https://doi.org/10.2352/lim.2024.5.1.18

Assessment of HDR-formats: challenges of perceptual evaluation and objective measurements of camera captured contents

Benoit Pochon, Scientific director DXOMARK Image Labs, France

1. Abstract

Recent cameras, especially smartphones, provide HDR formats for capturing videos and photos. For end-users, these formats hold great potential to enhance the visualization experience of captured content on supported displays. Consequently, there is a need to rigorously and objectively evaluate the content produced in HDR-Formats. In this article, we will address the current challenges in perceptual evaluation and objective measurement of camera footage in HDR formats, taking a practical perspective. Based on the results of a perceptual experiment conducted with HDR video formats, we will underline the importance of viewing conditions and signals levels, and list open questions about evaluating HDR still images. In a second part, we will provide an overview of objective measurements for HDR formats with the use of ICtCp.

2. Introduction

Technology improvements in both consumer and industrial cameras have enabled on-device capture and mastering of High Dynamic Range (HDR) content without user intervention.
Smartphones even provide HDR formats for their capture in the default setting. Several formats coexist for video (HDR10, HDR10+, Dolby Vision, Vivid HDR) and are standardized. For still pictures, *ISO standard 22028-5 - High dynamic range and wide color gamut encoding for still images* [1] - has recently been released and the format is expected to be available in cameras soon. Under the impulse of Apple and Adobe, formats incorporating gain maps to manage SDR and HDR in the same file are provided by several manufacturers. However, a universal file format is not yet standardized.

HDR Format and the HDR Experience

The HDR experience is not easy to define, and a proper definition is still under discussion within the industry.

One of the first difficulties relies in the multiple uses of the term High Dynamic Range through the different stages of an imaging pipeline.

Figure 1. Description of the HDR pipeline.

Even though no proper definition of *HDR scenes* exists, let us consider that a scene with a range of luminance of more than 7 EVs is called HDR. The arbitrary threshold of 7 EVs corresponds roughly to the range of luminance that the eye can distinguish instantaneously without adaptation [2].

HDR-Capture&Rendering refers to the set of techniques involved in cameras for decades to capture and render the range of luminance from the *HDR-scenes*. More specifically, *HDR-capture* encompasses hardware and software techniques such as multiframe exposure bracketing, or staggered HDR sensor. As of today, cameras capable of capturing more than 20 EVs of dynamic range are possible. *HDR-Rendering* relates to all techniques aiming at compressing the captured information into a smaller range. This last step, known as tone mapping, is inherent to the fact that the medium on which pictures or videos are displayed, has a limited dynamic range (paper 6-7 EVs, SDR monitors 8 EVs). As such, SDR formats such as sRGB, have benefited for years of *HDR capture&rendering* techniques and can store a large amount of dynamic range of the scene, at the expense of some information/ luminance range compression.
As we see, HDR-scenes and HDR-capture and rendering techniques are not required for the HDR experience, even though they are helpful to fully appreciate it.
HDR-Displays, however, are a required component for the HDR experience. Nevertheless, HDR displays as such are not officially defined nor certified in a general way that encompasses TV, Computers monitors and smartphones. Computer display manufacturers have agreed through the Video Electronics Standards Association (VESA) to define some HDR performance levels. Within the DisplayHDR-500 category, an “HDR” display must fulfill some constraints such as 500+ cd/m² of peak luminance, at least 11.6 EVs of contrast on a white/dark checkerboard, 10-bit inputs, and at least 8 bits of internal processing with some higher frame rate to simulate the two last bits (technically 8+2 FRC).

10-bit input is required to fully benefit from the performance of the display, and this is one of the reason for which HDR image and video formats were defined. *HDR-Format* files contain 10 bits of data, but also the necessary metadata data to help the playback system correctly interpret the content to be displayed, knowing the characteristics of the screen.

As we see, HDR displays deal in the first place with the displayed luminance levels. However, improvement in the color gamut capabilities of the displays has also pushed the HDR format to give the possibility of encoding wider color range than SDR format.

What is in the end the expectation of the HDR experience?

