Method for Evaluating Tone Mapping Operators for Natural High Dynamic Range Images

Mikko Kuhna, Mikko Nuutinen and Pirkko Oittinen

Department of Media Technology, Aalto University School of Science and Technology, P.O.Box 15500, FI-00076, Aalto, Finland

ABSTRACT

High dynamic range (HDR) imaging seems to have developed to a level of soon being a standard feature in consumer cameras. This study was motivated by the need for evaluating tone mapping operators especially for consumer imaging applications. A no-reference method based on ISO 20462-2:2005 triplet comparison was created for evaluating tone mapping operators. Multiple HDR test images were photographed and the method was validated by evaluating 25 tone mapping operators with five test images. Tone mapping operators were evaluated based on image naturalness and pleasantness. The results indicate that the method successfully ranked the method in terms of naturalness and pleasantness. The test image set could be improved for example based on an imaging photo space for HDR photography. The test images of this study are available for non-commercial research purposes.

Keywords: high dynamic range, image quality, naturalness, test image, subjective evaluation

1. INTRODUCTION

1.1 Background

The human visual system (HVS) functions over an illumination range of up to eight orders of magnitude and therefore the dynamic range of the HVS is roughly 10^8, but cameras are much more limited.[1] The dynamic range of modern digital cameras can be simplified based on the bit depth of the A/D converter, which often corresponds to a dynamic range of 10^3-10^4. For years photographers have been taking multiple images with different exposures to overcome this limitation. This kind of photography is often referred as HDR photography.

Presentation of HDR images is usually more difficult than their capture. In the film days, photographers used techniques such as dodging-and-burning to compress the dynamic range of the film to the lower dynamic range of the print. A similar process is also necessary in digital photography, where an HDR image needs to be compressed to the limited dynamic range of the display. This process is called tone mapping. The global tone mapping operators make modifications based on the image histogram or modify the tone reproduction curve, whereas the more complex local operators are often HVS-based appearance models.

HVS is a complex system that adapts to different illumination conditions. Without fully understanding the functioning of the HVS it is challenging to develop a tone mapping operator that would always produce natural looking images. The complex nature of the HVS is also a problem for evaluating tone mapping operators. A comparison between a HDR scene and a tone mapped image is not feasible without compromises because the HVS constantly adapts to the surrounding illumination. Therefore it is problematic to use a real reference even with a HDR display.

A natural HDR image is understood in this study as a scene with illumination which is challenging for regular cameras, taken by a typical consumer with a HDR camera. Instead of evaluating tone mapping operators based on overall image quality, often image quality attributes such as image naturalness, fidelity or pleasantness are used. Image naturalness is a high-level attribute that is usually evaluated using a memory-based reference. Image fidelity is understood as reproduction accuracy and usually evaluated with a real reference. Image pleasantness is understood as visually pleasing without necessarily being high in naturalness.

Further author information: (Send correspondence to M.K.)
E-mail: firstname.lastname@tkk.fi, Telephone: +358 9 47001

1.2 Related work

Many studies have been conducted for evaluating the performance of tone mapping operators. The study by Drago et al.[2] focused especially on the selection of parameters. In the study by Ledda et al.[3] a modified paired comparison method with a reference shown on a HDR display was used. The study by Ashikhmin and Goyal[4] consisted of multiple experiments. The first two experiments were not based on a reference, but in the third experiment, the participants were taken to the locations where the images were taken. Kuang et al.[5] have conducted multiple experiments with and without a reference. For their experiments with a real reference, they made three experimental HDR scenes and the participants were able to make comparisons.

The study by Cadik et al.[6] consisted of multiple experiments for evaluating different image quality attributes of HDR images. Some of the scenes were evaluated with a real reference (a view from the window) and some without a reference (printed samples). A study by Yoshida et al.[7] consisted of two experiments, with and without a reference. In their experiments the participants were able to adjust brightness, contrast and color saturation parameters for the purpose of creating a generic tone mapping operator. The study by Annighofer et al.[8] focused especially on selecting the best exposure for each scene.

In these studies 77 different HDR test images have been used for subjective evaluation of different image quality attributes. In only three of the images a person was clearly included in the scene and some previous studies have purposefully used test images without people.[7] In comparison a color chart was included in at least seven of the images. The lack of people in the test image can be explained with the fact that HDR images are usually taken with multi-exposure techniques and movement in the scene produces ghosting in the image. Nowadays modern deghosting algorithms are good enough so that consumer cameras with multi-exposure HDR techniques are emerging rapidly and HDR test images should be updated for evaluating consumer cameras and imagery.

1.3 Overview of the study

This study was motivated by the need of a method to evaluate tone mapping operators using natural scenes. A no-reference subjective evaluation method was developed using five natural HDR images. The method should support:

- large image size,
- multiple tone mapping operators,
- multiple test images,
- multiple image quality attributes,
- fast evaluation, and
- differentiating the operators.

Additionally colorimetric HDR images needed to be captured to be used as references for objective measurements.

The HDR test images captured for this study were 2160x1434 pixels and the tone mapped images were shown in the experiments with 1600x1050 pixels on 24" displays. A total of 25 tone mapping operators were evaluated using the five different test scenes. Tone mapping operators were evaluated based on two different image quality attributes, image naturalness and pleasantness. The experiment consisted of two parts. The first part was used to reduce the 25 tone mapping operators into seven based on overall image quality. The average duration of the main experiment was under 45 minutes (ISO 20462-1:2005[9] recommendation).

