Impacts of viewing conditions on HDR-VDP2

Maxime Rousselot, Éric Auffret, Xavier Ducloux
Harmonic Inc.
Rennes, France
maxime.rousselot@harmonicinc.com

Olivier Le Meur
Univ Rennes, CNRS, IRISA
Rennes, France
olivier.le_meur@irisa.fr

Rémi Cozot
Univ Rennes, CNRS, IRISA
Rennes, France
remi.cozot@irisa.fr

***Abstract*—HDR (High Dynamic Range) and WCG (Wide Color Gamut) increase significantly quality of viewing experience by rendering impressive images and videos. Automatic assessing the quality of these HDR WCG images is one crucial objective in broadcast process. Full-reference HDR metrics have been designed in the last years to achieve this objective:HDR-VDP2, HDR-VQM, PU-encoding metrics. Recent studies have pointed out that HDR-VDP2 is one of the best metric. Unfortunately, HDR-VDP2 is quite complex to use due to numerous and sometimes hard-to-know parameters such as display emission spectrum, surround luminance and angular resolution. In this paper, we show that HDR-VDP2 does not require an accurate knowledge of the viewing condition parameters. For that, we not only test the impact of these parameters on existing image databases of subjective quality scores, but also we propose a new and complementary image database made with a different HDR display.**

***Index Terms*—Image quality, Image database, Dynamic range**

I. Introduction

HDR (High Dynamic Range) and WCG (Wide Color Gamut) images/videos became in the last years an active field of research and standardization. HDR displays are capable of preserving more details in the bright and dark area of an image by increasing the peak brightness and lowering the black point brightness displayable. Some of these displays are also WCG, meaning that the available gamut is significantly bigger, allowing to display more saturated color. Assessing the quality of HDR and WCG content is a huge challenge. New metrics have been designed to this end. We can mention the HDR-VQM metrics [18] dedicated to HDR video and HDR-VDP2 [15] for still images. Some existing metrics used to assess the quality of SDR (Standard Dynamic Range) images have been also adapted to HDR by using a new encoding method called PU (Perceptual Unit) [1].

These quality metrics are crucial in order to ensure a good quality of broadcasting. Previous studies suggested that the full-reference HDR-VDP2 metric is one of the best existing metrics for compression artifacts. For example, in [5], the authors assessed the performance of 35 quality metrics over 240 images compressed with JPEG XT [19]. They concluded that HDR-VDP2 (version 2.2.1 [17]) and HDR-VQM metrics [18] were the best performing metrics. In [6], authors came to the conclusion that HDR-VDP2 (but in an earlier version 2.1.1) can be successfully used for predicting the quality of video pair comparison contrary to HDR-VQM. More recently, Zerman et al. [24], first, combined several existing image databases and, second, they found out that HDR-VQM is the best full-reference HDR quality metric, closely followed by the HDR-VDP2.2.1 metric which gives similar results when one particular database is discarded.

Although that HDR-VDP2 is often considered as one of the best HDR metric, it is also important to point out that, in some cases, simpler full-reference metrics perform as well as HDR-VDP2. For instance, in [4], authors showed that HDR-VQM, PU-encoding VIF [20] and PU-SSIM provide similar performances as HDR-VDP2. In [23], results indicate that PU-VIF and HDR-VDP2 have similar performances, although PU-VIF have a slightly better reliability than HDR-VDP2 for lower scores. Considering that HDR-VDP2 has similar results than much simpler metrics in some specific contexts, on one hand, and on the other hand, the complexity to use HDR-VDP2 which requires a number of parameters, we may ask whether it is worth to use this metric.

In this paper, we focus on HDR-VDP2 (version 2.2.1) metric. This metric is a bio-inspired metric that models accurately the early stages of the human vision such as intra-ocular scattering, spectral sensitivity of human eye photo-receptor and luminance masking. In addition, HDR-VDP2 requires several input parameters in order to define the visual environment in which image is seen. These important parameters are the surround luminance, the display spectral emission and the angular resolution. We investigate the sensibility to input parameters and the limitations of HDR-VDP2 by analyzing its performance both on existing HDR database image annotated with quality score and on a new and complementary database. The contributions of this paper are twofold:

- determining the influence of input parameters on the performance of HDR-VDP2;
- providing a new and complementary database for evaluating the quality of existing HDR metrics.

