Evaluation of HDR Image Rendering Algorithms using Real-world Scenes

Jiangtao Kuang, Changmeng Liu, Garrett M. Johnson, Mark D. Fairchild; Rochester Institute of Technology; Rochester, NY

Abstract

High-dynamic-range (HDR) image rendering algorithms are designed to scale the large range of luminance information that exists in the real world so that it can be displayed on a device that is capable of outputting a much lower dynamic range. Three real-world scenes with a diversity of dynamic range and spatial configuration were designed and captured for evaluation of rendering accuracy of seven HDR rendering algorithms. Observers were asked to directly compare the accuracy of the appearance of the physical scenes and the tone-mapped images. The purpose of this research was to propose a general psychophysical experiment based methodology for rendering accuracy evaluation of HDR rendering algorithms. This analysis has illustrated potential ways to improve and design more robust rendering algorithms for general HDR scenes in the future.

Introduction

Photographic systems often aim to capture and display images that have the same visual appearance as a viewer would have when looking at real-world scenes. However, a real-world scene often contains such a large range of luminance information that it can be easily lost in the details of highlight or shadow regions. Imaging technology has advanced such that the capture and storage of this broad dynamic range is now possible, but the output limitations of common desktop displays as well as hardcopy prints have significantly detracted from the advances made in image creation. In the last decade, many tone-mapping algorithms have been developed to scale this high dynamic range (HDR) to display devices that are only capable of outputting a low dynamic range. A thorough survey of many of these HDR rendering algorithms can be found in Devlin et at.[1]

While many HDR rendering algorithms have been proposed, fewer visual experiments have been completed to evaluate these algorithms' performance. Drago et al.[2] ran preference and naturalness evaluation experiments to measure the dissimilarity of tone-mapped images by different algorithms for various scenes. The two most salient stimulus space dimensions were found most predictive of the success of tone mapping. Based on the preference evaluation, they determined a preference point in this stimulus space, which they then used as a reference to determine what algorithms were most close to observers' preference. Kuang et al.[3] established a testing and evaluation methodology for tone-mapping algorithms based on psychophysical experiments. Paired-comparison experiments were developed to evaluate eight algorithms with the criterion of observers' preference. Kuang et al.[4] also described image preference modeling techniques for HDR image rendering as a continuation of this research. Ledda et al.[5] presented the results of a series of psychophysical trials to evaluate tone-mapping algorithms against linearly mapped high-dynamic-range scenes displayed on an HDR display. The convenience of using an HDR display instead of building HDR scenes for the evaluation is obvious. Overall spatial composition of a scene influences human perception and adaptation, and such a size effect is not available from this kind of evaluation. The simulation validity of using an HDR display verses an actual scene requires further investigation. Recently, Yoshida et al.[6] presented a first attempt to evaluate tone-mapping algorithms by directly comparing tone-mapped images with their corresponding real-world scenes.

Based on the previous work, we conducted a psychophysical experiment of direct comparison between three high-dynamic-range scenes and the tone-mapped images displayed on a low-dynamic-range LCD monitor. The experimental scenes were well designed and set up in the lab, providing the possibility for further investigation on how algorithm performance depends on scene configuration. The purpose of this research was not simply to find out the "best" algorithms, but to design a general psychophysical experiment based methodology to evaluate HDR image rendering algorithms on perceptual accuracy. This paper provides an overview of the many issues involved in an experimental framework that can be used for these evaluations.

Experimental Algorithms

HDR image rendering algorithms can be broadly classified by spatial processing techniques into two categories: global and local operators. Global operators[10] apply the same transformation to every pixel in the image based upon the global image content, while local operators[7,8,9,11,12,13] use a specific mapping tactic for each pixel, generally based on its local spatial content. From the view of design goals, some algorithms[9,10,11,12,13] aim for perceptual accuracy, attempting to simulate human visual effects, while some algorithms[7,8] aim for maximizing visual pleasure using photographic and digital image processing techniques. For this research, seven rendering algorithms were selected to represent different spatial processing techniques and design goals. Based on the preference evaluation results by Kuang et al.,[2] four of the most preferred test algorithms, Bilateral Filter,[7] Photographic Reproduction,[8] iCAM[9] and Histogram Adjustment,[10] were selected to investigate their corresponding rendering accuracy. Three more recent HDR image rendering algorithms, Local Eye Adaptation,[11] Rentix-Based Adaptive Filter[12] and Biggs' Modified iCAM,[13] were selected to evaluate recent developments for HDR rendering.

