https://doi.org/10.2352/ISSN.2470-1173.2018.14.HVEI-507

Perceived Dynamic Range of HDR Images with no Semantic Information

Vedad Hulusic, LTCI, Télécom ParisTech, Université Paris-Saclay
Giuseppe Valenzise, Laboratoire des Signaux et Systèmes (L2S, UMR 8506), CNRS - CentraleSupélec - Université Paris-Sud
Frédéric Dufaux, Laboratoire des Signaux et Systèmes (L2S, UMR 8506), CNRS - CentraleSupélec - Université Paris-Sud

Abstract

Computing dynamic range of high dynamic range (HDR) content is an important procedure when selecting the test material, designing and validating algorithms, or analyzing aesthetic attributes of HDR content. It can be computed on a pixel-based level, measured through subjective tests or predicted using a mathematical model. However, all these methods have certain limitations. This paper investigates whether dynamic range of modeled images with no semantic information, but with the same first order statistics as the original, natural content, is perceived the same as for the corresponding natural images. If so, it would be possible to improve the perceived dynamic range (PDR) predictor model by using additional objective metrics, more suitable for such synthetic content. Within the subjective study, three experiments were conducted with 43 participants. The results show significant correlation between the mean opinion scores for the two image groups. Nevertheless, natural images still seem to provide better cues for evaluation of PDR.

Introduction

The human visual system (HVS) is capable of perceiving a much wider range of luminance and color values than it is possible to capture or reproduce with the standard dynamic range (SDR) imaging systems. Therefore, in the past two decades there has been a plethora of research activities within the high dynamic range (HDR) field, aiming at overcoming these limitations by enabling capture, storage, transmission and display of such content, thus allowing more realistic and enhanced user experience [1, 2]. One important research question is how to predict the perceived dynamic range of HDR pictures. For instance, a perceptual dynamic range measure is needed for HDR content selection to conduct perceptual studies [3, 4], or to facilitate image aesthetic evaluation [5].

The dynamic range is generally computed as the ratio between the maximum and minimum luminance of a picture. However, this measure seems too simple to account for the rather complex perception of light by the human eye [6]. Studies on dynamic range perception in the field of display technologies have mainly focused on assessing the maximum span of luminance intensity human eye can sense in a brief temporal interval, also known as the steady-state dynamic range [7]. In this work, we are rather interested in measuring the dynamic range of a given picture as perceived by observers. This is related to the perception of lightness, defined as the relative brightness of objects in a scene. The perception of dynamic range depends then on the ratio between the lightest and darkest part of the picture. Early psychophysical experiments on lightness perception were based on simple stimuli, e.g., disks of varying intensity on uniform background, and led to the conclusion that lightness does not depend only on relative luminance ratios, but also on the relative area of the brightest patch [8]. More recent studies on this topic confirm that lightness is strongly context dependent also in HDR conditions [9].

A basic question in the perception of DR of a picture is whether the latter can be essentially explained with the distribution of light intensity of the picture, or if instead higher-order statistics play a significant role. The retinex theory [10, 11] suggests that lightness perception is a local process but is modulated holistically by the whole image appearance. However, so far these models have not been proved useful for predicting dynamic range in complex images. In an attempt to gain a better insight on DR perception for complex stimuli, we have recently collected a dataset of 36 real-world HDR pictures with subjective annotations [12]. The scores were obtained by asking observers to rate images based on the magnitude of the overall difference between the brightest and darkest region(s) of the picture. Later, we have leveraged this data to evaluate robust dynamic range (DR) measures [6], and to derive a DR predictor which takes into account also the area of highlight regions [13].

Although the formulated model [13] can predict well overall subjective DR scores, we also found significant exceptions and prediction failures. In fact, any DR predictor learned on such a small-size dataset might incur the risk of overfitting. Since, unfortunately, collecting much larger datasets to build more sophisticated data-driven DR measures is extremely time consuming and practically unfeasible, in this work we propose an alternative strategy to advance our understanding of DR perception. Instead of augmenting the dataset in [12] with more complex images, we collect subjective DR scores for lower-complexity stimuli, similar to Mondrian-like pictures [11]. Differently from previous studies, each Mondrian picture used in this work is directly obtained by, and shares some statistical characteristics with one of the real-world HDR complex stimuli of the dataset in [12]. Specifically, each Mondrian reproduces the same first-order statistics, i.e., histogram, of the corresponding complex picture, but is spatially uncorrelated with the original. This enables to directly compare how the perceived dynamic range changes when all the semantic information is removed from the stimulus, and only the light intensity distribution (and thus the simple max/min DR metric) is the same.