Section 2 describes different study stages from test scene design, image capturing and processing procedure, to selected tone mapping operators, subjective experiments and analysis of the subjective evaluations. See* for the objective measurements and more detailed information of this study. All the images and data related to this study are available for non-commercial research purposes†.

*http://mikkokuhna.com/publications/Mikko_Kuhna_Thesis_print.pdf

†http://mikkokuhna.com/research/naturalhdr/

2. METHODOLOGY

2.1 Test scene design

Test images used in subjective evaluations of image quality are referred to as pictorial test images.[10] Test images should be developed so that the test image differentiates the samples, in this case the tone mapping operators. For example when color reproduction is evaluated without a reference, memory colors such as skin tone needs to be included in the test image as a memory based reference.[11]

Besides the fact that most of the exterior scenes that can be captured with normal LDR cameras, the processing that many images will need for a natural reproduction will bring out the limitations of the LDR image capture. HDR image capture is necessary for many scenes like: sunset or sunrise, dim indoor scenes with relatively high illumination from windows or lamps, night scenes with illumination such as streetlight or advertisement lights, and many different types of scenes with direct illumination such as sunlight.[1]

One solution could be to cover as many different types of images as possible, but it would increase the number of samples so that subjective evaluations would be impractical. An extreme alternative could be to develop a single test image which includes as many critical image quality attributes as possible.[12] The development of a balanced single image[13] is not a trivial task and the common solution has been to use some kind of transitional form of these two alternatives.

One example of guidelines for test image development is the camera phone photo space by International Imaging Industry Association (I3A).[14] The photo space lists six clusters of common scenes for photographing with a mobile camera. For HDR images such a photo space has not yet been suggested. One step forward is the HDR photographic survey by Fairchild.[15] He photographed multiple HDR scenes and collected colorimetric data from the scenes. All the images and data have been made available for non-commercial research purposes[‡].

Test images captured for this study are shown in Figure 1. The test scenes were designed based on the guidelines by Salmi et al.[12] The test images were designed to be like the photographs consumers typically take. Multiple scenes were captured also in vertical orientation, but were forced to be left out for to take advantage of the full display screen size. It was seen most of the test images should include people in the image, but also a landscape image should be included. Overall 15 scenes were captured with a total 55 HDR images and in the end five images were selected: a landscape scene (**pond**), two indoor scenes with a person (**newspaper** and **office**) and two outdoor with a person (**bus stop** and **park**). The selections were made partly based on camera phone photo space and the fact that consumers typically take photographs of people. The photometric distribution of the selected test images are shown in Figure 2.

2.2 Image capturing procedure

Camera characterization measurements were made with a setup such as shown in Figure 3 (a) for light fall-off measurements. The camera was a Nikon D300 and the lenses were Nikon AF-S DX Nikkor 18-105 f/3.5-5.6G ED VR and Nikon AF-S DX Nikkor 35mm f/1.8G. Linearity of the camera and the RAW-conversion software (dcraw v.8.93) were validated from images taken of a OECF-chart (Image Engineering TE241) illuminated by an integrating sphere illuminator (Image Engineering spherical transparency illuminator LE6-100). Luminance of each patch was measured with a spectroradiometer (Photo Research PR-670). The image intensity values of the OECF patches increased linearly as a function of measured luminances from the patches ($R^2 = 0.9999$). Light fall-off was measured from images taken of a diffusor plate (Image Engineering TE255). Light fall-off was measured for both lenses. Aperture and focusing distance affects light fall-off significantly and therefore the calculations were made for multiple different aperture and focusing distance settings.

Images were captured only with the camera settings for which the camera characterization measurements were made. Other camera settings which were used in all images were 14-bit RAW file format and ISO-speed 200. Images were captured using a tripod to minimize vibrations. The image stabilization of the lens was turned off. White balance setting of was always set to 5000 K. In the image capture procedure, the auto focus was first used to focus the lens to the main object of the image and then switched to manual focus. All the images were captured using a wired remote. The exposure selection was based on capturing all possible colors reliably. A

