Evaluating HDR Rendering Algorithms

JIANGTAO KUANG,[1] HIROSHI YAMAGUCHI, CHANGMENG LIU, GARRETT M. JOHNSON, MARK D. FAIRCHILD

Munsell Color Science Laboratory, Rochester Institute of Technology, Rochester, New York

---

A series of three experiments has been performed to test both the preference and accuracy of HDR rendering algorithms in digital photography application. The goal was to develop a methodology for testing a wide-variety of previously published tone-mapping algorithms for overall preference and rendering accuracy. A number of algorithms were chosen and evaluated first in a paired-comparison experiment for overall image preference. A rating-scale experiment was then designed for further investigation of individual image-attributes that make up overall image preference. This was designed to identify the correlations between image attributes and the overall preference results obtained from the first experiments. In a third experiment, three real-world scenes with a diversity of dynamic range and spatial configuration were designed and captured to evaluate seven HDR rendering algorithms for both of their preference and accuracy performance by comparing the appearance of the physical scenes and the corresponding tone-mapped images directly. In this series of experiments, a modified Durand & Dorsey's bilateral filter technique consistently performed well for both preference and accuracy, suggesting that it a good candidate for a common algorithm that could be included in future HDR algorithm testing evaluations. The results of these experiments provide insight for understanding of perceptual HDR image rendering and should aid in design strategies for spatial processing and tone mapping. The results indicate ways to improve and design more robust rendering algorithms for general HDR scenes in the future. Moreover, the purpose of this research was not simply to find out the "best" algorithms, but rather to find a more general psychophysical experiment based methodology to evaluate HDR image rendering algorithms. This paper provides an overview of the many issues involved in an experimental framework that can be used for these evaluations.

Categories and Subject Descriptors: I.3.3 [**Computer Graphics**]: Picture/Image Generation – Display algorithms, viewing algorithms; I.4.0 [**Image Processing and Computer Vision**]: General – Image Display
Additional Key Words and Phrases: High-dynamic-range imaging, tone-mapping algorithms evaluation, psychophysical experiments

---

1. INTRODUCTION

High-dynamic-range (HDR) images typically contain a large range of luminance information and thus are represented by more than 8-bits per channel. As a general rule, an HDR scene can be thought of as needing more than 12-bits per channel when encoded in "linear-light" whereas images that need between 8-12 bits can be thought of as "extended dynamic range." Scenes in the "real-world" can cover a large absolute luminance range (up to 9 log units) between the highlights and the shadows, though typical scenes do not tend to span more than 5 log units [Jones and Condit 1941]. Imaging technology [Debevec and Malik 1997; Nayar and Mitsunaga 2000; Ikeda 1998; Ward Larson 1998; Ward 2005] has advanced such that the capture and storage of this

---

[1] Authors' addresses: Jiangtao Kuang, Munsell Color Science Laboratory, Rochester Institute of Technology, 54 Lomb Memorial Dr., Rochester, NY 14623.

broad dynamic range is now possible, but the output limitations of common desktop displays as well as hardcopy prints have not necessarily followed the same advances made in image creation. HDR 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. Although recent advances in display technology have suggested that HDR displays are on the horizon, for some applications such as hard copy printing the need for dynamic range reduction will always be there. Many tone-mapping algorithms have been developed for computer graphics and imaging application in the last decade. A thorough survey of many of these HDR rendering algorithms can be found in [Devlin et al 2002].

While many HDR rendering algorithms have been proposed, far fewer visual experiments have been conducted to evaluate these algorithms' performance. When a new rendering algorithm is proposed, the rendering performance is generally evaluated by comparing images against some of the previous HDR image rendering algorithms; however, a sound testing and evaluation methodology based on psychophysical experiment results has not yet been well established. This is especially true for the evaluation in terms of rendering accuracy compared to an original scene. In this context, the present article intends to provide detailed methods and results of a set of three psychophysical experiments for evaluating a number of previous published algorithms for their rendering preference and accuracy performance. The results of these experiments help to better understand existing tone-mapping techniques for HDR image rendering and can potentially serve as a starting point for the development of more robust HDR image rendering algorithms. It is important to note that besides finding out the "best" algorithms currently available, newly developed algorithms can be tested directly against existing algorithms using similar experimental setups. The goal of this article is also to provide a general psychophysics based evaluation framework for testing HDR rendering algorithms [Johnson 2005].

This general outline of this paper is divided into four parts: (1) a brief overview of tone-mapping algorithms used in HDR image rendering; (2) an outline of the experimental framework used in these evaluations; (3) the results of three psychophysical HDR rendering algorithms evaluation experiments; and (4) further analysis and suggestions for establishing the evaluation framework. More details on the background of this HDR imaging, implementation of the test algorithms, and definitions of image attributes have been previously published [Johnson 2005; Kuang et al. 2004; Kuang et al. 2005; Kuang et al. 2006].

2. HDR IMAGE RENDERING ALGORITHMS

HDR image rendering algorithms can be broadly classified by spatial processing techniques into two categories: global and local operators [Reinhard et al. 2006]. Global operator applies the same transformation to every pixel in the image based on the global image content, while for local operators a specific mapping tactic is used for each pixel based upon its spatially localized content. It is important to stress that for global operators it is not necessarily the same operator applied identically for every image, as the global operator can be a function of image information such as the histogram. Likewise the local operators take many different approaches for determining the spatial extent of the operator, such as low-pass filters, edge preserving low-pass filters, or multi-scale pyramids. There are strengths and weaknesses to both the global and local tone-mapping approaches. The global operators tend to be computationally simpler and as a result can be easier to implement and faster to perform. The spatial processing of the local operators tends to be computationally more expensive, but can allow for a more dramatic reduction in overall dynamic range. From the view of rendering intents or goals, some algorithms aim to produce images that are visual appealing, using photographic and digital image processing techniques to enhance rendering pleasantness, while other algorithms aim for perceptual accuracy. These algorithms attempt to mimic perceptual qualities in addition to compressing the range of luminance, resulting in an image which provokes the same visual response as a human may have when viewing the same real-world scene. Other algorithms are designed to maximize overall visibility in images; an example would be for use on HDR medical images. Since the local algorithms are capable of larger dynamic range compression and also tend to mimic the local adaptation behavior of the human visual system, more emphasis was put on testing more of these types of algorithms specifically. A brief introduction to the algorithms evaluated in this article is given below.

Sigmoidal Transformation

Braun [Braun and Fairchild 1999] presented an image lightness rescaling technique for gamut mapping of 8-bit images using a sigmoid contrast enhancement function. The form of the sigmoid functions was derived from a discrete cumulative normal function, given in Eq. 1, where x_0 and σ are the mean and variance of the normal distribution respectively. While this method was valuable for gamut mapping using the lightness channel in a color appearance space, such as CIELAB, it was not intended for the extreme dynamic range reduction. The original implementation was modified so that instead of "lightness", the logarithm of luminance, normalized from 0 to 100, was used to compress the HDR images.

$$s_i = \sum_{n=0}^{n=i} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_n - x_0)^2}{2\sigma^2}} \quad (1)$$

Histogram Adjustment

Ward [Ward Larson et al. 1997] presented a global operator to reproduce perceptually accurate tones in HDR scenes by discovering local luminance adaptation levels and modifying the luminance histogram. The human visual models of glare, spatial acuity and color sensitivity effects are incorporated into this model to reproduce imperfections in human vision and mimic the subjective viewing experience.