The paper is organized as follows. In Section II, we present existing image databases that are classically used for assessing the performance of HDR quality metrics. We also introduce our new database. Section III presents a new analysis of HDR-VDP2, revealing that HDR-VDP2 is not sensitive to a number of input parameters. The last section concludes the paper.

II. Databases presentation

A. Existing databases

In this section, three HDR image databases annotated with quality scores are presented. Table I presents the main char-

ISBN 978-90-827970-1-5 © EURASIP 2018

acteristics of these databases; a short description is also given below:

- Narwaria et al. [16][1]'s database is composed of 27 source images, which have been distorted by a backward compatible scheme, meaning that the HDR is first tone-mapped to a standard dynamic range (SDR) and compressed using a standard JPEG codec. Images are then tone-expanded before being displayed on screen. HDR-VDP2 is the best performing metric for this database.
- Korshunov et al. [13][2]'s database consists in images distorted using the backward-compatible JPEG-XT standard. HDR-VDP2 is also the best quality metric for this database, although without significant difference with HDR-VQM [5].
- Zerman et al. [24][3]'s database is partially composed of images from [22]. The distorted images are generated by using both backward-compatible distortion with codec JPEG, JPEG2000 and using a non backward-compatible scheme with the use of the Perceptual Quantizer (PQ) for the electro-optical transfer function (EOTF) first normalized in the SMPTE ST 2084 [21].

These three subjective experiments have been carried on using a SIM2 HDR47ES4MB display. This display is able to render values between 0.03 and 4000 cd/m^2.

As mentioned earlier, the most used HDR databases have been built with the same SIM2 HDR47ES4MB display. We believe that this is a first limitation of existing databases. Moreover, the gamut used to encode the images is always the legacy gamut, although a lot of HDR images and videos, even BT.709 [8] content, will be encapsulated in a wider gamut like BT.2020 [9]. Finally, as distortions included in current databases are mainly luminance-based distortions, it would be helpful to put more emphasis on color degradations. To overcome the aforementioned limitations, we propose a new database by using an other display, new kinds of distortions and BT.2020 encapsulation format.

B. A new database[4]

1) Content creation: Eight images were selected from 3 collections: two are from the MPEG HDR sequences (FireEater and Market) [14], one is from the Stuttgart HDR Video Database [3] and the remaining five images comes from HDR photographic survey [2]. Note that these images also belong to Zerman et al.'s database [24]. All these images have been encapsulated in the WCG gamut BT.2020 [9] instead of the standard gamut BT.709 [8].

Four kinds of distortions have been chosen:

- HEVC compression using the recommendation ITU-T H Suppl.15 [12]. This means a PQ encoding [10] and a 4:2:0 chroma sub-sampling using a luma-adjustment process. Four different quantizers (Qp) were selected for each image. The chroma Qp offset algorithm is used to overcome a compression issue with the chroma component Cr and Cb in HDR/WCG. In WCG, most of chroma values tend to be near their mean value (i.e. 512). This is even more true for BT.709 content encapsulated in a BT.2020 gamut. This kind of encapsulation creates color artifacts when the chroma components do not have enough bits allocated to it.
- HEVC compression without the chroma Qp offset algorithm as described above. Three Qp were selected for each image.
- Gaussian noise on the chroma components: 3 levels of noise were selected.
- Gamut mismatch: two degradations were created for this category. On one hand, the BT.709 images were considered as if they had been already encapsulated in a BT.2020 gamut creating more saturated images. On the other hand, we took images already encapsulated in a BT.2020 gamut and considered them as BT.709 images and re-encapsulated them in a BT.2020 gamut. This creates less saturated images.

