ON THE PERCEPTUAL SIMILARITY OF REALISTIC LOOKING TONE MAPPED HIGH DYNAMIC RANGE IMAGES

Marcus Barkowsky, Patrick Le Callet

Polytech' Nantes / IRCCyN CNRS UMR 6597, 44306 Nantes, France
Marcus.Barkowsky@univ-nantes.fr

ABSTRACT

High Dynamic Range (HDR) images are usually displayed on conventional Low Dynamic Range (LDR) displays because of the limited availability of HDR displays. For the conversion of the large dynamic luminance range into the eight bit quantized values, parameterized Tone Mapping Operators (TMO) are applied. Human observers are able to optimize the parameters in order to get the highest Quality of Experience by judging the displayed LDR images on a realism scale. In the study presented in this paper, two TMOs with three parameters each were evaluated by observers in a subjective experiment. Although the chosen parameter settings vary largely, the chosen images appear to have the same QoE for the observers. In order to assess this similarity objectively, three commonly used image quality measurement algorithms were applied. Their agreement with the preference of the observers was analyzed and it was found that the Visual Difference Predictor (VDP) outperforms the Structural Similarity Index and the Root Mean Square Error. A threshold value for VDP is derived that indicates when two LDR images appear to have the same Quality of Experience.

Index Terms— High Dynamic Range, Tone Mapping Operators, Quality of Experience, Objective Image Quality, Subjective Experiment

1. INTRODUCTION

High Dynamic Range (HDR) images provide a large range of new applications. For example, the higher precision may be used in Wide Color Gamut applications. On the other hand, the higher luminance range can be used for adjusting the contrast locally, similar to the human eye. There are no displays available today which allow displaying HDR images in a native representation. This would require the rendering of the luminance ranging from an object in moonlit night to direct sunlight. However, some displays allow a wider range of luminances and are thus capable of partly rendering HDR images.

In order to display HDR images on the currently available reference displays, a conversion to the frequently used eight bit range by Tone Mapping Operators (TMO) is necessary. A recent overview over the different classes of TMO algorithms and a subjective comparison of eight algorithms can be found in [1]. A subjective comparison of different TMOs with two real world indoor scenes has been presented in [2]. In [3] six TMOs were compared to the appearance of the same scene on an HDR capable monitor in a subjective experiment using an adaptation of the paired comparison method.

While the TMOs already exploit several properties of the Human Visual System (HVS), they still need manual tuning by adjusting the algorithm parameters. In the publications mentioned above, the standard parameters for the TMOs were used. In this study, the participants of a subjective experiment were asked to select the optimal parameter set. A set of images was prerendered that span the useful parameter range. The observers chose those images from the set which provided the most realistic impression. This was considered as a criterion for the best Quality of Experience in this application. No explicit reference by an HDR display or by a real world scene was provided to them.

As conducting a subjective experiment is very time consuming, a prediction of the result by an objective measurement would be advantageous. Towards this goal, a first analysis was performed which might help in reducing the number of observers and understanding the criterions which were used by the observers. The total set of rendered LDR images is split into two classes: Those which were chosen by the observers and those which were not. Taking any of the chosen images as a reference, it can be assumed that the other chosen images are perceptually closer than those which belong to the other class. Three full reference image quality metrics were tested for their ability to discriminate between the two classes.

The paper is organized as follows. In Sec. 2 the preparation of the LDR images and the conditions of the subjective experiment are presented which generated the two subsets. The analysis of the objective image quality algorithms is provided in Sec. 3 before conclusions are drawn in Sec. 4.

2. SUBJECTIVE EXPERIMENT

In the subjective experiment four synthetic and four natural HDR images were used. The images were scaled to a resolution of 1920×1080 pixels. These eight HDR images were processed by two TMOs. For both algorithms the implementation in Qtpfsgui was used [4].

The first TMO was published by Mantiuk et al. in [5]. It uses a contrast representation of the image and exploits the

contrast perception of the HVS. The implementation contains three parameters which are termed contrast, saturation and detail. The range of parameters was evaluated and 10 distinct choices for contrast, 8 for saturation and 3 for detail were chosen. Thus, for each of the eight HDR images, a total of 240 LDR images was created.

The second TMO models the characteristics of the cones in the retina and was proposed by Reinhard et al. in [6]. The parameter space was evaluated with 13 settings for brightness, 10 for chrominance, and 7 for the lightness parameter. This leads to a total of 910 LDR images for each of the eight HDR images.