Retinex

"A Retinex is all mechanisms from retina to cortex necessary to form images in terms of lightness" [McCann 2004]. The application of dynamic range compression of real images was initially described in a patent by [Frankle and McCann 1983]. We test the McCann99 version of Retinex with the public Matlab code made available by [Funt et al. 2000]. An automatic method [Funt et al. 2002] was used for the setting the number of iterations, a free parameter that controls contrast and dynamic range compression of the resulting image in Retinex.

iCAM

iCAM [Johnson and Fairchild 2003] is an image appearance model that has been extended to render high dynamic range images for display. iCAM attempts to predict complex human visual responses by combining traditional color appearance capabilities with properties of spatial vision.

Photographic Reproduction

Reinhard et al. [Reinhard et al. 2002] presented a tone reproduction technique for HDR rendering by simulating dodging-and-burning in traditional photography. This method automatically applies different luminance mapping scales to the relative highlight and shadow regions, where the local contrast are estimated using a center-surround function with different spatial extent.

Bilateral Filtering Technique

Durand and Dorsey [Durand and Dorsey 2002] proposed a rendering technique to reduce the overall contrast while preserving local details in the image. An edge-preserving spatial processing operator, called a bilateral filter, decomposes the image into two layers: the base layer, encoding large-scale variations, and the detail layer. The contrast of the base layer is compressed and they are combined again with the detail layer to produce the final image. This technique was modified slightly to use CIE XYZ colorimetry to calculate the luminance channel. Overall brightness and base contrast, two user-controlled parameters, were set to 1.8 and 1.4 respectively. A clipping to the 1st and 99th percentile

of the image rendering data was performed to remove any extremely dark and bright pixels prior to display.

Modified iCAM

The Bigg's modified iCAM [Biggs 2004] incorporates Reinhard and Devlin's tone mapping operator into iCAM framework, which is based on the physiology of the photoreceptors in the human eye. The operator output is some combination of a globally adapted value based on the image averages and a locally adapted value around each image pixel. This method aims to preserve the color fidelity and chromatic adaptation of iCAM while preserving the tone reproduction capabilities of Reinhard and Devlin's operator.

Local Eye Adaptation

The local-eye adaptation method [Ledda et al. 2004] compresses the dynamic range of the luminance channel by simulating the photoreceptor responses in the retina. The electrophysiological equation, known as the Naka-Rushton equation [Naka and Rushton 1966], was modified to a local operator that predicts the S-shaped response function of the rods and cones at localized adaptation intensities. The bilateral filter was used to compute the average local luminance to avoid halo artifacts common to other local operators.

Retinex-Based Adaptive Filter

Meylan [Meylan and Susstrunk 2005] proposed a center-surround Retinex model based adaptive filter for HDR rendering. The luminance channel is then processed by the surround-based Retinex methods, which is defined by the first component of a principal component analysis. Principal component analysis provides orthogonality between channels and thus reduces the chromatic changes caused by the modification of luminance. The weights of surrounding pixels are computed with an adaptive filter method by adapting the shape of the filter to the high contrast edges in the images to prevents halo artifacts, one of the important drawback of the previous surround-based Retinex methods.

Before proceeding, it is important to stress that the aim of these evaluations is to gain better understanding on how current HDR rendering algorithms perform in terms of image preference and rendering accuracy, not to select the best overall algorithm. The experimental results should be helpful for the investigations of HDR rendering techniques, specifically for spatial processing and tone mapping. The goal is to guide the development of more robust HDR rendering algorithms. The relationship between the image overall pleasantness or preference and visual accuracy was investigated in this study. The importance of understanding this relationship lies in determining whether separate algorithms are necessary for different rendering intents, such as accurate or

pleasing rendering. In addition, it would be beneficial to find a default or common algorithm that could serve as a baseline TMO for future algorithm testing experiments, such as the work of CIE Technical Committee 8-08 on testing spatial color appearance models [Johnson 2005].

3. EXPERIMENTAL FRAMEWORK

3.1 Experimental Techniques

Unfortunately no satisfying computational model exists for quantifying image preference of complex images. Likewise there is not a model for evaluating the accuracy of rendered images that are displayed on a monitor compared to the corresponding physical scenes. It is therefore necessary to use human observers and psychophysical experimentation to evaluate the performance of HDR rendering algorithms. The general aim in the development of most tone-mapping algorithms for high-dynamic-range digital photography is to reproduce an accurate visual appearance of the original scenes and to create a pleasant reproduction as well. Therefore, three psychophysical experiments were designed to evaluate HDR rendering algorithms judging both preference and accuracy.

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. The total display area was 1920 x 1200 pixels allowing images to be viewed in pairs with long-dimensions of approximately 800 pixels. The LCD display was characterized with the colorimetric characterization model presented by [Day et al. 2004]. 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 adapting white point.

3.2 Experimental Methods

Two psychophysical scaling methods were implemented in this research. The paired-comparison method is a powerful technique for generating interval scales of algorithm performance along a given dimension, which are derived using Thurstone's law of comparative judgment [Thurstone 1927]. In all paired-comparison experiments care was taken to randomize both the sequence in which the images were presented and their position on the screen (left or right). For each pair observers were asked to make a judgment as to which rendered image was preferred, or which was closer in appearance to the original scenes. In order to create an interval scale from these comparison data, every stimulus must be compared with every other stimulus. For n stimuli, this leads to $n(n-1)/2$ experimental trials. The total number of trials increases very rapidly as the

number of stimuli increase. Therefore, paired comparison experiments are generally most appropriate when there are a small number of experimental samples. An interval scale based on z-scores is calculated from observers' judgments under Thurstone's law, Case V. More thorough details of this type of analysis can be found in [Bartleson and Grum 1984].

The second experimental technique used was the rating-scale method. This method is a relatively simple way to estimate relationships amongst many stimuli. Graphical rating scales, with well-defined endpoints or anchors, are a commonly used rating-scale method. Observers are asked to graphically indicate where given stimulus lies on an attribute scale compared to the anchor images. Ratings can also be applied without the benefit of the graphical scale, as an observer may directly apply a numerical value to an image attribute. The rating-scale method is useful when dealing with a large number of samples, as it is faster for the observers and relatively easier to analyze compared with the method of paired-comparison. However, there are still many potential problems with rating-scale methods [Bartleson and Grum 1984]. In a rating-scale experiment, observers are asked to make a judgment requiring more subjective input, which makes the results more dependent on individual observers abilities and ranges, and thus the scales can be considered less precise than paired-comparison. By incorporating the rating-scale method in these evaluations, we were interested in determining whether a simple psychophysical method can provide as much (and as good) information as that resulting from a more laborious and more complex method.