We conduct a series of experiments to compare the DR scores from real-world images and the corresponding synthesized Mondrian-like stimuli. The results of our experiments show that, as expected, the DR perceived on the synthetic stimuli is well correlated ($r = .87$) with that conveyed by complex images. However, and more interestingly, we observe that confidence intervals

of these scores become larger for synthetic stimuli than for the natural ones. This result clearly confirms that higher-order statistics and other visual cues contribute to stabilize the perception of DR, and provides evidence and ground-truth observations to derive more perceptually-justified models of dynamic range.

Test material and apparatus

In this study we use both real-world, complex images from the dataset [12][1], and Mondrian-like stimuli synthesized from that dataset. Specifically, we generate synthetic stimuli starting from each of the 36 HDR images in the dataset [12] using the dead leaves model, a Mondrian-like representation without spatial correlation with the original content. The dead leaves model has been successfully utilized for reproduction of most of natural image statistics by using superposition of random closed sets and specific size distributions for objects [14, 15]. By utilizing such a model and constraining its luminous intensities with the extracted histogram from the original natural image, we preserve the first order statics of the reference, natural image.

In order to generate dead leaves stimuli, a publicly available Matlab toolbox[2] was used. The original *compute_dead_leaves_image* function, based on work by Lee et al. [15] and Gousseau and Roueff [17], was modified so that, once the dead leaves image is generated, its pixel intensity values are reassigned by exact histogram matching with the corresponding natural image. For each natural image from the dataset, 50 realizations of modeled, dead leaves (DL) images were generated.

All the experiments were conducted in a dark and quiet room. The stimuli were displayed at full HD (1920×1080 pixels) resolution on an HDR SIM2 HDR47ES4MB 47" screen. It was utilized in the DVI Plus (DVI+) mode, that allows for directly and independently controlling backlight LEDs and LCD pixel values, based on the dual-modulation algorithm [16]. The ambient illumination in the room, measured between the screen and participants, was 2.154 lux. The luminance of the screen when turned off was 0.03 cd/m^2. The distance from the screen was fixed to three heights of the display, with the eyes in the middle of the display, both horizontally and vertically.

Preliminary study

Since DL stimuli are randomly generated, we performed a preliminary study to determine: i) whether there is a significant effect of the specific DL realization on the perception of DR; and ii) if using DL stimuli introduces a systematic bias that would make impossible to compare their DR with natural stimuli. These two experiments and their results are discussed in the following sections.

The main aim of this study was to investigate if natural images can be substituted with the modeled images, using Mondrian-like representations, e.g. dead leaves model, in subjective evaluation of dynamic range of HDR images, where modeled images are generated with the same first order statistics, i.e. histogram, as in their natural counterparts. This was done by comparing the user scores obtained from subjective tests conducted on both natural and modeled images from the same dataset. Experimental method and the results for each test are described below, and the main study is presented in the next section.

Effect of DL realization

In the first experiment, we compare the DR of different DL realization of the same content, in order to assess whether the effect of using a different realization might impact the judgment of dynamic range. To this end, five natural HDR images from our dataset were selected, so that the PDR of the achromatic stimuli [13] is uniformly sampled. For each of these five images, five DL realizations (out of 50), with the highest variance of both pixel-based and topological measures (area and perimeter of the highlight regions[3], Euler characteristic, and contrast), were selected.

The test was conducted using a pairwise comparison method. Participants were asked to evaluate the perceived dynamic range as the overall impression of the magnitude of the difference between the brightest and the darkest parts of the image. 15 participants (9M/6F; age avg. 27.9) were presented 20 pairs of DL images, all within content, and asked to select the one with higher PDR. They could freely move between the compared images in each test pair with the arrow keys on the keyboard, and they had to select one by pressing the space button.

