PRACTICALITIES OF PREDICTING QUALITY OF HIGH DYNAMIC RANGE IMAGES AND VIDEO

Rafał K. Mantiuk

Computer Laboratory, University of Cambridge, UK

ABSTRACT

The paper discusses the use of existing metrics, such as HDR-VDP and extensions of MS-SSIM and PSNR, for prediction of quality in high dynamic range (HDR) images and video. The discussion is based on the experience in using those metrics to evaluate and improve image compression for the new JPEG XT standard, and video compression for the LumaHDR open source codec. The paper explains why existing non-HDR metrics perform very poorly on HDR data and how to improve their predictions. Since most HDR metrics require calibrated data, intended for an HDR display, such calibration step is explained. One of the popular HDR quality metrics, HDR-VDP, is briefly introduced with the update on the latest improvements. Finally, several studies comparing objective HDR metric performance are summarized.

***Index Terms*—** HDR quality, objective metrics, HDR-VDP, perceptual metrics

1. INTRODUCTION

The recent interest in compression of high dynamic range (HDR) images and video sparked renewed interest in objective quality metrics, which could be used for evaluation of new HDR compression methods. The lack of well established HDR metrics, however, led to some confusion as to which metrics are suitable for the task. As the requirements for quality assessment of HDR are different from standard (low dynamic range) images and video, both the metrics and the way they need to be used are different. This paper is intended to clarify those differences and provide practical remarks on using HDR objective quality metrics. The paper also reviews a selection of the available objective metrics, with the focus on the work by the author.

Early attempts of computing quality for high dynamic range content involved calculating quality indices at multiple exposures and averaging such predictions [1]. Although such approach gives better predictions than computing quality in linear luminance domain [2], it adds unnecessary complexity and computation, which can be avoided with perceptually uniform spaces, discussed in Section 2.2. In contrast to traditional image quality metrics, perceptually uniform color spaces make the quality prediction dependent on the brightness and contrast of a display. Because of that, they require input images to represent absolute photometric values, as discussed in Section 2.3. One of the first comprehensive image difference metrics intended for HDR images was HDR-VDP [3] by the author of this paper. The metric, however, was intended for predicting probability of detecting differences in different parts of an image and was not intended to predict the overall image quality. That functionality was added in HDR-VDP-2, discussed in Section 3. In this paper we do not discuss but acknowledge the latest quality metric, HDR-VQM [4], which employs perceptual uniform encoding, subband decomposition and spatio-temporal pooling to predict quality of HDR video. This paper is focused on image fidelity and it does not cover HDR metrics intended for tone-mapping [5, 6].

2. PRACTICALITIES OF HDR METRICS

2.1. LDR luma vs. HDR luminance

Although image quality could be, in principle, computed directly on HDR pixel values using existing LDR metrics, such approach is conceptually incorrect. This is because the majority of LDR metrics assume that the input values are approximately perceptually uniform. While *luma* values in LDR images have this property, the same cannot be said about *luminance* values found in HDR images.

Fig. 1 shows an example of this problem. The rectangles shown in this figure contain a random noise of the same amplitude in terms of luma values. Depending on the medium (paper or a display) on which this figure is seen, the visibility of noise is similar across all brightness levels, with perhaps slightly lower visibility for the brightest and the darkest patches. The PSNR computed from luma values (LDR-PSNR) is obviously the same for all 4 brightness levels. The rectangles can be transformed from LDR luma to HDR luminance by inverse gamma mapping; raising the luma values to the power 2.2. When the PSNR is computed from HDR luminance values (HDR-PSNR), the metric results differ widely between brightness levels: the brightest patch has the quality 20 dB worse then the darkest patch, which definitely does not reflect the perceived difference. This shows that computing PSNR and other LDR metrics directly on HDR lumi-

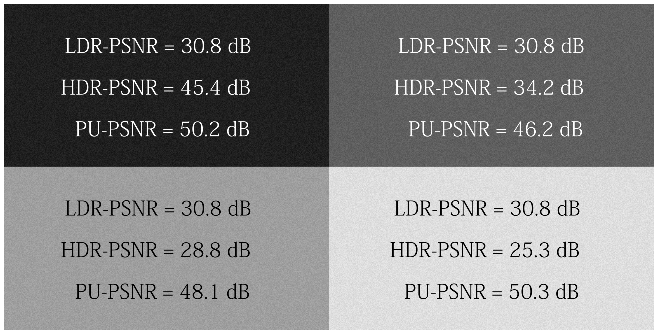

Fig. 1. PSNR values can differ substantially depending how they are computed. The four rectangular patches contain white noise on 4 background brightness levels. When PSNR is computed for LDR luma values (LDR-PSNR), the distortion (noise) has the same impact on PSNR. However, if the images are converted to HDR luminance values by inverse gamma function (2.2), the HDR-PSNR values vary widely between different brightness levels. The PSNR computed on perceptually uniform values (PU-PSNR) corrects for differences in noise visibility at different brightness levels and produces values that better correlate with perceived quality.

nance values results in overpredicted visibility of distortions for bright regions of an image.

