Journal of Electronic Imaging

JElectronicImaging.org

HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images

Manish Narwaria
Rafal K. Mantiuk
Mattheiu Perreira Da Silva
Patrick Le Callet

SPIE. IS&T imaging.org

HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images

Manish Narwaria,[a] Rafal K. Mantiuk,[b,*] Mattheiu Perreira Da Silva,[a] and Patrick Le Callet[a]
[a]LUNAM University, IRCCyN CNRS UMR 6597, Polytech Nantes, Rue Christian Pauc, La Chantrerie B.P. 50609 44306, Nantes Cedex 3, France
[b]Bangor University, School of Computer Science, Dean Street, Bangor, LL57 1UT, United Kingdom

Abstract. With the emergence of high-dynamic range (HDR) imaging, the existing visual signal processing systems will need to deal with both HDR and standard dynamic range (SDR) signals. In such systems, computing the objective quality is an important aspect in various optimization processes (e.g., video encoding). To that end, we present a newly calibrated objective method that can tackle both HDR and SDR signals. As it is based on the previously proposed HDR-VDP-2 method, we refer to the newly calibrated metric as HDR-VDP-2.2. Our main contribution is toward improving the frequency-based pooling in HDR-VDP-2 to enhance its objective quality prediction accuracy. We achieve this by formulating and solving a constrained optimization problem and thereby finding the optimal pooling weights. We also carried out extensive cross-validation as well as verified the performance of the new method on independent databases. These indicate clear improvement in prediction accuracy as compared with the default pooling weights. The source codes for HDR-VDP-2.2 are publicly available online for free download and use. [DOI: 10.1117/1.JEI.24.1.010501]

Keywords: high-dynamic range; objective quality; HDR-VDP-2.

Paper 14600L received Sep. 26, 2014; accepted for publication Dec. 29, 2014; published online Jan. 22, 2015.

1 Introduction

Human eyes have a remarkable ability to adapt and adjust to varying luminance conditions. As a result, humans can clearly visualize and see in lighting conditions ranging from a moonlit night to bright sunshine. In terms of physical luminance values, the former is in the range of about 10^{-2} cd/m^2, while the latter is more than 10^7 cd/m^2, a dynamic range in excess of 9 orders of magnitude. However, when it comes to scene capture and display, such large luminance ranges are beyond the capabilities of current standard dynamic range (SDR) imaging systems. Nevertheless, with the emergence of HDR imaging, it is now possible to capture and display scenes that can encapsulate much a higher dynamic range (HDR) than the traditional or SDR imaging techniques.[1] Particularly, typical SDR systems deal with signals up to 3 orders of magnitude. In contrast, with HDR imaging, scenes up to 5 orders of magnitude can be processed and displayed and it can also include SDR signals (e.g., tone-mapped signals). Therefore, it is logical to assume that the future video processing systems will have to deal with both SDR and HDR signals.

2 Background and Motivation

While human judgments of perceptual visual quality remain the most accurate, they cannot be employed in all situations. For instance, in a real-time video streaming application, it may be unfeasible to get human judgments of visual quality to continuously monitor the traffic from a quality aspect. In the light of such scenarios, objective quality measurement via the use of a computational model is more desirable. To that end, many objective methods have been proposed in the past. However, most of them have been designed for and tested only on SDR visual signals.[2] As mentioned before, with the emergence of HDR imaging, video processing systems may have to deal with both HDR and SDR signals. Thus, an objective quality measurement method that could potentially be applicable over a larger dynamic range (i.e., both SDR and HDR domains) is desirable. In that context, the HDR-VDP-2 algorithm[3] can be an attractive solution.

HDR-VDP-2 is a visibility prediction metric. It provides a two-dimensional map with probabilities of detection at each pixel point, which is obviously related to the perceived quality because a higher detection probability implies a higher distortion level at the specific point. However, in the case of supra-threshold distortions (i.e., distortions clearly visible to the eye), the error visibility will mostly be 1, and in such cases a single number denoting the visual quality is more desirable. This can be accomplished via the pooling of errors in the frequency bands. In the original implementation, the pooling weights were determined by optimization on an existing SDR dataset. There are, however, three limitations of that approach, especially in the context of dealing with SDR and HDR conditions. First, the original paper[3] used only an SDR image quality dataset which did not include any HDR images. Second, the optimization was done on a relatively small number of images. Finally, since the optimization was unconstrained, it lead to negative pooling weights that may not be easily interpretable.

