

Robust dynamic range computation for high dynamic range content

Vedad Hulusic, Giuseppe Valenzise, Kurt Debattista, Frederic Dufaux

▶ **To cite this version:**

Vedad Hulusic, Giuseppe Valenzise, Kurt Debattista, Frederic Dufaux. Robust dynamic range computation for high dynamic range content. Human Vision and Electronic Imaging Conference, IS&T International Symposium on Electronic Imaging (EI 2017), Jan 2017, Burlingame, United States. hal-01433779

HAL Id: hal-01433779
https://hal.science/hal-01433779

Submitted on 10 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Robust Dynamic Range Computation for High Dynamic Range Content

Vedad Hulusic, LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, Paris, France
Giuseppe Valenzise, Laboratoire des Signaux et Systèmes (L2S, UMR 8506), CNRS - CentraleSupélec - Université Paris-Sud, 91192 Gif-sur-Yvette, France
Kurt Debattista, WMG, University of Warwick, Coventry, UK
Frédéric Dufaux, Laboratoire des Signaux et Systèmes (L2S, UMR 8506), CNRS - CentraleSupélec - Université Paris-Sud, 91192 Gif-sur-Yvette, France

Abstract

High dynamic range (HDR) imaging has become an important topic in both academic and industrial domains. Nevertheless, the concept of dynamic range (DR), which underpins HDR, and the way it is measured are still not clearly understood. The current approach to measure DR results in a poor correlation with perceptual scores ($r \approx 0.6$). In this paper, we analyze the limitations of the existing DR measure, and propose several options to predict more accurately subjective DR judgments. Compared to the traditional DR estimates, the proposed measures show significant improvements in Spearman's and Pearson's correlations with subjective data (up to $r \approx 0.9$). Despite their straightforward nature, these improvements are particularly evident in specific cases, where the scores obtained by using the classical measure have the highest error compared to the perceptual mean opinion score.

Introduction

High dynamic range (HDR) imaging allows for capture, storage, processing and display of real-world content with a high range of luminance, and has potential impact on a wide variety of applications [1, 2]. The correct computation of dynamic range (DR) is important for HDR content selection in the use of algorithmic and metric validation [3, 4] as well as for image aesthetic evaluation [5].

In this paper, DR metrics for HDR content selection is the primary concern. While measuring DR of a scene seems trivial, and is usually done by finding a ratio between the highest and lowest luminance value, it is error prone and may not correspond to what humans perceive. This is mainly due to image noise or to the presence of (typically few) extreme luminance values that skew the results while having little or no impact on the observer's perception of DR. To avoid this, the ratio is usually calculated robustly, discarding a tiny percentage of the darkest and brightest pixels. However, there is no standardized way of performing this, i.e. various percentiles are used along with the clipping of extreme luminance values. While this usually improves the results and is a better predictor of the DR, there are many cases where it fails to provide accurate results. This is due to the complexity of the human visual system, and to the multi-dimensional nature of the perception of the DR attribute [6].

In this work, we consider a number of simple enhancements, of different computational complexity, over the classical DR measure. The considered measures can be broadly classified into two groups, as shown in Figure 1: i) robust DR computation; and ii) methods that consider the spatial characteristics of DR perception, as well as its intrinsic multiscale nature. In order to validate the accuracy of these improved DR measures, we correlate their predictions on 36 HDR images with the corresponding perceptual DR scores, gathered through a subjective experiment in which observers had to gauge the perceived overall impression of the difference between the brightest and darkest part(s) of an HDR image [7]. The results show that significant improvements can be achieved at a very low computational overhead with respect to the classical DR measure.

The main contribution of this paper is a systematic investigation and comparison of different approaches for calculating the DR, and a discussion on the possible implications when using them. To the authors' knowledge, there has not been any previous study comparing computational approaches to measure the perceived dynamic range. Furthermore, we single out and discuss specific cases where the classical DR measure fails, and show how prediction errors for those cases can be substantially reduced using low-cost and more robust DR computation approaches.