Experimental Framework

HDR Image Creation

To evaluate the rendering accuracy of tone mapping algorithms, the method of direct comparison of real-world scenes and their corresponding tone-mapped images on a common desktop display was chosen in this research. An important requirement for this evaluation is that experimental scenes should

be invariant during the experiment process. Therefore, it is a good choice to build up scenes in fully controlled conditions for this evaluation as to ensure the constant illumination and configuration. Objects used for scene designs were chosen to represent a large variety of typical image contents. Besides indoor objects that are easily set up in the lab, it is desirable to include other important photographic contents, such as landscapes and skin tone. Scenes were designed based on a criteria that they should have a variety of dynamic ranges and spatial configurations in order to test two of the most important features in a HDR rendering algorithms: tone-mapping and spatial processing.

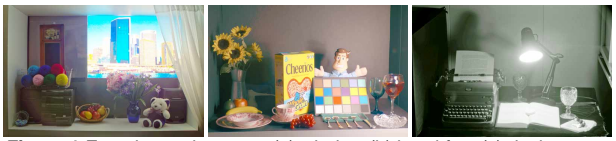

Figure 1 *Experimental scenes: (a) window (b) breakfast (c) desk*

Three HDR scenes were designed for the accuracy evaluation, as shown in Figure 1. The first scene, designated *window*, was built to simulate a window scene including a translucent print attached to a large light booth serving as a bright cityscape outdoor scene, and a black stereo with fine dark details, together with colorful objects such as wool yarns, fruits, flowers, a toy bear and some decorations. This scene has a large lightsource/highlight area, and the absolute luminance is close to a real natural scene (20000 cd/m^2). The second scene, *breakfast*, was designed to incorporate highly chromatic colors, such as a Gretag Macbeth Color Checker, a bright yellow cereal box, artificial fruits and shinny dinnerware. An important feature of this scene is to test algorithms' skin tone rendering by including a doll. Reflections of the light sources from glasses and silverware provides small spot lights, while the cereal box, Color Checker and doll are in the mid-luminance range, and the tablecloths behind the cereal box provides fine shadow details. The third scene, *desk*, has very high dynamic range of luminance, consisting of mostly black-and-white objects, such as a black typewriter, a black table lamp, a book, a white silk napkin, keys, glasses, and a Halon (pressed PTFE), serving as a white point in the scene. The scenes' statistics are summarized in Table 1.

Table 1 Test HDR scenes statistics

	window	breakfast	desk
Max. Lum. (cd/m2)	20,000	30,000	99,800
Min. Lum. (cd/m2)	11.8	1.02	0.742
Dynamic Range	1,700:1	29,500:1	135,000:1
White Point Lum. (cd/m2) & Chromaticity	16,800, x=0.346, y=0.381	775, x=0.448, y=0.403	432, x=0.416, y=0.369

A specially designed Fuji S2 digital camera was used to capture HDR scenes. A monochrome sensor replaced the normal CCD with a color filter array, and three external filters were instead installed in a color-wheel in front of the camera. The spectral transmittances of the filters are close to color-matching functions $\bar{x},\bar{y},\bar{z}$ of the 1931 CIE standard observer, which make it possible for accurate colorimetric reproduction under different illuminants. The camera was first colorimetricly characterized to recover the response curve and the color transform matrix.[14] The camera aperture was fixed to f/8 during the capturing with different shutter speeds ranging from 1/2000 to 8.0 seconds. All captured images were stored with 12-bit raw data for the construction of HDR images with camera response curve using the multiple exposure method proposed by Robertson et al.,[15] and then saved in the Radiance RGBE format. Each HDR image was created using 15 static images. The red, green and blue channels of the HDR images are linear to physical luminance. The characterized transform matrix was applied to convert RGB images to XYZ images for algorithms that require colorimetric input, such as iCAM.

