View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Apollo

TRAINED PERCEPTUAL TRANSFORM FOR QUALITY ASSESSMENT OF HIGH DYNAMIC RANGE IMAGES AND VIDEO

Nanyang Ye, María Pérez-Ortiz and Rafał K.Mantiuk

Department of Computer Science and Technology, University of Cambridge, UK

ABSTRACT

In this paper, we propose a trained perceptually transform for quality assessment of high dynamic range (HDR) images and video. The transform is used to convert absolute luminance values found in HDR images into perceptually uniform units, which can be used with any standard-dynamic-range metric. The new transform is derived by fitting the parameters of a previously proposed perceptual encoding function to 4 different HDR subjective quality assessment datasets using Bayesian optimization. The new transform combined with a simple peak signal-to-noise ratio measure achieves better prediction performance in cross-dataset validation than existing transforms. We provide Matlab code for our metric [1].

***Index Terms*—** Image quality assessment, high dynamic range, perceptually uniform encoding

1. INTRODUCTION

Quality metrics are necessary to develop robust compression and processing algorithms for high dynamic range (HDR) imaging. However, given that the perception of linear red, green, blue or luminance values, found in HDR content, is strongly non-linear, standard low-dynamic-range quality metrics, such as Peak Signal-to-Noise Ratio (PSNR) or Structural Similarity Index (SSIM), cannot be directly used with HDR images and video. Linear HDR pixel values can be made more perceptually uniform by transforming them into the logarithmic domain [1, 2]. However, such a logarithmic transform does not account for the absolute brightness of an HDR display. A content shown on a brighter display will reveal more distortions than the same content shown on a darker display. Therefore, most widely used HDR metrics are display-referred and require HDR values to be adjusted by a display model so that they represent absolute luminance values (in cd/m^2) emitted from an HDR display. Such adjustment usually involves multiplying pixel values by a constant and clipping the values above or below the dynamic range of a particular display.

The display-referred HDR quality metrics include those that were specifically designed to handle HDR content, such as HDR-VDP [3, 4], HDR-VQM [5], or DRIQM [6], and those that were adapted from standard-dynamic-range (SDR) metrics to process HDR content. The adaptation involves a perceptual transform (PT) that converts linear HDR pixel values into perceptually uniform units, which can be directly used with SDR metrics [7], such as PSNR or SSIM. Figure 1 illustrates the typical processing blocks of PT-based metrics. The original HDR images are first transformed by a display model to simulate an HDR display and to obtain absolute display-referred color values. Then, a perceptual transform converts both distorted and reference images into perceptually uniform units, which could be input directly into an SDR quality metrics. Such an approach may not provide the best predictive performance but it leads to a simple, fast and differentiable quality metric, which could be easily used as a cost function in image processing algorithms. The property of being differentiable is especially important when the metric is used as a loss function in optimization-driven problems.

In this paper, we extend previous work on PT-based metrics [7], proposing a new version that improves the predictive performance of the PU-PSNR metric. Instead of deriving PT from contrast detection models (contrast sensitivity function), we use existing HDR subjective image quality datasets [8, 9, 10, 11] to fit the parameters of a new PT function. We provide Matlab code for our metric.

2. PREVIOUS WORK

Quality metrics for HDR images are traditionally based on the models of low-level vision, accounting for the limitations of the visual system. The very first metric, HDR-VDP [12] was designed to predict a map representing visibility of differences between a pair of images, rather than a quality score that would be correlated with mean opinion scores (MOS). The prediction of quality was added in HDR-VDP-2 [3] and then improved in HDR-VDP-2.2 [5] by calibrating metric parameters on HDR quality datasets. The Dynamic Range Independent Quality Metric (DRIQM) [6] extended HDR-VDP with a set of rules for predicting loss, amplification and reversal of visible contrast to predict objectionable changes between images of different dynamic range, for example a tone-mapped image and its HDR counterpart. The metric for high dynamic range video, HDR-VQM [5], simplified spatial processing but added temporal pooling to offer quality predictions for video.