Results. In this subjective test we wanted to see for which pairs the user preference was at the 75% rate or higher, that is, where it was not a result of a random selection (50% rate). The preference matrix with the color coded rate is shown in Figure 2. The same preferences were found as significant by the binomial test at $p < .05$. As can be observed from the figure, not many significant differences were found between the compared images. The only image that has been found to have significantly lower PDR, compared to other four DL realizations with the same histogram is *EC_41*. In addition, *LT_37* and *OP_37* have been evaluated as significantly higher dynamic range than 3 out of four compared images.

These results suggest that each natural content can be safely compared with a random DL realization, without biasing the results for most of the cases. In particular, each observer can be presented a different DL realization, and averaging the viewers' scores will be generally consistent.

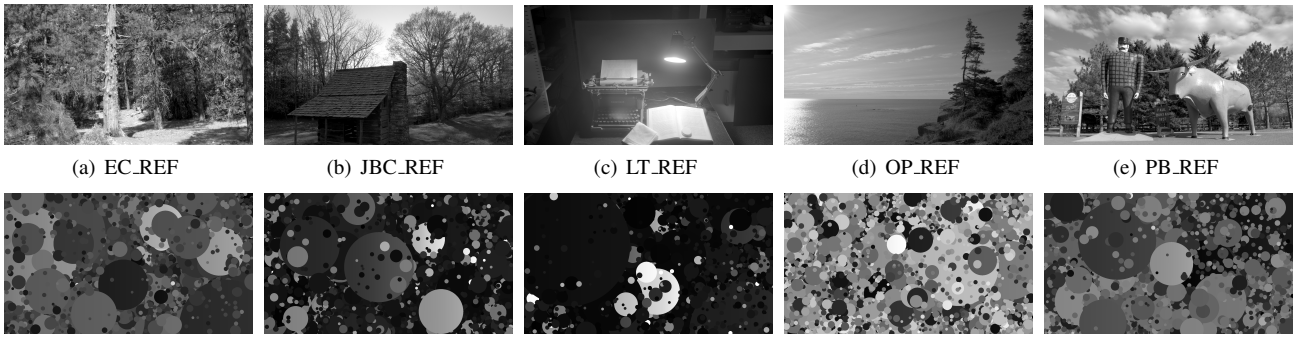
Systematic bias of DL stimuli

The goal of this experiment is to verify whether using DL stimuli introduces a systematic bias in the perception of dynamic range, i.e., whether DL images display always larger or smaller DR than the corresponding natural image *independently* of the content. This is particularly important as our DL stimuli do not reproduce second-order statistics of the corresponding original pictures, and in particular they might introduce new edges and increase the perception of simultaneous contrast. For this experiment, the same DL images as in the Experiment 1, plus the 5 original natural images, were used. Using the same methodology, participants compared the natural images with their five DL realizations using repetitions (each image in each test pair appearing

[1] http://pdr.lefca.net/

[2] https://fr.mathworks.com/matlabcentral/fileexchange/16201-toolbox-image

[3] The highlight regions are defined in [13] as those pixels having luminance larger than 2400 cd/m^2.



(a) EC_REF (b) JBC_REF (c) LT_REF (d) OP_REF (e) PB_REF

(f) EC_17 (g) JBC_40 (h) LT_37 (i) OP_37 (j) PB_14

Figure 1. *Tone mapped representations of the images used in the preliminary study. Left to right: ElCapitan_bottom (EC), JasseBrownsCabin (JBC), LabTypewriter(LT), OtterPoint (OP), PaulBunyan (PB); top: natural images (REF); bottom: corresponding DL images.*