Psychophysical Evaluation

The aim of development of tone-mapping algorithms for high-dynamic-range digital photography is often to reproduce the accurate visual appearance of the original scenes. Therefore, the aim of this experiment is evaluate the accuracy of a rendering when the original scenes are present.

Viewing Techniques

All psychophysical experiments were performed in a dark surround. The rendering results were displayed on a 23-inch Apple Cinema HD LCD Display with the maximum luminance of 180 cd/m^2 and a 1920 by 1200 pixels resolution. The LCD display was characterized with colorimetric characterization model presented by Day.[16] Observers sat at approximately 60 cm from the display. The experimental images were presented on a gray background with a luminance of 20% of the adopted white's luminance.

Experimental Methods

The paired comparison method derived from Thurstone's law of comparative judgment[17] was used in the experimental design for this research. Both the sequence in which the images were presented and their position on the screen (left or right) were randomized. For each pair, observers were asked to make a judgment as to which rendered image was closer in appearance to the original scenes. An interval scale based on z-scores is calculated from observers' judgments under Thurstone's law, Case 5.[18]

Experiment Procedure

Observers were asked to compare the appearance of tone-mapped images with their of corresponding real-world scenes, which were separately set up in an adjoining room to avoid interaction. When looking at the scenes, the participants were asked to stand in a position where the viewing angles for the physical scenes were the same as those for the tone-mapped images on the display. Image attributes investigated in this research were: highlight contrast, shadow contrast, highlight colorfulness, shadow colorfulness, and overall contrast. The scene *desk* was designed to test luminance tone-mapping performance and only included achromatic objects. Hence, colorfulness was ignored in the evaluation for this scene. The subjects were instructed to judge only the single test image attribute with respect to the physical scene and avoid the influence from other image attributes. In addition, observers evaluated the overall rendering

accuracy comparing to the overall appearance of the real-world scenes. As the white points and luminance ranges of the display and the physical scenes were very different, observers were obligated to have at least 30 seconds of adaptation time. The observers were required to remember the appearance of the physical scenes in the after viewing for at least 30 seconds and return to the display to make their evaluation after a second 30 seconds adaptation. They were allowed to examine one image attribute for all 7 algorithms in one time based on their memory before they were obligated to look at the scenes again, and they could go to back to the scenes anytime they felt it necessary. By enforcing repeated viewings of the original scene it was intended to ensure that observers made their judgment based on the rendering accuracy instead of their own preference.

A paired comparison experiment consisting of 378 comparisons (7 algorithms, 3 scenes and 6 image attributes) was first conducted. It took approximately 90 minutes to complete. 19 observers took part in the experiment. All of them were either staff or students at RIT with different culture backgrounds and with varying imaging experience.

Results and Discussion

The accuracy of the rendering algorithms was first evaluated using a paired comparison method. Figure 2 shows the average accuracy scores of the overall judgments made for the three test scenes, which were obtained from the results for individual scenes. The interval scale along with 95% confidence limits was generated using Thurston's Law of Comparative Judgments, Case V. These results show how algorithms reproduce the appearance of their corresponding physical scenes for overall accuracy together with other five image attributes. We can see that bilateral filter generated significantly more accurate rendering results than other algorithms. The results for individual algorithms are not significantly different over the test image attributes, having similar relative performance pattern with the overall accuracy. The bilateral filter shows consistently the most accurate results for all image attributes, while the Retinex-based filter and modified-iCAM are always the worst two algorithms.

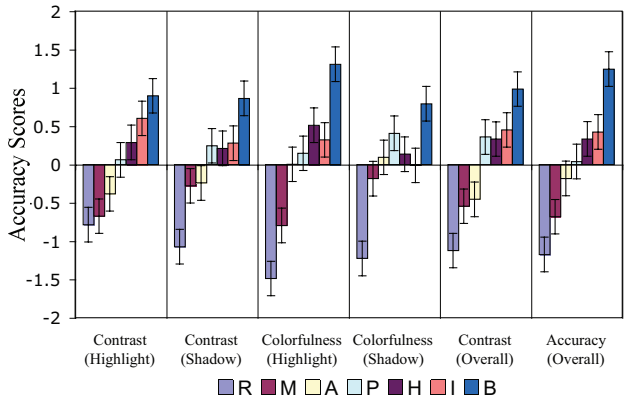

Figure 2 *Overall accuracy scores for HDR rendering algorithms (The algorithms are labeled as Retinex-based filters (R), Modified iCAM (M), Local eye adaptation (A), Photographic reproduction (P), Histogram adjustment (H), iCAM (I) and Bilateral filter (B). The same labels are used in this article).*