[1] https://github.com/ynyCL/T-PT-metric

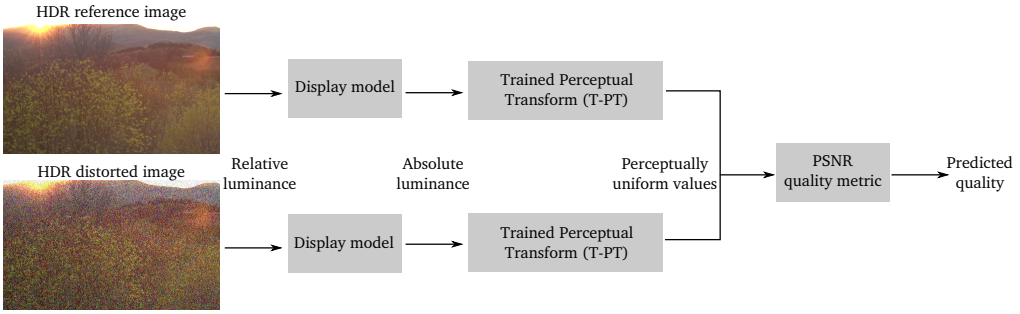


Fig. 1. The existing SDR quality metrics, such as PSNR, can be adapted to handle HDR content by transforming display-referred color values into perceptually uniform values.

Aydin et al. [7] proposed a perceptually uniform (PU) transform to convert absolute display-referred HDR color values into perceptually uniform units[2], which could be used with existing SDR metrics, such as PSNR or SSIM. The transform is derived to ensure that the change in PU units is relative to just-noticeable-differences in luminance, as predicted by the contrast sensitivity function. The transformation is further constrained so that the range of luminance values typically reproduced on SDR monitors (0.8-80 cd/m^2) is mapped to a range 0-255, so that the resulting quality values for SDR images corresponded to those produced by SDR quality metrics. Some authors started to apply PQ EOTF [13] to achieve a similar goals as the PU transform. The main difference between PU and PQ transforms is that the former was derived from the HDR-VDP-2 CSF function while the latter from Barten's CSF [14]. It should be noted that a perceptual transform is one of the first processing steps of all advanced HDR quality metrics, including HDR-VDP, HDR-VDP-2, DRIQM and HDR-VQM.

While simple quality metrics based on a perceptual uniform transform do not achieve as high predictive performance as more advanced HDR quality metrics, they offer many benefits. They are much less complex, fast to compute and differentiable, making them a suitable candidate for a perceptual loss function in optimization problems. The obvious limitation of the PU transform is that it does not account for more complex visual phenomena, such as contrast masking. In this paper we explore whether such more complex effects can be partially accounted for by training the PU transform on HDR quality datasets.

3. TRAINED PERCEPTUAL UNIFORM ENCODING

Perceptual transform functions (PT), such as PU and PQ, were derived and optimized from contrast detection models intended for simple patterns, such as sinusoidal gratings and Gabor patches. This results in some of the drawbacks found with the existing PT encodings: i) the used models predict visibility but visibility may not be directly related to quality and ii) those models do not take complex semantic information in images into account, thus, their may not perform well on complex scenes. These two reasons motivate us to consider fitting the PT using HDR image quality datasets with real-world complex images.

To train such a PT, we first need to determine which transform to use. In practice, both PU [7] and PQ [13] PT encodings have very similar function shapes. However, after analyzing their results we can see that the PU function's performance is better than the PQ function. This is shown in Table 2, where it can be seen that the Spearman Rank Correlation Coefficient (SROCC) of PU-PSNR is higher than that of PQ-PSNR. Because of this, we use the function used for the PU encoding. The PU transform is defined as an integral of inverse of detection thresholds:

$$P(L) = \int_{L_{min}}^{L} \frac{1}{T(l)} dl \tag{1}$$

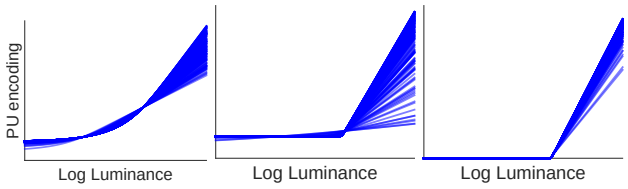
where L_{min} is the minimum luminance to be encoded. The detection thresholds $T(L)$ are modeled as a function of absolute luminance L:

$$T(L) = S \cdot \left(\left(\frac{C_1}{L} \right)^{C_2} + 1 \right)^{C_3} \tag{2}$$

Where S is the absolute sensitivity constant, L is the luminance, C_1, C_2, C_3 are scaling parameters. We further linearly rescale the $P(L)$ values so that $P(0.8) = 0$ and $P(80) = 255$. Because of the rescaling, the parameter S does not influence the shape of the function and the only three adjustable parameters are C_1, C_2 and C_3. To illustrate how the curve changes with regard to C_1, C_2 and C_3, in Figure 2 we plot the PU curves when each parameter is varied individually in the range of [0.1-10].

