Objective image quality evaluation of HDR videos captured by smartphones

Cyril Lajarge, François-Xavier Thomas, Elodie Souksava, Laurent Chanas, Hoang-Phi Nguyen, Frédéric Guichard
DXOMARK, Boulogne-Billancourt, France

Abstract

High Dynamic Range (HDR) videos attract industry and consumer markets thanks to their ability to reproduce wider color gamuts, higher luminance ranges and contrast. While the cinema and broadcast industries traditionally go through a manual mastering step on calibrated color grading hardware, consumer cameras capable of HDR video capture without user intervention are now available. The aim of this article is to review the challenges found in evaluating cameras capturing and encoding videos in an HDR format, and improve existing measurement protocols to objectively quantify the video quality produced by those systems. These protocols study adaptation to static and dynamic HDR scenes with illuminant changes as well as the general consistency and readability of the scene's dynamic range. An experimental study has been made to compare the performances of HDR video capture to Standard Dynamic Range (SDR) video capture, where significant differences are observed, often with scene-specific content adaptation similar to the human visual system.

Introduction

Recent technology improvements in both consumer and industrial cameras have enabled on-device capture and mastering of High Dynamic Range (HDR) videos without user intervention. Several formats coexist (HDR10, HDR10+, Dolby Vision, ...) encoding an increased range of luminance (greater or equal to $1\,000\,\mathrm{cd/m^2}$), color (through the use of wide-gamut color spaces), and contrast (greater or equal to $1\,000:1$) compared to the more common Standard Dynamic Range (SDR) formats. When viewed on a compatible display, these videos can reproduce real-life scenes more realistically for human observers.

The evaluation of HDR video capture is tested on camera systems where recording is fully automatic and intended for direct playback (without editing), on a compatible display, targeted at human viewers. We therefore evaluate HDR contents on the mastering and color grading of the device, independently of the optional tone/gamut mapping for rendering on any display.

The end goal of our work is to adjust our existing setups and metrics to be able to evaluate HDR videos captured by smartphones both objectively and without reference in similar ways to SDR content, and in doing so to compare the spatial and temporal performances of the device in both SDR and HDR formats.

Our previous works

To study the behavior of devices in HDR formats, we used two existing setups from our previous work on photo and video measurements.

The first setup is an SDR scene containing a *ColorChecker* Classic chart (figure 1.a), lit uniformly by a controllable illuminant. While an SDR scene by itself does not reach the limits of an HDR-capable device, this lets us compare both static and dynamic recording behavior for different scenarios on both SDR and HDR formats with existing measurements [3].

The second setup, named *HDR Composite* (figure 1.b), was initially introduced in [1] and [2] to measure a device's capacity to capture and render static HDR scenes into SDR images. The setup contains two back-lit panels, each covered by a transmissive target that contains a color chart, a texture chart, and a gray scale of 63 uniform patches spanning a transmittance range of approximately 8EV. This setup is used to simulate a simple HDR scenes, with one panel set at constant luminance while the other can be increased up to $\Delta\mathrm{EV} = 7$, bringing the maximum luminance range of the scene up to 15EV (equivalent to 90 dB).

Other works

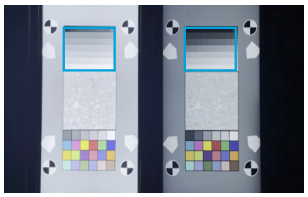
Several metrics and standards exist on evaluating some aspect of HDR content but each method are not well adapted to evaluate capturing video HDR format. For example, studies have been conducted to predict visible quality differences between pairs of reference and distorted HDR images, either for still images with HDR-VDP-2 [9] or videos with HDR-VQM [8]. The limitation of HDR-VDP-2 [9], is that it is only predict luminance difference per image pair and does not take into account the temporal aspect. HDR-VQM [8] propose an objective and subjective full-reference video quality evaluation based on spatio-temporal analysis related to human video viewing. However, both require existing "ideal" HDR content, which make them unsuitable for building no-reference objective metrics to estimate the quality of the original HDR video captured by a device.