2) Protocol: Participants scored the quality of images by using the Double-Stimulus Impairment Scale (DSIS) variant I methodology [7] with a side-by-side comparison. A continuous scale form 0 to 100 was used: 100 meaning imperceptible, 75 perceptible but not annoying, 50 slightly annoying, 25 annoying and 0 very annoying. Because of the side-by-side comparison, the images were cropped to 944×1080 with 32 black pixels in the middle.

Fifteen naive subjects participated in this test (11 male, 4 female) with an average age of 25.8. All declared normal or corrected-to-normal vision. One participant was removed from the analysis using the methodology described in [7].

Image pairs were displayed on Sony BVM-X300 monitor with a peak brightness measured at 1000 cd/m^2. The luminance of a black pixel was too low to be measured by our equipment ($<$0.2 cd/m^2). The distance between the participant and the screen was 3.2 times the picture height, leading to an angular resolution of approximately 60 pixels per degree.

Ninety-six pairs of images were presented to the viewers, one side being always the reference. 50% of the participant had the reference always on the right-hand side, 50% always on the left-hand side. To avoid a bias with the order of presentation, the pairs of images were randomized for each participant with the condition that the same content was never shown consecutively. Each image pair was shown 10 seconds and voting time was 5 seconds. The test session lasts 35 minutes (including instructions and training time) with a 5 minutes pause in the middle of the test.

The distribution of the collected subjective scores, illustrated on Fig. 1, is almost uniform, indicating an appropriate choice of the different levels of distortion.

3) Objective metric: Because of the limitation in luminance of our equipment, we cropped the luminance range of the images between 0 and 1000 cd/m^2.

[1]available at http://ivc.univ-nantes.fr/en/databases/JPEG_HDR_Images/
[2]available at http://mmspg. epfl.ch/jpegxt-hdr
[3]available at http://webpages.l2s.centralesupelec.fr/perso/giuseppe.valenzise/
[4]available at http://www-percept.irisa.fr

ISBN 978-90-827970-1-5 © EURASIP 2018

TABLE I: Existing HDR databases and the proposed one.

Name	#Obs	#Img	Protocol	Distortion	Display	Gamut	Surround luminance	Angular resolution
Narwaria et al. [16]	27	140	ACR-HR	JPEG	SIM2 HDR	BT.709	$130cd/m^2$	60 pix/deg
Korshunov et al. [13]	24	240	DSIS (side by side)	JPEG-XT	SIM2 HDR	BT.709	$20cd/m^2$	60 pix/deg
Zerman et al. [24]	15	100	DSIS (side by side)	JPEG, JPEG-XT JPEG2000	SIM2 HDR	BT.709	$20cd/m^2$	40 pix/deg
Proposed one	15	96	DSIS (side by side)	HEVC, Gaussian noise Gamut mismatch	Sony BVM-X300	BT.2020	$40cd/m^2$	60 pix/deg

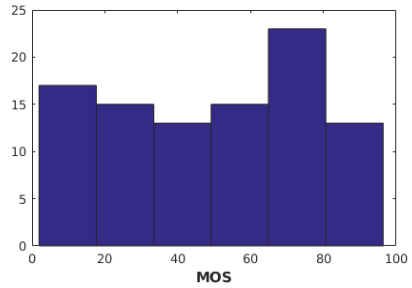

Fig. 1: MOS Distribution for the proposed database.

TABLE II: Performances on the proposed database.

Metric	PCC	SROCC	OR	RMSE
HDR-VDP2	0.89	0.87	0.46	12.5
HDR-VQM	0.85	0.83	0.56	14.7
PU-VIF	0.77	0.75	0.58	17.7
PU-MSSSIM	0.89	0.87	0.46	12.2
PU-SSIM	0.71	0.70	0.59	19.7
PU-PSNR	0.80	0.78	0.52	16.7

The performances were evaluated using four metrics: the Pearson Correlation Coefficient (PCC), The Spearman Rank Order Coefficient (SROCC), the Outlier Ratio (OR) and The Root Mean Square Error (RMSE). These performance metrics were calculated after mapping the objective metrics on the subjective score. This is done by applying a non-linear regression on the objective metric scores using a logistic function describe by Equation 1.

$$Y = a + \frac{b}{1 + e^{-(cX+d)}} \tag{1}$$

where Y is the objective score map on the subjective score, X the objective score and a, b, c and d the parameters determined by the regression.