Most of the definitions found in recommendation and standards are relative in the sense that High Dynamic Range provides viewers with an *enhanced* visual experience *compared* to standard dynamic range.

More explicitly, ITU BT2100 [3] describes HDR-TV as an experience that provides images that have been produced to look correct on brighter displays, that provide much brighter highlights, and that provide improved detail in dark areas.
For ISO 22028-5 [1], High dynamic range images allow a greater range of shadow and highlight detail to be conveyed, with sufficient precision and acceptable artifacts, including sufficient separation of diffuse white and specular highlights.

At DXOMARK, and in general in the industry, the use of image quality attributes is a common practice, as they allow us to describe more precisely what the final user perceives.
In those terms, one may summaries the HDR experience benefits as follows:

- Better contrast with brighter highlights (thanks to higher luminance displays
- Fewer artifacts and less quantization in dark tones (thanks to 10 bits encoding)
- More faithful and pleasant colors due to wider gamut encoding.

We may add that brighter highlights do not mean overall brighter images, but only less highlight roll-off compared to SDR images. We shall go back to this later.

Comparing HDR experience performance provided by devices capturing HDR formats can be interpreted from different points of view: one may consider the performance of the system defined by the camera and the output file format, letting the display as a floating variable. However, nowadays, most of the cameras are smartphones, and incorporate a display. One may therefore consider the HDR experience to be inseparable from the display and consider the system to evaluate as the union of camera and display. Both are possible but present different challenges.

3. Perceptual Evaluation of smartphone HDR Video captures

Let us consider here the problem of evaluating the performance of Cameras providing video in HDR format as output. Conducting a perceptual experiment requires rigorous methodology. We describe here the example of a survey conducted by DXOMARK.
First, setting a reference HDR display, and optimal viewing conditions is needed. We use Recommendation ITU-R BT.2408-7 [4] for HDR display settings, and ITU-R BT.2100 for environmental conditions.
Explicitly, the monitor used is an Apple XDR display set at 1000 cd/m² peak brightness and 203 cd/m² SDR white level, with a PQ EOTF. Surround luminance is set to 5 cd/m² thanks to adjustable light source behind the display and appropriate grey background. Viewing distance to the display is carefully monitored.
The perceptual experiment consists of different stages:

- a pairwise evaluation where pairs of short videos captured with different cameras are shown side by side to users, filling 75% of the screen surface. Background gray of the interface is set to 5 cd/m². The user must choose his favorite video based on a general guideline: “Select the video you prefer, based solely on video quality”. The videos of the devices are all anonymized and proposed following an Active Sample Pairwise comparison algorithm (ASAP [5]). Following [6], the output of this stage is a metric per device and per scene, on a JOD scale.
- a rejection stage where the users must select none, one or more videos they deem as unacceptable. All videos are displayed side by side in a random order at a size of half the screen. A video is considered unacceptable if the user would opt to delete it or refrain from sharing it with friends, family, or on social media platforms due to poor image quality. The user also provides further details or reasons for their rejection, choosing among a list of criteria, such as Brightness is too high / low, Contrast is too high / low, Colors are not natural, etc.
- an additional question stage where, for each scene, the video with the highest JOD score and no rejections is displayed. In a first question, the user selects one or more reasons for the preference, and in a second one the user is asked about what he would like to be improve in the video, among a list of possibilities.

A total of 90 scenes were shot with 4 different cameras: 3 flagships smartphones of 2023, and a semi-professional DSLR camera recording raw footage. The scenes all include people, carefully chosen to include a large variety of skin tones. Scenes are covering outdoor, indoor, lowlight, and night use cases, with a mix of SDR, HDR and backlit conditions.
The RAW footages are manually processed in a HDR workflow by a professional grader. Three versions are produced ranging from a pure cinematographic standard style sticking to ITU-recommendation, to a look-and-feel more comparable to smartphone outputs.
A total of 41 people conducted the perceptual evaluation, among which all the models present in the videos. Those were consumers recruited for the purpose of the experiment and without expertise in image quality. The rest of annotators were photographers and image quality experts, representing 40% of the panel.
We present here a short extract of the results, without focusing on which device performs better than the others.
As can be seen in **figure 1a,b,c**, overall brightness is generally the most prevalent criteria cited by annotators for explaining the rejection and the preference. Among the priorities to improve on the preferred videos, we note that people cited both higher overall brightness level and skin tones rendering ex-aequo. Balance between highlights and dark tones (an important feature of the HDR experience) is secondary and did not show a consensus as almost the same proportion of people find the contrast to be too strong or too low.

