation of a single image from a sequence of images raises several problems. These problems and artifacts are mostly due to motion in the scene or camera. For a review of methods for dealing with these issues we refer to [5, 8] and references therein. A different type of artifacts concern the tone mapping of HDR scenes. Since the images must be displayed on screens with limited range, the choice of tone mapping becomes a critical part of the MI system [16]. This process is more 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 [2, 14].

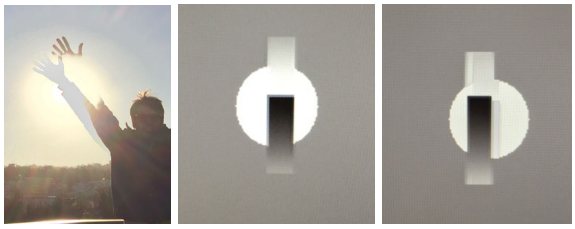


(a) (b)

Figure 1. *Picture (a) shows the proposed setup for measuring the capacity of MI devices to capture HDR scenes. The target contains a color chart, a texture chart, and a grayscale of 80 uniform patches linearly distributed between 0 and 100% of transmittance. The luminance of the left light source is always 170 cd/m^2. The luminance of the right light source varies between 170 and 17000 cd/m^2. In (b) a real scene is displayed in order to correlate the measurement with a perceptual analysis.*

This article is concerned with the artifacts resulting from the MI technology used in modern photographic devices and proposes objective metrics for evaluating some of those artifacts. Currently, evaluating the quality of MI algorithms is very important as this technology is more and more present in different devices. Apart of being present, this technology is often transparent to the user. Nowadays, the MI capture is mostly judged through subjective evaluations [15, 3]. Other authors or assume access to the complete stack of input images which is often hand annotated [4]. But there is an obvious need of end-to-end objective evaluations that do not require access to the intermediate images. On the other hand, the traditional lab setup (tripod, flat shooting targets) is particularly favorable to MI systems, more favorable than real life, so that traditional evaluation methods tend to overestimate the real-world quality of MI systems. The goal of this paper is to propose lab setups and objective metrics so as to simplify the evaluation process of MI technology.

Ghosting. The most common artifact related to MI acquisition is ghosting [8, 6]. When combining images acquired at different instants, because of motion, there is no warranty of observing the exact same image. When the underlying MI algorithm fails at detecting or compensating this motion, "ghosts" appear in the combined image as result of combining incoherent frames (see Figure 2).



(a) (b) (c)

Figure 2. *Reproduction of ghosting artifacts using a synthetic video of a moving pattern. (a) shows a real ghosting artifact. (b) is an image taken with the iPhone 6+ with HDR OFF, while (c) is taken with HDR ON. We can observe the same kind of ghosting artifacts. As we do not know what happens when HDR is off (we are not necessarily looking at a "single-source" image, Maybe the device did some other (non-HDR) multi-image fusion without telling us), it makes sense to test this mode.*

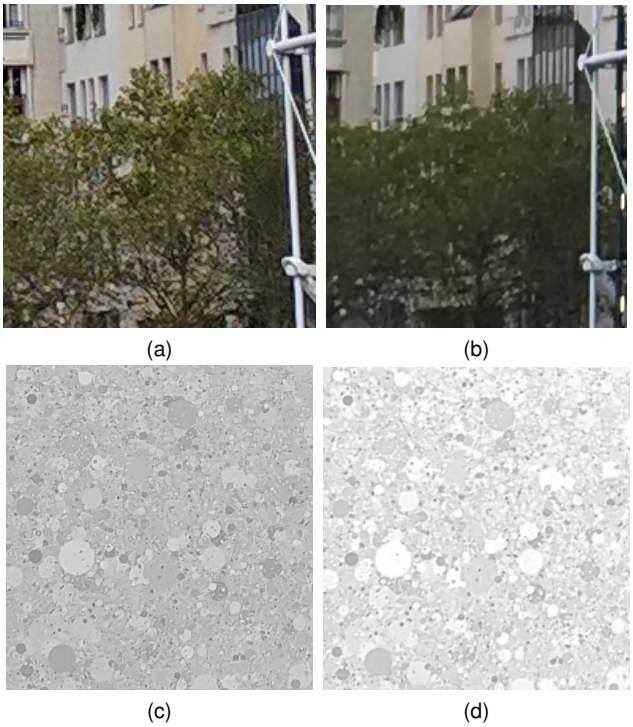
Contrast and texture loss. Very present defects are *halos* and *contrast and texture loss* resulting from the tone mapping of HDR images (see Figure 4, 3, 5). We observe that local contrast perception, along with a spatial/intensity coherence in the image is an important factor to account for a good HDR image. Tone mapping algorithms allow to display an HDR image on a support with limited dynamic range. The tone mapping algorithm must produce a pleasant rendering of the image while preserving low contrast details [1]. This is usually done by local contrast adaptations, which are inspired on perceptual principles [2] (i.e. humans do not perceive absolute intensities but rather local contrast changes).

Color appearance. Color is a very sensitive matter when dealing with HDR scenes. The human visual system (HVS) adapts differently in front of a real HDR scene and in front of a typical LDR display. Even with an accurate HDR image of a scene, the tone mapped and displayed image may appear different from the real scene due to a different adaptation of the HVS [16]. Understanding the HVS and the models created to predict how colors are be perceived under such different lighting conditions [18] will have a direct impact on color fidelity. Moreover, color treatment in HDR tone mapped images can be used to recover contrast and texture [19], which influence directly the color fidelity. This implies that color appearance is a key factor to a good final image.

Noise. Noise adds a new dimension of complexity to the MI problem. When stitching multiple images, the risk is that the produced image may have incoherent noise (see Figure 6). As noise is perceived as part of the image texture, its incoherence results in an apparent boundary. Although the analysis below is not centered on noise coherence, we still observe its impact in the proposed metrics.