[‡]http://www.cis.rit.edu/fairchild/HDR.html

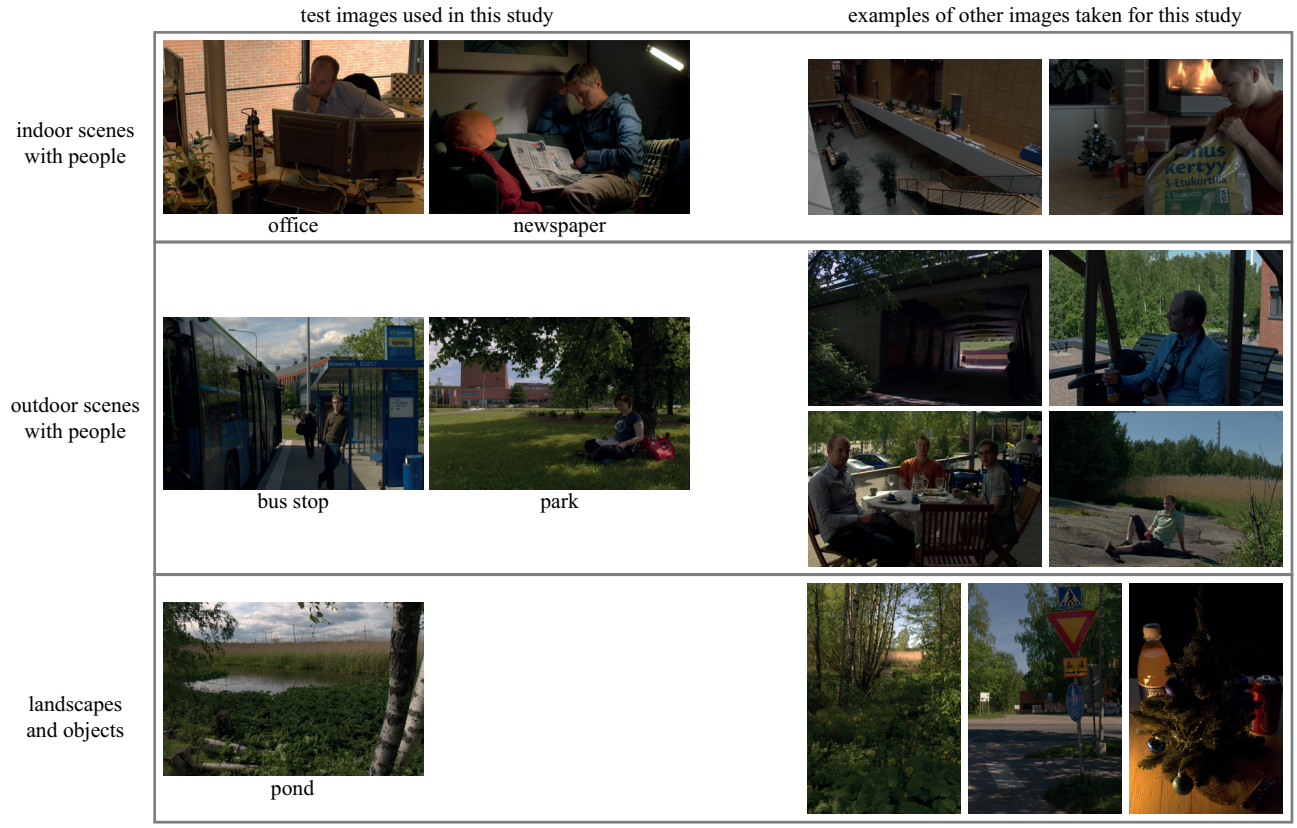


Figure 1. Tone mapped images of the designed test scenes.

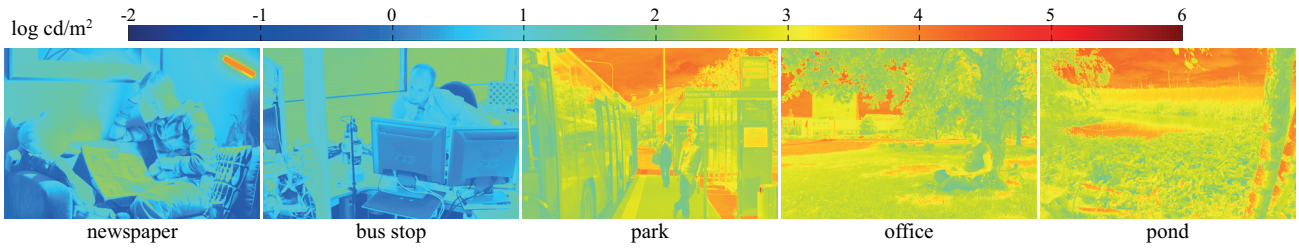


Figure 2. Photometric distribution of the test images.

(a) (b) (c)

Figure 3. Image capture procedure: light fall-off measurements (a), camera calibration using the color charts (b) and close-up of the color charts (c).

RGB histogram was used to evaluate the base exposure for exposure bracketed images. Manual exposure mode was used and all the HDR images were captured with 9 exposure bracketed shots with 1 EV step.

Ultimately the colorimetric camera calibration needed to be versatile, simple and fast as the test images were also photographed in outdoor locations with uncontrollable illumination conditions. Therefore a simple two color chart procedure shown in Figure 3 was used. One chart (X-Rite ColorChecker) was used as a training set chart for the characterization and another chart (X-Rite ColorChecker Semi-Gloss) as a test set chart in measuring the characterization accuracy. Polynomial modeling by Hong et al.[16] was used at the beginning of the study, but it often produced errors in the characterized images (example of an overfitting problem). All the images were converted using the following linear transform by minimizing the Euclidean distance in CIE XYZ color space

$$[XYZ]^T = [M][RGB]^T, \quad (1)$$

where RGB is the device-dependent source signal, M is a 3x3 linear transform matrix, and XYZ is the device-independent target signal (CIE XYZ).

The reflectances from all the patches of the training set chart were measured in every scene. All the patches of the test set chart were not measured in every scene, but the illumination spectrum at the level of the test chart was measured and the expected reflectances from the patches were calculated by multiplying the illumination spectrum and the reflectance of the field measured in a controlled environment. A minimum of 25 reflectance spectra were measured with a spectroradiometer in each scene. The scene illumination and spectral reflectances of the color chart were measured using a spectroradiometer. This spectroradiometer measures the spectral range from $380nm$ to $780nm$ at every $2nm$ and thus meets the ISO 17321-1:2006[17] recommendations. The 24 patch color chart used also meets the ISO 17321-1:2006 minimum recommendations.

Overall image processing procedure consisted of:

- Conversion of nine exposure bracketed RAW files (14-bit NEF) to linear 16-bit TIFF files using a RAW-conversion software. The RAW software parameters were adjusted so that the images were not interpolated with demosaicing, but instead four CFA pixels were combined to a single RGB pixel.
- The bracketed images were processed in MATLAB. The light fall-off correction was based on the camera and lens aperture as well as the focusing distance parameters. The images were converted to the same exposure domain. All the images were averaged with blown-out highlights and noisy pixels excluded. The averaged image and all the individual exposure images were saved in Radiance RGBE image format and used to manually mask areas with visible artifacts.
- The image was converted to colorimetric XYZ color space using the transformation matrix derived from the composite chart image and the spectral measurements.