The results obtained for the accuracy performance for individual scenes can be viewed in the same way as the averaged results, illustrated in Figure 3-5. As only tone-mapping processing was designed to evaluate for *desk*, colorfulness in highlight and shadow are ignored in these plots. Generally, the results show a strong correlation between the accuracy scores of algorithms for *breakfast* and *desk*, whereas *window* has different overall patterns. For example, iCAM is among the best two rendering algorithms for *breakfast* and *desk*, but among the worst for *window*. This suggests that iCAM might not do well for scenes that have a large area light-source. It is interesting to find that the Modified iCAM has the opposite pattern as iCAM, which suggests that it may be possible to combine the two versions of iCAM to get better rendering results for general scenes. Histogram adjustment, on the contrary, has the opposite trend by performing much better in *window* than the other two scenes. Local eye adaptation doesn't perform as well for *desk*, a very high dynamic range scene, as the other two scenes, suggesting a linkage between a scene's dynamic range and its rendering performance. Other algorithms have more consistency across the different image contents.

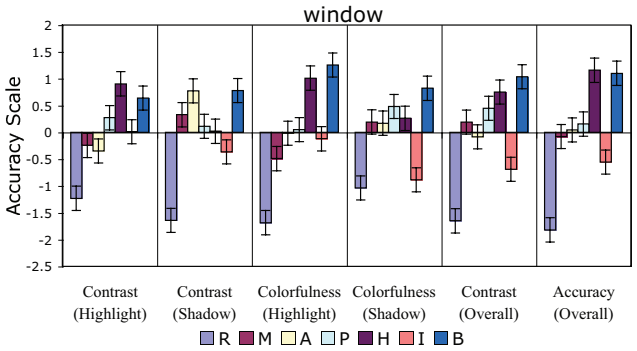

Figure 3 *Accuracy scores for window by image attribute*

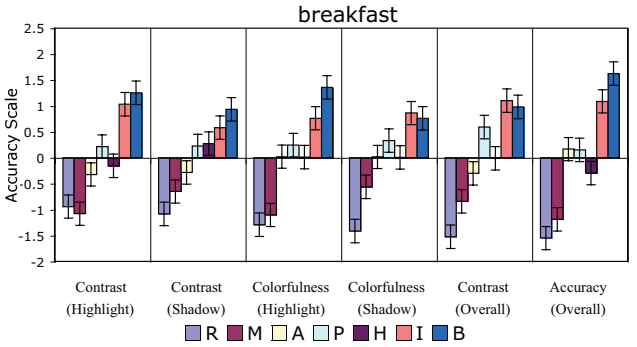

Figure 4 *Accuracy scores for breakfast by image attribute*

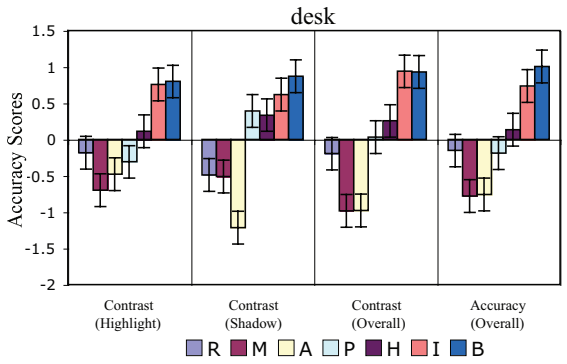

Figure 5 *Accuracy scores for desk by image attribute*

Conclusion

A psychophysical experimental framework has been developed to evaluate the rendering accuracy of HDR rendering algorithms. Three HDR real-world scenes with varied characteristics were designed and constructed to test seven algorithms representing different spatial processing techniques and design goals. The scenes were captured with a characterized digital camera for accurate photometry and colorimetry. A paired comparison psychophysical experiment was developed for accuracy evaluation of seven test algorithms. The bilateral filter consistently performed well, making it a good candidate for an obligatory algorithm that could be included in future algorithm testing experiments. It should be noted that the bilateral filter algorithm used was modified slightly from the original publication, to use photometric luminance and different parameters settings. This could explain the performance differences seen by Ledda.[5] The evaluation results for individual image attributes have illustrated ways to test algorithms to improve their overall performance.