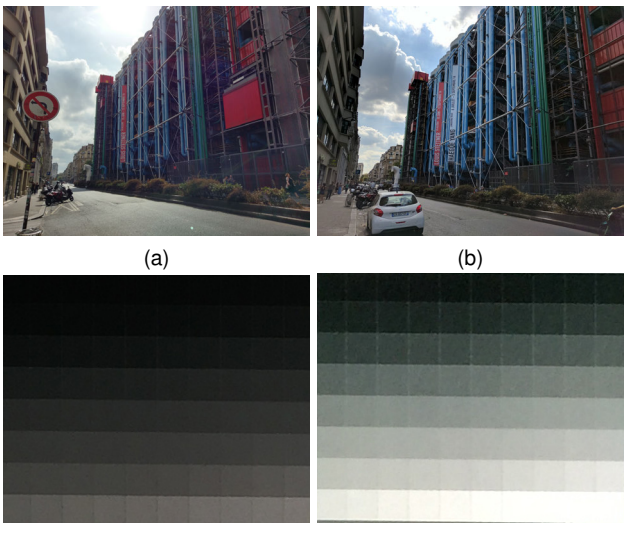
Evaluation of multi-image systems. There is an obvious need to evaluate these new technologies. Several papers in the literature address the evaluation of MI artifacts, they usually focus on HDR deghosting but their observations can be applied to most of the MI applications. These papers can be roughly divided into two categories.

In the first category we find perceptual evaluations carried on by trained observers [3] instructed to look for precise artifacts



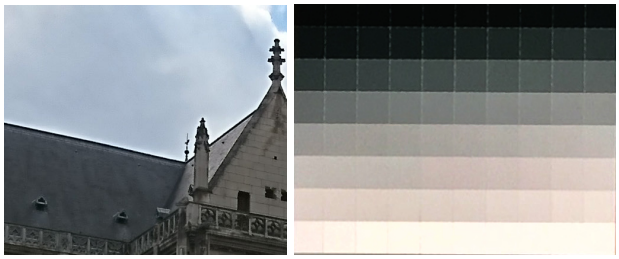
(a) (b) (c) (d)

Figure 3. *An important aspect of HDR rendering is texture preservation.The picture (b) is less textured in comparison with (a). The texture target of the HDR setup can highlight the preservation of texture. The picture (d) is less textured than (c), even if it has higher contrast.*



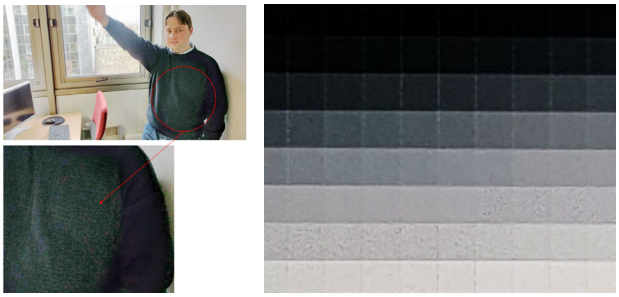
(a) (b) (c) (d)

Figure 4. *An important aspect of HDR rendering is perceptual contrast preservation. The picture (a) is less contrasted and some colors are lost in comparison with (b). The gray scale target of the HDR setup can highlight the preservation of local contrast. The picture (c) shows a poorly contrasted grayscale in comparison with (d).*



(a) (b)

Figure 5. *(a) is an example of halo artifacts between the roof and the sky, which can be reproduced by our setup (b). The homogeneous patches are brighter on their upper edge than on their lower edge.*



(a) (b)

Figure 6. *In (a) is shown an example of noise artifact due to bad stitching, which can be reproduced by our setup (b). Notice in the 6th and 7th rows the rupture of noise consistency.*

such as noise, blur, ghosts. However, these evaluations are difficult (and expensive) to put in place and the evaluation is inherently noisy as different observers may have different opinions despite of the instructions.

The second category of evaluation methods allow to compute objective metrics for different artifacts. However, these methods assume a setup in which the stack of input images is available along with the HDR deghosted result [7, 4]. These methods compute different artifact maps based on perceptual metrics. These artifact maps are then combined to yield a single quality score. These studies are usually accompanied by subjective studies that validate the agreement of the subjective scores with the proposed metrics.

Our contributions. In this paper we put forward several quantitative evaluations for multi-frame images. What distinguishes our approach from other evaluations:

- We evaluate objectively the system as a whole, the proposed metrics aim at measuring quantitatively the properties of the final image.
- Existing metrics take as input a stack of exposures. However, intermediate images are often not available, as for instance in most of the existing smartphone devices. Therefore in this paper we propose laboratory setups that allow to observe and measure the different MI artifacts without using intermediate images. For example, we use synthetic videos and high frame-rate screens that allow to evaluate ghosting artifacts.
- The measured properties are precisely defined, they reflect proven aspects of the human vision such as texture sensitivity.
- Whereas most of the metrics proposed today depend on the content of the scene, the measurements through the proposed setup are independent of the content of the scene. We are interested by objective and repeatable metrics. The evaluation is performed in a controlled and repeatable environment.
- The metrics we propose are based on the dynamic of the scene. We propose a laboratory setup that creates a reproducible high dynamic range scene with the use of two diffuse light sources with precisely adjustable intensity and printed transparent charts. Using the two adjustable light sources gives the possibility to measure and trace the contrast and color gains due to MI technology for scenes with increasing dynamic range through predefined stops.
- We expose the question of what is the definition of the exposure time for a single reconstructed image through a MI acquisition and how we can measure this specific time.

In the next section we focus on the most used application of multi-imaging: HDR. We will present a laboratory setup conceived so as to evaluate the percentage of contrast and color saturation gain we can get from a multi-exposure acquisition. Then, we will present an exploratory laboratory setup that gives the possibility to observe one of the multi-imaging related artifacts, ghosting, and some ideas about how to measure it. In the third part, we will share our observations about the exposure time of the final reconstructed image through the combination of the intermediate acquired images. Finally, we will describe our work on correlating the observations on natural images with the results of the proposed objective metrics.