- The image was white balanced (Bradford chromatic adaptation transformation).
- The image was saved in XYZE-format and another version was converted to linear RGB using the sRGB matrix[18] and saved in RGBE image format.

The used illumination white point values were selected based on the overall image quality by the authors. Test images **office** and **pond** proved to be problematic. By selecting a natural looking white point for **office** based on sRGB rendering, *iCAM06* produced a bluish toned image. A compromise white point was selected so that the sRGB and most of the operators produced slightly warmer toned images. The sky and clouds in image **pond** consisted reddish tones with many white points and had to be fine tuned manually.

2.3 Tone mapping operators

A total of 24 tone mapping operators (and sRGB with manual clipping as a reference) were selected for the study. The purpose of selecting so many operators was to validate the functioning of the method. Most of the methods have been evaluated in previous studies and can be therefore used for comparison. Most of the operators include user controllable parameters. All parameters were tested and the best fit for each image was selected by the authors based on overall image quality. For practical reasons, all the operators were not individually examined, but multiple parameters were tested as batch processing and the parameter value producing the highest image quality was selected.

The following operators were selected:

- *iCAM06*[19,§]
- *Meylan*[20] *and Tamburrino*[21,¶]
- *Ashikhmin, Chiu, Drago, Bilateral, Ferschin, Ferwerda, GDC, Histadj, Log, Oppenheim, Pattanaik, MOM, Retinex, Reinhard, Photoreceptor, Schlick, Trilateral, TR, Ward and Yee*[‖]

The photographic tone reproduction operator by Reinhard et al.[22] has the option to be used with only simple global processing and with local processing that resembles dodging-and-burning. Both options were used in this study. *Reinhard* refers to simple global processing and *Reinhard** to the same operator with local processing included.

2.4 Subjective experiments

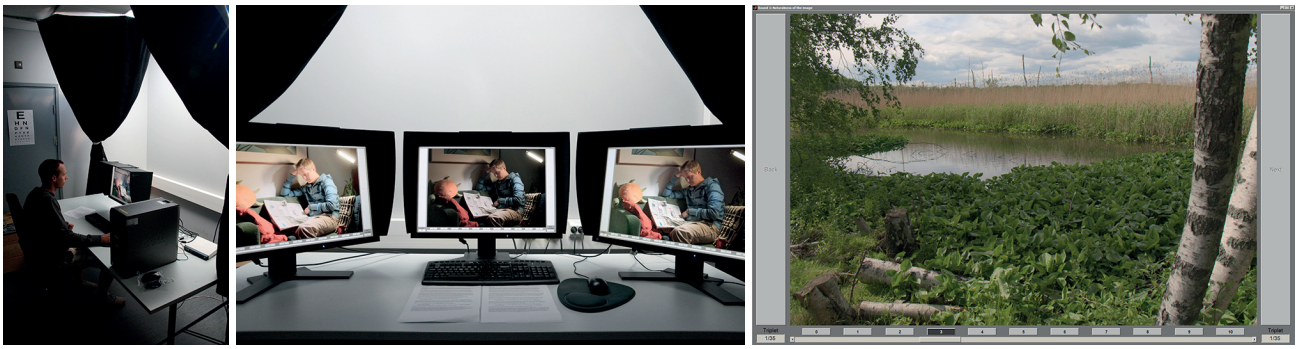
The triplet comparison method was selected for the subjective evaluation based on ISO 20462-1:2005[9] recommendations. The triplet comparison method is described in more detail in ISO 20462-2:2005.[23] The standard recommends a preliminary categorical sort experiment to be conducted for the purpose of reducing the number of samples. Persons with normal or corrected color vision were recruited for the experiments. MATLAB-program was made for each experiment. Some examples of the experimental setup are shown in Figure 4.

Two types of 24" displays (Eizo ColorEdge CG241W & CG242W) were used in the experiments. The displays were profiled using Eizo's own software (Eizo ColorNavigator 5.3) and a spectrophotometer (GretagMacbeth Eye-One Monitor). The profile was created using the suggested sRGB reference viewing environment.[18] Display white point luminance was set to 80 $\frac{cd}{m^2}$, 6504 K (D65) and gamma 2.2. The illuminance level measured 200 lx (5000 K) at the level of the gray wall behind the display. The displays were positioned so that the participants observed each display perpendicularly and the distance to the display was kept at $65-70$ cm. The participants were advised to adjust the height of the chair to the top level of the display and not to lean forward. The average display calibration error (CIEDE2000) was 0.49 in the categorical sort experiments and $0.43-0.54$ in the triplet comparison experiments.

[§]implementation: http://www.cis.rit.edu/mcsl/icam06/

[¶]implementation: http://ivrgwww.epfl.ch/supplementary_material/index.html

[‖]implementation and detailed information: Reinhard[1]



(a) (b) (c)

Figure 4. Examples of the test setup: categorical sort experiment (a), triplet comparison experiment (b) and the MATLAB application developed for the triplet comparison experiment (c).