Table II gives the performance of HDR-VDP2, HDR-VQM, PU-VIF, PU-MSSIM on the proposed database. HDR-VDP2 and PU-MSSIM performs the best even if these performances are not significantly better than HDR-VQM (for all metrics except for RMSE). The significance was estimated using the methods described in [11]. Both of them have a higher correlation than PU-VIF, PU-SSIM and PU-PSNR.

III. Comprehensive analysis of HDR-VDP2

In this section, we elaborate on the performance of HDR-VDP2 on the proposed database. Then, we evaluate the influence of the user-defined parameters on the ability of the metric to predict quality scores.

TABLE III: HDR-VDP2 scores on compression degradations.

Metric	PCC	SROCC	OR	RMSE
HDR-VDP2	0.90	0.87	0.45	9.3
HDR-VQM	0.85	0.83	0.61	11.45
PU-VIF	0.90	0.88	0.36	9.25
PU-MSSSIM	0.92	0.90	0.46	11.0
PU-SSIM	0.65	0.63	0.75	21.7
PU-PSNR	0.83	0.82	0.54	15.9

A. Detailed analysis of HDR-VDP2 on the proposed database

In the previous section II-B3, Table II gives the overall performances of HDR-VDP2 on the proposed database. In order to go further in the analysis, we first plot the MOS in function of the corrected HDR-VDP2 scores as illustrated by Figure 2. Several observations can be made. First, the HDR-VDP2 metric behaves quite well for the different kinds of degradations. When considering the compression artifacts only (blue and red dots in Figure 2), this metric performs well. Table III presents the performance of HDR-VDP2 as well as HDR-VQM and PU-VIF for the proposed compression artifacts. We observe that PU-VIF and HDR-VDP2 have similar performances. HDR-VQM is the worst performing metric. The difference is significantly lower in term of OR with PU-VIF and in term of RMSE with both HDR-VDP2 and PU-VIF.

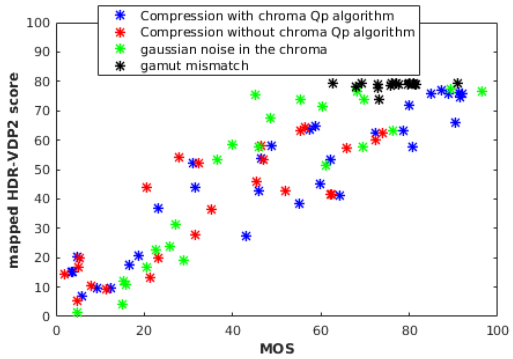

Fig. 2: HDR-VDP2 scores in function of the MOS.

We also observe that the MOS scores obtained for compressed images with and without Qp offset method are strongly correlated (pcc=0.95). This suggests that observers are consistent in their quality scoring whether or not the Qp offset method is used. One exception concerns the RedwoodSunset image, for which the use of Qp offset brings a significant

ISBN 978-90-827970-1-5 © EURASIP 2018

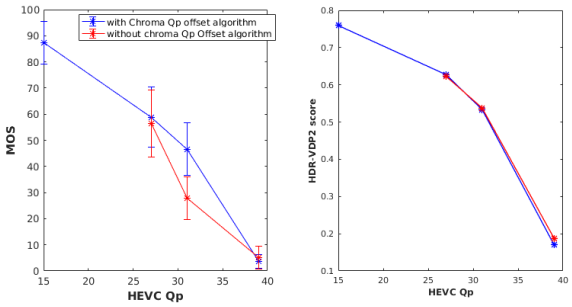

Fig. 3: Subjective and objective scores in function of HEVC Qp for the RedwoodSunset image.

add-value, as illustrated by Figure 3. In this case, HDR-VDP2 metric provides exactly the same score.