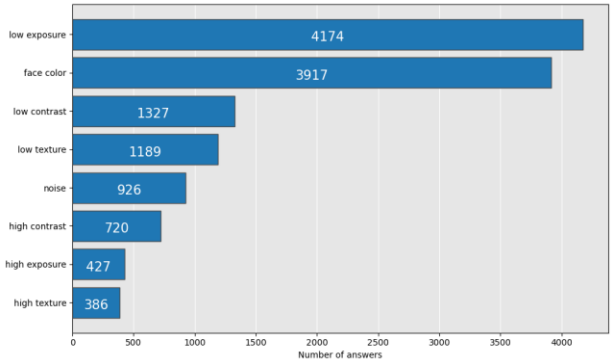


Figure 1a. *Specific reasons for rejecting a scene rendering*

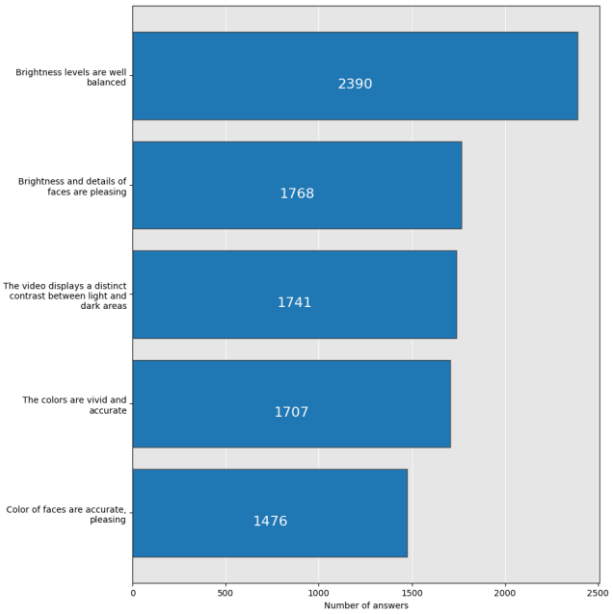


Figure 1b. *Specific reasons for preferred choice of scene renderings*

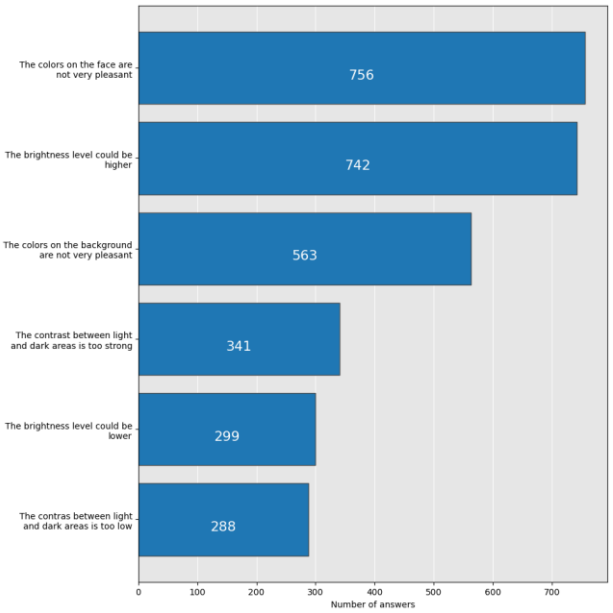
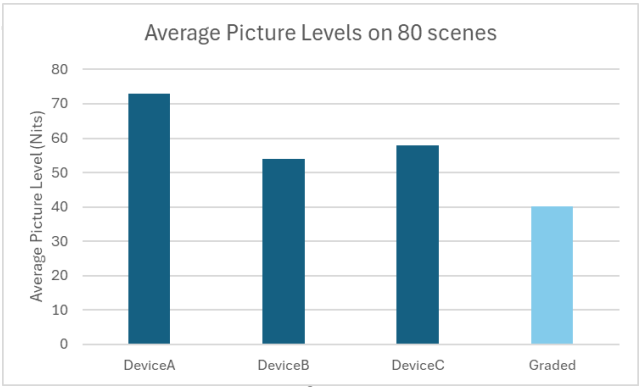