2.5 Categorical sort experiment

The categorical sort experiment was a short experiment in where the participants classified 125 samples into three categories based on overall image quality. The samples consisted of five test images processed with 24 tone mapping operators plus standard sRGB processing. The categories were named as unacceptable, acceptable and favorable. Based on the experiences of a pilot experiment, a delay of one second was added when a new sample was loaded. The instructions were shown to the user in a guide dialog before each stage. Test instructor was present during the whole experiment.

The participants in the categorical sort experiments consisted of media technology students and researchers from the Aalto University School of Science and Technology. All the participants can be considered as experts in image quality. The purpose of the categorical sort test was to reduce the number of samples for the triplet comparison experiment. The classifications were consistent after 10 participants.

2.6 Triplet comparison experiment

Triplet comparison experiment was a more complex experiment. Image naturalness and pleasantness were used as the evaluated attributes. Seven operators were selected based on the results of the categorical sort experiment so that 35 triplets per attribute were evaluated. The different image content samples were evaluated as separate triplets and the order of image content was randomized. The evaluation thus consisted of five different triplet comparisons mixed together per evaluated attribute.

The experiment began with instructions shown in printed form. In the next stage seven fixed example images were shown in the center display. The purpose of the example images was to guide the participant to the quality differences of the samples in order for them to be able to use the full scale. After the participant browsed through all the example images, the program continued to a training phase. The training phase was exactly the same procedure as the actual triplet comparison. Five randomly selected sample triplets rotated in the training phase. A minimum of two triplets was made mandatory for the participant in the training phase before the program allowed continuing to the next phase. During the training phase, the instructor of the experiment advised the participant in using the program and made sure that the participant had understood the instructions.

The triplet comparison stage started after the training stage. First naturalness was evaluated with 35 triplets appearing in random order. After this a five minute break was held. The participant was able to relax and leave the experiment room. The program included a timer so that the next stage could not be begun until the 5 minutes had elapsed. After the break pleasantness was evaluated in a similar manner.

The participants in the triplet comparison experiments consisted on mostly students of various nationalities. Their level of knowledge on image quality estimation was non-expert. The tests were conducted with 28 participants (19 male and 9 female). The goal was to keep the test duration under 45 minutes as ISO 20462-1:2005[9] recommends. Average experiment duration was 40 minutes (maximum 70 minutes).

2.7 Categorical sort experiment analysis

The analysis of the categorical sort test data was relatively simple. Participants gave a score of either unacceptable, acceptable and favorable for each sample. The classifications were analyzed so that unacceptable equaled to -1, acceptable to 0 and favorable to $+1$. 95% confidence intervals were calculated from the classifications.

2.8 Triplet comparison experiment analysis

The responses in the triplet comparison were analyzed using the ISO 20462-2:2005[23] recommendations. The triplets were divided to paired comparisons and the corresponding scale values were compared so that a higher scale value corresponded to a preferred sample. The frequencies of preferred samples were put into a matrix described as cumulative frequency distribution. Cumulative frequencies n were converted to probabilities p

$$p = \frac{N+n}{2N}\,, \tag{2}$$

where N is the number of samples.

Probabilities p were converted to the amount of differentiation Q

$$Q = \frac{12}{\pi} arcsin(\sqrt{p} - 3)\,, \tag{3}$$

Amount of differentiation Q-values formed a matrix in where one cell represented the differentiation between samples in JND units. The quality scale values were formed by summing all the Q-values of an operator column. A separate scale was processed for each individual image as well as a combined scale for all the test images. A confidence interval CI suggested by Montag[24] was used. CI is based on an empirical formula for standard deviation σ_{obs}

$$\begin{aligned} CI &= 1.96\sigma_{obs} \\ \sigma_{obs} &= 1.76(n+3.08)^{-0.613}(N-2.55)^{-0.491}\,, \end{aligned} \tag{4}$$

Mean opinion scores (MOS) were processed from the magnitude scale $(0-10)$ values. Each sample was evaluated three times during the experiment as three triplets were formed to make paired comparison between all 7 tone mapping operators per test image.

3. RESULTS

The averaged classifications (MOS) for the tone mapping operators are illustrated in Figure 5. Error bars represent 95% confidence intervals. The results of the categorical sort test reflects, as expected, the extensive range of quality difference between samples. Most of the operators were more often classified as unacceptable than favorable. Clearly a group of nine operators can be distinguished from *Schlick* to *Reinhard**.

The operators for triplet comparison were selected based on MOS and CI in Figure 5. The requirements of using triplet comparison are for example that the selected samples should not differentiate too much in the attribute measured. The maximum scale that can be measured is -3 to $+3$ JNDs and ISO 20462-2:2005[23] recommends that the maximum difference between samples should be less than 1.5 JND. Based on these facts, the *Reinhard** was excluded of possibly differentiating too much. *sRGB* was also excluded as it was an LDR sRGB rendering of the scene with blown out details and it was only included in the categorical sort experiment because it was interesting to see how high it would be ranked.

Naturalness and pleasantness scores from the triplet comparison experiments for all the test images are presented in Table 1. The table includes also the MOS values for comparison. The score values for all the test images are also presented in Figure 6 and for each test image in Figure 7. As can be seen the naturalness and pleasantness scores, *Reinhard* operator was evaluated highest in naturalness and pleasantness overall. The individual image scores show the there is still a lot variation. *Photoreceptor* is the only operator that has been evaluated below average for every image and respectively *Reinhard* above average.