Data Set and Procedures

Test images and subjective ground-truth

In this study, a set of 36 luminance-only HDR images was used. All pictures have HD resolution (1920×1080 pixels). In the subjective study by Hulusic et al. [7], an associated mean opinion score (MOS), describing its perceived dynamic range, was obtained for each image. 20 participants observed pictures on a SIM2 HDR47 display [8] and were asked to evaluate the overall impression of the difference between the brightest and the darkest part(s) of the observed image. The images in the data set were selected based on the variability in the three objective measures: pixel-based DR, image key (IK) [9] and spatial information (SI)[10]. Details are available in the original paper [7].

Dynamic Range Measures

In order to systematically investigate different aspects of DR, we considered three kinds of measurements: classical (C) calculation; robust (R) calculation using percentiles; and filtering/interpolation (F). Several DR measures have been formulated by combining principles from these three families, as shown in Figure 1. These approaches are described in detail in the following paragraphs.

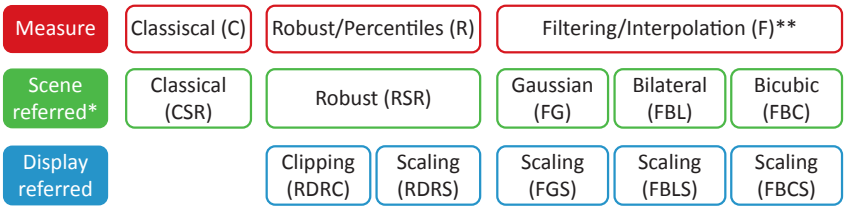


Figure 1: Measures used in the study. **Due to existence of zero values in many images, even the scene referred images were clipped to* 10^{-5} *at the low end. **All filtering/interpolation was done on the* $\log_{10}$ *encoded values.*

Classical Approach (C)

The first category uses the ratio between the maximum and minimum luminance value in the image:

$$DR = \log_{10}(L_{\max}/L_{\min}). \quad (1)$$

This method is know to be very sensitive to extreme values. In addition, images with completely black pixels result in infinite DR, which was the case in 19 out of 36 images. Therefore, pixel values in all the images were clipped to $10^{-5} cd/m^2$ during preprocessing.

Robust Calculation Using Percentiles (R)

The dynamic range of the images in the second category is computed using (1), where L_{min} and L_{max} were the minimum and maximum relative luminance values of the image, respectively, computed after *excluding a given percentage of the darkest and brightest pixels*. The low and high percentage values used for robust calculations are: 0.1%, 0.5%, 1%, 2%, 3%, 5% and 10%, presented as α and β indices respectively next to the abbreviated name of the measure.

These calculations are performed on both scene-referred ($\text{RSR}_{\alpha,\beta}$) and displayed-referred pixel values. The former assumes HDR images with real-world luminance values[1], while the latter uses luminance values within the range of the display.

Furthermore, display-referred images can be obtained either by *clipping* ($\text{RDRC}_{\alpha,\beta}$) the HDR scene-referred luminance I to the display luminance range $[I_{\min}, I_{\max}]$, using

$$I' = \min(\max(I, I_{\min}), I_{\max}); \quad (2)$$

or *scaling* ($\text{RDRS}_{\alpha,\beta}$) HDR luminance values to the display luminance range, such that

$$I' = \frac{I - \min(I)}{\max(I) - \min(I)} \cdot (I_{\max} - I_{\min}) + I_{\min}, \quad (3)$$

where $I_{\min} = 0.03\, cd/m^2$ and $I_{\max} = 4250\, cd/m^2$ in this paper.

Image Filtering and Multiscale Representations (F)

In order to take into account the possible effects of *surface* in the perception of lightness and dynamic range [6], we compute dynamic range on low-frequency and multiscale representations of the image. This approach is supposed to reduce the impact of small image regions with extreme luminance values in the prediction of perceived DR. We consider three low-pass filtering approaches, calculated both on scene-referred images, and display-referred images after scaling. In each approach, DR is calculated for each combination within the predefined range of the clipped percentiles of luminance values, pyramid levels and spatial and range parameters for bilateral filtering. However, only the combinations and their corresponding parameters, for which the computed DR had the highest correlations with the MOS values, are reported in the paper.