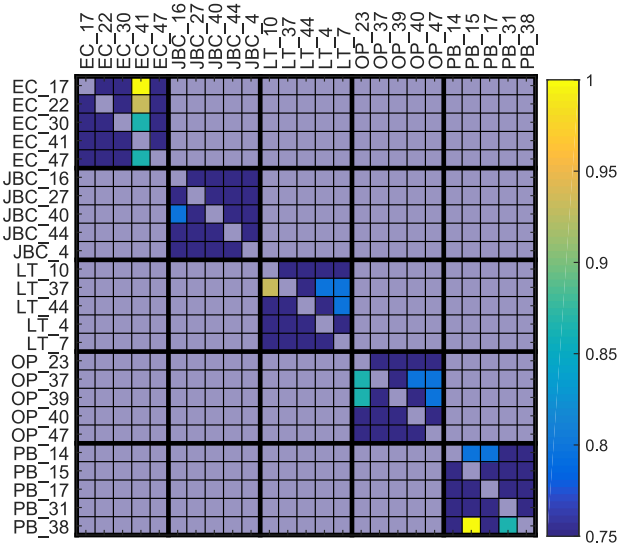


Figure 2. *A preference matrix for the scores of the first preliminary test. Only the pairs in the 5-by-5 blocks on the main diagonal were compared and the results for the preference rates over 75% are color coded. The same preferences are found as significant at $p < .05$ running the binomial test.*

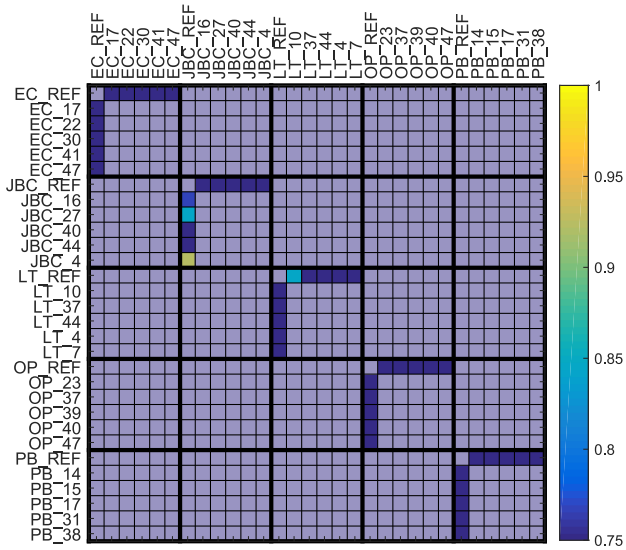


Figure 3. *A preference matrix for the scores of the second preliminary test. Only the pairs in the first row and column of the 6-by-6 blocks on the main diagonal were compared and the results for the preference rates over 75% are color coded. The same preferences are found as significant at $p < .05$ running the binomial test.*

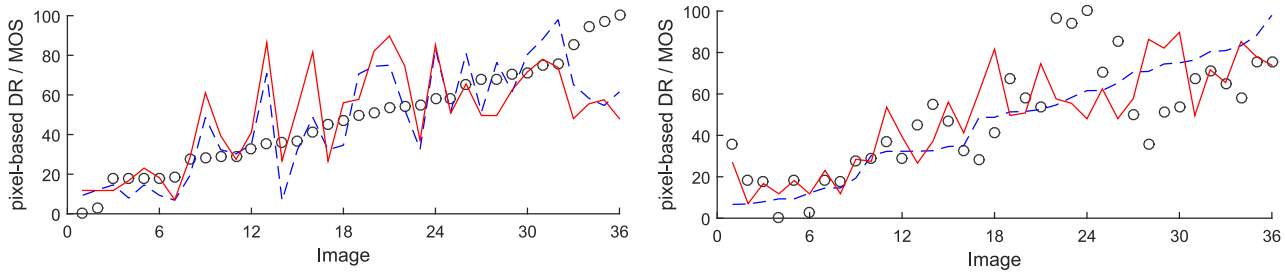
both as first and as second). 13 participants (8M/5F; age avg 29.6) took part, providing 26 opinion scores per pair.

Results. Using the same analysis as in the first experiment, a significant preference ($p < .05$) was found for only a few cases in two out of five contents, see Figure 3. The natural (REF) image for the *JBC* content has been rated as significantly lower dynamic range than 3 out of 5 DL realizations of that image. In addition, *LT_10* was found to have a significantly lower dynamic range compared to the corresponding natural image. All the other comparisons did not show any significant difference.

These results show that DR perception is different from natural stimuli to DL images only in some cases, which supports our main study aimed at identifying the relationship between perceived dynamic ranges for the two complete set of stimuli.