3.3 Test Images and scenes

An ideal HDR rendering algorithm should be image independent, or capable of performing the desired tone-mapping task regardless of the input image content. Therefore, it is important to have a wide variety of images available for testing various algorithms. These scenes should include both indoor and outdoor photographs covering wide range of luminance histograms. Johnson [Johnson 2005] demonstrated examples of typical histograms of several different types of HDR images. Overall dynamic-range as well as mean luminance of the scenes are probably the two most salient factors that can determine the rendering performance of a given algorithm. The dynamic range of natural scenes, which is typically defined as the absolute luminance range between the highlights and shadows, describes the overall amount of compression necessary for low-dynamic-range outputs. The average luminance, often described as the "key" of an image, can indicate whether a scene is subjectively light, normal, or dark. A quantitative evaluation system was previously developed [Kuang et al. 2004] to describe these two factors for

characterizing HDR images using the zone system [Reinhard et al. 2002], which is widely used in traditional photography. Generally, test images should include a large variety of image content, such as landscape, architecture, human portraiture, and images with differing light-source sizes. Using a variety of images can help identify both strengths and weaknesses of individual rendering algorithms.

Based on the guidance for test image selection described above, twelve HDR images (Fig. 1) were chosen for HDR rendering algorithms image preference evaluation in the first and second experiments. They covered a range of image content types, including landscape scenes (bristolb and tahoe1), indoor scenes (lamp_up and church), architecture (garage, clockbui, split_cute2), both interior and outdoor scenes (colorcube and belgium), human portraits (ashi01 and ashi05) and computer-generated renderings (lamp_pete). These images have a large diversity of image characteristics in histogram distributions, dynamic ranges and average luminances. Links to many of these can be found at http://www.colour.org/tc8-08/Links/.

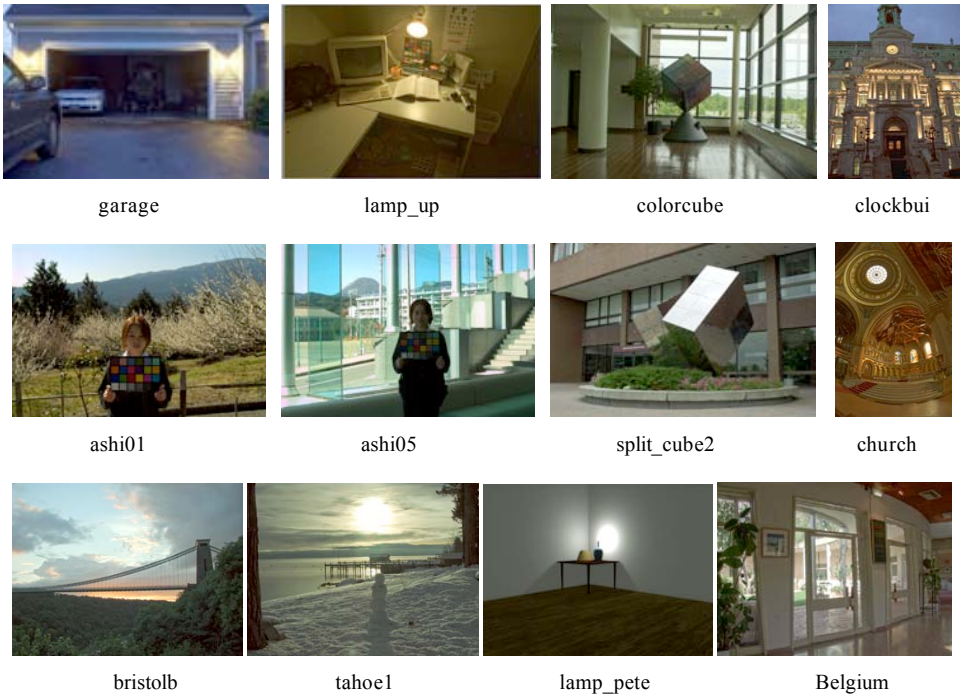

Fig. 1 Thumbnails of experimental images

The third experiment was a direct evaluation of the rendering accuracy of the test algorithms, so it was necessary for the rendered results need to be directly compared against the corresponding real-world scenes. Therefore, an important requirement for this evaluation is that experimental scenes should be invariant during the experiment process. It was important to construct scenes with fully controlled conditions to ensure the

constant illumination and scene configuration. Objects used for the scene designs were chosen to represent a large variety of typical image content. Besides indoor objects that are easily set up in the lab, it was desirable to include other important photographic contents, such as landscapes and skin tones. Scenes were designed based on the same 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 algorithm: overall contrast tone-mapping and spatial processing.

Three HDR real-world scenes (Fig. 2) were designed for Experiment 3. 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 light source/highlight area, and the absolute luminance is close to a real natural scene, with the maximum and minimum luminances of 20,000 cd/m^2 and 11.8 cd/m^2 respectively. 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 shiny dinnerware. An important feature of this scene was the inclusion of a doll to test algorithms' potential skin tone rendering. Reflections of the light sources from glasses and silverware provides small spot highlights, while the cereal box, Color Checker and a doll are in the mid-luminance range, and the tablecloths behind the cereal box provide fine shadow details. The luminance range of this scene is from 1.02 cd/m^2 to 30,000 cd/m^2. The third scene, *desk*, has very high dynamic range of luminance from 0.74 cd/m^2 to 99,800 cd/m^2, 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 disk (pressed PTFE) which served as the white point in the scene. This scene was designed to test algorithms' luminance tone-mapping performance exclusively by excluding chromatic content.

Fig. 2 Experimental scenes: (a) window (b) breakfast (c) desk

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 were designed to be close to linear transforms of the color-matching functions $\bar{x},\bar{y},\bar{z}$ of the 1931 CIE standard observer. These filters 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 [Murphy et al. 2005]. 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. 1999], 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, within the non-clipped regions. The characterized transform matrix was applied to convert RGB images to XYZ images for algorithms that require colorimetric input, such as iCAM.

3.4 Observers

Varying numbers of observers took part in each of the three experiments in this study. All of them were either staff or students at RIT with different cultural backgrounds and with varying imaging experience. Details of the total number and age range are given in Table I.

Table I. Observer statistics

	Number of observers	Age range
Experiment 1 (section 1)	33	22-60
Experiment 1 (section 2)	23	23-40
Experiment 2	19	23-60
Experiment 3	19	23-60

4. EXPERIMENT 1: PREFERENCE EVALUATION

The experiments described in this section aimed to test the performance of HDR rendering algorithms in regards to image preference. Six rendering algorithms were chosen from the literature, which represent different tone mapping and spatial processing techniques. The sigmoid transformation [Braun and Fairchild 1999] was selected to examine the performance of using classic 8-bit image contrast enhancement techniques for rendering higher-dynamic-range images. The histogram adjustment technique proposed by [Ward Larson et al. 1997], which incorporates a human perceptual model, is generally considered one of the best global operators. Local operators often have better compression performance from empirical studies; for this reason, four algorithms are included in our experiment: Retinex [Funt et al. 2000], iCAM by [Johnson and Fairchild 2003], the

bilateral fast filter by [Durand and Dorsey 2002], and the photographic tone reproduction by [Reinhard et al. 2002]. Note that two test algorithms described in the paper by [Kuang et al. 2004] are not described in the discussion in this paper in order to be consistent with the test algorithms in Experiment 2.