The obvious solution is to convert perceptually non-uniform luminance into perceptually uniform luma. But unfortunately the "gamma correction" formula cannot be used to convert HDR luminance values to LDR luma. This is because "gamma" well approximates the non-linearity of luminance perception only within a small range of values, restricted to the luminance range of CRT displays, for which it was intended. Outside that range, "gamma correction" greatly overestimates the visibility of very bright features. Similarly, CIE Lab or similar uniform color spaces are not suitable as they were not intended for HDR images.

The simplest approach to transform HDR pixel values into approximately perceptually uniform units is to compute the logarithms of pixel luminance. Such approach is commonly used in many tone-mapping operators but also in some HDR pixel encodings [7]. The logarithm unifies contrast differences assuming the Weber-Fechner law of contrast perception [8], making the resulting values better aligned with the perceived brightness of HDR pixels. However, the Weber-Fechner law is only a rough approximation, which overpredicts distortion visibility at low luminance levels. A better uniformity can be achieved with perceptual uniform encoding, discussed in the next section.

2.2. Perceptual uniform encoding

Aydin et al. [9] proposed a simple luminance encoding that makes it possible to use existing LDR metrics, such as

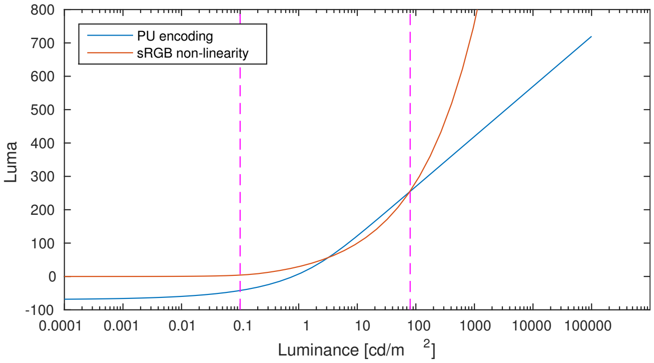

Fig. 2. Perceptually uniform (PU) encoding for evaluating quality of HDR images. The absolute luminance values are converted into luma values before they are used with standard image quality metrics, such as MSE, PSNR or SSIM. Note that the PU encoding is designed to approximately match the magnitude of sRGB non-linearity within the range $0.1 - 80$ cd/m^2 so that the results for low dynamic range images are consistent with those computed in the sRGB color space. This, however, requires that some PU-encoded values are negative.

PSNR or SSIM, with HDR images. The encoding transforms physical luminance values (represented in cd/m^2) into an approximately perceptually uniform (PU) representation (refer to Figure 2). The transformation is derived from luminance detection data. It is further constrained so that the luminance values produced by a typical CRT display (in the range 0.1–$80\,cd/m^2$) are mapped to the 0–255 range to mimic the sRGB non-linearity. This way, the quality predictions for typical low-dynamic range images are comparable to those calculated using pixel values while the metric is also able to operate on a much greater range of luminance.

Our earlier example of rectangular noise patches in Figure 1 includes also the results for PSNR computed for PU-encoded values. The PU-PSNR is actually larger for the darkest patch and the brightest patches, as the noise is slightly less visible for those (though this may depend on the display medium). The PU-PSNR predictions arguably better corresponds with the visibility of the noise on those patches, especially when compared to HDR-PSNR values.

The PU encoding shown in Figure 2 is a revision from the original work [9]. The revision uses a more recent Contrast Sensitivity Function from [10] instead of the historical t.v.i. measurements to derive the curve. The source code for the encoding can be found at `http://goo.gl/rpfkB9`. PU encoding is not the only option for transforming HDR images into perceptual space and some recent papers used instead Perceptual Quantizer (PQ) [11]. PQ was originally proposed as a transfer function used for encoding HDR video. The function is conceptually very similar to PU encoding with the difference that other contrast sensitivity function was used for its derivation.

2.3. Display-referred metrics

One important feature of PU encoding and most HDR metrics is that they are sensitive to absolute luminance levels. An image shown on bright display is more likely to reveal artefacts than the same image shown on a dark display, so a metric should reflect that. Most LDR metrics do not account for the differences between displays as they base their predictions on pixel values, which are the same regardless of the contrast or brightness of a display. In contrast to that, HDR metrics often operate on the output luminance produced by a given display.