Gaussian Pyramid This is a classical multiscale construction [11] that consists in decimating an image using a Gaussian filter. Specifically, the original $N \times M$ pixels image I_0 is convolved with a 5×5 Gaussian filter, and the low-pass filtered result is downsampled by a factor of 2, yielding and $N/2 \times M/2$ image I_1. The downsampling here is used to reduce the computational complexity of the filtering at higher levels of the pyramid, while it does not affect the spectra of filtered images. This procedure can be repeated recursively to produce an image pyramid, where image I_J at level J has a size $N/2^J \times M/2^J$. In this paper we consider $1 \leq J \leq 7$.

Polynomial interpolation We also build an image pyramid representation using a polynomial interpolation kernel instead of a Gaussian as low-pass filter. We consider polynomial kernels of degree 1, 2 or 3, i.e., bilinear, biquadratic and bicubic interpolation. Due to space constraints, we report only the results with bicubic interpolation, which was the one giving the best performance in our experiments.

Bilateral Filter In this case, the filtered image in the pyramid is obtained as:

$$I^{filtered}(x) = \frac{1}{W_p} \sum_{x_i \in \Omega} I(x_i) f_\theta(\|I(x_i) - I(x)\|) g_\sigma(\|x_i - x\|) \quad (4)$$

where f_θ and g_σ are two Gaussian kernels of variance θ and σ, respectively; Ω is a neighborhood of pixel x and W_p is a normalization constant [12]. As opposed to Gaussian filtering, the bilateral filter does not smooth out edges.

Results and Analysis

The scores for all tested measures were calculated in MATLAB R2014b. The Spearman's and Pearson's correlation along with the ℓ_2-error between the obtained results and the MOS for the dataset were investigated. Spearmans r_s and Pearson's r both range between 0 (no correlation) and 1 (complete agreement between tested values). The former is automatically calculated on ranks, while the latter uses actual values and thus depicts the linear relationship between variables. The ℓ_2-error is the ℓ_2 norm of the difference between predicted and subjective DR values and

[1] In order to prevent singularities, pixel values were clipped as in the classical approach.

Table 1: Spearman's r_s and Pearson's r coefficients and ℓ_2-error for different measures (Figure 1). Only the configurations with the best empirically found parameters are reported. In $\mathrm{DR}_{\alpha,\beta}$, the indices denote the lowest and highest percentiles of luminance values that have been clipped. In general, $\beta = 1-\alpha$, unless the * is used, which denotes the optimal combination found over the Cartesian product of β and α. J is the Gaussian or polynomial pyramid level. θ and σ are the range and spatial variances of the bilateral filter. $I_{range} = \max(I) - \min(I)$ is the range of the (possibly scaled/clipped) HDR image.

	CSR	$\mathrm{RSR}_{2,98}$	$\mathrm{RDRC}_{2,98}$	$\mathrm{RDRS}_{1,99}$	$\mathrm{RDRS}^*_{3,99.5}$	$\mathrm{FG}_{2,98}$ ($J=4$)	$\mathrm{FGS}_{0.1,99.9}$ ($J=3$)	$\mathrm{FBL}_{2,98}$ ($\sigma=11$ $\theta=I_{range}/6$ $J=1$)	$\mathrm{FBLS}_{1,99}$ ($\sigma=11$ $\theta=I_{range}/6$ $J=1$)	$\mathrm{FBC}_{2,98}$ ($J=6$)	$\mathrm{FBCS}_{0.1,99.9}$ ($J=3$)
r_s	0.652	0.820	0.823	0.868	0.895	0.827	0.869	0.815	0.859	0.833	0.871
r	0.614	0.661	0.825	0.863	0.887	0.659	0.874	0.658	0.868	0.659	0.869
ℓ_2-error	28.790	25.736	16.321	17.529	14.742	23.634	15.492	26.646	15.639	26.487	16.175

serves to give a measure of distance from the prediction to the ground truth data. Before calculating the ℓ_2-error, the calculated DR values were scaled to the MOS range. A summary of the results with the best combination of parameters for different measures and their corresponding r_s, r and ℓ_2-error values are presented in Table 1.

Improvement Over Classical DR Measure

From the results in Table 1, it is apparent that the prediction accuracy and monotonicity of the CSR measure ($r_s = .652$ and $r = .614$) can be substantially improved using both robust approaches based on removal of extreme luminance or multiscale computation. This can be seen graphically by looking at the scatter plots in Figure 2, where all metrics have been scaled accordingly to MOS range to make meaningful comparisons.