Limitations

Existing metrics and evaluation protocols, including those referenced above, often lack applicability to our use case in at least one way among:

- The image acquisition protocol is ill-adapted to HDR content (e.g. reliance on common SDR behaviors such as target exposure that are not necessarily consistent in the case of HDR formats).
- The metrics themselves are ill-adapted to HDR content (*i.e.* reliance on measurement spaces that either have known limitations or have not been validated in the case of HDR videos, such as CIE-L*a*b*) ([1], [2] and [3]).
- They have only been tested on still images content, and we lack data on their behavior when confronted with a dynamic scene ([1], [2]).

(a) ColorChecker Setup



(b) HDR Composite Setup

(c) Portrait HDR Setup

Figure 1. *Different setups evaluated in this paper ; Evaluation of luminance reproduction in blue for each setup and the exposure target selection in purple on the 18% gray patch (a) and on the forehead of the realistic mannequin (c).*

Additionally, a lot of previous works available in the literature concern perceptual evaluation of whether an existing HDR video is reproduced correctly on the target display (after potentially lossy encoding, transmission or decoding stages), but do not evaluate whether the capture of a scene into that HDR content was done in an optimal manner in the first place.

Proposed setups

The *ColorChecker* setup can be extended to HDR videos directly to compare baseline behavior of the device under test over a simple scene. The *HDR Composite* setup can be extended to a) dynamic scenes similar to what was done for the *ColorChecker* and b) HDR format video quality assessment instead of SDR still images.

In addition to this setup, we explore possible improvements by combining multiple charts and/or elements to get closer as possible to real life HDR scenes. The advancement for the new *Portrait HDR* setup is to add a realistic mannequin bust next to an emissive surface with strong luminance (figure 1.c).

The fake face is produced by a society specializing in the business of post-production of motion pictures, video and television programs. We wanted to produce a fake face as real as possible, as if we had a real person in front of the camera to make our measurements automatic and repeatable. The realistic heads have a reflectance spectrum close to that real human skin.

A previous work exists to evaluate and rank tone mapping algorithms for images containing faces [10]. The conclusion of this paper is that the best tone algorithm is the one that manages to have a mean luminance level L* around 50 on the crop of the face. Finally, our new *Portrait HDR* setup can be constructed to evaluate exposure decisions in presence of recognizable content by replacing one of the panels with a recognizable portrait. The aim of this new setup is to evaluate how the device behave when a face is in the field with illuminance changes. We compare the average luminance behavior on the forehead of the realistic mannequin and on the 63 gray patches.

We therefore collected a database of both HDR-encoded and SDR-encoded videos (in several different encodings, such as HDR10, HDR10+, Dolby Vision with HLG or PQ EOTF) using several smartphone devices in the following setups:

- S1: HDR/SDR illuminant transitions and ramps with SDR lab scene (ColorChecker based on [3])
- S2: HDR/SDR illuminant transitions and ramps with HDR lab scene (based on [1] and [2])
- S3: HDR/SDR illuminant transitions and ramps with HDR portrait scene (new scene)

Each scene lets us evaluate the choice of target exposure and chromatic adaptation in respectively a SDR reflective surface, 2 emissive surfaces with very different luminances, and an emissive surface with strong luminance next to recognizable diffuse SDR scene content (a human face).

Adaptation of metrics to HDR content

Display luminances and tone-mapping

Analyzing images in a reliable way requires defining how the encoded digital RGB values are intended to be displayed, which in large part depends on the definition of an EOTF (Electro-Optical Transfer Function) converting these values to display luminances (in cd/m²). Transfer function encodings come in 2 distinct flavors:

- *Scene-referred encodings* primarily define an OETF, and therefore require the additional specification of the EOTF used to display them.
- *Display-referred encodings* primarily define an EOTF, and therefore can be displayed directly.