A subjective experiment was setup that allowed the participants to evaluate the three parameters by using the slider device shown in Fig. 1. Only the first three sliders were used

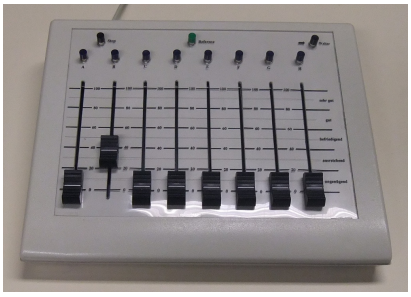

Fig. 1. Slider device used in the subjective experiment

and each corresponded to one parameter. The LDR image that corresponded to the slider position was displayed immediately on a TVLogic LVM-401W reference screen. The room setup and viewing distance corresponded to ITU-R BT.500 and the HDTV testplan of the Video Quality Experts Group (VQEG)[7]. For the two TMO non-overlapping groups of naïve observers were invited, 20 for Mantiuk and 21 for Reinhard. The viewers were screened for acuity and color vision. They were asked to choose the image that provided the most realistic impression as this can be expected to give the highest Quality of Experience for a longer presentation. At least five changes of the sliders and at least one change per slider was required by the assessment program before the observer was allowed to vote. The finally chosen image was seen on the screen while the observer confirmed his choice. A training session preceded the subjective experiment. The chosen image was recorded for each observer and each of the eight HDR images.

3. PERFORMANCE ANALYSIS FOR IMAGE QUALITY ALGORITHMS

Three image quality algorithms were tested whether they could predict the observers choice. The inputs to all algorithms are two images and the output is a single value which indicates the similarity of the two images. Firstly, the simple Root Mean Square Error (RMSE) measure is used. As the images are stored in red, green, and blue component, the mean squared error of all three planes was calculated. Please note, that this differs from the calculations that are usually performed for PSNR calculations in that no conversion to the Y, Cb, Cr color space was performed [8]. The second algorithm is the Structural Similarity Index (SSIM) [9]. This measurement is based on local statistics. The freely available Matlab implementation was used. As a third algorithm the Visual Difference Predictor (VDP) was chosen in order to include an algorithm which incorporates the modelling of the Human Visual System [10]. It was specifically written to predict a visible difference and accepts HDR images as well as LDR images. In this evaluation, only the LDR part was used. The algorithm indicates the probability of change detection for each pixel individually. The freely available implementation also outputs the number of pixels in the image that were alerted to exceed a certain visibility threshold. In this evaluation, the percentage of pixels that exceed the 75% visibility threshold was chosen.

Table 1. Properties of the analysis for each of the eight HDR images

TMO Algorithm	Mantiuk	Reinhard
Number of images presented	240	910
Number of observers	20	21
Number of “equal” conditions	380	420
Number of “dissimilar” conditions	4.400	18.669

The performance analysis compares the difference measured by the three algorithms to the observers' opinion. The analysis is based on two assumptions. The first assumption is that images chosen by the observers are equal. In our subjective experiment, the term “equal” refers to a notion of “realistic appearance” which is shared by the observers. The second assumption is that images which were not chosen by the observers differ significantly from those which were chosen. This assumption may be weak because only 20 out of 240 for Mantiuk and 21 out of 910 images could be selected. So it has to be considered that a certain number of non-chosen images does not differ significantly. This problem is addressed by the type of analysis performed.

In Table 1 an overview of the number of tested conditions is given. The number of “equal” conditions refers to comparisons in which both images were chosen by observers while the term “dissimilar” refers to comparisons in which only one image was chosen by an observer. The RMSE and SSIM algorithms are commutative, e.g. testing image A versus B reveals the same result as image B compared to image A. The VDP is not commutative so the full matrix of combinations needs to be considered.

For the performance comparison, a Receiver Operating Characteristics (ROC) analysis is performed. The notation follows the one presented in [11]. In our case, there are two alternatives: Either the images are equal or they are dissimilar.

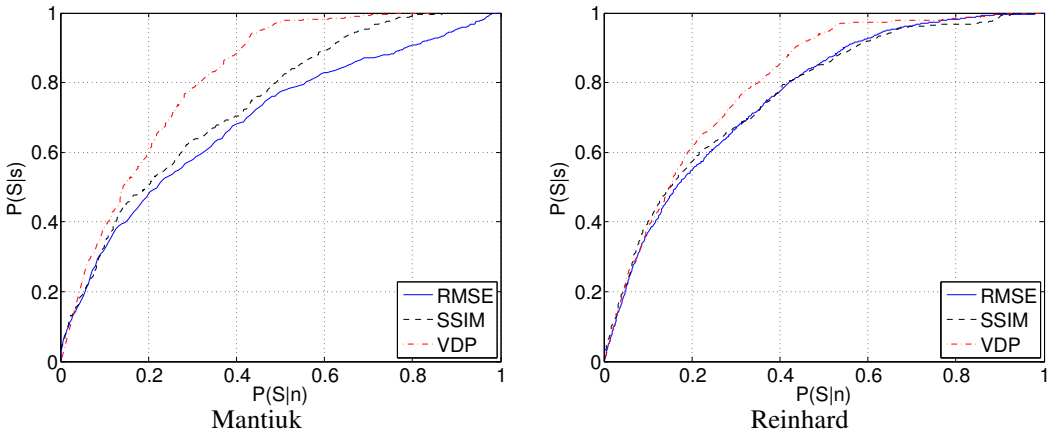

Fig. 2. ROC analysis displaying the percentage of correct decisions on the y-axis and false alarms on the x-axis.