This property of HDR metrics poses certain difficulty as most HDR images and video are represented in relative units, which do not directly correspond to physical absolute luminance produced by a display. Ideally, the content intended for an HDR metric should be mapped to a target display, so that the HDR pixel values represent absolute color and luminance values emitted from that display. In the simplest case such mapping may involve just multiplication by a constant, so that the peak image value is mapped to the display peak luminance, for example 4,000 cd/m^2 for a bright HDR display. If the content must be processed automatically, it is possible to tone-map both images and video using a display adaptive tone-mapping[1], which can tone-map for both LDR and HDR displays.

2.4. RGB vs. luminance-only metrics

The PU-encoding is typically used on the image luminance channel, while ignoring color information. An obvious extension is to compute PU-encoded values for red, green and blue color channels and then compute aggregate PSNR for all of them, as commonly done in case of LDR images. Such approach should be able to detect distortion in color, which can be missed by a luminance-only metric.

In practice, however, PU-PSNR computed for all color channels may perform substantially worse than for luminance alone. Table 1 shows the ranking of HDR image metrics according to several metric performance indices from [12]. In almost all cases, the color RGB metrics (with *_RGB* suffix) perform significantly worse than luminance-only counterparts (*_Y* suffix). A better performance of luminance-only metrics was also demonstrated in [2]. One potential reason for worse performance of RGB metrics is that they do not differentiate between highly visible luminance distortions from much less visible chroma distortions. At the same time luminance-only metrics do not perform poorly for color distortions because most color distortions affect both luminance and chroma channels.

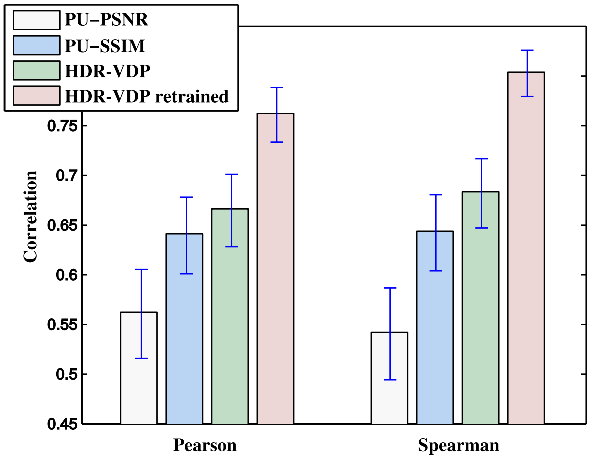

Fig. 3. Comparison of HDR-VDP performance before (HDR-VDP v2.1.3) and after training (HDR-VDP v.2.2) with the dataset of 2932 images. The results for PU-SSIM and PU-PSNR (see Section 2.2) are also included. The error bars denote 95% confidence interval for cross-validation results (70/30 split). Figure adapted from [4].

3. HDR-VDP-2

HDR-VDP-2[2] is the visibility (discrimination) and quality metric capable of detecting differences in HDR images [10]. The metric originates from the classical Visual Difference Predictor [13], and its extension — HDR-VDP [3]. What makes HDR-VDP-2 different from other quality metrics, such as HDR-VQM [14], or MSSIM [15], is that it is based on a comprehensive model of detection and discrimination, which has been calibrated and validated on a large dataset of psychophysical data, including ModelFest [16], historical Blackwell's t.v.i. measurements [17], and newly measured contrast sensitivity data [18].

Most quality metrics are trained with quality data alone. Such data usually consists of a set of distorted images and the corresponding mean-opinion-scores (MOS) — one number per image. Since quality datasets contain at most a few thousands of images (more often a few hundreds), the data used to train such metrics is rather limited, especially when compared to datasets used in computer vision and machine learning, which can contain millions of images. The best evidence that quality datasets are often inadequate in terms of sample size is demonstrated by a number of papers comparing the performance of quality metrics with subjective data, in which the reported metric performance can vary widely between the datasets.

In the absence of very large quality datasets, especially

[1] Display adaptive tone-mapping is provided in *pfstools* software: `http://pfstools.sourceforge.net`

[2] HDR-VDP-2 source code is available at `http://hdrvdp.sourceforge.net`

Table 1. Accuracy (PLCC and RMSE) and consistency (OR) indexes for several objective metrics computed for the dataset of 80 images. No evidence for statistical difference was found for metrics whose performance indexes are underlined. The table adapted from [12].