HDR scene rendering evaluation

In HDR imaging, the aim is to capture a scene with higher dynamic range than the camera is capable of capturing with a single exposure. This is interesting as many natural scenes have a dynamic range larger than the dynamic range of a camera image sensor. Some of the existing camera phone devices give the choice to acquire an image by activating or not the HDR mode. For our HDR laboratory setup, we chose a static scene composed of two diffuse light sources with precisely adjustable intensity (Kino Flo LED 201 DMX devices, DMX for short) that can provide 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 charts. Our final image contains the two DMX devices as it can be seen in Figure 1. The two DMX devices are then programmed. They begin with the same luminous emittance (170 cd/m^2), and the right one increase its luminous emittance with predefined values: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20, 40, 60, 80, 100% of the maximum DMX luminous emittance. By stretching the intensities of the two DMX devices we intend to create scenes with increasing dynamic range, where the left part of the scene has low light (LL) conditions while the right part has bright light (BL) condition, as can be observed in a strong backlit scene. For each dynamic range setting we acquire 4 images:

- Two with auto exposure (AE), one with HDR ON and one

with HDR OFF.

- Two with forced exposure (FE), one with HDR ON and one with HDR OFF. In most devices exposure can be "forced" so that a point of interest is well exposed (by tapping on it). In this setting we force the exposure on the low light part of the HDR scene. This can be explained as most pictures of HDR scenes involve a bright background (i.e. sky or bright light), and a low light main subject (a person, a building, etc).

In addition to the charts we use for quantitative measurements, we consider a target with a natural scene as shown in Figure 1b, which permits to perceptually validate the consistency of our measures. This setup allows to:

- Compare the performance of different devices.
- Analyze the performance of a MI device in auto and forced exposure mode.
- Compare the performance in a high dynamic scene between the low light part (left DMX) and the high light part (right DMX).
- Compare the performance with HDR ON and OFF, to analyze the gain of using HDR, and to highlight the trade-offs made by the device maker.
- Show artifacts such as bad stitching, halo, glare, noise reduction (Figure 4, 5, 6).

The characteristics we want to highlight are the preservation of local contrast, texture and color. 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 [16]. We want to quantify how the device compresses the HDR scene to fit the display range while preserving details and local contrast. 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.

It is worth noting that most devices are by default in mode HDR AUTO, which means that the device chooses whether or not to activate HDR mode. It could be interesting to evaluate this mode, in addition to HDR ON and OFF, to analyze if the device takes the optimal decision, in the sense that HDR AUTO attains a higher score than both HDR ON and HDR OFF. This question will be addressed in future works.

Local contrast preservation

The grayscale part of the target is composed of 80 uniform different patches with linearly increasing transmission. Having two grayscales with two different dynamic ranges on the same scene allows to measure how a device preserves the local dynamic range of each. The metric used is the entropy of the normalized grayscale histogram $hist_{gs}$.

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

It can be seen as the quantity of information contained in the grayscale. 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. A gray scale with evenly distributed values will have an entropy equal to the dynamic of the grayscale (8 bits maximum).

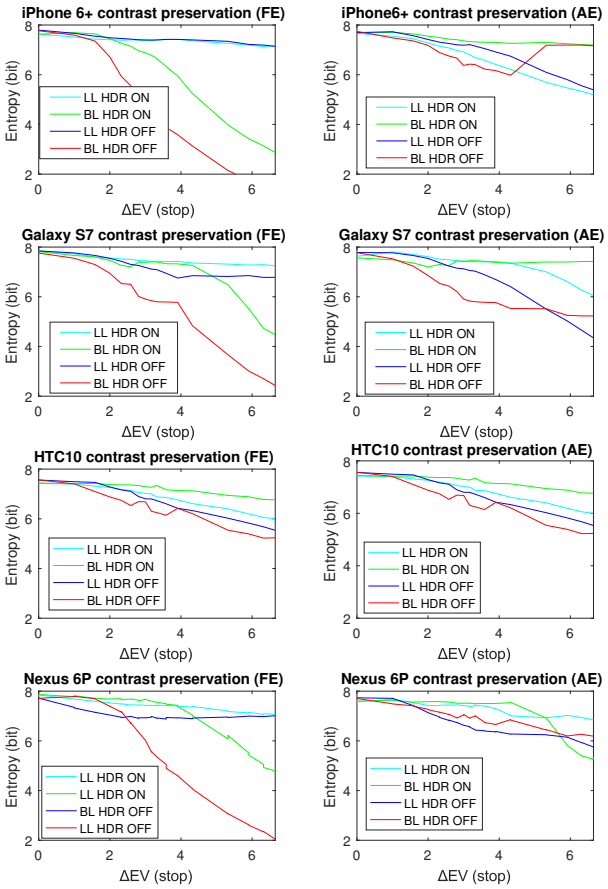


Figure 7. *Local contrast preservation measurement results for 4 devices (from top to bottom: iPhone 6+, Galaxy S7, HTC 10, Nexus 6P). The first column shows the result with exposure forced on the left DMX texture. The second column is the result with auto exposure. The abscissa is the difference of luminance between the two DMX in photographic stops.*

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, the experiments show that this choice seems to provide a good indicator of the perceived contrast. As for the spatial artifacts mentioned above, they can be part of a different metric that will complete the local contrast perception, emphasizing the trade-off relation between reducing halo artifacts and preserving image contrast. The results for four devices, with auto and forced exposure, are presented in Figure 7. For each device it shows the preservation of contrast for the left and right DMX with HDR ON and OFF (4 curves).