The subjective experiment is used as ground truth. The two alternatives are denoted as s and n, corresponding to "signal" and "no-signal". In our case, "signal" means that the two images are equal, thus each chosen by an observer. On the contrary, "no-signal" means that only one image was chosen by an observer. Correspondingly, the objective measure should provide the answer whether the images are equal ("S") or dissimilar ("N").

The objective part will be exemplified with the RMSE measure. For any two images given, the RMSE algorithm returns a value that measures the distance between the images. By choosing a threshold which is denoted as t, two classes of images can be generated. If the RMSE value is smaller than the threshold, the images are considered equal, otherwise they are classified as dissimilar. By evaluating the difference between the images chosen by the observers, the probability $P(S|s)$ can be obtained. This probability specifies the number of correct decisions, e.g. when the RMSE is less than the threshold. The same threshold is applied to the RMSE values where the images were found dissimilar in the subjective experiment. The probability $P(S|n)$ is obtained. This is the false alarm ratio: the objective algorithm identifies the images as equal while they were judged to be dissimilar. Please note, that in this particular evaluation, the false alarm ratio also includes the drawback of the second assumption mentioned above. Some conditions which are meant to be equal are classified as dissimilar by the subjective experiment because of the limited number of observers. Thus, all algorithms are affected in the same way by the fact that the measured false detection rate is higher.

In the optimal case, $P(S|s) = 1$ and $P(S|n) = 0$. In practice, the values depend on the threshold t. The smaller the value of t, the less images are classified as equal and the lower the two probabilities and vice versa. The relationship can be plotted in a ROC curve.

The results for our experiment are shown in Fig. 2. On the left side, the results for Mantiuk are presented. For each percentage of correct decisions chosen on the y-axis, a certain percentage of false alarms has to be accepted. It can be seen that the best performance is provided by the VDP algorithm as it has the lowest false alarm rate in the relevant area which usually ranges from 50% to 90%. The same is true for the Reinhard TMO. The RMSE and SSIM perform slightly better when compared to Mantiuk.

Table 2. Results for the analysis of the image quality algorithms in terms of threshold value (t), correct decision (CD) $P(S|s)$, and false alarm (FA) $P(S|n)$ in percent

Algorithm		Mantiuk		Reinhard		Combined	
	CD	FA	t	FA	t	FA	t
RMSE	50	23	0.047	16	0.126	25	0.072
	75	47	0.078	37	0.251	54	0.155
	90	79	0.126	55	0.352	71	0.281
SSIM	50	20	0.972	15	0.887	21	0.949
	75	44	0.928	38	0.635	44	0.842
	90	61	0.837	57	0.435	69	0.591
VDP	50	14	0.024	15	0.019	15	0.022
	75	28	0.316	30	0.187	29	0.251
	90	41	1.040	44	0.845	43	0.980

Typically, a correct decision percentage of 50%, 75%, or 90% is requested. The resulting false alarm rates and the threshold values that need to be chosen for each algorithm are provided in the first two columns of Table 2. The choice of the threshold depends on the requirements for a particular purpose. For example, when it is preferred to have a low false alarm (FA) ratio, it may be agreed to have a chance of 50% of missing two images that are actually similar. On the opposite, when a manual inspection assures that the falsely alarmed images are sorted out, a large correct detection percentage (CD) might be used, e.g. 90%.

A practical application of these results would require the specification of a single threshold value for the output of the image quality measurement algorithm. This poses a problem which shall be explained with the RMSE algorithm. In Table 2 the threshold value for 90% correct detection and the Mantiuk TMO is given as 0.126. When using the same threshold when assessing images generated with the Reinhard TMO

we get only a correct detection of 50%. In order to achieve 90%, a threshold of 0.352 would be necessary. Please note, that this threshold is very high: it indicates that the average difference in pixel values is 35%. Correspondingly, an average difference of 90 for each pixel on an eight bit image is accepted as similar.