Twelve HDR images (Fig. 1) were rendered by all test algorithms, and the results were evaluated in two paired-comparison experiments. Observers were presented with the task of choosing which of the two images they preferred from a given pair. The general rendering performance, the overall impression on image contrast, colorfulness, image sharpness, and natural appearance, etc. were compared in the first experiment, while only tone mapping performances using grayscale images were compared in the second experiment. The grayscale images used in the second section were converted from the CIE Y luminance channel of the color images, discounting by the white point of LCD display to minimize the color shift in the images.

4.1 Overall Results

The paired comparison data were analyzed using Thurstone's Law, Case V. This analysis results in an interval scale of image preference. Thrustone's law relies on the assumption of a one-dimensional scale. Ideally, the uni-dimensional preference scale constructed from paired comparison data should not have intransitive judgments (e.g., A is preferred to B, B to C, and C to A). For section 1, the interval scale along with 95% confidence limits [Montag 2004] for the average preference scores of twelve scenes are shown in Fig. 3, and the results for individual test images are shown in Fig. 4. In both Fig. 3 and Fig. 4 each algorithm is shown along the ordinate in the order of the average scores across all scenes, from the worst to the best. A test of Average Absolute Deviation on the interval scores results in the error of 0.042, indicating that Case V model fits the data well. From the results, the bilateral fast filter has the best overall rendering performance among the test algorithms. It is clear that there are distinct image dependencies from the rendering image preference. No single algorithm consistently performs well for all images, indicating that like gamut mapping algorithms there may be a strong image dependency. The bilateral filter has good results in almost all images except image "lamp_pete", while the photographic tone reproduction performed well on average, but did significantly poorer for image "ashi05", with lower average preference score than the bilateral fast filtering. iCAM and Ward's histogram adjustment are located in the second tier of the algorithms, with good results in most of the test images. Retinex has largest preference variance, resulting in the best rendering performance for the image "lamp_up" and the worst for the image "garage". The fact that the global sigmoid

function has low average preference indicates that traditional 8-bit image enhancement technique may have limited application in HDR image rendering. In Fig. 4, the individual images are sorted by their overall image dynamic ranges, while the order of the algorithms are identical to Fig. 3. From this ordering, no significant trend is identifiable, indicating that there is no strong correlation between preference performance and overall image dynamic range for these particular test algorithms.

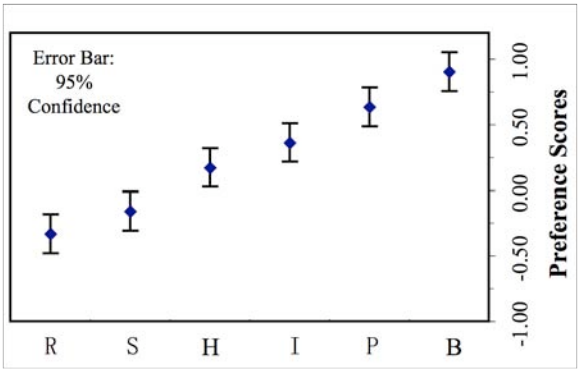

Fig. 3 Average preference scores for 12 scenes (color images) (The algorithms are labeled as Retinex-based filters (R), Sigmoid function (S), Histogram adjustment (H), iCAM (I), Photographic reproduction (P), and Bilateral filter (B). The same labels are used in this article).

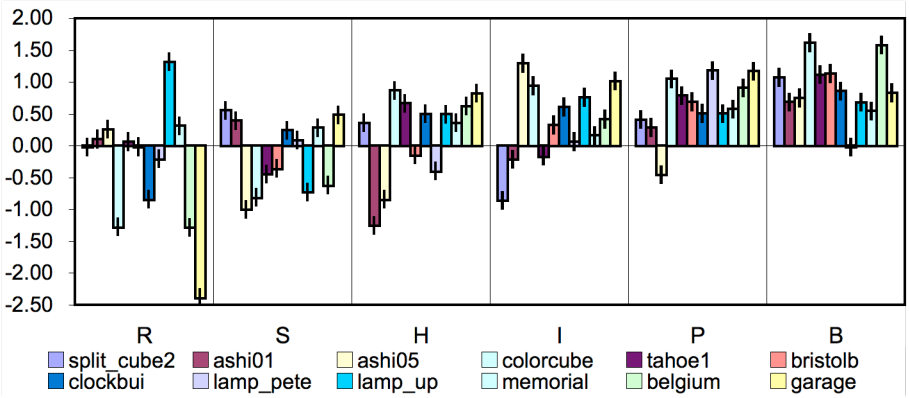

Fig. 4 Preference scores by scene (color images)

In the second experiment, the HDR rendering algorithms were tested only for their tone-mapping performance by using the grayscale version of the rendered images. Fig. 5 presents a graph of the average preference scale obtained from the grayscale tone mapping performance for ten test images, as image “ashi01” and “ashi05” were not included in this experiment. Again, each algorithm is shown along the ordinate axis in the order of average preference scale value. Generally, these results show a similar pattern as those in Fig. 3. The Average Absolute Deviation calculated is 0.046, indicating that Thurstone’s

Law provided a good fit for the data. To determine the relationship between tone mapping and overall rendering performance, the interval scales from Experiment 1 (average preference for ten test images) were plotted against those from Experiment 2, as shown in Fig. 6. The tone mapping performance on just the grayscale correlates with the overall rendering quite well, with a correlation coefficient of 0.98. The grayscale tone mapping performances are consistent with those in the overall rendering results, and often identical. This suggests that tone-mapping performance may be a dominant factor at evaluating the overall rendering performance, when scaling image preference.

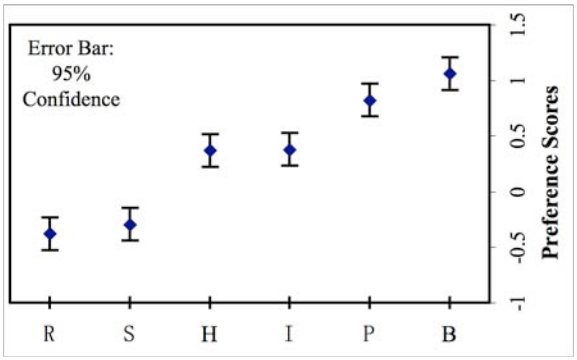

Fig. 5 Average preference scores for 10 scenes (grayscale images)

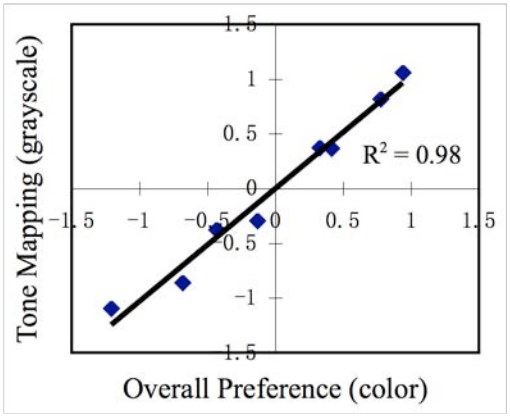

Fig. 6 Tone mapping preference scores vs. overall image rendering preference scores