Texture preservation

The texture measure is designed to evaluate how fine details are preserved after tone mapping and denoising have been applied [20, 21, 26]. The Dead Leaves pattern [20] is used to simulate a texture with natural image properties, which are hard for post processing to enhance. Let us define the texture spatial frequency response (SFR) [21] as the measured power spectral den-

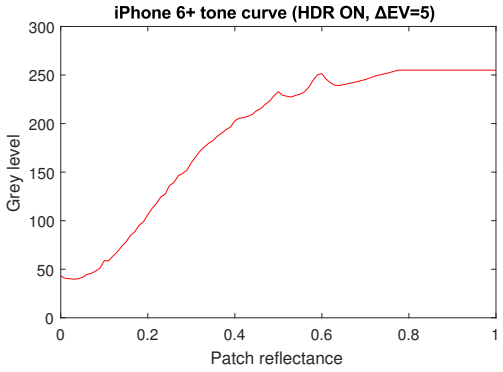


Figure 8. *This is an example of a tone curve extracted from the gray patches shot by the iPhone 6 with HDR ON. We see the non-monotony due to halo artifacts.*

sity (PSD) of the texture divided by the ideal target power spectral density

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

where PSD_{noise} denotes the noise power spectral density in the image, measured on uniform patches. Then, the acutance metric A is defined as the weighted average of the texture SFR with the contrast sensitivity function (CSF), which represents the sensitivity of the HVS to different frequencies

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

The acutance gives information about how texture is preserved, however it is contrast dependent. So in order to compute it, a preprocessing step is required, a normalization that linearizes and scales the gray levels of the observed image. This is done by estimating a tone curve using the gray patches surrounding the texture target (see Figure 1a).

It is worth noting that a tone curve would not undo the local adaptation effects of HDR tone mapping, so this step is very sensitive to tone mapping. Moreover, glare induced by the brighter DMX causes a gradient of intensity on the low light DMX. These two problems imply that there is no guarantee that the estimated tone curve is monotone (as seen in Figure 8) and that the tone curve measured on the gray patches is valid on the texture.

To lessen these effects we forced the monotonicity of the estimated tone curve and make the hypothesis that this curve is applied uniformly over the texture. The perceptual validation confirm that this setup can measure the effects of texture loss. Nevertheless, this measure is not in its final state. For instance, this measure can be wronged when strong noise is present in the image. An efficient denoising will result in an underestimated noise power spectrum, texture sharpening will result in a false amplification in high frequencies. This would improve wrongly the acutance. Those problems can be corrected by phase information [26, 27]. Some work on normalization may have to be done if local tone mapping becomes an issue.

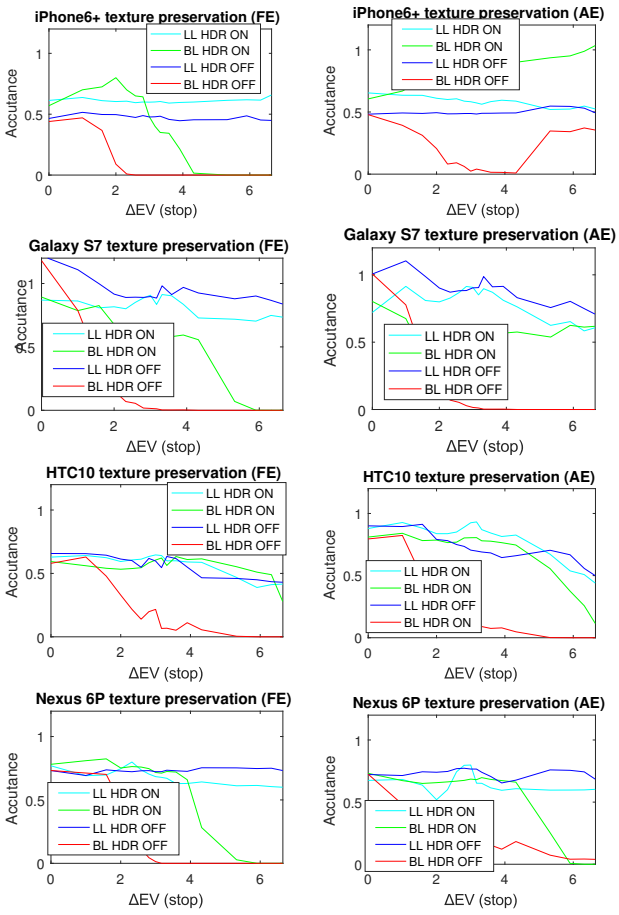


Figure 9. *Texture preservation measurement results for 4 devices (from top to bottom: iPhone 6+, Galaxy S7, HTC 10, Nexus 6P). The first column shows the result with exposure forced on the left DMX texture. The second column is the result with auto exposure. The abscissa is the difference of luminance between the two DMX in photographic stops.*

Color preservation

The classic image color reproduction evaluation [24, 23] can be extended to MI devices.

Color preservation can be described as the ability of a camera to preserve colors across different dynamic ranges. The target chart in Figure 1 contains a set of 24 representative colors, inspired by the *Macbeth ColorChecker*.

The color coordinates of each uniformly colored patch are measured in RGB and then converted to L*a*b* coordinates (assuming that the color space of the shot is sRGB). For evaluating the color preservation, in this article, we use the classic color difference metric $\Delta ab*$, which, by removing the lightness L^*, minimizes the dependency on exposure. Considering two colors expressed in L*a*b* coordinates the measure is computed as:

$$\begin{aligned} Color_{ref} &= (L^*_{ref}, a^*_{ref}, b^*_{ref}) \\ Color_{test} &= (L^*, a^*, b^*) \\ \Delta ab* &= \sqrt{(a^*_{ref} - a^*)^2 + (b^*_{ref} - b^*)^2}. \end{aligned} \tag{4}$$

This metric aims at removing most of the differences resulting from the difference in brightness (neglecting nevertheless sat-