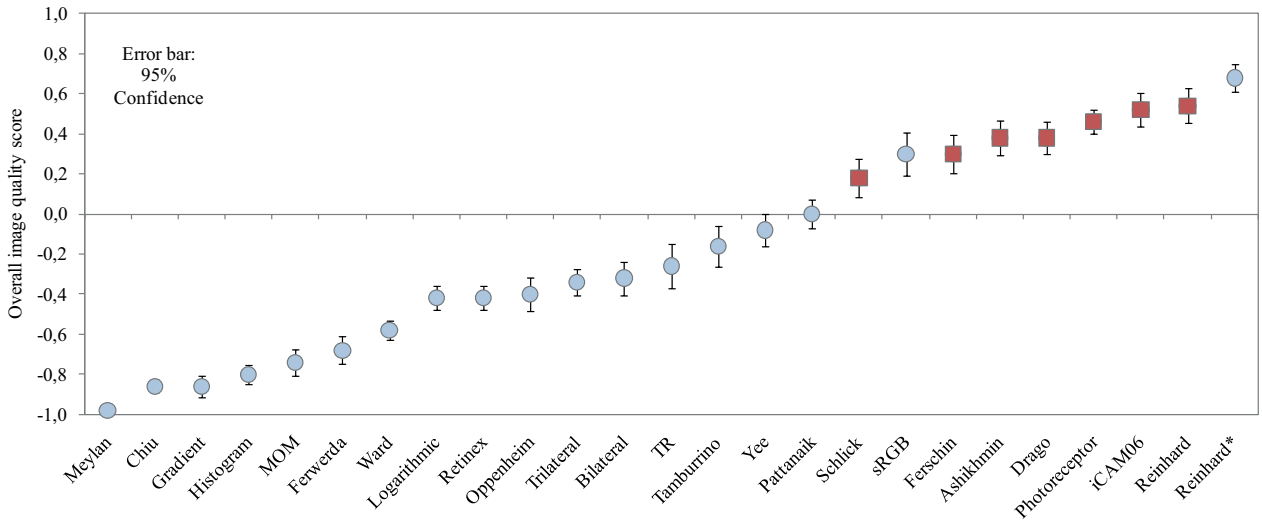


Figure 5. Overall image quality scores from the categorical sort experiment. Operators with square symbols were evaluated in the triplet comparison experiment.

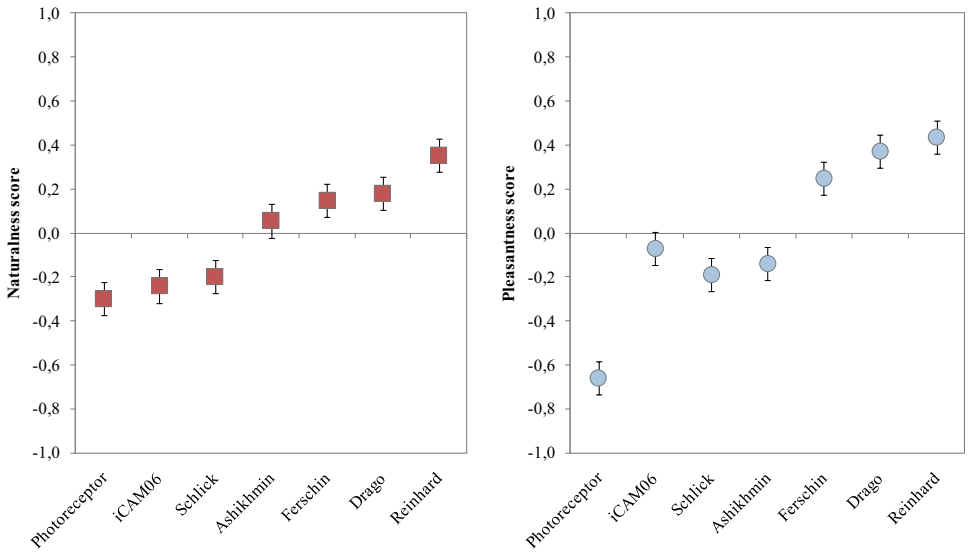


Figure 6. Naturalness and pleasantness scores from the triplet comparison experiment. Naturalness scores on the left (square) and pleasantness scores on the right (round). Error bars $CI = 1.96\sigma_{obs}$.

Table 1. Naturalness and pleasantness scores derived from the triplet comparison experiments. Score (JND) values are used in this study and MOS values shown here for comparison. In parenthesis $CI = 1.96\sigma_{obs}$ for JND and $CI_{95\%}$ for MOS.

Operator	**Naturalness score, JND**	**Pleasantness score, JND**	**Naturalness score, MOS**	**Pleasantness score, MOS**
Reinhard	0.35 (0.08)	0.44 (0.08)	6.26 (0.16)	6.28 (0.15)
Drago	0.18 (0.08)	0.37 (0.08)	5.92 (0.18)	6.11 (0.16)
Ferschin	0.15 (0.08)	0.25 (0.08)	5.78 (0.19)	5.86 (0.18)
Ashikhmin	0.06 (0.08)	-0.14 (0.08)	5.44 (0.21)	5.17 (0.20)
Schlick	-0.20 (0.08)	-0.19 (0.08)	5.12 (0.19)	5.22 (0.20)
iCAM06	-0.24 (0.08)	-0.07 (0.08)	4.83 (0.22)	5.27 (0.22)
Photoreceptor	-0.30 (0.08)	-0.66 (0.08)	4.96 (0.20)	4.26 (0.18)