4.2 Estimating Dimensions Used In Preference Judgments

The paired comparison data were further analyzed using dual scaling [Nishisato 1994], a multidimensional technique that can delineate relations among variables, linear or nonlinear, from multivariate categorical data. The dual scaling analysis attempts to determine the number of independent dimensions that characterize observers' preference, and the percent of the variance that each dimension accounts for in the resulting scale. This can be thought of as another method for validating the use of Thurstone's Law for

the paired-comparison data. The results of the dual scaling analysis for overall preference of all scenes (Fig. 3) are shown in Fig. 7. From the percentage of the variance shown in Fig. 7, we can see that the first dimensions of all scenes are dominant, accounting for over 90 percent of the variance in the preference judgments, and that the remaining dimensions are by comparison marginal. This singular dimensionality supports the assumption used when constructed the interval scale generated using Thurston's law.

4.3 Summary of Experiment 1

A two-part preference evaluation experiment indicated that the bilateral filter performed the best overall, although there was a distinct image dependency for all the test algorithms' performance. The second part of the experiment showed that grayscale tone mapping performance correlates very well with the overall preference, suggesting that tone mapping is a very important factor in the overall HDR image rendering performance. Dual scaling analysis indicated that there was a single perceptual dimension that accounted for most of the variance of preference judgments. This supports the assumptions for using Thurstone's law.

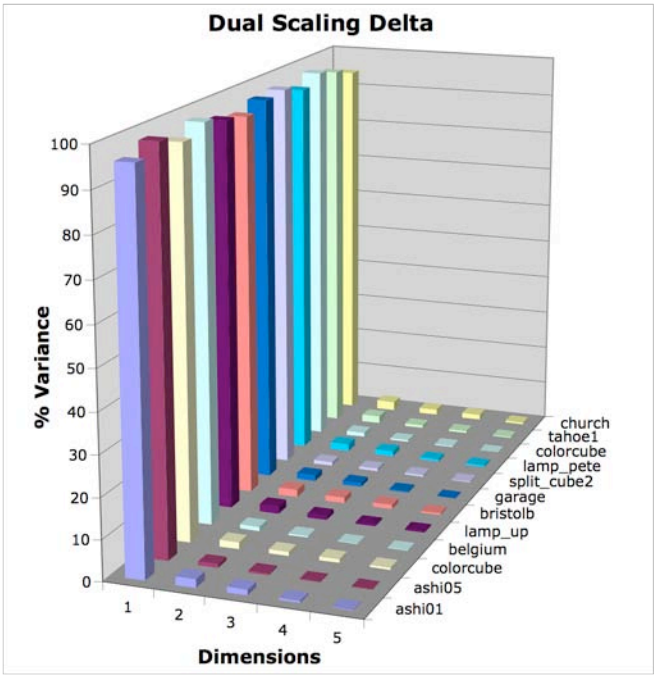

Fig. 7 The percentage of variance accounted for each dimension of dual scaling analysis

5. EXPERIMENT 2: IMAGE PREFERENCE MODELING

While the paired-comparison psychophysical experiment and analysis described above shows the observers' preference for the HDR rendering algorithms, the experiments do

not provide total insight into the actual criteria observers used to scale image preference. This information is important for future research to understand what specific areas are most important for the overall image preference performance of HDR rendering algorithms. By measuring the preference of various individual image attributes from the rendered images, we can compare the performance of a specific attribute for each of the tone-mapping algorithms. The dual scaling analysis discloses that one dimension is dominant to account for the variance, though this does not necessarily correlate to a single image appearance attribute. This result suggests that observers hold a single criterion for overall preference, though that criterion may involve any number of individual attributes, such as overall contrast, colorfulness, or naturalness. It is of interest to determine what, if any, are the most important individual attributes that determine observers' overall image preference.

The image attributes investigated in this experiment were: highlight details, shadow details, overall contrast, sharpness, colorfulness and the appearance of artifacts. The definitions of these image attributes were summarized in the article by [Kuang et al. 2005]. This experiment used the same rendered images and algorithms that were discussed above in Experiment 1, totaling 72 images. A rating-scale method was used to handle the large number of experimental samples in this experiment. The subjects were asked to evaluate their preference for each of the 6 image attributes, making each judgment individually. The rating scale was generated by comparing the rendering to their internal representation of a "perfect" image in their mind. A rating scale number from 0 to 10 was used to express their preference. The attribute "artifacts" was replaced with "lack of artifacts", which made a rating of 0 always mean the least preferred and 10 mean the most preferred for all attributes.

5.1 Overall Results

As no anchor points were provided to the preference scale in this rating experiment, the consequence of observers using the scale arbitrarily is that each observer's ratings are on what can be considered a "rubber band" compared to other observers' ratings, and the rubber band may be shifted or stretched about some origin. The obtained rating scales were first normalized by subtracting the mean value from each observer's rating and dividing the result by the observer's rating scale standard deviation. In this way, all observers have a mean scale value of zero and a standard deviation of unity [Engeldrum 2000]. The normalized rating scales along with 95% confidence intervals for each image attribute over the 12 scenes are shown in Figure 8. For each attribute shown in Fig. 8 the algorithms are ordered by average preference rank, as shown in Fig. 3 above.

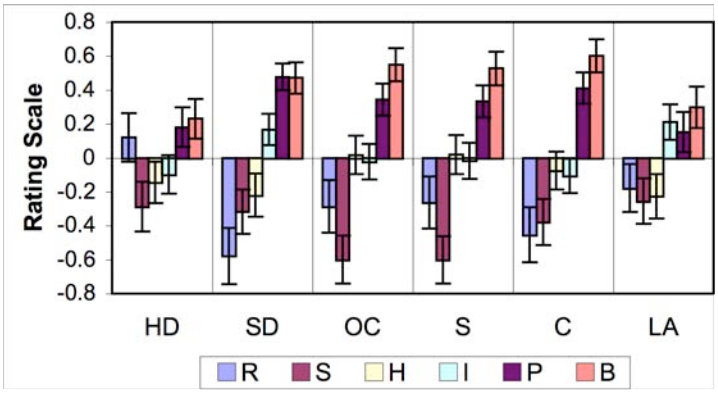

Fig. 8 Rating scales of 6 image attributes over 12 scenes (The image attributes are labeled as Highlight Details (HD), Shadow Details (SD), Overall Contrast (OC), Sharpness (S), Colorfulness (C), and Lack of Artifacts (LA). The same labels are used in this article)

The results show a homogeneous trend for the performance across all the individual image attributes. It can be seen that the bilateral filter and photographic reproduction algorithms have good rendering preference for all the image attributes, whereas Retinex, the sigmoid function and histogram adjustment perform poorly (with the exception of Retinex for highlight details), and the performance of iCAM varied across the image attributes. At first glance this would suggest that if an algorithm performs well with a single attribute, it probably does well for all attributes; in other words, bad performance in one image attribute would strongly affect other attributes. This finding is comparable with the well-known degradation rule in image quality [Keelan 2002].