Main study: Comparison of perceived DR for natural and DL stimuli

This experiment was a replication of the one conducted in our previous work [13] for grayscale images, except for the stimuli, where random DL realizations of natural stimuli, with the same histogram, were tested. For this test, we used random DL realization (different for each participant) of all 36 natural HDR images. 15 participants (12M/3F; age avg 26.7) evaluated the DL images using the same SAMVIQ-like methodology proposed in [12]. They were displayed 12 image thumbnails at a time. They had to click on each image to see the full screen presentation, and give it the score on a continuous scale with a 0-100 range. They could evaluate images in any order, take as much time as they needed and re-evaluate any image at any time. After evaluating all 12 images, they could proceed to the next subset of 12 images.



(a) Sorted by the pixel-based DR

(b) Sorted by the MOS (natural images)

Figure 4. *A comparison between the pixel-based DR values (circles) and MOS scores for the natural (blue dashed line) and DL (red solid line) images, sorted by the pixel-based DR (a) and MOS for natural images (b).*

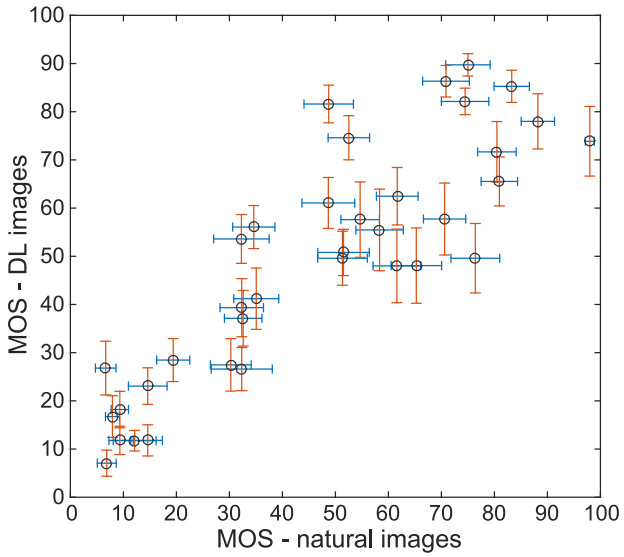


Figure 5. *A comparison of the MOS values with the corresponding confidence intervals (CIs) for the natural (blue) and DL (red) images.*

Results. Comparing the obtained scores and the confidence intervals from the two experiments, we wanted to validate these findings and provide more conclusive results. Looking at the confidence intervals (CIs) of the MOS scores for both natural and DL images (Figure 5), we can infer the level of reliability of the user scores. The mean of the CIs for these two types of stimuli were $\mu_{nat} = 7.41$ and $\mu_{DL} = 10.173$. In most cases CIs were much lower for the natural images whose PDR were between 58 and 88. Only in two cases it occurred at the far ends of the PDR scale, as shown in Table 1. Looking further at the cases where the differences in both MOS values and the CIs between the natural and DL images were the highest (see Figure 6), revealed that two of these images, *OCanadaNoLights_b* and *OtterPoint*, are those that had the highest discrepancy between the pixel-based DR and PDR in the original study with natural images [13]. For these two cases, the subjective scores obtained for the DL images were much closer to the pixel-based DR, than it was the case for the natural images.

We also wanted to see the correlations between the MOS scores for both sets of stimuli and the pixel-based DR values. The

Scene	MOS_DL	MOS_nat	CI_DL	CI_nat
BloomingGorse(1)	16.733	7.933	8.659	2.633
Carousel	71.667	80.467	12.587	7.255
Flamingo	57.600	54.667	15.655	7.268
HdrMark	78.000	88.200	11.456	6.280
LabTypewriter	73.867	98.000	14.452	1.704
OCanadaNoLights_b	26.800	6.667	11.147	3.883
OtterPoint	57.733	70.600	14.930	7.972
URChapel(1)_t	49.600	76.400	14.409	9.192
WaffleHouse	48.000	61.600	15.258	8.989
Zentrum	55.467	58.333	16.925	8.950

Table 1: The stimuli where the difference between the confidence intervals between the natural and DL images were the highest. In most of the cases, these were the images with the MOS_{nat} values between 58 and 88.