In order to check for the significance of the improvement of the measures, a test for significance of the difference between Spearman's correlation coefficients of the robust (R) and filtering (F) measures against the CSR was performed, using the Fisher z transformation [13]. A significant improvement of the RDRS ($p = .013$), RDRS* ($p = .003$), FGS ($p = .013$), FBLS ($p = .019$), FBC ($p = .045$) and FBCS ($p = .012$) measures over the CSR was found.

Interestingly, the RDRS* measure, i.e., the RDRS with the optimal combination of lower and higher percentiles, is the best performing method with the three performance indicators. Similar results can be obtained if the same principle of removing extreme values is applied to the filtered images. The test of significance of the difference between Spearman's correlation coefficients for these two or any other combination of robust (R) and filtering (F) measures showed no substantial difference ($p > .05$), i.e., the simplest percentile clipping gives similar results with no significant difference, compared to the relatively more computationally expensive filtering techniques.

Another interesting conclusion can be drawn by observing that display-referred measures are systematically more accurate than scene-referred ones. A significant difference of Pearson's coefficients has been found between the RSR and RDRS ($p = .019$), FG and FGS ($p = .012$), FBL and FBLS ($p = .015$) and FBC and FBCS ($p = .014$). This is somehow expected, as the subjective ground truth has been collected on an HDR display which comes with physical limitations in the actual range of possible emitted luminance.

Dynamic Range Prediction Failures

In order to analyze specific cases where the classical DR computation fails, five images from the dataset with the highest prediction error were singled out and displayed as red triangles in Figure 2. Two of these images are displayed in Figure 3, while the normalized DR predictions for the five pictures with different measures, as well as their MOS values, are reported in Table 2. Those cases mainly correspond to images where the distribution of highlights is concentrated in small or tiny regions, whose surface is too small to generate a significant perceptual response in the human visual system. The scatter plots (Figure 2) for the RDRS and FGS methods show that those cases are correctly handled by these measures.

There are also some cases where the robust DR computation

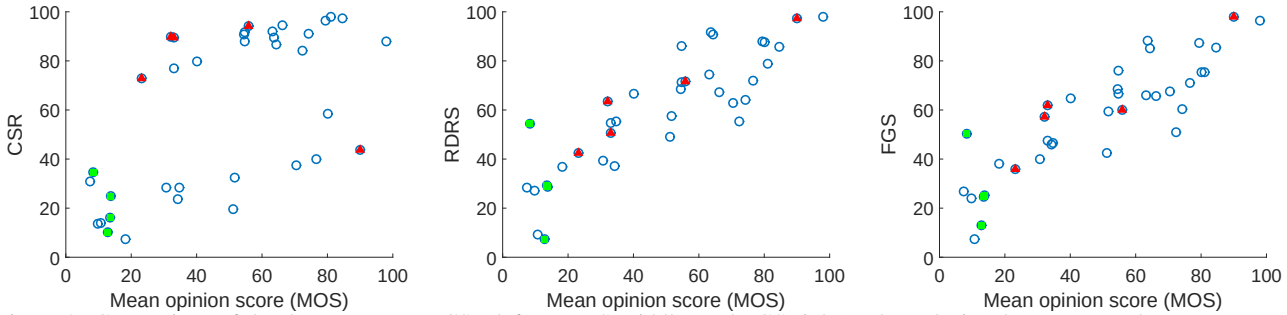


Figure 2: Comparison of the three measures: CSR(left), RDRS(middle) and FGS(right). The red triangles represent the extreme cases with the highest difference between the CSR and MOS scores (see Table 2). The green squares show the examples where the DR predictions result in higher error compared to CSR (see Table 3).



Figure 3: Two extreme case images: HannockKitchenOutside (CSR=89.837; MOS=32.200)(left) contains a light bulb with a very bright but very small area; and URChapel(2)_middle (CSR=89.548; MOS=33.150)(right) which contains many very small areas with extremely low pixel values.

Table 2: Extreme cases where the difference between the CSR and MOS were the highest, showing the improvement of the scores in tested measures. All computed scores are scaled to MOS range.