This letter seeks to address the specific issues raised with regards to pooling in HDR-VDP-2. To that end, we reoptimized the pooling weights on a combined dataset of subjectively rated HDR and SDR images. As a result, the newly calibrated model is expected to be more effective across both HDR and SDR test conditions. Second, we also reformulated the said optimization as being constrained, due to which the resultant weights can be computed in a bounded manner, leading to better interpretability. Finally, we verified the prediction performance of the new weights via extensive cross-validation studies on a collection of nearly 3000 images (including HDR and SDR content and their corresponding subjective quality ratings).

3 Method Calibration

In HDR-VDP-2, the following equation is used to predict the quality score Q_{hdrvdp} for a distorted image with respect to its reference:

*Address all correspondence to: Rafal K. Mantiuk, E-mail: mantiuk@gmail.com

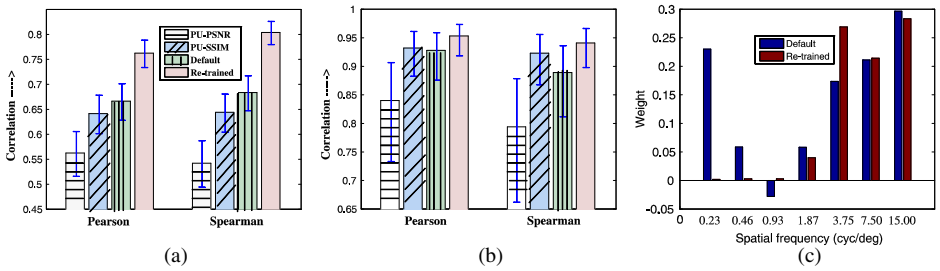

Fig. 1 Comparative prediction performance: (a) cross-validation tests (exp 1), 900 test images, (b) new HDR dataset (exp 2), 50 test images, and (c) plot of retrained and default weights. Error bars indicate 95% confidence intervals. The same colors of the bars are used for (a) and (b).

$$Q_{\text{hdrvdp}} = \frac{1}{F.O}\sum_{f=1}^{F}\sum_{o=1}^{O}\mathbf{w}_f \log\left(\frac{1}{I}\sum_{i=1}^{I}\mathbf{D}_p^2[f,o](i)+\varepsilon\right), \quad (1)$$

where i is the pixel index, $\mathbf{D}_p$ denotes the noise-normalized difference between the f'th spatial frequency ($f = 1$ to F) band and o'th orientation ($o = 1$ to O) of the steerable pyramid for the reference and test images, $\varepsilon = 10^{-5}$ is a constant to avoid singularities when $\mathbf{D}_p$ is close to 0, and I is the total number of pixels. In the above equation, $\mathbf{w}_f$ is the vector of per-band pooling weights, which can be determined by maximizing correlations with subjective opinion scores. However, unconstrained optimization in this case may lead to some negative $\mathbf{w}_f$. Since $\mathbf{w}_f$ determines the weight (importance) of each frequency band, a negative $\mathbf{w}_f$ is implausible and may indicate overfitting. Therefore, in this letter, we introduce a constraint on $\mathbf{w}_f$ during optimization.

Let $\mathbf{Q}_{\text{hdrvdp}}$ and $\mathbf{S}$, respectively, denote the vector of objective quality scores from HDR-VDP-2 and subjective scores for a given set of N images. Then, the aim is to maximize the Spearman rank-order correlation between the two vectors with $\mathbf{w}_f$ being the optimized variables. To that end, we first rank the values in $\mathbf{Q}_{\text{hdrvdp}}$ and $\mathbf{S}$ from 1 to N and obtain new vectors $\mathbf{R}_{\text{hdrvdp}}$ and $\mathbf{R}_{\text{subjective}}$, which consist of the respective ranks. Further, define $\mathbf{E} = \mathbf{R}_{\text{hdrvdp}} - \mathbf{R}_{\text{subjective}}$ as the rank difference vector. Then, the optimization problem can be denoted as

$$\underset{\mathbf{w}_f}{\text{maximize}}\left(1 - \frac{6\sum_{i=1}^{N}\mathbf{E}_i^2}{N(N^2-1)}\right), \quad \text{subject to } \mathbf{w}_f \geq \mathbf{0}. \quad (2)$$