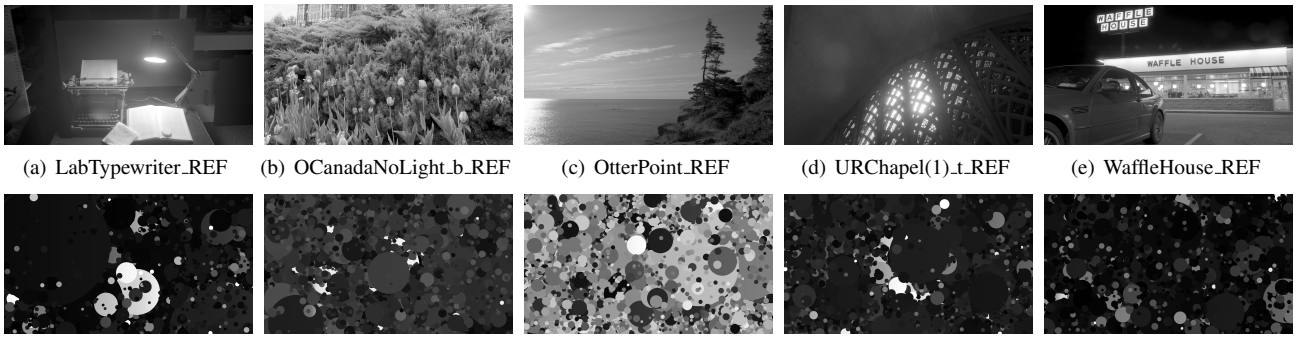
results show that there is a higher correlation between the pixel-based DR and MOS values for the natural images ($r = .853$) than for the DL images ($r = .741$). Furthermore, the Pearson's and Spearman's correlations between the MOS values for the natural and DL images were found as significant ($p < .001$) with the $r = .87$ and $r_s = .845$ coefficients respectively, see Table 2. These comparisons are presented visually in Figure 4. All this indicates that using only first-order statistics might not be enough for modeling natural HDR images in evaluating perceived dynamic range.

		MOS_nat	MOS_DL
Pearson (r)	**DR**	0.853	0.741
	MOS_nat		0.870
Spearman (r_s)	**DR**	0.862	0.700
	MOS_nat		0.845

Table 2: Pearson's and Spearman's correlation coefficients r and r_s between the pixel-based DR and MOS vales, and between the MOS values obtained from natural and DL images. All correlation coefficients are found as significant at $p < .001$.

Conclusions and future work

In this paper the perceived dynamic range has been investigated on synthetically generated images using the dead leaves model. The test stimuli were generated from the natural images from the publicly available perceptually annotated HDR image dataset with MOS values [12]. For each natural image 50 DL image realizations were generated, each preserving the histogram of



(a) LabTypewriter_REF (b) OCanadaNoLight_b_REF (c) OtterPoint_REF (d) URChapel(1)_t_REF (e) WaffleHouse_REF

(f) DL_LabTypewriter (g) DL_OCanadaNoLight_b (h) DL_OtterPoint (i) DL_URChapel(1)_t (j) DL_WaffleHouse

Figure 6. *Five cases where the difference in the MOS values and CIs between the natural and DL images were the highest, see Table 1.*

the luminance values from the input image. Three main objectives of the study were:

1. to see if there is a significant difference in PDR between different DL realizations with the same histogram,
2. to verify whether using DL stimuli introduces a systematic bias in the perception of dynamic range, and
3. to compare the PDR scores obtained for the natural images and their corresponding DL realizations.

Three subjective test were conducted using 43 participants in total. In the preliminary study, the focus was on the first two objectives. The results indicate that there is neither significant difference between the different realizations of the DL images from the same natural image nor systematic bias in comparing the PDR of the natural images and their DL realizations. This implicitly supports the finding from the first experiment. The main study investigated whether the same dynamic range is perceived when observing natural and DL images with the same histogram (objective 3). The results showed that, although there is a high correlation between the obtained scores ($r = .87$, $r_s = .845$), the natural images gave more stable results with lower confidence intervals, and were higher correlated with the pixel-based DR ($r = .853$), compared to the DL images ($r = .741$).

From the results we can see that preserving first-order statistics might not be enough for a truthful representation of the natural HDR images in such perceptual tasks. Therefore, in the future, we would like to extract other, higher order statistics, from the reference, natural images and use them in the modeled image representations. Harnessing this correlation will allow us to extract other image features, e.g. topological, that might be used for better modeling of perceived dynamic range of HDR images.