A major challenge when using multiple image quality datasets is that each dataset represents quality scores using a different scale. For example, the quality score for two images in two different datasets could be very similar, but the

[2]The code for the PU transform can be found at `https://sourceforge.net/projects/hdrvdp/files/simple_metrics/`



(a) C_1 (b) C_2 (c) C_3

Fig. 2. Plausible PT encoding functions, where C_1, C_2 and C_3 follow uniform distributions in [0.1-10] with 100 samples.

actual quality of both images could be very different. To use all datasets together, those datasets have to be realigned and put into a common quality scale. Zerman at el. [15] realigned four HDR image quality datasets by using multiple well-known quality metrics. We argue, however, that this may not be an appropriate approach when trying to test the performance of quality metrics as the realignment may bias the quality scores to offer better prediction for the metrics used in the alignment.

To address the alignment problem we optimize the PT by maximizing Spearman Rank Correlation Coefficient (SROCC) between the predicted quality and dataset's MOS scores. SROCC is computed individually for each dataset and the values are averaged. SROCC is invariant to any monotonic scaling function and thus avoids the need for quality realignment. We use a Bayesian optimization method to optimize the parameters that was proven effective in hyper-parameter fine-tuning in many applications [16]. The Bayesian optimization method uses Gaussian process regression to estimate the landscape of the loss function and determine the next parameter set for evaluation.

4. RESULTS

This section presents the experiments performed to test the behavior of the new trained PU encoding function.

4.1. HDR image quality datasets

For our experiments, we selected Narwaria's 2013 dataset[3] [8] and 2014 dataset [9], the dataset by Korshunov[4] [11] and the latest HDR image quality assessment dataset [15] by Zerman et al.[5]. These are, to the best of our knowledge, all the datasets that can be found for HDR image quality assessment. A summary of the main characteristics of these datasets can be found in Table 1, including the number of observers, the subjective quality measurement method, the number of conditions and scenes, the distortion type and display type. All datasets contain images in absolute display-referred units, which corresponds to physical luminance and color emitted from the display used in the original experiments. However, due to the differences in implementation of Radiance HDR format, the values from Narwaria2013 and Narwaria2014 datasets need to be multiplied 179 when reading images with *pfstools* software [6].

4.2. Trained perceptually uniform encoding

In order to derive the trained PT encoding function and validate the consistency of the results on different datasets, we train our metric using three datasets and test it on the remaining dataset, repeating the procedure four times. Parameters are initialized in our optimization procedure to the original parameters for the PU function. The results are shown in Table 2, which includes the final trained parameter C_1, C_2 and C_3 (T-C_1, T-C_2 and T-C_3) and our optimized result (T-PT-PSNR), the original PQ encoding's result (PQ-PSNR) and the original PU encoding's result (PU-PSNR).

Resulting T-PT encoding functions are shown in Figure 3. The name of the dataset in the legend indicates the test dataset. From this figure we can observe that despite training on different datasets, the curves show a similar trend. The biggest difference in the shape of curves can be observed for low luminance, where the T-PT curves have steeper slope. This suggests that the visibility of distortions is higher than predicted by the simple detection models (CSFs) used to derive PU and PQ. From Table 2, we can also conclude that T-PT functions achieved better results than PU and PQ functions on all datasets. Note that despite improved performance, the new T-PT-PSNR metric is still worse than HDR-VDP2.2 and VQM metrics. However, the new T-PT-PSNR can compute quality in a fraction of the time required by these two complex metrics.

Test Dataset	T-PT-PSNR
#1 Narwaria2013[8]	0.6186
#2 Narwaria2014[9]	0.5230
#3 Korsunov2015[11]	0.8906
#4 Emin2017 [15]	0.8669

Table 3. T-PT-PSNR SROCC results when training on all datasets.

To derive our final proposed PT encoding function, we use all the datasets for training. In the training, we optimize the mean of the SROCC values for each dataset to achieve better performance on all datasets. The proposed function curve is shown in Figure 4. The final T-C_1, T-C_2 and T-C_3 on all datasets are 0.14249, 2.192 and 0.30499. The SROCC results for this final PT encoding, shown in Table 3, indicate further improvement in prediction performance. To further evaluate the results, we take the final PT encoding function and use it as a transfer function for SSIM. The results in Table 4, indi-