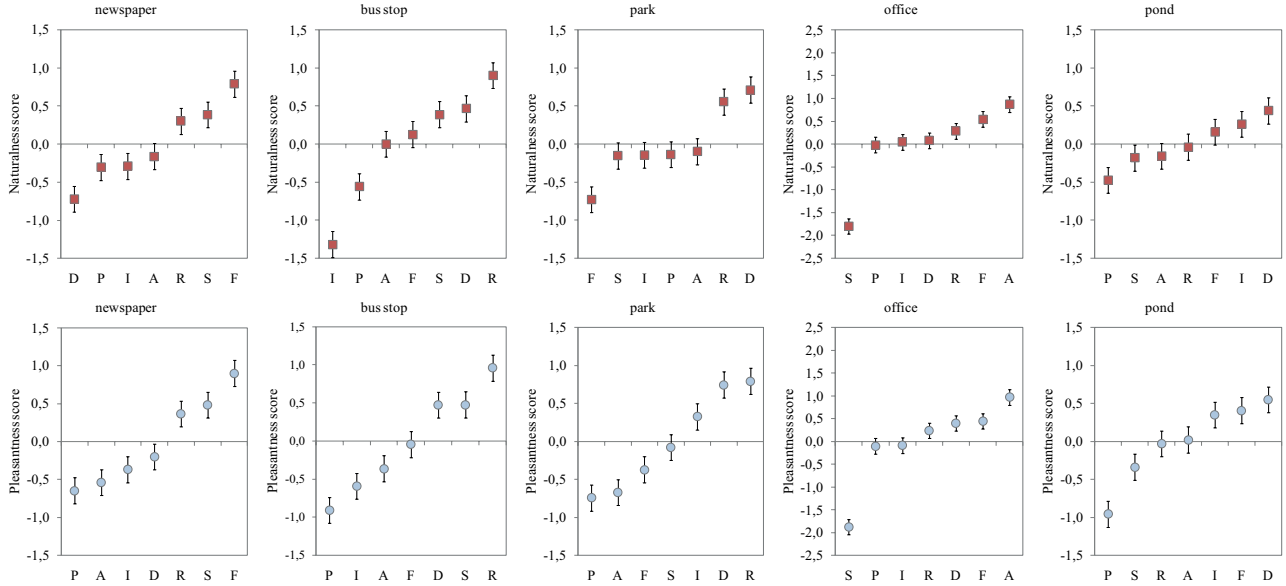


Figure 7. Naturalness and pleasantness scores from the triplet comparison experiment for each test image. P = *Photoreceptor*, I = *iCAM06*, S = *Schlick*, A = *Ashikhmin*, F = *Ferschin*, D = *Drago* and R = *Reinhard*. Naturalness scores on top (square) and pleasantness scores below (round). Error bars $CI = 1.96\sigma_{obs}$.

Standard deviation of naturalness and pleasantness was used to indicate how well different test images differentiated the operators and can be also seen from the naturalness and pleasantness scores in Figure 6. The **pond** image clearly differentiated the operators the least. The range of scores for **office** image is the highest, but that image only differentiated one operator from the others. The **newspaper**, **bus stop** and **park** images were successful and differentiated the samples well with naturalness in the **park** image being an exception.

4. DISCUSSION

The image **pond** has the lowest standard deviation and therefore differentiated the operators the worst. It could mean that the differences between the samples for that image were almost non-distinguishable. The test image may have been too easy for the tone mapping operators. The dynamic range of the image is relatively low and not enough important details are located in the dark regions of the image. Other test images could also have been improved. Test image **office** has the highest standard deviation, but it only differentiated one operator

from the others. The test image feels a bit chaotic and the window scene in the background could have have been a nature or city scene instead of the brick wall.

The image **bus stop** differentiated the operators the best, especially based on naturalness. Interestingly the *iCAM06* operator was evaluated worse in naturalness than pleasantness. The overcolorful red brick wall was mentioned in the *iCAM06* **bus stop** image more than once by the participants. The operator controls colorfulness as a function of the luminance level based on the Hunt effect. The image in question was shot in relatively high illumination.

As many tone mapping operators as possible were included in this study with the limitation of only operators with available implementation. Multiple retinex and bilateral filter based operators have been published and the selection of used operators was based on the fact that other implementations were also used from the same source. Only HDR tone mapping operators were used due the idea of the same HDR source image and therefore interesting methods like the exposure fusion by Mertens et al[25] were forced to be left outside of the study.

The iCAM06 operator includes white balancing with chromatic adaptation procedure which was also performed for all the test images beforehand. The tone mapped images by iCAM06 could have been somewhat different if the input HDR image would have been pure colorimetric and not white balanced beforehand. The retinal operators by Meylan et al.[20] and Tampurrino et al.[21] are designed for non-demoisacked CFA images, but also support HDR RGB images.

The *Retinex*[26] and *Bilateral*[27] operators used did not perform very well, which is surprising, as especially the bilateral operator has been ranked high in many tone mapping studies. This might mean that the implementation or the parameters used have not been optimal. Reinhard points out in the implementation documentation that the *Retinex* implementation might be incomplete due to a lack of sufficient details in the original papers.

5. CONCLUSION

A method was created for evaluating natural HDR images. The functioning of the method was validated by conducting subjective evaluating for 25 tone mapping operators using five test images. The method differentiated the operators well and photographic tone reproduction operator by Reinhard et al.[22] was found to perform the best for the used images. The same operator has been found to perform well in other studies as well.[3,4,6,8]

Multiple HDR test images were developed in this study and have been made available for non-commercial research purposes. These images are provided in colorimetrical CIE XYZ color space and are especially useful for evaluating tone mapping operators or HDR cameras for consumer applications. The need for a HDR photo space was seen important for future development of the HDR test image set.