Figure 1c. *Improvements suggested of scene renderings*

Graded footage was systematically among the least preferred, except for a few very challenging backlit scenes.
As we can see on **figure 2**, smartphone videos are indeed significantly brighter than graded footage, confirming a difficulty in the evaluation of performance of HDR format: brighter is often preferred, and might eclipse more subtle difference, especially in a pairwise experiment.

One might argue that this problem is marginal as people rarely look at videos side-by-side in practice, and that a different methodology than pairwise should be used to avoid the brightness bias in the evaluation of the quality of HDR format. Nevertheless, correct brightness level for videos of smartphones is currently an open question and there is a risk that inhomogeneity of levels troubles the final user.



11. ***Figure 2.*** *Average picture levels of footage for different cameras*

12. Challenges for HDR stills perceptual evaluation

We do not look at pictures the same way we look at videos. Videos are more immersive: being in a cinema theater, at home in front of a TV, or on a smartphone, videos are most often seen in full screen. Photos on the contrary are often browsed within an interface, which can be a photo player, a social media app, or simply a web-browser. Displaying multiple pictures side-by-side and simultaneously is common.
This introduces new challenges for the perceptual evaluation of HDR formats.

It is well known that the Human Vision System works much more in a relative way than an absolute way. Therefore, when visible, the background of an image can have a strong influence on our perception: when a picture is displayed on a dark background, there is no other reference luminance than the picture. On the contrary when displayed above a white background – sufficiently large and visible – the perceived brightness is relative to that background.
This is the reason why Standards define the reference white, a crucial parameter for the HDR experience, especially for perceptual evaluation of still-HDR format. The reference white usually coincides with the graphical white of the interface and is set to 203 cd/m² in the reference or mastering display configuration. This value is to be compared to the peak white at 1000 cd/m² for this reference display.

This leads to a list of open questions related to perceptual evaluation of still HDR formats:

- *Should one display a reference white when comparing side by side HDR-format pictures from the same scene?*

As we saw with the video experiment, brighter is preferred. If no reference is provided the brightest overall image might be preferred, regardless of more subtle differences such as brighter specular highlights, wider color gamut, or reduced artifacts in the dark tones. Focusing on preference for tuning camera may lead to a constant one-upmanship, or even a brightness war, analog to the loudness war in audio, and potentially to the loss of one of the HDR promises, which is less highlight roll-off compared to SDR. On the other hand, one might argue that providing a reference white is a fuzzy and arbitrary practice: what should be the size of the reference white area, and can we really be sure that people will take it into account, knowing the complexity of the HVS?

- *When comparing side by side pictures, should a normalization of the brightness be performed?*

This solution is appealing but raises the question of the type of normalization. One may opt for a normalization of the overall brightness of the image, such as the Average Light Level. Studies showed that other estimates are more correlated with the perception of brightness [12]. An alternative option, inspired by ITU recommendation is to normalize according to the displayed luminance of a diffuse white patch in the scene. The diffuse white patch in the scene could be adjusted to the reference white level of 203 cd/m². However this method requires an estimate of the white patch in the scene, an ill-pose problem since ambiguity may arise for scenes with non-well-defined subject or multiple light condition.