When the peak luminance output by the selected EOTF is too high for the target display (when for example displaying HDR content on a SDR screen), it is necessary to select an appropriate *tone-mapping* to reduce it to the acceptable range while preserving as much quality as possible, for example using BT2390 [4].

In the results presented here we assume that this is not the case, and that the target display is both calibrated and able to perfectly reproduce specified display luminances. Unless specified otherwise, we assume the following defaults:

- Scene-referred BT.709 [5]-encoded SDR content is displayed using the EOTF from BT.1886 [6] with a peak luminance of 100 cd/m².
- Scene-referred HLG-encoded HDR content is displayed according to the EOTF from BT.2100 [7] with a peak luminance of 1 000 cd/m².
- Display-referred sRGB (SDR) and PQ (HDR) content are displayed as-is.

Target Exposure

The term *target exposure* refers to the fact that a given scene luminance L emitted by a known object (usually based on a target diffuse reflectance R relative to an illuminant) will be mapped around a non-linear encoded signal value Y', whose choice depends on multiple factors ; this value is usually chosen to make the best use of the available encoded signal range. For example, a common reference chosen for diffuse SDR scenes is a diffuse Lambertian surface with reflectance $R = 18\%$, whose encoded (scene-referred) BT.709 signal is mapped close to $Y' = 50\%$.

The use of both HLG and PQ in common HDR formats complicates this decision. Being scene-referred, HLG encodes relative scene luminances in similar ways to BT.709, and as such it is possible to define some notion of a target exposure relative to the scene; however, this is not available directly to PQ-encoded formats which are already altered to be displayed on a given mastering display.

The scene reference used for HDR formats is typically the cutoff point between "HDR" and "SDR" ranges, i.e. a diffuse white Lambertian surfaces with reflectance $R = 100\%$. The corresponding encoded luminance is either 100 cd/m² or 203 cd/m² depending on which recommendation you look at, but a) this has different meanings in HLG and PQ (scene versus display luminance) and b) tested devices can and do choose different references anyway.

Normalization

The last two points bring to light the main challenge in evaluating and comparing image data on both SDR and HDR formats: the different formats don't have the same luminance dynamic range and don't even use it the same way.

A normalization is therefore necessary to compare between these formats; for example, this is the case in the CIE-XYZ space before converting to the color space CIE-L*a*b*. It is common in SDR formats to sidestep the difference between different encodings, by assuming that the scene normalization point is the same as the display normalization point ($Y' = 100\%$ corresponds to the normalization reference on both the scene and the display), but this is no longer possible in HDR encodings.

Normalizing in the middle of the dynamic avoids being affected by non-linearities near 100% diffuse reflectance even though it is a more natural reference for HDR scenes ; as a result, the two main normalization references we tested are:

18% reflectance normalization for example on the luminance of the 18% gray patch of the ColorChecker setup.

100% reflectance normalization for example on the luminance of the white patch (or peak luminance) of the ColorChecker setup.

Analyzing videos requires additional specification of the time frame on which the normalization reference is evaluated. Although more complex methods could be conceived, the two simple methods we tested are:

At the beginning of a sequence assuming the luminance adaptation state of a human observer of the same scene would not change for the duration of the sequence.

On each frame independently assuming the adaptation state is changing rapidly.

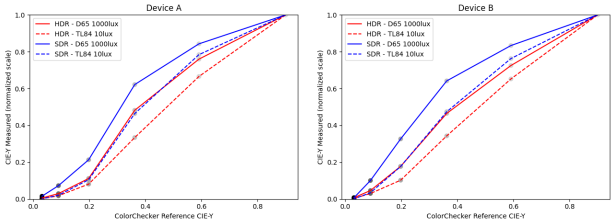
Results

Description of the devices

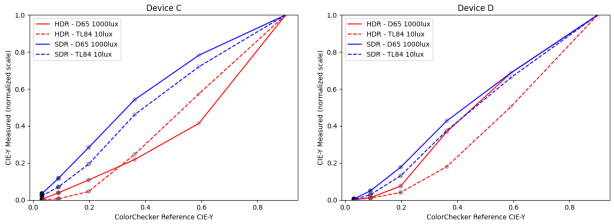
For the tests, we have selected 4 smartphones that can capture both SDR and HDR contents. The specifications of the devices used are defined in the figure 2.