The strengths and weaknesses of all the algorithms can be ascertained from these data. For instance the bilateral filter has significant higher rating scales than the other algorithms in regards to overall contrast, sharpness and colorfulness. iCAM has comparable rendering performance in regards to not introducing artifacts into the images, and Retinex can provide appealing results in highlight details. The two global operators, histogram adjustment and sigmoid transform, were found to always be less preferred for most attributes. The tone mapping algorithms differ most in the shadow details, overall contrast, sharpness and colorfulness, while they have similar performance in highlight details.

Mahalanobis distances analysis was performed on the individual rating scales to show the similarity of the tone mapping algorithms for each image attribute. This analysis was performed for the average ratings across all the scenes. The Mahalanobis distances among tone mapping algorithms are visualized in Fig. 9 as dendrogram plots of the hierarchical binary cluster trees. The results show that the similarity patterns are very close to the overall preference ranks. The bilateral filter and photographic reproduction techniques have very similar preference for almost all image attributes. iCAM and histogram

adjustment also have high similarity except in shadow details and introduction of artifacts, where iCAM has better performance as shown in Fig. 8.

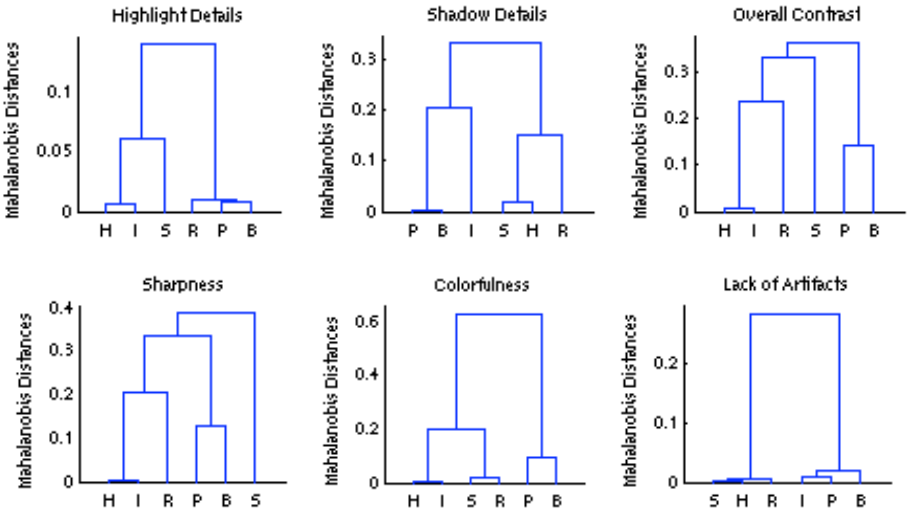

Fig. 9 Hierarchical cluster trees of Mahalanobis distances among tone mapping algorithms for each image attribute

The correlations between the overall preference and all image attributes were tested using Pearson's correlation coefficients. The analysis was performed for all the scenes with the overall preference scales and attribute rating scales for the 6 image attributes. The correlation values are shown in Table II. The results show that overall preference has relatively strong correlations with 4 attributes: shadow details, overall contrast, sharpness and colorfulness. It seems that for these test particular images the highlight details and introduction of artifacts have less of a contribution towards observers' image preference judgments. The correlation coefficients between contrast and sharpness and colorfulness are over 0.9. It indicates that the perceived contrast, sharpness and colorfulness have significant interaction with each other as well as overall image preference, which was also shown in Calabria's experiment [Calabria and Fairchild 2002].

Table II Correlation values among overall preference and image attributes

	Highlight Details	Shadow Details	Overall Contrast	Sharpness	Colorfulness	Lack of Artifacts
Preference	0.45	0.75	0.82	0.81	0.77	0.47
Highlight Details		0.38	0.55	0.53	0.51	0.61
Shadow Details			0.79	0.74	0.82	0.48
Overall Contrast				0.93	0.91	0.57
Sharpness					0.90	0.59
Colorfulness						0.58

A stepwise regression [Draper and Smith 1981] was performed on the overall preference interval scales with the rating scales of the image attributes. This analysis attempts to model the overall preference as a function of linear combination of predictor variables using the subset of the image attributes ratings. This analysis was performed for the average of all the scenes, and independently for each scene. For the average scenes, the overall preference scales can be fit very well just based upon the colorfulness scale with a Pearson's Correlation of 0.98. This is illustrated in Fig. 10.

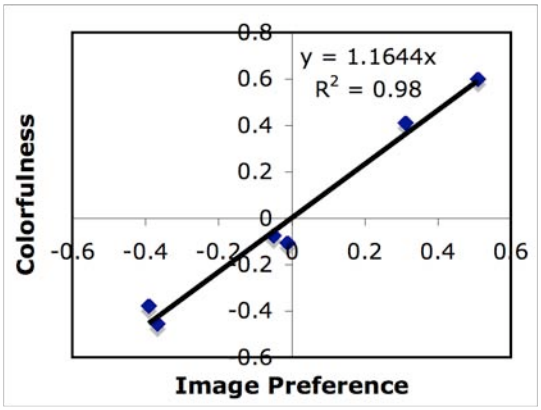

Fig. 10 Average preference scale estimation using colorfulness rating scales

While the single attribute of colorfulness might be capable of predicting the overall preference of the average scenes, it is of interest to determine whether that is the case for all the individual scenes. It is also important to reiterate that colorfulness itself can be highly correlated with overall contrast, as well as sharpness. The image attributes necessary to fit the preference scales for each scene from the stepwise regression analysis are listed in Table III. Other than the church scene, the preference for all the scenes could be predicted with the rating of a single attribute, although that attribute differs from scene to scene. As the correlations among the scales of the individual image attributes are also very strong, such as those for contrast, sharpness and colorfulness, many of these attributes were able to fit the preference results almost as well. These results seem to agree with the conclusion from the dual scaling analysis in Experiment 1, which suggested that the overall image preference scale could be explained in a single dimension of preference, and that dimension may be a single image attribute.

5.2 Summary of Experiment 2

The experimental results of overall image preference show a consistency with a given algorithms performance for the individual image attributes. The bilateral filter

consistently performs well with significantly higher rating scale than the other algorithms for most of the attributes. The overall image preference was highly correlated with the rating scales for the shadow details, overall contrast, sharpness and colorfulness. A stepwise regression analysis showed that the rating scale of a single image appearance attribute is often capable of predicting the overall preference.