- *How to compare HDR and SDR format?*

Although such a comparison is not recommended, it is difficult to prevent people from doing it. As we saw, the influence of brightness in a comparison can easily introduce biases. For a fair comparison, the SDR content might be normalized to similar light level as HDR. Nevertheless, there are many ways to do it, and it is not easy to define the limit in complexity for such a normalization. Content based local tone mapping would be considered as intrusive with respect to the artistic intent. Global non-linear tone mapping algorithms are available such as in ITU BT2446 [7], but none of them seems to be universal. A simple scaling in linear display light is potentially the best tradeoff between complexity and performance. But choosing the right scaling factor might be content dependent and so not obvious to choose.

Objective Measurements of HDR-Format Contents

Perceptual analysis is important to compare cameras but is in essence relative. There is also a need for characterization and measurements of HDR-format contents in a non-relative manner.

Since HDR-Format defines the way videos should be displayed, the measurement should be done in display linear light.

But what display should be considered to do the measurement?

Fortunately, HDR-Format always contains metadata to define the mastering display on which the content should be displayed. These metadata are always available in the base layer, and therefore do not require to interpret proprietary metadata. The *de facto* solution to do measurement is therefore to simulate linear light coming out of the mastering display and operate on this signal.

HDR contents viewed on such a mastering display and in optimal viewing conditions trigger rather different adaption states of the human vision system. This makes Lab, classically used in SDR objective measurements, unsuitable for HDR-Format measurements. The reasons are multiple:

- CIE Lab supposes a fixed steady state of adaptation.
- This state of adaptation is determined relatively to a white point, reflective diffuse point.
- CIE Lab is validated for levels below the white point and for contrast not exceeding 1:100.

As SDR displays were designed as emulation of print, the white point of CIE Lab for SDR measurement is set to the peak white of the display. This is not possible for HDR display since levels around the peak white of the display are dedicated to render specular or emissive high lights. The alternative idea is to set the white point of CIE Lab to the reference/graphical white of the display, but that would generate value of L* above 100, an area where the color space was never designed to work.

Many color appearance models and associated uniform color spaces are available to quantify perception in the high range of luminance and large color gamut achievable with a mastering display in the optimal viewing conditions. It seems nevertheless that a consensus in the industry has been reached to use ICtCp color space [8][3]. ICtCp has several advantages:

- It provides an exact mapping with absolute luminance, which is convenient for linear display light output.
- It is approximately uniform locally (with the necessary scaling of Ct)
- It reaches the optimal accuracy, since the derived JND at each point of the color space is obtained in the optimal state of the HVS (this state being potentially different for different point in the color space).

Let's see from a very high point of view how objective measurements of different image quality attributes could be performed by using ICtCp.

Brightness levels target acceptance

Doing objective measurements of brightness output levels of HDR formats and setting related acceptance targets is not as easy as it seems, especially when auto-exposure is involved.

First, one needs to choose a stimulus, or what we call a lab scene. It could be a simple color checker on a grey background, but it could also be laboratory scenes with much higher dynamics, such for instance scenes with backlit panels to simulate controllable highlights. In this case, we recommend inserting some objects with semantic interpretation such as faces.

Once stimuli are chosen, the next challenge is the design of the acceptance targets. To avoid any arbitrary choice, these should be explainable. This leads to an open question about the correct brightness level that a smartphone camera should follow: broadcast industry provides recommendations to achieve brightness homogeneity among HDR programs. Cinema industry provides much less guidelines and relies mostly on artistic intent for the choice of brightness levels during a movie grading. Smartphone industry stands in a blurred area: captured videos are assumed to be ready to post on social media/online video sharing platforms, which are neither broadcast, nor cinema but may require some sort of homogeneity. The widespread choice of HLG as a transfer function in the smartphone industry suggests that smartphones manufacturers want to follow an approach like broadcast industry, with homogeneity of brightness levels in mind.

It seems however that the levels of footage are different from the one of the broadcast industry. For a purely reflective scene, such as a color checker chart over a gray background, we would expect the reflectance of white patches to be below the reference white levels. As a matter of fact, it is not the case, as smartphone flagships in 2023 have white patch levels ranging from 260 to 520 cd/m² in photo-HDR format, and up to 840 cd/m² in video-HDR formats, quite above the 203 cd/m².