	Device A	Device B	Device C	Device D
Transfer Function	HLG	PQ	HLG	PQ
Peak Transfer Function Luminance	$1000\,cd/m^2$	$10000\,cd/m^2$	$1000\,cd/m^2$	$10000\,cd/m^2$
Peak Normalize Luminance	$1000\,cd/m^2$	$1000\,cd/m^2$	$1000\,cd/m^2$	$4000\,cd/m^2$
Year	2020	2021	2018	2021
Price	> 1000 €	~ 1000 €	~ 400 €	> 1000 €

Figure 2. *Devices specifications*



(a) Device A (b) Device B



(c) Device C (d) Device D

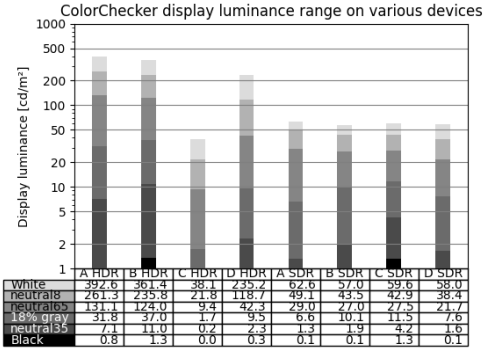
Figure 3. *Luminance tone curves normalized by 100% reflectance for different illuminant on the gray scale patches of the ColorChecker setup*

Luminance in HDR

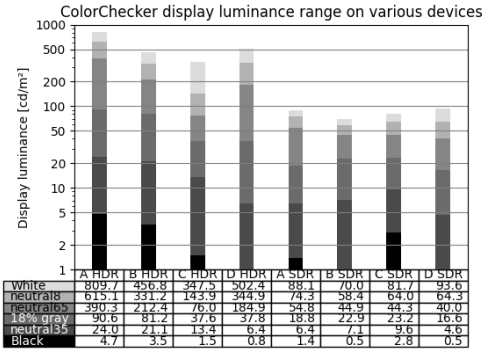
Tone Curves

Figure 3 illustrates the devices luminance tone curves behavior computed on the gray scale patches of the ColorChecker setup. The tone curves are normalized by white diffuse patch to be able to compare the transfer function behavior dependant on the different illuminance (bright light scene at 1000 lux and low light scene at 10 lux) and formats.

Devices A and B have approximately the same behavior by conserving the contrast perception in both formats. However, the luminance of the device C for bright patches in HDR format is lower compare to the other devices. This behavior for the device C in HDR format can also be observed in the figure 4. Indeed, in low light scene at 10 lux, the highlight display luminance of the device in HDR format is lower than in SDR format. Otherwise, we can notice difference up to around 10 ratio between the two formats in midtones and highlights scales. Unlike in shadows tones, there are more significant differences for both HDR and SDR formats.



(a) TL84 10 lux



(b) D65 1 000 lux

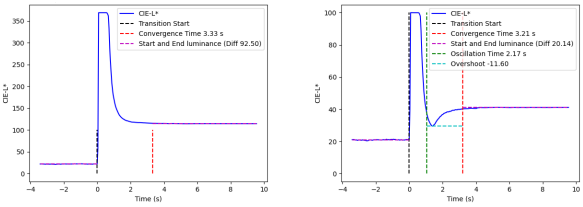
Figure 4. *Gray scale patches display luminance of the ColorChekcer setup.*

Exposure Convergence

The exposure convergence measurement evaluate the temporal exposure behavior of the device during a luminance transition.