Table III image attributes necessary to fit the overall preference scales for individual scenes

	Image attributes needed
belgium	Contrast
bristolb	Colorfulness
church	Shadow details, Colorfulness
colorcube	Artifact
garage	Colorfulness
lamp_pete	Colorfulness
lamp_up	Contrast
tahoe1	Sharpness
clockbui	Shadow details
split_cube2	Contrast
ashi01	Sharpness
ashi05	Sharpness

6. EXPERIMENT 3: ACCURACY EVALUATION

The third experiment aimed to evaluate HDR rendering algorithms for their perceptual accuracy of reproducing the appearance of real-world scenes. This was accomplished by direct comparison between three high-dynamic-range real-world scenes (Fig. 2) and their corresponding rendered 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 as to how algorithm performance may depend on the scene configuration. The design purpose and scene characteristics were described above. Based upon the preference evaluation results in Experiment 1, four of the most preferred test algorithms, the bilateral filter, photographic reproduction, iCAM and Ward's histogram adjustment, were selected to investigate their corresponding rendering accuracy. Three more recent HDR rendering algorithms, Biggs' modified iCAM,[0] local eye adaptation,[0] and Meylan's Retinex-based adaptive filter[0] were selected to evaluate recent developments in HDR rendering.

Experiment 3 was then conducted with the following five goals in mind:

1. To present a sound psychophysical experimental framework to evaluate the rendering accuracy of HDR rendering algorithms.

2. To evaluate algorithms for their overall accuracy and rendering accuracy of individual image attributes.
3. To test whether the accuracy of a rendering is correlated with its preference when the original scenes are pleasant.
4. To investigate the extent the experimental results depend on the experimental techniques used for obtaining them.
5. To study the influence of HDR scenes' characteristics on the performance of the selected algorithms.

To fulfill aims 2, 3 and 4, three experiments using two psychophysical techniques, paired-comparison and rating-scale, were developed to evaluate observers' preference and rendering accuracy for the test algorithms.

A paired-comparison experiment was first conducted to evaluate image preference of HDR rendering algorithms, similar to Experiment 1 described above. There were a total of 63 comparisons for the three image scenes and seven algorithms. Without viewing the original scenes, observers were first presented with the task of selecting which of the two simultaneously displayed images they preferred for a given image attribute, such as overall contrast, colorfulness, image sharpness, and overall natural appearance.

Observers were then asked to compare the appearance of rendered images with their corresponding real-world scenes, which were separately set up in an adjoining room to avoid interaction. When viewing 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, to avoid changes in viewing extent. The image attributes investigated in this experiment were: highlight contrast, shadow contrast, highlight colorfulness, shadow colorfulness, and overall contrast. The scene *desk, shown on the right side of Figure 2,* was designed to test luminance tone-mapping performance and only included achromatic objects. Hence, colorfulness was ignored in the accuracy 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 for both the real-world scenes, and the LCD display. The observers were asked to remember the appearance of the physical scenes after viewing for at least 30 seconds (though as long as they wished) and return to the display to make their evaluation after a 30 seconds adaptation period. The observers were allowed to scale one image attribute for all 7 algorithms in one sitting based upon their memory

before they were obligated to look at the original scenes again, though they could go back to view the scenes anytime they felt it necessary. By enforcing repeated viewing 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. The second rating-scale experiment contained 126 rating judgments. The entire rating procedure for one participant took approximately 45 minutes, approximately half the time of the paired-comparison experiment. A graphical rating scale using three verbal anchors with lines between them was designed to evaluate the accuracy of the image attributes, as illustrated in Fig. 11 (a). The left of the line was denoted "too low", the right "too high", and the middle "good match", with a total of seven levels in this equal-interval scale. An observer was asked to choose a rating on this scale to indicate how well the image attributes of the rendered images matched those in the real scenes. For the overall accuracy, a monotonic numerical rating scale with a range of 1 to 7 was used instead, as shown in Fig. 11 (b).

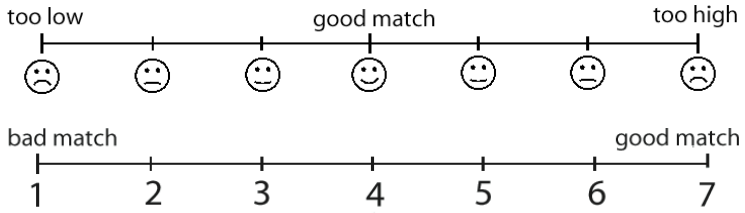

Fig. 11 (a) rating scale for evaluation of image attributes, and (b) rating scale for evaluation of overall image accuracy

6.1 Results for Image Preference Evaluation

The results shown in Fig. 12 represent the overall preference results of this experiment in the order of the averaged scale value, from the worst to the best. Again we can see that on average the bilateral filter is significantly better than other algorithms, and the Retinex-based filter is the worst with regards to preference. The other algorithms have relatively similar preference performance to those described in Experiment 1 above. The results obtained for the performance of HDR rendering algorithms for the three individual scenes is shown in Fig. 13. The results show similar overall patterns between the preference scores of the algorithms for *breakfast* and *desk*, whereas *window* shows a different pattern. For example, iCAM is amongst the best two rendering algorithms for *breakfast* and *desk*, but amongst 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 out that the Modified

iCAM shows the opposite pattern as iCAM, which indicates that it may be possible to combine the two versions of iCAM to get better rendering results for general scenes.

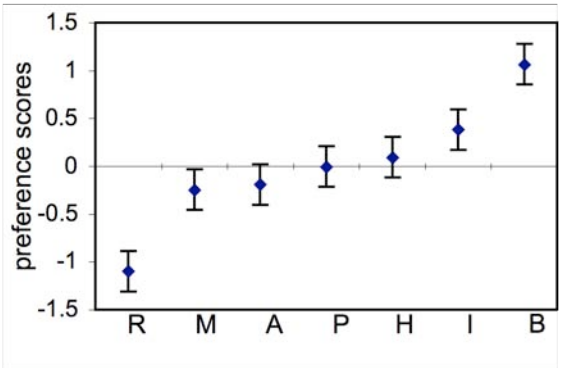

Fig. 12 Overall preference scores of HDR rendering algorithms for 3 scenes (The algorithms are labeled as Retinex-based filters (R), Modified iCAM (M), Local eye adaptation (A), Photographic reproduction (P), Histogram adjustment (H), iCAM (I) and Bilateral filter (B). The same labels are used in this article).

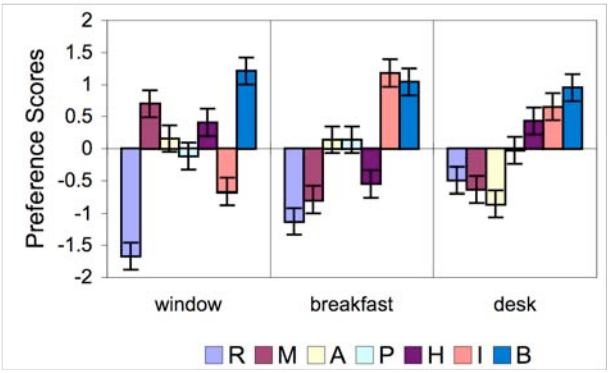

Fig. 13 Preference scores for HDR rendering algorithms by scene

6.2 Results for Paired Comparison Evaluation of Rendering Accuracy

Fig. 14 shows the average overall accuracy scales for the three test scenes. These results indicate how well the algorithms reproduce the appearance of the corresponding physical scenes. The overall accuracy scores show that on average the bilateral filter generated significantly more accurate renderings than other algorithms. The results for individual algorithms are not significantly different across the individual image attributes, showing similar relative performance patterns with the overall accuracy. The bilateral filter produced consistently the most accurate results for all the image attributes, while the Retinex-based filter and modified-iCAM were always the least accurate two algorithms.