ACKNOWLEDGMENTS

We would like to thank everyone who participated in our subjective experiments and special word of thanks to all the picturesque people agreeing to model in our test images.

REFERENCES

[1] Reinhard, E., Ward, G., Pattanaik, S., and Debevec, P., [*High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting (The Morgan Kaufmann Series in Computer Graphics)*], Morgan Kaufmann Publishers Inc., San Francisco (2005).

[2] Drago, F., Martens, W., Myszkowski, K., and Seidel, H., "Perceptual evaluation of tone mapping operators," *Proc. ACM SIGGRAPH 2003 Sketches & Applications*, 1–1 (2003).

[3] Ledda, P., Chalmers, A., Troscianko, T., and Seetzen, H., "Evaluation of tone mapping operators using a high dynamic range display," *ACM Trans. Graph.* **24**, 640–648 (2005).

[4] Ashikhmin, M. and Goyal, J., "A reality check for tone-mapping operators," *ACM Trans. Appl. Percept.* **3**, 399–411 (2006).

[5] Kuang, J., Yamaguchi, H., Liu, C., Johnson, G., and Fairchild, M., "Evaluating hdr rendering algorithms," *ACM Trans. Appl. Percept.* **4**, 1–9 (2007).

[6] Cadik, M., Wimmer, M., Neumann, L., and Artusi, A., "Evaluation of HDR tone mapping methods using essential perceptual attributes," *Computers & Graphics* **32**(3), 330–349 (2008).

[7] Yoshida, A., Mantiuk, R., Myszkowski, K., and Seidel, H., "Analysis of reproducing real-world appearance on displays of varying dynamic range," *Computer Graphics Forum* **25**(3), 415–426 (2006).

[8] Annighofer, B., Tajbakhsh, T., and Grigat, R., "Prediction of results from subjective evaluation of real-time-capable tone-mapping operators applied to limited high-dynamic-range images," *Journal of Electronic Imaging* **19**(1), 011015 (2010).

[9] ISO, "Photography – psychophysical experimental methods for estimating image quality – part 1: Overview of psychophysical elements," ISO 20462-1:2005, International Organization for Standardization (2005).

[10] Field, G., "Test image design guidelines for color quality evaluation," *Seventh Color Imaging Conference: Color Science, Systems, and Application*, 194–196 (1999).

[11] Keelan, B., [*Handbook of image quality: characterization and prediction*], CRC Press (2002).

[12] Salmi, H., Halonen, R., Leisti, T., Oittinen, P., and Saarelma, H., "Development of a balanced test image for visual print quality evaluation," *Proc. SPIE* **7242**, 72420B–11 (2009).

[13] Halonen, R., Nuutinen, M., Asikainen, R., and Oittinen, P., "Development and measurement of the goodness of test images for visual print quality evaluation," *Proc. SPIE* **7529**, 752909–10 (2010).

[14] I3A, "Camera phone image quality (CPIQ) initiative group: Phase 1 white paper," CPIQ, International Imaging Industry Association (2007).

[15] Fairchild, M., "The HDR photographic survey," *Proc. 15 th Color Imaging Conference: Color Science and Engineering Systems, Technologies, and Applications*, 233–238 (2007).

[16] Hong, G., Luo, M., and Rhodes, P., "A study of digital camera colorimetric characterization based on polynomial modeling," *Color Research & Application* **26**(1), 76–84 (2001).

[17] ISO, "Graphic technology and photography – colour characterisation of digital still cameras (DSCs) – part 1: Stimuli, metrology and test procedures," ISO 17321-1:2006, International Organization for Standardization (2006).

[18] Stokes, M., Anderson, M., Chandrasekar, S., and Motta, R., "A standard default color space for the internet - sRGB," (1996). http://www.w3.org/Graphics/Color/sRGB.html.

[19] Kuang, J., Johnson, G., and Fairchild, M., "iCAM06: a refined image appearance model for HDR image rendering," *J. Vis. Comun. Image Represent.* **18**(5), 406–414 (2007).

[20] Meylan, L., Alleysson, D., and Susstrunk, S., "Model of retinal local adaptation for the tone mapping of color filter array images," *Journal of the Optical Society of America A* **24**(9), 2807–2816 (2007).

[21] Tamburrino, D., Alleysson, D., Meylan, L., Susstrunk, S., DiCarlo, J., and Rodricks, B., "Digital camera workflow for high dynamic range images using a model of retinal processing," *Proc. SPIE* **6817**, 68170J–12 (2008).

[22] Reinhard, E., Stark, M., Shirley, P., and Ferwerda, J., "Photographic tone reproduction for digital images," *ACM Trans. Graph.* **21**(3), 267–276 (2002).

[23] ISO, "Photography – psychophysical experimental methods for estimating image quality – part 2: Triplet comparison method," ISO 20462-2:2005, International Organization for Standardization (2005).

[24] Montag, E., "Empirical formula for creating error bars for the method of paired comparison," *Journal of Electronic Imaging* **15**, 0502 (2006).

[25] Mertens, T., Kautz, J., and Reeth, F. V., "Exposure fusion," *Proceedings of the 15th Pacific Conference on Computer Graphics and Applications*, 382–390 (2007).

[26] Rahman, Z., Jobson, D., and Woodell, G., "Multiscale retinex for color rendition and dynamic range compression," *Proc. SPIE* **2847**, 183–191 (1996).

[27] Durand, F. and Dorsey, J., "Fast bilateral filtering for the display of high-dynamic-range images," *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 257–266 (2002).