Despite the fact that HDR-VDP2 is based on the luminance channel only, HDR-VDP2 turns out to be quite resilient to chromatic distortions, except for the default gamut mismatch. This default was reported quite hard to evaluate by naïve observers. A rather strong color distortion that does not change the global aesthetic of the scenes can be assessed positively by observers.

B. Simplifying the usage of HDR-VDP2

1) Sensitivity to the screen spectral emission.: In this section, we will discuss about the sensitivity of HDR-VDP2 to the spectral emission of the display. For this purpose, we measured the spectral emission of 5 HDR displays, two Led LCD displays (Sony KD-75X9405C and SIM2 HDR47ES4MB) and three OLED displays (LG OLED 65 E6V,Sony BVM-X300 and Loewe Bild 7.55). Three patches of pure blue, green and Red were measured on displays using the probe X-rite Eye One Pro 2. We then measured a white patch to assess the spectral additivity of the components. The measured spectrums are plotted in Figure 4.

We then tested the performance of HDR-VDP2 when changing this parameter with this five spectrum and with a white spectrum (so we don't take into account the display characteristic), with a Dirac spectrum where the value of the Diracs correspond to the primary wavelengths define by the CIE and with a spectrum with only the blue component.

Results are reported in Table IV for Zerman et al. and the proposed database. Surprisingly, the spectrum modification has almost no effect on the performance metrics. Note that we observed a similar result for Narwaria et al. and Korshunov et al.'s databases. The variations of all the values aren't significant for all real spectrum. On the proposed database, the performance slightly decreases when using a white spectrum. When the spectrum is reduced to the blue component, the performance drops for the proposed database but stays stable for Zerman et al.'s database. Overall, this test suggests that only a coarse estimation of the display spectrum is required for evaluating the quality score with HDR-VDP2.

2) Sensitivity to the surround luminance: In this section, We measure the performance when using different values of

TABLE IV: Performance of HDR-VDP2 on Zerman et al. and on the proposed database.

ZERMAN ET AL.				
Spectrum	PCC	SROCC	OR	RMSE
SONY BVM-X300	0.94	0.93	0.45	10.4
LG OLED 65 E6V	0.94	0.93	0.47	10.4
Sony KD-75X9405C	0.93	0.93	0.47	10.4
Loewe Bild 7.55	0.93	0.93	0.47	10.6
SIM2 HDR47ES4MB	0.93	0.92	0.46	10.6
white spectrum	0.93	0.91	0.55	11.8
Diracs spectrum	0.91	0.92	0.50	11.0
blue-only spectrum	0.88	0.87	0.55	13.9
PROPOSED DATABASE				
Spectrum	PCC	SROCC	OR	RMSE
SONY BVM-X300	0.89	0.87	0.48	12.5
LG OLED 65 E6V	0.90	0.87	0.47	12.3
Sony KD-75X9405C	0.90	0.87	0.46	12.4
Loewe Bild 7.55	0.90	0.87	0.48	12.3
SIM2 HDR47ES4MB	0.90	0.87	0.48	12.2
white spectrum	0.84	0.87	0.59	14.9
Diracs spectrum	0.90	0.87	0.46	12.4
blue-only spectrum	0.67	0.67	0.68	20.5

the surround luminance. The PCC and SROCC are reported in Figure 5 (Top). OR and RMSE show similar trends. The surround luminance has a very small impact on HDR-VDP2 performance, which slightly decreases for low luminance for our database and Narwaria et al..

3) Sensitivity to the angular resolution: As in the previous section, we estimate the sensibility of HDR-VDP2 to the angular resolution. The PCC and the SROCC are given in Figure 5 (Bottom). Performances are stable in the range 30 to 80 pixel/degree. Below 30 pixel/degree performances steadily drop down.

However, it is important to emphasize that, if the values of the metric did not significantly change in previous sections, changing the angular resolution parameter shifts HDR-VDP2 scores without losing in correlation. For example, the mean value of HDR-VDP2 scores on our database is 60.3 when the parameter is 30 pixel/degree and 48.9 when the parameter is equal to 60 pixel/degree. These two phenomena are due to the pooling phase of HDR-VDP2, in which the weights associated with the lowest frequency sub-bands are almost null (when the angular resolution diminish, the viewing distance decreases and the amount of low-frequency increases).