To determine the correct output brightness levels, an alternative to the recommendation is to set the acceptance around the average of the industry, but as we can see, a consensus has not

been reached yet. A third alternative would be to keep a wide acceptance target since acceptability depends on the state of adaptation of the user, which in turn depends among others on the presence or absence of a reference white when looking at the video.

Objective Color Characterization

In our previous work[9] we propose a framework to analyze objectively color rendering of HDR formats and SDR formats. The methodology relies on the hypothesis that free linear scaling (e.g. in CIEXYZ) results in the same perception of an image when adaptation reaches steady-state. Color differences are analyzed in two steps: 1. scale the reference color in the scene in the XYZ domain so that it will be comparable to the luminance on a display. 2. Calculate the color difference with the scaled references using ΔE_{ITP} or ΔC_{ITP}, derived from ITPJND. The ITPJND scale defines "potentially visible" color differences in the most sensitive adaptation state, unlike CIELAB which uses a fixed known adaptation state. Thresholds of acceptability are therefore an order of magnitude higher than ΔE_{76}, as shown in Table 1.

ΔE_{ITP}	Description
≤ 5	Very good, strict color match
≤ 10	Good match for achromatic patches
≤ 20	Good match for color patches
= 50	Luminance increments/decrements of ≈ 1 to 1.5 EVs

Table 1. *Indicative color accuracy thresholds for* ΔE_{ITP}

This methodology makes the hypothesis that an adaptation, locally to a patch or a small area, is performed. To estimate color differences with multiple areas, ie larger charts, such as a color checker, a parametric model of color rendering is proposed. [9] propose to fit a 6 parameter model by minimizing ΔE_{ITP} on patches, and open the possibility to describe the color rendering in attributes parameters such as brightness, saturation, casts... with the intention to simplify the problem. However, acceptance thresholds in this space of parameters are still an open problem. If the fitting does not match, it may be that the model is too simple, or that some specific colors do not follow the consensus. In the later case, one may roll back to a per patch analysis.

Texture and Noise objective measurements.

The texture objective measurements as described in ISO/TS 19567-2 [10] compare the capture of a stochastic pattern chart (deadleaves charts) to a reference description using Fourier analysis in the linear domain. Linearization could be "scene referred" by measuring the OECF of the camera or "display-referred" by using the tone curve information of the output colour space. This last configuration should be used for HDR-format. Since Deadleaves patterns have a limited range of contrast (4:1), the method is a good approximation around a given region of the luminance range, if one considers the user to be adapted to this zone. Since HDR-format covers a large range of luminance, it could be wise to apply the method around different levels instead of only the center of 8 bits dynamic as in SDR. However, the method may fall short if one wants to characterize objectively the texture perception in areas of an HDR-image when a viewer that can observe dark areas and bright areas at the same time. This use case, which was not considered with SDR format because it is unlikely to happen, is an open question as far as we know.

ISO/TS 15739 Annex B [11] describes a procedure for computing an output-referred noise metric called Visual Noise, which uses a human visual model that aims to predict the perceived quality of the image. Although sRGB is explicitly mentioned to retrieve the display linear light, the same method could be applied for HDR formats with application of the proper EOTF. In the standard, signals are converted to normalized XYZ with a SDR white point at 80 cd/m², before being converted to AC_1C_2 for CSF application, and then CIELAB color space with D65 white point. Standard deviations are computed in L*a*b* space with appropriate weight for a* and b* for the final calculation. The same ambiguity as for brightness and color objective measurements arises for the choice of the white point normalization with HDR-Formats. Using the HDR display peak white would probably results in inaccurate measurements for low levels, while using the reference white would trigger invalids Lab values for highlights. An approach to explore would be to transpose the visual noise calculation in ICtCp. The weight for Ct and Cp used in the final formula could be estimated on a set of SDR images, by least square minimization to give visual noise value like the one computed in the Lab space. However, as for texture, one may find noise more or less objectionable depending on the state of adaptation: a small noise in a dark frame viewed in a dark room is more visible and objectionable than the same noise in a frame with high contrast. Adapting Visual Noise measurement to with new CSF and ICtCp needs to be studied.