(a) Pearson Linear Correlation Coefficient (PLCC).

LOG_PSNR_RGB	PSNR	SNR	PU2PSNR_RGB	W_RMSE_RGB	MRSE	W_RMSE_Y	PU2PSNR_Y	LOG_PSNR_Y	PU2SSIM	PU2MSSIM	HDRVDP_Q
0.6548	0.6800	0.7128	0.7340	0.7386	0.7527	0.8812	0.8839	0.8881	0.9231	0.9447	0.9510

(b) Root-Mean-Square Error (RMSE).

LOG_PSNR_RGB	PSNR	SNR	PU2PSNR_RGB	W_RMSE_RGB	MRSE	W_RMSE_Y	PU2PSNR_Y	LOG_PSNR_Y	PU2SSIM	PU2MSSIM	HDRVDP_Q
0.9487	0.9204	0.8805	0.8526	0.8466	0.8266	0.5941	0.5873	0.5770	0.4831	0.4133	0.3882

(c) Outlier ratio (OR).

PSNR	W_RMSE_RGB	SNR	MRSE	LOG_PSNR_RGB	PU2PSNR_RGB	W_RMSE_Y	PU2PSNR_Y	LOG_PSNR_Y	PU2SSIM	PU2MSSIM	HDRVDP_Q
0.7625	0.7375	0.7208	0.7167	0.7000	0.6917	0.6208	0.5958	0.5833	0.5583	0.5250	0.3500

for HDR images and video, one strategy is to build a metric that relies on low-level vision, which is mostly well understood and can be tested against a substantial amount of psychophysical data. The assumption here is that when the brain makes quality judgments, it operates on a signal that has been distorted by early vision. Therefore, a quality metric should mimic this behavior and also model early vision. This may lead to metrics, which are potentially more robust to new data even if trained with limited datasets.

For those reasons, HDR-VDP-2 accounts for many early vision phenomena, such as scattering of the light in the eye and optics (glare), rod and cone vision, local adaptation, luminance masking, spatial contrast sensitivity, contrast masking, neural noise in visual channels and contrast constancy. Such early vision processing makes the metric sensitive to many factors, which most metrics ignore. HDR-VDP-2 predictions will be different depending on display brightness, viewing distance and spectral emission of the color primaries used in the display. But, also because of that, the data supplied to HDR-VDP-2 needs more effort to prepare. The images need to be calibrated in absolute units of $\mathrm{cd/m^2}$, as discussed in Section 2.3. As the metric is sensitive to the viewing distance, it is necessary to provide image angular resolution in pixels per visual degree. The spectral emission curves for red, green and blue primaries can be optionally specified, or one of the default curves, for CRT and LCD displays with different backlight, can be selected.

The metric has undergone several software revision, from which the latest improves quality predictions. The quality prediction for earlier releases (prior to 2.2) was trained solely with the low dynamic range quality datasets (LIVE and TID2008), as no datasets for high dynamic range were available. From version 2.2, HDR-VDP-2 is calibrated and cross-validated with over 2900 distorted images, coming from four different datasets: CSIQ and TID2008 for LDR images, and datasets from image compression and tone-mapping studies [19, 20] for HDR images. Interestingly, the improvement in quality predictions only required altering metric parameters without any changes to the model. We found the model derived from the LDR image data performs equally well on the HDR image data. The performance of HDR-VDP v2.2, compared to HDR-VDP v2.1.3, PU-PSNR and PU-SSIM is shown in Figure 3.

4. METRIC PERFORMANCE

Table 1 and Figure 3 are the examples of two studies demonstrating good performance of HDR-VDP-2 for compressed images. HDR-VDP-2 was also shown to correlate well with subjective studies in [21, 2, 22]. It must be noted, however, that studies performed before 2015 used the earlier version of the metric (v2.1.3). [23] found HDR-VDP-2 (v2.1.3) to well predict compression artifact, however VIF in PU-space performed much better for other types of distortions, including intensity shifting, salt and pepper noise and low-pass filtering. The study in [14] demonstrated that HDR-VQM and SSIM (in PU space) performed better than HDR-VDP-2 (v2.2) for video content. But the study in [22] showed a better performance of HDR-VDP-2 (v2.2) for content compressed with HDR extension of HEVC. Given that most studies have been performed on datasets of less than 100 images or video sequences, it is difficult to draw general conclusions on the performance of HDR metrics and larger datasets are needed to better assess their predictive power.