Acknowledgments

We would like to thank all the participants who volunteered in the experiments.

References

[1] Francesco Banterle, Alessandro Artusi, Kurt Debattista, and Alan Chalmers. *Advanced high dynamic range imaging: theory and practice*. CRC Press, 2011.

[2] Frédéric Dufaux, Patrick Le Callet, Rafal Mantiuk, and Marta Mrak. *High Dynamic Range Video. From Acquisition, to Display and Applications*. Academic Press, 2016.

[3] Manish Narwaria, Claire Mantel, Matthieu Perreira Da Silva, Patrick Le Callet, and Soren Forchhammer. An objective method for high dynamic range source content selection. In *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*, pages 13–18. IEEE, 2014.

[4] Allan G Rempel, Matthew Trentacoste, Helge Seetzen, H David Young, Wolfgang Heidrich, Lorne Whitehead, and Greg Ward. Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs. In *ACM Transactions on Graphics (TOG)*, volume 26, page 39. ACM, 2007.

[5] Tunc Ozan Aydin, Aljoscha Smolic, and Markus Gross. Automated aesthetic analysis of photographic images. *Visualization and Computer Graphics, IEEE Transactions on*, 21(1):31–42, 2015.

[6] Vedad Hulusic, Giuseppe Valenzise, Kurt Debattista, and Frederic Dufaux. Robust dynamic range computation for high dynamic range content. In *Human Vision and Electronic Imaging Conference, IS&T International Symposium on Electronic Imaging (EI 2017)*, 2017.

[7] Timo Kunkel and Erik Reinhard. A reassessment of the simultaneous dynamic range of the human visual system. In *Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization*, pages 17–24. ACM, 2010.

[8] Xiaojun Li and Alan L Gilchrist. Relative area and relative luminance combine to anchor surface lightness values. *Perception & Psychophysics*, 61(5):771–785, 1999.

[9] Ana Radonjić, Sarah R Allred, Alan L Gilchrist, and David H Brainard. The dynamic range of human lightness perception. *Current Biology*, 21(22):1931–1936, 2011.

[10] Edoardo Provenzi, Massimo Fierro, Alessandro Rizzi, Luca De Carli, Davide Gadia, and Daniele Marini. Random spray retinex: A new retinex implementation to investigate the local properties of the model. *IEEE Transactions on Image Processing*, 16(1):162–171, 2007.

[11] Edwin H Land and John J McCann. Lightness and retinex theory. *JOSA*, 61(1):1–11, 1971.

[12] Vedad Hulusic, Giuseppe Valenzise, Edoardo Provenzi, Kurt Debattista, and Frederic Dufaux. Perceived dynamic range of HDR images. In *Proceedings 8th International Workshop on Quality of Multimedia Experience (QoMEX'2016)*. IEEE, 2016.

[13] Vedad Hulusic, Kurt Debattista, Giuseppe Valenzise, and Frédéric Dufaux. A model of perceived dynamic range for hdr images. *Signal Processing: Image Communication*, 51:26–39, 2017.

[14] Daniel L Ruderman. Origins of scaling in natural images. *Vision research*, 37(23):3385–3398, 1997.

[15] Ann B Lee, David Mumford, and Jinggang Huang. Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model. *International Journal of Computer Vision*, 41(1):35–59, 2001.

[16] Emin Zerman, Giuseppe Valenzise, Francesca De Simone, Francesco Banterle, and Frederic Dufaux. Effects of display rendering on HDR image quality assessment. In *SPIE Optical Engineering+ Applications*, pages 95990R–95990R. International Society for Optics and Photonics, 2015.

[17] Yann Gousseau and François Roueff. The dead leaves model: general results and limits at small scales. *arXiv preprint math/0312035*, 2003.

Author Biography

Vedad Hulusic is a post-doctoral researcher at Télécom ParisTech in Paris, France. He has a PhD in Engineering from the University of Warwick, UK and a first degree in Computer Science from the University of Sarajevo, Bosnia and Herzegovina. He has a long-standing interest in the computer graphics, high dynamic range imaging, visual perception and attention, image and video quality assessment, image processing, virtual museums, cross-modal interaction and serious games in which he has been a published author.