Image name	CSR	RDRS	FGS	FBLS*	MOS
Flamingo	94.082	71.487	59.899	66.386	56.00
HancockKitchenOut	89.837	63.351	57.322	41.300	32.20
HdrMark	43.638	97.406	98.100	84.187	89.90
NorthBubble	72.899	42.584	35.775	32.438	23.25
URChapel(2)_m	89.548	50.557	61.931	55.572	33.15
ℓ_2-error	50.116	19.715	18.432	12.731	

does increase the prediction error, as illustrated in Figure 2 by green squares and reported in Table 3. By conducting a further investigation, it was found that the images with the lowest MOS scores are the most error-prone with the robust measures. However, the prediction error in this range of MOS is relatively low compared to the extreme cases found at higher DR, where CSR error can be extremely high and more misleading in practice.

Conclusions

Accurately and consistently calculating DR of HDR images and its relation to the HVS is still an unresolved problem. There are some common procedures, generally accepted in the HDR community, but these, to the best of authors' knowledge, have never been validated.

In this paper, the scores obtained from different DR measures, including classical calculation (CSR) of the ratio between the maximum and minimum values, robust computation using percentiles, and methods based on filtering and multiscale representations were systematically compared with the perceived DR (measured by MOS). The Spearman's and Pearson's correlations with the MOS were computed, along with the prediction error. The results show that the robust measures can significantly improve over the CSR measure. Interestingly, all these methods achieve comparable performance, although at a different computational complexity, which leans toward adopting simple percentile-based clipping in practical applications.

Table 3: Images where the scores obtained by the tested measures are worse than CSR. All the images are the ones with the lowest MOS scores in the dataset. All computed scores are scaled to MOS range.

Image name	CSR	RDRS	FGS	FBLS*	MOS
DelicateFlowers	16.095	29.445	24.539	21.128	13.55
OCanadaNoLights_b	34.720	54.362	50.281	41.967	8.35
TupperLake(1)	10.225	7.400	13.047	7.400	12.85
URChapel(2)_b	24.823	28.531	25.248	27.325	13.75
ℓ_2-error	12.895	22.882	20.056	16.742	

The reason why those measures can better describe the perceived DR is mainly found in the characteristics of the content, and especially in the fact that very small but bright regions, which are taken into account in CSR computation, are not sufficiently extended to stimulate an equivalent response in the HVS. Nonetheless, it is hard to make final conclusions that apply to all the images, due to the complexity and multidimensionality of the perception of the DR attribute.

While more advanced studies are necessary in order to build a comprehensive model that will work on any image, this study shows that simple adjustments of the basic CSR measure are already able to substantially increase the accuracy of DR computation, without the need to resort to sophisticated and application-specific techniques as in [3].

In the future, we plan to further investigate this phenomenon by focusing on extreme cases that have the highest variance across measures. We will also take into account other features, such as area or distance between the peak values in the image, in a more comprehensive way and embedding explicitly well-known psychophysical laws.

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] Alan Gilchrist, Christos Kossyfidis, Frederick Bonato, Tiziano Agostini, Joseph Cataliotti, Xiaojun Li, Branka Spehar, Vidal Annan, and Elias Economou. An anchoring theory of lightness perception. *Psychological review*, 106(4):795, 1999.

[7] 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 (QoMEX2016)*. IEEE, 2016.

[8] SIM2. http://www.sim2.com/hdr/, June 2014.

[9] Ahmet Oğuz Akyüz and Erik Reinhard. Color appearance in high-dynamic-range imaging. *Journal of Electronic Imaging*, 15(3):033001–033001, 2006.

[10] Giuseppe Valenzise, Francesca De Simone, Paul Lauga, and Frederic Dufaux. Performance evaluation of objective quality metrics

for hdr image compression. In *SPIE Optical Engineering+ Applications*, pages 92170C–92170C. International Society for Optics and Photonics, 2014.

[11] Peter J. Burt and Edward H. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983.

[12] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *6-th International Conference on Computer Vision*, pages 839–846. IEEE, 1998.

[13] Ronald A Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

Author Biography

Vedad Hulusić is a post-doctoral researcher at Mines-Télécom Institute in Paris, France. He worked as an Assistant Professor at the University Sarajevo School of Science and Technology, Bosnia and Herzegovina. He has a PhD in Engineering from the University of Warwick, UK. He has a long-standing interest in the computer graphics, high dynamic range imaging, image and video quality assessment, visual perception and attention, and cross-modal interaction in which he has been a published author.