5. CONCLUSIONS

The paper discussed several practical issues to be considered when using HDR quality metrics, including the need to convert HDR pixel values into the perceptually uniform space, requirement for display-referred data, and the advantage of luminance-only metrics. A brief overview of HDR-VDP-2 gives a rationale for the metric that is based on the model of early human vision. Several independent studies confirmed good performance of HDR-VDP-2, though in some studies simple metrics based on PU-encoding, or the specialized video metric HDR-VQM, proved to be better.

6. REFERENCES

[1] J. Munkberg, P. Clarberg, J. Hasselgren, and T. Akenine-Möller, "High dynamic range texture compression for graphics hardware," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 698, jul 2006.

[2] M. Rerabek, P. Hanhart, P. Korshunov, and T. Ebrahimi, "Subjective and objective evaluation of HDR video compression," in *Int. Workshop on Video Proc. and Quality Metrics for Consumer Electronics (VPQM)*, 2015, number EPFL-CONF-203874.

[3] R. Mantiuk, S.J. Daly, K. Myszkowski, and H.-P. Seidel, "Predicting visible differences in high dynamic range images: model and its calibration," in *Human Vision and Electronic Imaging*, 2005, pp. 204–214.

[4] 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. 010501, jan 2015.

[5] T.O. Aydin, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Dynamic range independent image quality assessment," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 69, 2008.

[6] H. Yeganeh and Z. Wang, "Objective Quality Assessment of Tone-Mapped Images," *IEEE Trans. Img. Proc.*, vol. 22, no. 2, pp. 657–667, 2013.

[7] G. Ward-Larson, "LogLuv Encoding for Full-Gamut, High-Dynamic Range Images," *Journal of Graphics Tools*, vol. 3, no. 1, pp. 15–31, jan 1998.

[8] R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Lossy Compression of High Dynamic Range Images and Video," in *Human Vision and Electronic Imaging*, 2006, p. 60570V.

[9] T.O. Aydn, R. Mantiuk, and H.-P. Seidel, "Extending quality metrics to full luminance range images," in *Human Vision and Electronic Imaging*, 2008, pp. 68060B–10.

[10] R. Mantiuk, K.J. Kim, A.G. Rempel, and W. Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Transactions on Graphics*, vol. 30, no. 4, pp. 40:1–40:14, jul 2011.

[11] S. Miller, M. Nezamabadi, and S. Daly, "Perceptual Signal Coding for More Efficient Usage of Bit Codes," *SMPTE Motion Imaging Journal*, vol. 122, no. 4, pp. 52–59, may 2013.

[12] A. Artusi, R. K. Mantiuk, T. Richter, P. Hanhart, P. Korshunov, M. Agostinelli, A. Ten, and T. Ebrahimi, "Overview and evaluation of the JPEG XT HDR image compression standard," *Journal of Real-Time Image Processing*, vol. in print, 2015.

[13] S. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, Andrew B. Watson, Ed., pp. 179–206. MIT Press, 1993.

[14] 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.

[15] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, apr 2004.

[16] A.B. Watson and A.J. Ahumada Jr, "A standard model for foveal detection of spatial contrast," *Journal of Vision*, vol. 5, no. 9, pp. 717–740, 2005.

[17] H.R. Blackwell, "Contrast thresholds of the human eye," *J. Opt. Soc. Am*, vol. 36, pp. 624–632, 1946.

[18] K.J. Kim, R. Mantiuk, and K.H. Lee, "Measurements of achromatic and chromatic contrast sensitivity functions for an extended range of adaptation luminance," in *Human Vision and Electronic Imaging*, 2013, p. 86511A.

[19] 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, no. 10, pp. 102008, oct 2013.

[20] M. Narwaria, M.P. Da Silva, P. Le Callet, and R. Pépion, "Impact of tone mapping in high dynamic range image compression," in *Proc. of VPQM*, 2014, pp. pp. 1–6.

[21] P. Hanhart, M.V. Bernardo, P. Korshunov, M. Pereira, A.M.G. Pinheiro, and T. Ebrahimi, "HDR image compression: A new challenge for objective quality metrics," in *QoMEX*. sep 2014, pp. 159–164, IEEE.

[22] P. Hanhart, M. Rerábek, and T. Ebrahimi, "Towards high dynamic range extensions of HEVC: subjective evaluation of potential coding technologies," in *SPIE App. of Digit. Img. Proc.*, 2015, p. 95990G.

[23] M. Azimi, A. Banitalebi-Dehkordi, Y. Dong, M.T. Pourazad, and P. Nasiopoulos, "Evaluating the performance of existing full-reference quality metrics on high dynamic range (HDR) video content," in *Int. Conf. on Multimedia Signal Proc. (ICMSP)*, 2014.