Also note that in our case, the said optimization is solved using the Nelder–Mead method, which does not require computing gradients. This is because our objective function is not continuous and differentiable as we use the Spearman rank-order correlation. Our aim was to calibrate the metric so that it can handle both HDR and SDR conditions. Thus, we computed the optimized $\mathbf{w}_f$ based on a set of subjectively rated SDR and HDR images. In particular, our study used two recent HDR datasets,[4,5] in which there is a total 366 subjectively rated compressed HDR images. In contrast to the HDR case, there are several SDR datasets that are publicly available, and we selected the two biggest ones in terms of the number of images (TID2008[6] and CSIQ[7]). Note that these datasets use different rating methodologies. Therefore, for the HDR (scale of 1 to 5) and TID2008 (scale of 1 to 9) datasets, which report mean opinion scores (MOSs), we first converted the MOSs to difference MOS. On the other hand, the CSIQ dataset reports difference mean opinion score (DMOS). Finally, we rescaled all the DMOSs between 0 and 100. This enabled a more consistent scale of rating scores during optimization.

4 Cross-Validation Results

In this section, we outline the method used to verify the performance of new weights. Recall that we used four datasets: two each for the HDR and SDR cases. The former has a total of 10 source (reference) contents, while the latter has a total of 55 (30 in CSIQ and 25 in TID2008) source contents. So, there were 65 source contents in total and 2932 distorted contents (obtained by applying different distortion types and levels to the source content). For the cross-validation studies, we selected all the distorted images from 45 (this corresponds to about 70%) source contents as the training set to find the optimal $\mathbf{w}_f$ vector, and the remaining images from 20 source contents were used as the test set. To enable a more robust estimate of the prediction performance, we randomly repeated the said division into training and test sets over 1000 iterations, and it was ensured that the two sets were different in terms of the source content. Hence, in each of the 1000 iterations, the prediction performance was assessed only for untrained content, thus providing a reasonably robust approach toward content-independent verification. The reader may also be informed that with the stated data partition (45 source contents as training set and remaining as test set), there were an average of 2032 and 900 images, respectively, in the training and test sets, during each iteration.

The experimental results for this case (exp 1) are shown in Fig. 1(a), where the performance is measured in terms of mean (over 1000 iterations) values of Pearson and Spearman correlation values (a higher value implies better for these measures). Recall that the existing LDR methods cannot be directly used for HDR. Nevertheless, we also employed peak signal to noise ratio (PSNR) and structural similarity index measure (SSIM) (both are LDR methods) on a perceptually uniformly (PU)[8] transformed HDR signal to compute objective quality and to provide a base line for comparison. One can notice that the prediction performance using the weights obtained from the training set is better than the default weights as well as the two modified LDR methods. We have also plotted in the same figure, the 95% confidence intervals (using error bars) to provide an indication of uncertainty in the measured values. As can be seen, the confidence intervals do not overlap indicating a better performance with the trained weights from statistical considerations. It was also found that the retrained weights lead to a larger improvement in case of HDR images but did not jeopardize the prediction accuracy for the LDR case, and this improved the overall

prediction performance. Finally, to verify the consistency in prediction, we computed the number of outliers over 1000 iterations based on box plots (which are convenient tools to visualize data variability and detect points outside the quartiles). We found that for all the cases, the number of outliers was less than 1% of the total points, indicating good consistency.

5 Validation on Independent Dataset

The results in the previous section revealed that the trained weights lead to better prediction accuracies with content-independent training and test sets. This section provides further evidence of that by using another independent HDR dataset reported in Ref. 9. Note that this dataset includes source content and a set of distortions that did not appear in any of the other datasets we used for calibration. To perform the experiment (exp 2), we first obtained the optimal $\mathbf{w}_f$ by using all the datasets used in the previous section. The resultant $\mathbf{w}_f$ (referred to as the new weights) was then used to predict the quality of HDR images in the new dataset. The comparative results along with those from PU-PSNR and PU-SSIM are shown in Fig. 1(b), from which we can see that the new weights improved the prediction accuracies over default weights as well as the two modified LDR methods. Note that the statistical differences are not apparent because of the much smaller size of the dataset: 50 images versus 900 used in Sec. 4.