References

[1] K. Devlin, A Review of Tone Reproduction Techniques, Technical Report CSTR-02-005, Department of Computer Science, University of Bristol (2002).

[2] F. Drago, W.L. Martens, K. Myszkowski, H.P. Seidel, Perceptual evaluation of tone mapping operators, In ACM SIGGRAPH Conference Abstracts and Applications (2003)

[3] J. Kuang, H. Yamaguchi, G.M. Johnson, M.D. Fairchild, Testing HDR image rendering algorithms, IS&T/SID 12th Color Imaging Conference (2004).

[4] J. Kuang, G.M. Johnson, M.D. Fairchild, Image Preference Scaling for HDR image Rendering, IS&T/SID 13th Color Imaging Conference (2005).

[5] P. Ledda, A. Chalmers, T. Troscianko, H. Seetzen, Evaluation of tone mapping operators using a high dynamic range display, Proceeding of ACM SIGGRAPH 2005, pg. 640-648 (2005).

[6] A. Yoshida, V. Blanz, K. Myszkowski, H. P. Seidel, Perceptual evaluation of tone mapping operator with real-world scenes, Proceedings of the SPIE, Volume 5666, pp. 192-203 (2005).

[7] F. Durand and J. Dorsey, Fast Bilateral Filtering for the Display of High-Dynamic-Range Image, In *Proceedings of ACM SIGGRAPH 2002*, Computer Graphics Proceedings, Annual Conference Proceedings, pg. 257-266 (2002).

[8] E. Reinhard, M. Stark, P. Shirley and J. Ferwerda, Photographic Tone Reproduction for Digital Images, In *Proceedings of ACM SIGGRAPH 2002*, Computer Graphics Proceedings, Annual Conference Proceedings, pg. 267-276 (2002).

[9] G.M. Johnson and M.D. Fairchild, Rendering HDR Images, IS&T/SID 11th Color Imaging Conference, Scottsdale, pg. 36-41 (2003).

[10] G.W. Larson, H. Rushmeier and C. Piatko, A Visibility Matching Tone Reproduction Operator for High Dynamic Range Scenes, IEEE Transactions on Visualization and Computer Graphics, pg. 291-306 (1997).

[11] P. Ledda, L. P. Santos, A. Chalmers, a local model of eye adaptation for high dynamic range images, in Proceedings of the 3rd International Conference on Computer Graphics, Virtual Reality, Visualization and Interaction in Africa, AFRIGRAPH2004 (2004).

[12] L. Meylan, S. Susstrunk, High dynamic range image rendering using a Retinex-based adaptive filter, IEEE Transactions on Image Processing (2005).

[13] W. Biggs, Perceptual accuracy of tone mapping algorithms, MS. Thesis, Dalhousie University, (2004).

[14] E. Murphy, L.A. Taplin, R.S. Berns, Experimental evaluation of museum case study digital camera systems, Proc. IS&T Second Image Archiving Conference, (2005)

[15] M.A. Robertson, S. Borman, R.L. Stevenson, Dynamic range improvement through multiple exposures, IEEE International Conference on Image Processing, (1999).

[16] E.A. Day, L.A. Taplin, and R.S. Roy, Colorimetric Characterization of a Computer-Controlled Liquid Crystal Display, Color Res. Appl. in press (2004)

[17] L.L. Thurstone, a Law of Comparative Judgment, Psychological Review, 34:273-286, 1927

[18] P. Engeldrum, Psychometric Scaling: a Toolkit for Imaging Systems Development, Imcotek Press, Winchester, 2000, 93-108

Author Biography

Jiangtao Kuang is a PhD candidate at the Munsell Color Science Lab, Rochester Institute of Technology. He received his B.S. degree in optical engineering from Zhejiang University, China. His research interests include image color appearance, high-dynamic-range digital photography, gamut mapping and image quality.

.