IV. CONCLUSION

In this paper, we investigated the sensibility of HDR-VDP2 metric to three user-defined parameters, namely the angular resolution, the spectrum emission and the surround luminance. Our results suggest that only an approximate estimation of the viewing conditions is required. Indeed, in the context of this study, HDR-VDP2 quality scores are not sensitive to moderate variations of viewing condition parameters. It should be noted however that this is valid only for the quality scores and not for the distortion map.

To go deeper into the analysis, we proposed a new database, including new kinds of distortion. We found out that HDR-VDP2 has a good resilience to chromatic artifacts, although this is a luminance-only metric (a similar comment can be

ISBN 978-90-827970-1-5 © EURASIP 2018

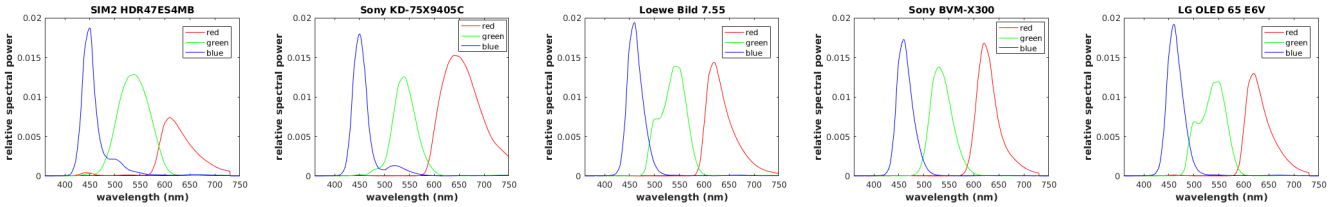

Fig. 4: Spectral emission of displays

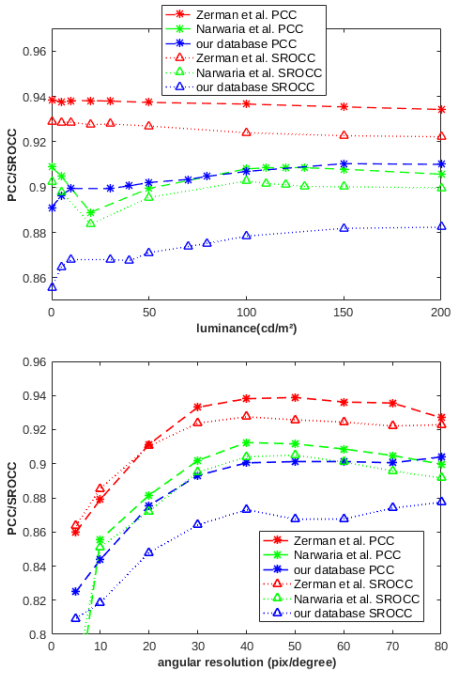

Fig. 5: HDR-VDP2 sensitivity to the surround luminance (Top) and the angular resolution (Bottom).

made for simpler metrics such as PU-MSSSIM). This observation raises some questions about HDR images and their quality evaluation. As HDR images are still quite confidential, we may ask the following question: are naïve observers able to evaluate small to medium distortions on this new format? We believe that more experienced viewers could have been more sensible to the color distortions, which are much more complex to assess than luminance artifacts.

References

[1] T. O. Aydın, R. Mantiuk, and H.-P. Seidel, "Extending quality metrics to full luminance range images," in *Human Vision and Electronic Imaging XIII*, vol. 6806. International Society for Optics and Photonics, 2008, p. 68060B.

[2] M. D. Fairchild, "The hdr photographic survey," in *Color and Imaging Conference*, vol. 2007, no. 1. Society for Imaging Science and Technology, 2007, pp. 233–238.