The choice of the normalization can have an impact on the amplitude of the transition. The figure 5 presents the temporal behavior of the exposure during a transition for the device C in HDR mode with both normalization methods. The graph on the left presents the CIE-L* behavior of the 18% gray patch, normalized by a factor computed from 18% gray patch average on all frames, then applied identically on all frames. This led to increase the luminance range (L* above 100), hence to the difference in the start and end transition luminance. The graph on the right present the 18% gray patch normalized by a factor computed on white patch for each frame independently. We can observe that the luminance range is well between 0-100 and the difference of luminance between the start and end of the transition has reduced. Nevertheless, the normalization doesn't impact on the convergence times.

The figure 6 presents the convergence time for up (10 lux to 1 000 lux) and down (1 000 lux to 250 lux) transitions for both formats. Most of the devices have convergence times under 1 second in both SDR and HDR formats which corresponds to the Human Visual System (HVS) maximum adaptation time [11], except for the device C. Indeed, the device C has convergences times larger than 1 s or even 2 s which starts to be a long convergence time.



(a) 18% reflectance normalization on average of all frames

(b) 100% reflectance normalization on each frame independently

Figure 5. *Device C exposure convergence graphs in HDR mode for up transition (10 lux to 1 000 lux)*

	Luminance Transitions			
	Inc		Dec	
	HDR	**SDR**	**HDR**	**SDR**
Device A	0.60 s	0.50 s	0.70 s	0.40 s
Device B	0.84 s	0.88 s	0.70 s	0.72 s
Device C	3.21 s	2.13 s	3.83 s	1.70 s
Device D	0.53 s	0.89 s	0.40 s	0.33 s

Figure 6. *Convergence time results for up (10 lux to 1 000 lux) and down (1 000 lux to 250 lux) luminance transitions for HDR and SDR formats.*

Moreover, the convergence times in HDR format are larger than in SDR format which shows a lack of performance in HDR of the device C.

Exposure Stability

Figure 7 presents CIE-L* temporal behavior of the 18% gray patch of the ColorChecker setup for two luminance ramps:

Bright light Ramp D65 800 lux to 50 lux in 20 seconds
Low light Ramp TL84 50 lux to 25 lux in 20 seconds

A white diffuse normalization is necessary to be able to compare the luminance exposure stability between the two formats. There is no significant temporal difference behavior between the formats for all the devices in both ramps. We can notice a small reduction of the CIE-L* for the device C in low light ramp in HDR mode while in SDR it remains constant. But overall, the devices have the same luminance stability behavior in HDR and SDR contents.

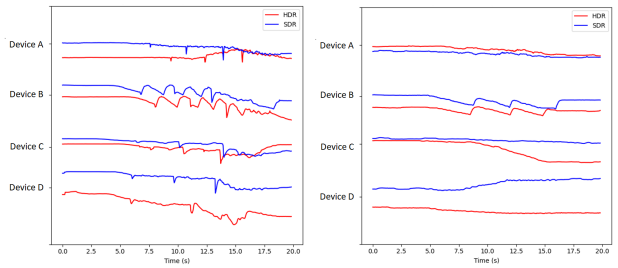
HDR Scenes

HDR Composite

We have tested the HDR composite setup with three lighting conditions:

0EV 10 cd/m^2 on both panels
4EV 10 cd/m^2 on the left panel and 160 cd/m^2 on the right panel
7EV 10 cd/m^2 on the left panel and 1 280 cd/m^2 on the right panel

Figure 8 presents the luminance tone curves computed on the gray scales of both panels for 4 devices, both in HDR and SDR modes. Whereas most devices reach saturation at 7EV, the device A in HDR and SDR modes and the device C in HDR mode do not reach saturation on the tested conditions. However, the constant panel is underexposed but still visible for the device A, so the device is able to render a wider dynamic range in both formats, which is not the case for the device C for which the constant panel is completely black.