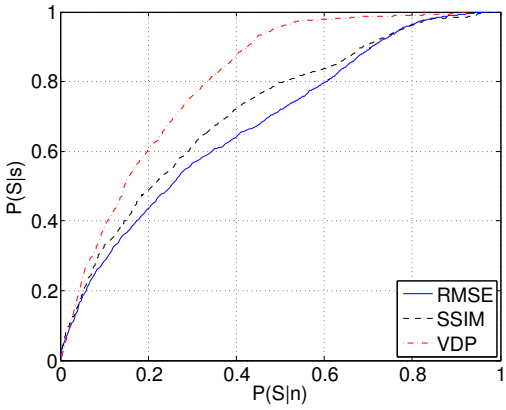

Fig. 3. ROC analysis for the combined datasets of the TMOs Mantiuk and Reinhard.

Exactly the same reasoning is true for the SSIM algorithm. The threshold value of 90% for Mantiuk corresponds to 0.837 which is closest to the 50% threshold for the Reinhard TMO. For obtaining 90% correct decisions, a SSIM threshold of 0.435 would be necessary. This value is usually the result of comparing the reference image to largely degraded images which are voted as "poor" in subjective experiments.

The VDP appears to be more stable as it shows a value of 1.04 for the threshold in Mantiuk and 0.845 in Reinhard. This threshold value is also more comprehensible: About 1% of the pixels in the image exceed a detection threshold of 75%.

In order to further evaluate the robustness of the algorithms and to determine a common threshold, the datasets of chosen and nonchosen data were merged for the two TMOs. The resulting ROC curves are displayed in Fig. 3 and the threshold values are provided in the last column of Table 2. It is apparent, that VDP outperforms RMSE and SSIM. A threshold value of 0.98 for distinguishing between chosen and nonchosen images at 90% correct decision is indicated for the VDP value. The number of false alarms would then be 43%.

4. CONCLUSIONS

Previously, the development of subjective and objective assessment of image quality was mainly focused on characterizing transmission systems. However, in recent times, there is an increasing demand for judging the visual equivalence of images. One of the main areas is the High Dynmic Range Imagery where the improved accuracy is either used to span a larger range of brightness or to enlarge the color gamut. In this contribution, a subjective assessment methodology was introduced to locate the best parameters for two Tone Mapping Operators. Furthermore, the question was investigated whether the subjective experiment can be predicted by using the current image quality algorithms. The ROC analysis was adapted and applied to the problem. It was shown that the Visual Difference Predictor outperforms RMSE and SSIM, probably due to its modelling of the Human Visual System. The threshold value of 1% for the VDP value is proposed as a criterion when judging whether two image of the same TMO appear similar.

5. REFERENCES

[1] J. Kuang, H. Yamaguchi, G.M. Johnson, and M.D. Fairchild, "Testing HDR Image Rendering Algorithms," in *Color Imaging Conference*, 2004, pp. 315–320.

[2] A. Yoshida, V. Blanz, K. Myszkowski, and H.-peter Seidel, "Perceptual evaluation of tone mapping operators with real-world scenes," in *Human Vision & Electronic Imaging X, SPIE*. 2005, pp. 192–203, Spie.

[3] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen, "Evaluation of tone mapping operators using a High Dynamic Range display," *ACM Trans. Graph.*, vol. 24, pp. 640–648, 2005.

[4] Open Source Community, "Qtpfsgui Project Homepage," http://qtpfsgui.sourceforge.net/index.php.

[5] Rafal Mantiuk, Karol Myszkowski, and Hans-Peter Seidel, "A perceptual framework for contrast processing of high dynamic range images," *ACM Trans. Appl. Percept.*, vol. 3, no. 3, pp. 286–308, 2006.

[6] E. Reinhard and K. Devlin, "Dynamic range reduction inspired by photoreceptor physiology," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 11, no. 1, pp. 13–24, Jan.-Feb. 2005.

[7] Greg Cermak, Leigh Thorpe, and Margaret Pinson, "Test Plan for Evaluation of Video Quality Models for Use with High Definition TV Content," *Video Quality Experts Group (VQEG)*, 2009.

[8] NTIA / ITS, "A3: Objective Video Quality Measurement Using a Peak-Signal-to-Noise-Ratio (PSNR) Full Reference Technique," *ATIS T1.TR.PP.74-2001*, 2001.

[9] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, 2004.

[10] R. Mantiuk, S. Daly, K. Myszkowski, and H.-P. Seidel, "Predicting Visible Differences in High Dynamic Range Images – Model and its Calibration," in *SPIE Human Vision and Electronic Imaging X*, B.E. Rogowitz, T.N. Pappas, and S.J. Daly, Eds., 2005, vol. 5666, pp. 204–214.

[11] David M. Green and John A. Swets, *Signal Detection Theory and Psychophysics*, Peninsula Publishing, Los Altos Hills, 3 edition, 1966.