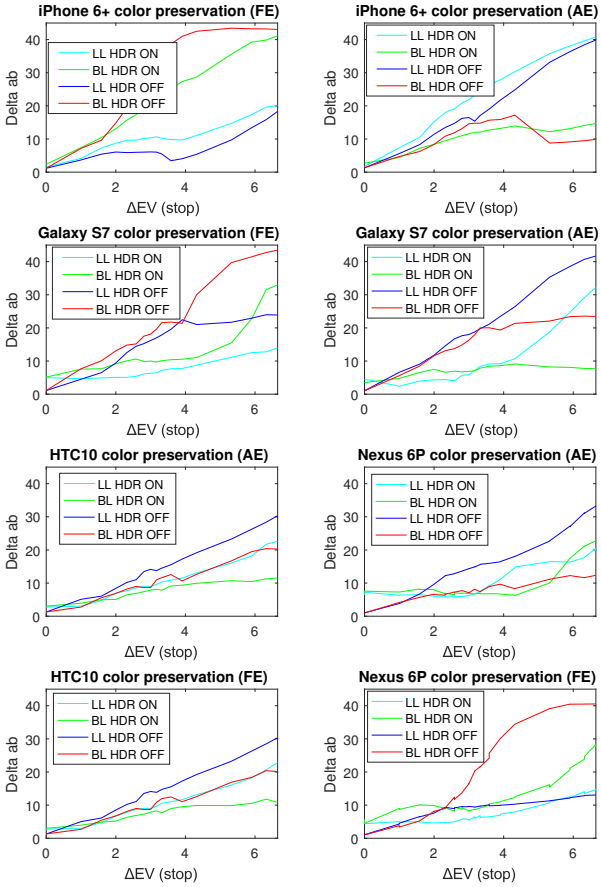


Figure 10. *Color preservation measurement results for 4 devices (from top to bottom: iPhone 6+, Galaxy S7, HTC 10, Nexus 6P). The first column shows the result with exposure forced on the left DMX texture. The second column is the result with auto exposure. The abscissa is the difference of luminance between the two DMX in photographic stops.*

uration effects). To prevent the measurement to be penalized by any color bias that a device maker could apply for aesthetic appearance, each reference value of (a^*_{ref}, b^*_{ref}) is computed with HDR OFF on a scene with the two DMX at the same intensity. The results are presented in Figure 10 for four devices with auto and forced exposure. For each device it shows the $\Delta ab*$ for the left and right DMX with HDR ON and OFF (4 curves).

Ghosting evaluation

In this part of our paper, we explore artifacts related to the deghosting algorithms present in MI technology. These artifacts are often due to bad blending and lack of correspondence between the different images and can be observed as merging artifacts, color artifacts, motion artifacts, blurring or noise artifacts (as seen in Figure 11). We propose a setup for objectively evaluate those ghosting artifacts.

To have a reproducible setup, we explore the use of a high refresh rate desktop monitor (*Asus Rog Swift PG278QR 27*) as shooting target. We have created different synthetic videos of moving patterns (see Figure 12 for one example) that allow us to reproduce the observed artifacts (Figure 11). The videos of the

(a)

(b)

Figure 11. *(a) Ghost artifacts in natural scenes. From left to right and top to bottom: iPhone 6+, HTC 10, Nexus 6P, Galaxy S7. The iPhone 6+ produces a strong discontinuity (an arm appears in the sky and the hand and the electric pole are confounded). The HTC 10 shows classic ghosts (the ghost of the arm). The Nexus 6P has a strong noise inconsistency and the hand and the foliage are confounded. The Galaxy S7 has limited ghosting on the edge of the arm. (b) Ghost artifacts highlighted by our setup. From left to right: iPhone 6+, HTC 10,Nexus 6P, Galaxy S7. The iPhone 6+ produces also strong discontinuities (the moving target and the building are merged. The HTC 10 shows also classic ghosts (the moving target is multiplied). The Nexus 6P also mixed the black part of the moving target with the window fence. The Galaxy S7 has also a limited ghosting. The correlation between artifacts present in natural scenes and in our setup for each device is good.*

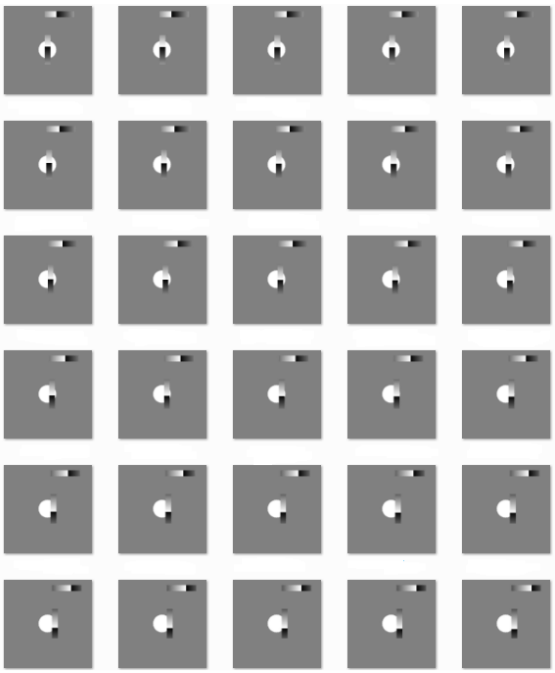


Figure 12. *Frames of the moving pattern and a fix circle. The fix pattern of our video, allows us to reproduce artifacts related to occlusion of different moving objects in the scene.*

moving patterns are played on the desktop monitor. Displays have well known artifacts like motion blur, pixel persistence, asymmetric pixel transitions, inverse ghosting due to overshoot of the pixel response, pulse-width modulation artifacts etc. Some are intrinsic to a specific technology.

So as to isolate the blur and other artifacts related to the display technology, we characterized the frame rate and pixel response time of the device with the help of a high frame rate and high shutter speed device (*Sony RX100 IV*) on a tripod at a known distance to the screen. In this way, we are able to qualitatively distinguish the display artifacts from those related to the HDR fusion algorithms. In a second part, a validation was made of the setup. We link the type and occurrence rate of ghosting artifacts that a device can produce in a natural scene with the artifacts that our setup is able to reproduce (see Figure 11).

This shows that the proposed setup has the potential to systematically reproduce ghosting artifacts as they appear on real-world scenes. This opens the door for defining a quantitative and reproducible measure for such artifacts, which we will address in the future.