(a) 800 lux to 50 lux Ramp (b) 50 lux to 25 lux Ramp

Figure 7. *CIE-L* curves of the 18% gray patch for a luminance ramps*

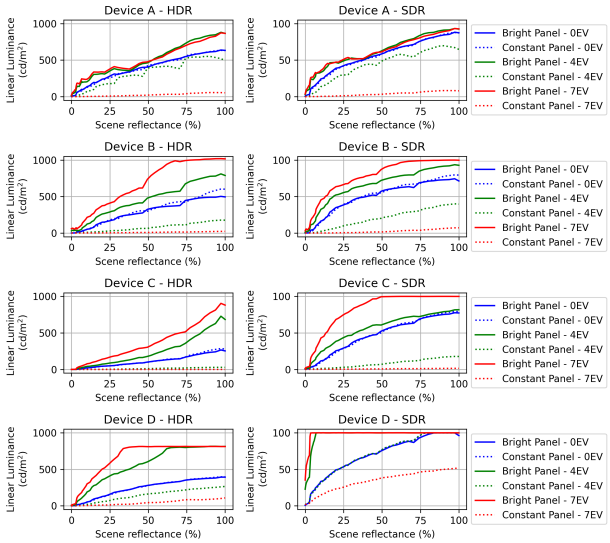


Figure 8. *Luminance tone curves for different illuminant for the HDR composite setup.*

There is no major difference between HDR and SDR for the device A. However for the other devices the highlights are less saturated and the luminance of the constant panel is lower in HDR mode than in SDR mode compared to the maximum luminance. These devices are able to use the HDR mode to better render the dynamic of the scene.

Portrait HDR

We have tested the portrait HDR setup with the same lighting conditions as the portrait HDR setup:

0EV 10 cd/m^2 on the forehead of the mannequin and 10 cd/m^2 on the 18% gray gray patch on the right panel

4EV 10 cd/m^2 on the face (forehead) and 160 cd/m^2 on the 18% gray gray patch on the right panel

7EV 10 cd/m^2 on the face (forehead) and 1 280 cd/m^2 on the 18% gray gray patch on the right panel

Figure 9 presents the luminance tone curves computed on the panel and the face luminance for 4 devices, both in HDR and SDR modes. The device A that preserves the luminance of the face when the lighting on the panel increases, both in HDR and SDR. For all the other devices the face luminance is always lower in HDR mode than in SDR mode compared to the maximum luminance. It means that the devices use the full available dynamic in HDR mode to render the dynamic of the scene.

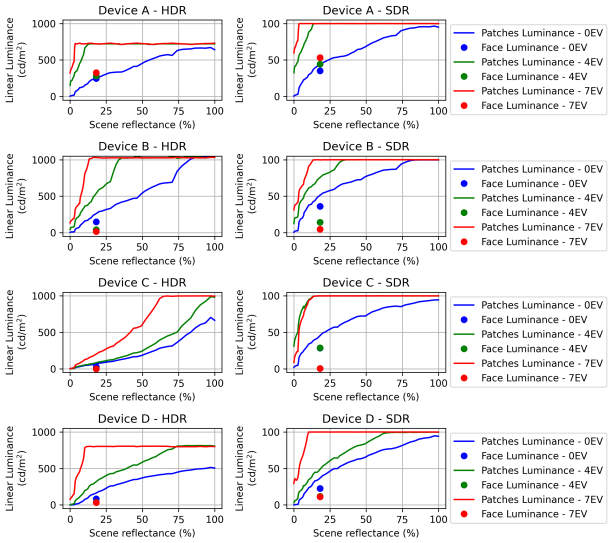


Figure 9. *Luminance tone curves for different illuminant for the portrait HDR setup.*

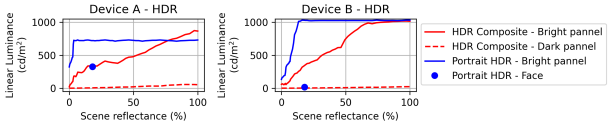


Figure 10. *Comparison of the Portrait HDR (blue) and HDR composite (red) setups at 7EV for 2 devices*

Comparison

The comparison of the portrait HDR and HDR composite setups highlights the impact on the luminance of the presence or not of a face in the scene. For a fair comparison the same lighting conditions have been applied to both setups.