[3] J. Froehlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling, and H. Brendel, "Creating cinematic wide gamut hdr-video for the evaluation of tone mapping operators and hdr-displays," in *Digital Photography X*, vol. 9023. International Society for Optics and Photonics, 2014.

[4] P. Hanhart, M. Řeřábek, and T. Ebrahimi, "Subjective and objective evaluation of hdr video coding technologies," in *International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6.

[5] P. Hanhart, M. V. Bernardo, M. Pereira, A. M. Pinheiro, and T. Ebrahimi, "Benchmarking of objective quality metrics for hdr image quality assessment," *EURASIP Journal on Image and Video Processing*, vol. 2015, no. 1, p. 39, 2015.

[6] P. Hanhart, M. Řeřábek, and T. Ebrahimi, "Towards high dynamic range extensions of hevc: subjective evaluation of potential coding technologies," in *Applications of Digital Image Processing XXXVIII*, vol. 9599. International Society for Optics and Photonics, 2015.

[7] *Methodology for the subjective assessment of the quality of television pictures*, ITU-R Rec BT.500-13, 2012.

[8] *Parameter values for the HDTV standard for production and international programme exchange*, ITU-R Rec BT.709-6, 2015.

[9] *Parameter values for ultra-high definition television systems for production and international programme exchange*, ITU-R Rec BT.2020-2, 2016.

[10] *Image parameter values for high dynamic range television for use in production and international programme exchange*, ITU-R Rec BT.2100-1, 2017.

[11] *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*, ITU-T Rec P.1401, 2012.

[12] *Conversion and coding practices for HDR/WCG Y'CbCr 4:2:0 video with PQ transfer characteristics*, ITU-T Rec H-Suppl.15, 2017.

[13] P. Korshunov, P. Hanhart, T. Richter, A. Artusi, R. Mantiuk, and T. Ebrahimi, "Subjective quality assessment database of HDR images compressed with JPEG XT," in *Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, 2015, pp. 1–6.

[14] S. Lasserre, F. LeLéannec, and E. Francois, "Description of hdr sequences proposed by technicolor," *ISO/IEC JTC1/SC29/WG11 JCTVC-P0228], IEEE, San Jose, USA*, 2013.

[15] 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," in *ACM Transactions on graphics (TOG)*, vol. 30, no. 4, 2011, p. 40.

[16] M. Narwaria, M. P. Da Silva, P. Le Callet, and R. Pepion, "Tone mapping-based high-dynamic-range image compression: study of optimization criterion and perceptual quality," *Optical Engineering*, vol. 52, 2013.

[17] M. Narwaria, R. K. Mantiuk, M. P. Da Silva, and P. Le Callet, "HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images," *Journal of Electronic Imaging*, vol. 24, no. 1, pp. 010 501–010 501, 2015.

[18] M. Narwaria, M. Perreira Da Silva, and P. Le Callet, "HDR-VQM: An objective quality measure for high dynamic range video," *Signal Processing: Image Communication*, vol. 35, pp. 46–60, 2015.

[19] T. Richter, "On the standardization of the jpeg xt image compression," in *Picture Coding Symposium (PCS), 2013*. Ieee, 2013, pp. 37–40.

[20] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[21] *High dynamic range electro-optical transfer function of mastering reference displays*, SMPTE Std. ST.2084, 2014.

[22] G. Valenzise, F. De Simone, P. Lauga, and F. Dufaux, "Performance evaluation of objective quality metrics for hdr image compression," vol. 9217, 2014, pp. 9217 – 9217 – 10.

[23] T. Vigier, L. Krasula, A. Milliat, M. P. Da Silva, and P. Le Callet, "Performance and robustness of hdr objective quality metrics in the context of recent compression scenarios," in *Digital Media Industry & Academic Forum (DMIAF)*. Ieee, 2016, pp. 59–64.

[24] E. Zerman, G. Valenzise, and F. Dufaux, "An extensive performance evaluation of full-reference hdr image quality metrics," *Quality and User Experience*, vol. 2, no. 1, 2017.

ISBN 978-90-827970-1-5 © EURASIP 2018