A proposal for defining MI exposure time

The exposure time, also called shutter speed (usually denoted as E), is the effective time interval during which light fills up the sensor photosites. Classical measurements [25] assume that all pixels of the sensor have the same integration time. With MI technology that assumption does no longer hold true as some pixel may be composed from different source images than other pixels. Besides, classical timing measurement may fail because of ghost removal algorithms, as seen in Figure 13.

We observe that device makers do not agree on the value

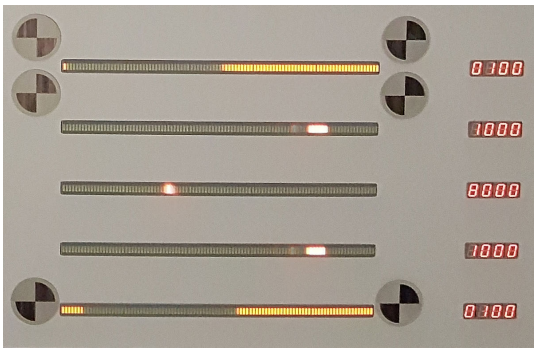


Figure 13. *Ghost observed during classic timing measurement. Exposure time is computed by counting the number of lit LEDs. As the LEDs are lighted on and off at a known speed, we can deduce the resulting exposure time. Thus Ghosting can result in a wrong exposure time.*

of exposure time to be reported in the image metadata (EXIF) in the case of MI processing. We see in Figure 14 for the Nexus 6P that the exposure time with HDR ON is underestimated by the EXIF. On the contrary the exposure time measured for the iPhone 6+ with HDR ON is consistent with the exposure time in the EXIF. However, other MI techniques like temporal noise reduction might be used even with HDR set to OFF. As the value provided for the iPhone 6+ with HDR OFF in the EXIF is about four times larger than one measured may raise questions (61 ms measured, 250 ms reported in the EXIF).

In this section we want to expose the challenges in extending the notion of exposure time to the MI case. The objective of this discussion is to come to a consensus about what should be stored in the EXIF field corresponding to exposure time in the MI case. The ideal knowledge to be extracted from an HDR MI device is: how many pictures have been combined, their individual settings (exposure time, ISO...), and the spatially varying contributions of each image to the final result. However, besides being unrealistic to store so much information along with the image, this level of detail is against the intuitive concepts behind exposure time.

For this reason we start by recalling the effects of exposure time in the single-image case and see if these properties can be preserved by a multi-image generalization.

- First of all, increasing the exposure time increases exposure. So the image becomes brighter.
- Increasing the exposure time and decreasing the ISO **reduces noise**. However, noise is not a reliable reference as denoising may greatly reduce it. Adjusting the MI exposure time so that noise is preserved would inevitably lead to questions about the texture preservation as mentioned in the previous sections.
- Longer exposure times lead to **more motion blur**. This property is exploited in [25] to measure the exposure time and could be easily adapted to the MI case. However, devices may freeze the motion of some parts of the image, while blurring others, leading to a confuse statement.

The *Exposure time Envelope* (EE). By defining the exposure time as the absolute time interval between the beginning of light integration in the first line in the first image and the end of light

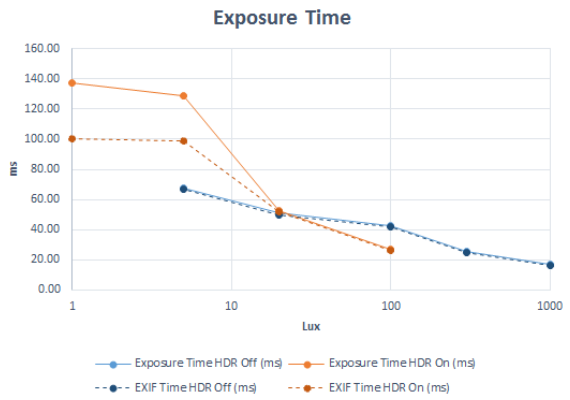


Figure 14. *Exposure time reported in EXIF compared with exposure time measured for the Nexus 6P with HDR ON and OFF.*

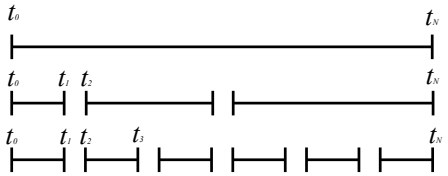


Figure 15. *The exposure time of a traditional single-image device (represented as a line in the first row) can be used by a multi-image device to allocate multiple exposures in different ways (second and third rows).*

integration in that same line in the last image used by the fusion algorithm, we preserve the intuitive properties of this measure (see Figure 15).

- If the fusion simply averages the images (possibly compensating for gain differences), then the result tends to behave as a standard long exposure photograph, both in terms of noise and motion, and the envelope EE reflects this.
- If the algorithm picks only a single image (a "lucky frame"), then the exposure time of the selected image would be $E_i = EE$, which is exact in terms of noise and motion.
- If the fusion combines the images using spatially varying weights, then both noise and motion blur can severely differ from an equivalently exposed single-image shot. Motion can be frozen by multi-image devices and noise could be reduced unevenly across the image. Nevertheless, a large EE value permits to inform the user about the manipulation that the image has undergone, which may explain processing artifacts.

Clearly the exposure time envelope cannot capture the full complexity of a multi-image process, so in order to complement its information we also propose two additional exposure time metadata: the minimum E_{min} and the maximum E_{max} exposure times in the burst. The three values EE, E_{min}, and E_{max} can summarize the multi-image process and the effects that may result from it.