[3] http://ivc.univ-nantes.fr/en/databases/JPEG_HDR_Images/

[4] http://mmspg.epfl.ch/jpegxt-hdr

[5] http://webpages.l2s.centralesupelec.fr/perso/giuseppe.valenzise/download.htm

[6] http://pfstools.sourceforge.net/

Dataset	Observers	Method	Conditions	Scenes	Distortion type	Display type
#1 Narwaria2013[8]	27	ACR-HR	140	10	JPEG	SIM2 HDR47E S 4K
#2 Narwaria2014[9]	29	ACR-HR	210	6	JPEG 2000	SIM2 HDR47E S 4K
#3 Korsunov2015[11]	24	DSIS	240	21	JPEG-XT	SIM2 HDR
#4 Emin2017 [15]	15	DSIS	100	11	JPEG, JPEG2000, JPEG-XT	SIM2 HDR47E S 4 K

Table 1. Summary of the characteristics of the datasets used in the experiments.

Test dataset	T-C_1	T-C_2	T-$C3$	T-PT-PSNR	PQ-PSNR	PU-PSNR	HDR-VDP2.2	HDR-VQM
#1 Narwaria2013[8]	0.10568	4.7378	0.10824	**0.6024**	0.58478	0.5898	*0.8911*	*0.8874*
#2 Narwaria2014[9]	0.10078	8.8794	4.405	**0.4887**	0.38043	0.3605	*0.5727*	*0.8126*
#3 Korsunov2015[11]	0.1135	3.3663	0.22871	**0.8908**	0.8751	0.8833	*0.9503*	*0.9572*
#4 Emin2017 [15]	0.10054	9.794	2.2137	**0.8673**	0.81347	0.8249	*0.9298*	*0.9193*

Table 2. The trained T-C_1 – T-C_3 parameters and SROCC results for cross-dataset validation. Each row corresponds to different test dataset. Bold font indicates the best result excluding complex metrics (HDR-VDP2.2 and HDR-VQM).

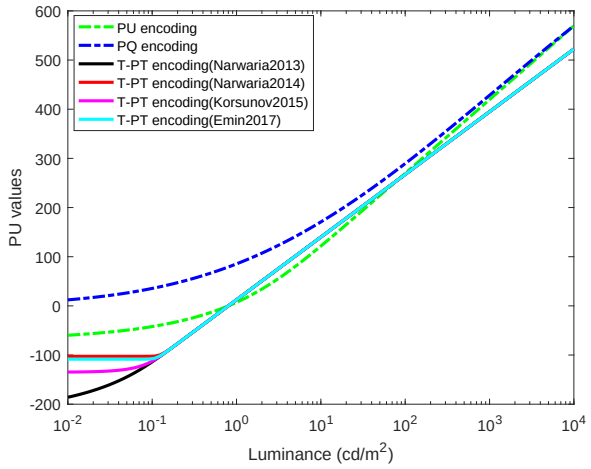


Fig. 3. T-PT encoding function results from the cross-dataset validation experiment, where the dataset name indicates the one missing in training. The PQ curve is rescaled to have the same maximum value as PU encoding curve.

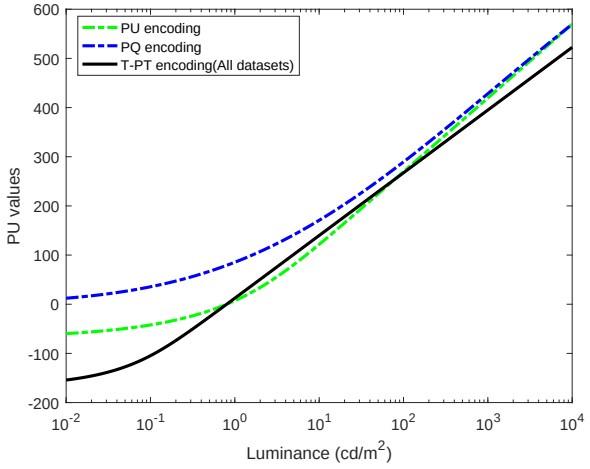


Fig. 4. T-PT encoding function trained on all datasets. The PQ curve is rescaled to have the same maximum value as PU encoding

Test Dataset	T-PT-SSIM	PU-SSIM	PQ-SSIM
#1 Narwaria2013[8]	0.6838	0.6969	**0.7348**
#2 Narwaria2014[9]	0.6145	0.5149	**0.8292**
#3 Korsunov2015[11]	**0.9268**	0.9239	0.8728
#4 Emin2017 [15]	**0.8864**	0.8430	0.8022

Table 4. T-PT-SSIM, PU-SSIM and PQ-SSIM results.

cate that T-PT-SSIM offers better performance than PU and PQ alternatives for datasets # 3 and # 4, but not for datasets # 1 and # 2, for which PQ-SSIM provides better predictions. We are not sure what could be causing this difference, but we also observe that PQ-SSIM outperforms PU-SSIM for this pair of datasets. Since T-PT is based on the PU function, it is also likely to share worse performance for that particular combination of metric and datasets. It must be noted that T-PT-SSIM was not trained using SSIM metric and it is likely that a transfer function needs to be trained separately for each metric.