The comparison between Figure 8 and Figure 9 shows that excepted for device D in SDR mode, all devices have more saturation of the bright panel with the portrait HDR setup than with the HDR composite setup. This means that when there is a face in the scene, the devices identify it as the main subject and they preserve the face luminance and saturate other parts of the image. Without a face on the image, the devices compress the tones to prevent any saturation in the image.

Figure 10 highlights the comparison of the tone curves of the 2 setups for 2 devices at 7EV. We can see that for the 2 devices the bright panel in the portrait HDR setup is saturated whereas it is better preserved in the HDR composite setup. We can also see that the device A has a better face preservation than the other devices. The face in portrait HDR mode has the same luminance as the bright panel of the HDR composite setup, whereas for the other devices the face is as underexposed as the dark panel of the HDR composite mode.

Conclusion and future work

Our work has led us to adapt metrics described in our previous work on both SDR and HDR videos on several devices and the development of new setups and measurements for captured HDR videos contents. This new work extends to evaluate objectively with no reference the quality of HDR video captured by any consumer video camera with any HDR format and to compare their performances with SDR videos.

The results of our work shows that there are generally no significant difference between HDR and SDR contents for temporal metrics. A normalization is required to be able to compare the luminance range in several HDR formats which is not necessary in SDR format. Moreover, depending on what is in the scene, some HDR devices prioritize homogenizing luminance across the entire scene, and other devices decide to adjust exposure based on objects in the scene.

Future works can be done to complete the study of the comparison between HDR and SDR format on color. Especially, complete our objective evaluation with a perceptual evaluation of comparing HDR and SDR. Further research can be done to evaluate other color models than CIE-L*a*b* that would better modulate HVS. Moreover, although not shown in this article, we have experimented that flare can be a huge limitation to HDR scene capture. Therefore, evaluating a device's ability to capture an HDR scene without being limited by the flare is necessary.

References

[1] M. Renaudin et al. Towards a quantitative evaluation of multi-imaging systems. Electronic Imaging 2017.

[2] G. Facciolo et al. Quantitative measurement of contrast, texture, color, and noise for digital photography of high dynamic range scenes. Electronic Imaging 2018.

[3] E. Baudin et al. DXOMARK Objective Video Quality Measurements. Electronic Imaging 2020.

[4] Rep. ITU-R BT.2390-9. High dynamic range television for production and international programme exchange. 2021.

[5] Rec. ITU-R BT.709-6. HDTV system with square pixel common image format. 2015.

[6] Rec. ITU-R BT.1886. Reference electro-optical transfer function for flat panel displays used in HDTV studio production. 2011.

[7] Rec. ITU-R BT.2100-2. Image parameter values for high dynamic range television for use in production and international programme exchange. 2018.

[8] Manish Narwaria, Matthieu Perreira da Silva, Patrick Le Callet. HDR-VQM: An Objective Quality Measure for High Dynamic Range Video. Signal Processing: Image Communication, Elsevier, 2015, 35, pp.46-60.

[9] R.Mantiuk, K. J. Kim, A. G. Rempel and W. Heidrich. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. ACM Transactions on Graphics, article no. 40, 2011.

[10] P. B. Delahunt, X.Zhang, D. H. Brainard. Perceptual image quality: Effects of tone characteristics. J Electron Imaging, 2005.

[11] S. Oh, C. Passmore, B. Gold, T. Skilling, S. Pieper, T. Kim, and M. Belska. A framework for auto-exposure subjective comparison. Electronic Imaging, 2017(12):202–208, 2017.