Evaluation of four well known 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

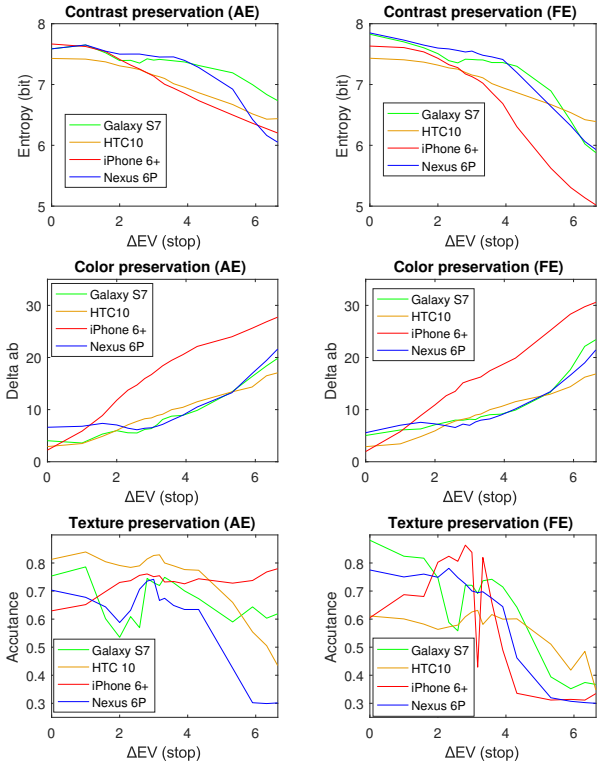


Figure 16. *Results of the HDR setup for the Galaxy S7 (green), HTC 10 (orange), iPhone 6+ (red) and Nexus 6P (blue). Each measurement is computed as the mean of the measurement of the bright light DMX and the low light DMX on HDR ON picture. The abscissa is the difference of luminance between the two DMX in photographic stops.*

metrics, which will eventually be combined into a single one. For that purpose, the laboratory setup and the metrics presented above are tested and compared against subjective evaluations conducted on natural scenes captured with four devices. The four devices are:

- iPhone 6+
- HTC 10
- Galaxy S7
- Nexus 6P

HDR scene rendering metrics

The laboratory results of the HDR scene rendering evaluation are summarized in Figure 16. For the sake of simplicity, we present the average of the metrics computed on the two DMX (low and high light), only for the case HDR ON. Averaging the measures of the two DMX means that we do not prefer any of them. For instance, a device with good contrast (8 bits) in low light and poor contrast (4 bits) in bright light is seen as equivalent to a device with medium contrast (6 bits) in each range.

Local contrast. All devices have roughly 7.5 bits of local contrast in a low dynamic range scene (up to 3 stops of luminance difference) both in AE and FE. For higher dynamic range, according to the results shown in Figure 16, in AE, the Galaxy S7

Figure 17. *Natural HDR scene. (a) HDR scene shot with the Galaxy S7. (b), (c), (d) and (e), from left to right: Galaxy S7, Nexus 6P, HTC 10 and iPhone 6+. Compared to the sky, (b) has 2 stops of luminance difference, (c) and (d) have 4 to 5 stops of luminance difference, (e) has 7 stop of luminance difference.*

leads the competition. In FE, the Galaxy S7 and Nexus 6P have the best performance until 5 stops. Beyond 5 stops we see that the HTC 10 becomes better. If we look in details (see Figure 7), we can see that the HTC 10 gain against the Galaxy S7 is in the bright light.

In FE and HDR OFF, we identified two different strategies adopted by device makers. The first, like the iPhone 6+, chooses to correctly expose the select part of the scene (the low light DMX). That saturates the bright light DMX, yielding a poor local contrast preservation score on this DMX (see Figure 7). On the other hand, devices like the HTC 10 underexpose the selected part of the scene. This allows to better expose the other DMX, and also to gain local contrast in both DMX when HDR is ON, while for the iPhone 6+ the gain can only be on the bright light DMX. This explains why the HTC 10 comes out best in forced exposure. The Nexus 6P and the Galaxy S7 also adopt this strategy, but less marked. But such behavior should be penalized to a certain extent. When exposure is forced on the dark scene, it is to see the details of the dark scene. If it is underexposed in a way that contrast is lost, this is not appreciated.

Color preservation. The iPhone 6+ is clearly the worst in preserving its color in both AE and FE when HDR is ON. For the other three devices the color preservation is approximately the same.

Texture. In AE and HDR ON, the best of the 4 devices for scenes before 5 stops is the HTC 10. After 5 stops, the iPhone 6+ is the best. It compensates its poor gain in local contrast with a huge gain in texture preservation (see Figure 7).

In FE, the HTC 10 is the only one who recovers texture in bright light, which makes it the best for dynamics superior than 5 stops. The iPhone 6+ is not good after 4 stops because it looses all its texture by doing a good exposure on the scene manually selected (the low light DMX). Variation in acutance can be of several reasons: change of focus in AE (Nexus 6P, HTC 10), activation of denoising (iPhone 6+), glare (iPhone 6+ and HTC 10).

Validation on natural scenes

For our validation, several natural HDR scenes were shot in auto exposure mode with the four devices. The scenes 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. For the scene of Figure 17, we observe 8 stops of exposure difference between the sky and the darkest brush. While the maximum ΔEV that could be attained with our setup using the Kino Flo LED 201 DMX is about 12 stops, our initial laboratory setup only explored a ΔEV of ~ 6.5 stops. For this reason we extrapolate the curves of Figure 7 by 1.5 stops.

Contrast. Our measurements for high-light local contrast in a 7 to 8 stop dynamic scene rank the four devices as:

1. Galaxy S7
2. iPhone 6+ $\approx$ HTC 10

3. Nexus 6P

The crops (c), (d), and (e) of Figure 17 correspond to low light parts of the scene. They are respectively 5,5 and 7 stops below the exposure of the sky. Crops (c) and (e) correlate well with the local contrast metric except for the Galaxy S7 which has better local contrast on the HDR scene than our setup would suggest according to Figure 7. To be sure, the setup must be extended to 8 stop.