5. CONCLUSION

In this paper, we have proposed a trained perceptually uniform transform for fast quality assessment for HDR images and videos by fitting a perceptual encoding function to a set of subjective quality assessment datasets. We have shown that when combined with SDR metrics, such as PSNR and SSIM, better performance can be achieved compared to original perceptually uniform transforms. The new transfer function offers a better alternative for low-complexity HDR quality metrics, which are used in the applications for which computational cost is a significant factor.

6. ACKNOWLEDGEMENTS

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 725253–EyeCode) and Cambridge-China Joint Scholarship.

7. REFERENCES

[1] Rafał K. Mantiuk, Karol Myszkowski, and Hans-Peter Seidel, *High Dynamic Range Imaging*, John Wiley and Sons, Inc., 1999.

[2] Rafał K. Mantiuk, "Practicalities of predicting quality of high dynamic range images and video," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sept 2016, pp. 904–908.

[3] Rafał K. Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang 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.

[4] Manish Narwaria, Rafał K. Mantiuk, Mattheiu Perreira Da Silva, and Patrick 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] Manish Narwaria, Matthieu Perreira Da Silva, and Patrick Le Callet, "Hdr-vqm: An objective quality measure for high dynamic range video," *Signal Processing: Image Communication*, vol. 35, pp. 46 – 60, 2015.

[6] Tunç Ozan Aydin, Rafał. K Mantiuk, Karol Myszkowski, and Hans-Peter Seidel, "Dynamic range independent image quality assessment," *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, vol. 27, no. 3, pp. 69, 2008.

[7] Tunc O Aydin, Rafał K. Mantiuk, and Hans-Peter Seidel, "Extending quality metrics to full luminance range images," *Proceedings of SPIE*, vol. 6806, pp. 68060B–68060B–10, 2008.

[8] Manish Narwaria, Matthieu Perreira Da Silva, Patrick Le Callet, and Romuald Pepion, "Tone mapping based hdr compression: Does it affect visual experience?," *Signal Processing: Image Communication*, vol. 29, no. 2, pp. 257 – 273, 2014, Special Issue on Advances in High Dynamic Range Video Research.

[9] Manish Narwaria, Da Silva Matthieu Perreira, Le Callet Patrick, and Pepion Romuald, "Impact of tone mapping in high dynamic range image compression," 2014.

[10] Giuseppe Valenzise, Francesca De Simone, Paul Lauga, and Frederic Dufaux, "Performance evaluation of objective quality metrics for HDR image compression," in *Applications of Digital Image Processing XXXVII*, San Diego, United States, Aug. 2014, SPIE.

[11] Pavel Korshunov, Phillippe Hanhart, Thomas Richter, Alessandro Artusi, Rafał. K. Mantiuk, and Touradj Ebrahimi, "Subjective quality assessment database of hdr images compressed with jpeg xt," in *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, May 2015, pp. 1–6.

[12] Rafał. K Mantiuk, Scott. J Daly, Karol Myszkowski, and Hans-Peter Seidel Seidel, "Predicting visible differences in high dynamic range images: model and its calibration," in *Human Vision and Electronic Imaging*, 2005, pp. 204–214.

[13] Scott Miller, Mahdi Nezamabadi, and Scott Daly, "Perceptual signal coding for more efficient usage of bit codes," in *The 2012 Annual Technical Conference Exhibition*, Oct 2012, pp. 1–9.

[14] Peter. G. J. Barten, "Formula for the contrast sensitivity of the human eye," in *Image Quality and System Performance*, Y. Miyake and D. R. Rasmussen, Eds., Dec. 2003, vol. 5294, pp. 231–238.

[15] Emin Zerman, Giuseppe Valenzise, and Frederic Dufaux, "An extensive performance evaluation of full-reference HDR image quality metrics," *Quality and User Experience*, 2017.

[16] Jasper Snoek, Hugo Larochelle, and Ryan P Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 2951–2959. Curran Associates, Inc., 2012.