HDR-format evaluation in practical use cases

The approach presented up to now considers performing perceptual evaluation or objective measurements in optimal viewing conditions, which is in the end a similar approach to what was done for SDR. However, with SDR format, a quite large deviation from the optimal conditions would not make the perception of the image very different.

This is not the case for HDR: as soon as one deviates from the optimal conditions, evaluation of HDR can change significantly.

Optimal viewing conditions are difficult to reach and far from the actual use case. One could emulate different playback and viewing conditions. For objective evaluation, this requires interpreting all metadata, including the potential licensed ones, and to introduce more parameters to model the display, the playback system and the viewing environment.

With the advent of smartphones providing high performance displays, people watch HDR contents in environments that are very uncontrolled. Smartphones manufacturers develop algorithms aggregating information from light sensor and distance of viewing to adjust at the playback level the brightness of the display and potentially its EOTF.

To evaluate the HDR experience on smartphones, there is also a need to take the whole pipeline into account, for instance by comparing footage captured and displayed on the same smartphones. Although this approach mixes a lot of aspects and does not dispense to evaluate capture separately, it probably aligns more closely with practical use cases of end users.

Conclusion

Evaluation of HDR content, being perceptual or objective, requires a normalization of brightness levels for a fair evaluation. As of now, levels are not homogenous and standards are loose on this aspect, mainly because it depends on many factors, such as viewing conditions, state of adaptation of the user, and artistic intent of the grader or smartphone processing. Once normalization of levels is done, SDR-image quality evaluation methods can be applied in a first approximation for local IQ attributes, if we are not looking simultaneously at areas with big contrast differences. An additional question is to quantify the performance of smartphone devices offering HDR capture under practical usage scenarios, including viewing the content on the device. This requires a glass-to-glass evaluation, which assesses not only the camera's performance but also the adjustments that other elements in the chain, such as the playback system, must provide to adapt to the practical viewing conditions of end users.

13. References

[1] ISO/TS 22028-5:2023 - High dynamic range and wide colour gamut encoding for still images

[2] Darmont, Arnaud. "High dynamic range imaging: sensors and architectures." *SPIE* (2013).

[3] International Telecommunication Union. Recommendation ITU-R BT.2100-2 - Image Parameter Values for High Dynamic Range Television for Use in Production and International Programme Exchange. 2018.

[4] International Telecommunication Union. Recommendation ITU-R BT.2408-7 - Guidance for operational practices in HDR television production. 2023.

[5] A. Mikhailiuk, C. Wilmot, M. Perez-Ortiz, D. Yue and R. K. Mantiuk, 2020. "Active Sampling for Pairwise Comparisons via Approximate Message Passing and Information Gain Maximization", International Conf on Pattern Recognition (ICPR)

[6] M Perez-Ortiz and R K Mantiuk. "A practical guide and software for analyzing pairwise comparison experiments". 2017.

[7] International Telecommunication Union. Recommendation ITU-R BT.2446-1 - Methods for conversion of high dynamic range content to standard dynamic range content and vice-versa. 2021.

[8] Dolby Laboratories. ICtCp White Paper. URL: https://professional.dolby.com/siteassets/pdfs/ictcp_dolbywhitepaper_v071.pdf.

[9] Anna Tigranyan, Paul Mathieu, Corentin Nannini, Franc¸ois-Xavier Thomas, "Objective color characterization of HDR videos captured by smartphones: laboratory setups and analysis framework". In: Electronic Imaging 2024

[10] ISO/TS 19567-2:2019 Photography — Digital cameras — Part 2: Texture analysis using stochastic pattern

[11] ISO 15739:2023 - Photography — Electronic still-picture imaging - Noise measurements

[12] Quantitative Evaluation and Attribute of Overall Brightness in a High Dynamic Range World DOI:10.5594/JMI.2019.2940860 (2019)

Author Biography

Benoit Pochon *received his Master's degree in engineering from Centrale Supelec (2001) and his Master's degree in Electrical Engineering from GeorgiaTech University (2001). After several years working in the signal processing domain, he joined DXOMARK Image labs in 2017, as image science director.*