The picture (b) of Figure 17 corresponds to the high light part of the scene, which is 1/2 stop under the the exposition of the sky. In that picture, the Galaxy S7 has the best contrast, the iPhone 6+ and HTC 10 has relatively similar contrast, but the Nexus 6P is quite under estimated by our setup.

Color. Regarding the colors, our setup extrapolation classes the devices as:

1. Nexus 6P $\approx$ HTC 10 $\approx$ Galaxy S7
2. iPhone 6+

It is difficult to draw conclusions about the preservation of colors, but it seems that the HTC 10 is over evaluated by our setup as its performance is closer to iPhone 6+ than the others.

Texture. Regarding texture, our setup in the case of low light texture classes the devices as (see Figure 9):

1. Nexus 6P
2. Galaxy S7 $\approx$ iPhone 6+
3. HTC 10

The crops (c) and (d) in Figure 17 correlate well with the metric. In (d), the Galaxy S7 is similarly textured than the iPhone 6+, and it highlights well that Nexus 6P is better than HTC 10. The HTC 10 texture is very uneven between (c) (more sharp) and (d) (flatten), but always worse than the Nexus 6P and the iPhone 6+.

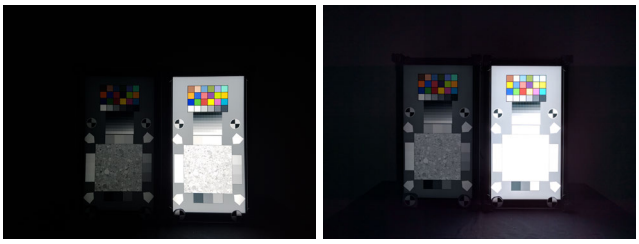
On the other hand in the bright parts of the image our setup would class the devices as:

1. iPhone 6+
2. Galaxy S7
3. HTC 10 $\approx$ Nexus 6P

The iPhone 6+ and the Galaxy S7 have a good texture, and the HTC 10 a bad one, but then again the Nexus 6P is quite under evaluated by our setup. This is because Nexus 6P privileges allocating contrast in the low dynamic range of the image, thus loosing texture in the bright parts. Because of this behavior in Figure 16 the Nexus 6P has a low score. This may point to a limitation of our setup, which is discussed next.

Limitations of our current setup

The constant underestimation of the Nexus 6P in the highlights may be due to the framing choice in our setup (see Figure 18. More generally, the reparation of high light and low light in the scene may have consequences that we have not taken into account. We see in that figure that the Nexus 6P chooses to better expose the table in front of the DMXs than the Galaxy S7. This implies that the Nexus 6P focuses more on the low lights, and therefore lowers its contrast in high lights. As we see our setup



(a) (b)

Figure 18. *Our setup for Δ_{EV}=6.5. (a) Galaxy S7. (b) Nexus 6P.*

results in high lights for the Nexus 6P diverge from natural HDR scene, we must investigate towards that direction. This argument is sustained when looking at our setup target with a natural scene natural image (fig 1 .b). The ranking of contrast perception for $\Delta EV = 5$ by a DxO Mark perceptual analyst change when only the target are taken into account (ranking: 1-Galaxy S7, 2-Nexus 6P, 3- iPhone 6+ $\approx$ HTC 10) or when the contrast of the whole image is perceived (ranking: 1-Nexus 6P, 2-Galaxy S7, 3-HTC 10, 4-iPhone 6+).

Conclusion

In this paper, we put forward three setups with associated metrics for the quantitative evaluation of multi-image system properties which are shown to be consistent with human visual system. We evaluate objectively and repeatably the system as a whole, without using intermediate images, independently of the content of the scene. We also brought to light the issue of the definition of exposure time in the context of MI acquisitions and proposed a new notion better adapted to this type of devices. The proposed setups are consistent with real world conditions and can reproduce the artifacts observed in multi-imaging systems. Overall, the experiments showed a good correlation between the quantitative and subjective measurements, nevertheless some will be further optimized in future works.

References

[1] 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.

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

[3] Karaduzovic-Hadziabdic, K., Telalovic, J. H., and Mantiuk, R. (2014). Expert evaluation of deghosting algorithms for multi-exposure high dynamic range imaging. In HDRi2014-Second International Conference and SME Workshop on HDR imaging.

[4] Tursun, O. T., Akyüz, A. O., Erdem, A., and Erdem, E. (2016). An Objective Deghosting Quality Metric for HDR Images. In Computer Graphics Forum (Vol. 35, No. 2, pp. 139-152).

[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.

[6] Gryaditskaya, Y., Pouli, T., Reinhard, E., Myszkowski, K., and Seidel, H. P. (2015). Motion aware exposure bracketing for HDR video. In Computer Graphics Forum (Vol. 34, No. 4, pp. 119-130).

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

[8] 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).

[9] Aguerrebere, C., Delon, J., Gousseau, Y., and Muse, P. (2013). Simultaneous HDR image reconstruction and denoising for dynamic scenes. In Computational Photography (ICCP), 2013 IEEE International Conference on (pp. 1-11). IEEE.

[10] 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.

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

[12] 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).

[13] 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.

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

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

[16] 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.

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

[18] Akyüz, A. O., Reinhard, E. (2006). Color appearance in high-dynamic-range imaging. Journal of Electronic Imaging, 15(3), 033001-033001.

[19] Smith, K., Krawczyk, G., Myszkowski, K., Seidel, H. P. (2006). Beyond tone mapping: Enhanced depiction of tone mapped HDR images. In Computer Graphics Forum (Vol. 25, No. 3, pp. 427-438). Blackwell Publishing, Inc.

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

[21] McElvain, J., Campbell, S. P., Miller, J., 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.

[22] Kirk, L., Herzer, P., Artmann, U., 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.

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

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

[25] Masson, L., Cao, F., Viard, C., Guichard, F. (2014). Device and algorithms for camera timing evaluation. In IST/SPIE Electronic Imaging (pp. 90160G-90160G). International Society for Optics and Photonics.

[26] Kirk, L., Herzer, P., Artmann, U., 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.

[27] 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.