Finally, we compare the retrained and default weights via the frequency versus weight plot shown in Fig. 1(c). The frequency is expressed in cycles per degree (cyc/deg), and the left and right bars at each cyc/deg, respectively, indicate default and retrained weights. We notice that the retrained weights reduce the importance of low frequency bands. However, it may also be mentioned that they need not be related to the contrast sensitivity function because the goal of pooling is to quantify quality (or annoyance level) which may not always be at the level of visibility thresholds. Also note that the negative weights found in the original HDR-VDP-2 could cause an increase of quality with a higher amount of distortion. This situation is valid only in very specific cases such as denoising and contrast enhancement (where visual quality may be enhanced). However, since this condition is not included in any of the datasets that we used, the retrained weights lend to better physical interpretability (since all of them are positive, the quality will decrease with an increased level of distortion).

6 Concluding Remarks

Visual quality assessment is a useful tool in many image and video processing applications. In addition, the recent interests of the multimedia signal processing community in HDR imaging have lead to activities toward development and standardization of HDR image and video processing tools (e.g., extension of the JPEG standard to support HDR image compression). In such scenarios, an objective quality prediction tool is needed to validate such tools from the view point of visual quality benchmarking with both SDR and HDR signals. In that context, the contribution of this letter can be summarized as follows:

We identified and addressed the specific issue of feature pooling in HDR-VDP-2, and thus proposed the extension HDR-VDP-2.2. Specifically, we computed the pooling weights via constrained optimization on a set of subjectively rated SDR and HDR images, in order that the resultant metric would be effective across a large luminance range of the visual signal. This represents a clear advantage over existing SDR metrics that may not be directly applicable in the case of HDR signals.

We verified the performance of the new weights by way of extensive cross-validation and also on an independent HDR dataset. In this way, the prediction performance of HDR-VDP-2 (both with new and default pooling weights) has also been verified and benchmarked on a test bed with nearly 3000 HDR and SDR images.

With regards to the practical implications of the work reported in this letter, we note that the new version HDR-VDP-2.2 is a more accurate objective visual quality estimator for both HDR and SDR conditions. Hence, it is expected to be useful in standardizing HDR and SDR visual signal processing tools with regards to their impact on visual quality and can also be employed as a standalone quality predictor. While no objective quality method can entirely replace subjective opinion, nevertheless the proposed improved version HDR-VDP-2.2 can still be useful in certain scenarios and applications. A software implementation of HDR-VDP-2.2 is freely available for download at Ref. 10.

It should also be stressed that the retrained pooling weights are related to the characteristics of perceptual noise introduced in different frequency bands as a result of the processing considered in the datasets used. Thus, in the current work, we considered a standard processing method (including compression, tone mapping, and inverse tone mapping). Hence, HDR-VDP-2.2 is expected to be more accurate with the current use cases of HDR deployment in the HDR delivery chain. For other applications, the pooling weights may need to be revisited and possible profiles may be added to HDR-VDP-2.2.

Acknowledgments

This work was supported by COST Action IC1005 and the NEVEx project FUI11 financed by the French government.

References

1. F. Banterle et al., *Advanced High Dynamic Range Imaging: Theory and Practice*, AK Peters (CRC Press), Natrick, Massachusetts (2011). ISBN: 978-156881-719-4.
2. W. Lin and C. Kuo, "Perceptual visual quality metrics: a survey," *J. Visual Commun. Image Represent.* **22**, 297–312 (2011).
3. R. Mantiuk et al., "HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Trans. Graphics* **30**, 40 (2011).
4. M. Narwaria et al., "Tone mapping based high dynamic range image compression: study of optimization criterion and perceptual quality," *Opt. Eng.* **52**(10), 102008 (2013).
5. M. Narwaria et al., "Impact of tone mapping in high dynamic range image compression," in *Proc. Eighth Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)* (2014).
6. N. Ponomarenko et al., "Color image database for evaluation of image quality metrics," in *Proc. Int. Workshop Multimedia Signal Process*, Cairns, Queensland p. 403408, IEEE (2008).
7. E. Larson and D. Chandler, "Categorical image quality (CSIQ) database," 2010, http://vision.okstate.edu/csiq (September 2014).
8. T. Aydin et al., "Extending quality metrics to full luminance range images," *Proc. SPIE* **6806**, 68060B (2008).
9. G. Valenzise et al., "Performance evaluation of objective quality metrics for HDR image compression," *Proc. SPIE* **9217**, 92170C (2014).
10. http://hdrvdp.sf.net/.