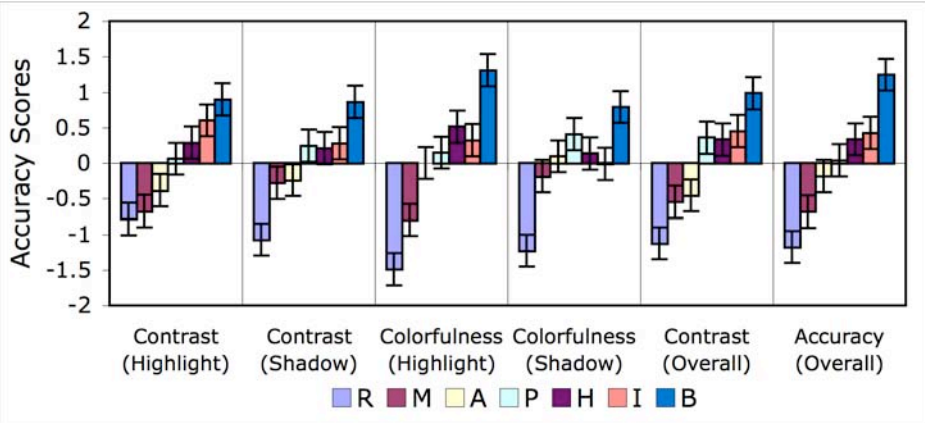

Fig. 14 Overall accuracy scores for HDR rendering algorithms

The results obtained for the accuracy performance of the individual scenes can be illustrated in the same way as the averaged results, shown in Figs. 15-17. As only luminance tone mapping was evaluated for *desk*, colorfulness in the highlights and shadows are ignored in these figures. Generally, the results show a strong correlation between the accuracy scores of algorithms for *breakfast* and *desk*, whereas *window* has different overall patterns. Again iCAM has good performance for *breakfast* and *desk* (among the top two), but significantly worse performance for *window*, showing the same trend as the preference results. The histogram adjustment technique shows the opposite trend, performing much better in *window* than the other two scenes. The local eye adaptation does not perform as well for *desk*, a scene with a very high dynamic range, as the other two scenes, suggesting a linkage between a scene's dynamic range and its rendering performance. The remaining algorithms have more consistency across the different scene contents.

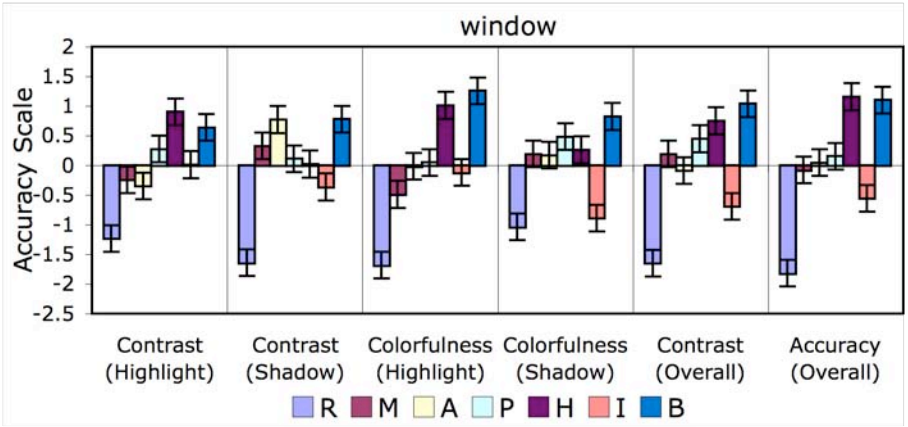

Fig. 15 Accuracy scores for window by image attribute

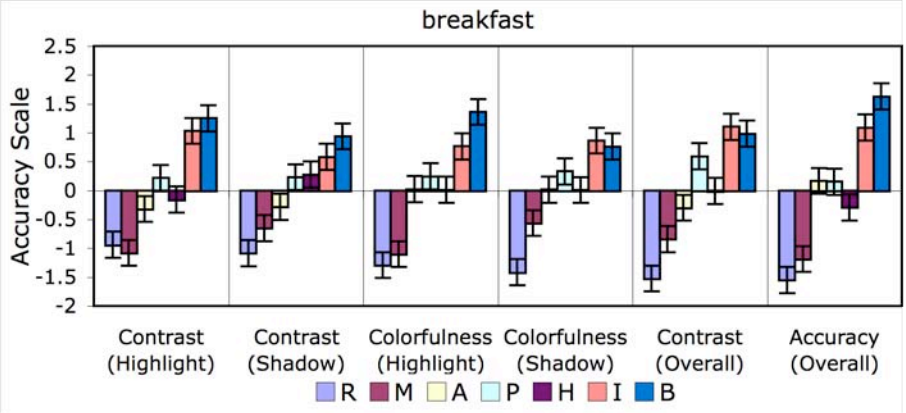

Fig. 16 Accuracy scores for breakfast by image attribute

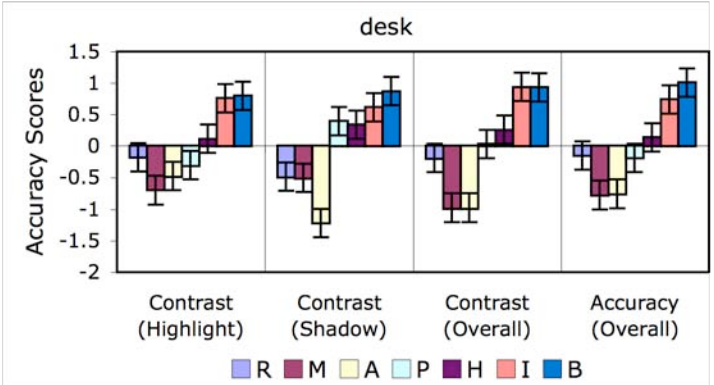

Fig. 17 Accuracy scores for desk by image attribute

6.3 Results for Rating-scale Evaluation of Rendering Accuracy

The rating-scale method can provide an absolute scale as well as the relative performance scales that are generated from the method of paired comparison. The overall rating results (Fig. 18) were calculated by averaging observers' scales from the three individual test scenes, using the rating scales shown in Fig. 11. Numerical weights were assigned to adjectives in the scale shown in Fig. 11 (a), with "4" indicating a good match, with ratings lower or higher than "4" are in the direction of "too low" or "too high" respectively. The dashed lines on the diagrams of the five image attributes denote the scales corresponding to accurate matches with the physical scenes. The closer the individual scales are to these lines, the more accurate the rendered images were considered. As the range for the observers' ratings across the individual image attributes could vary, the 95 percent confidence intervals were also different.

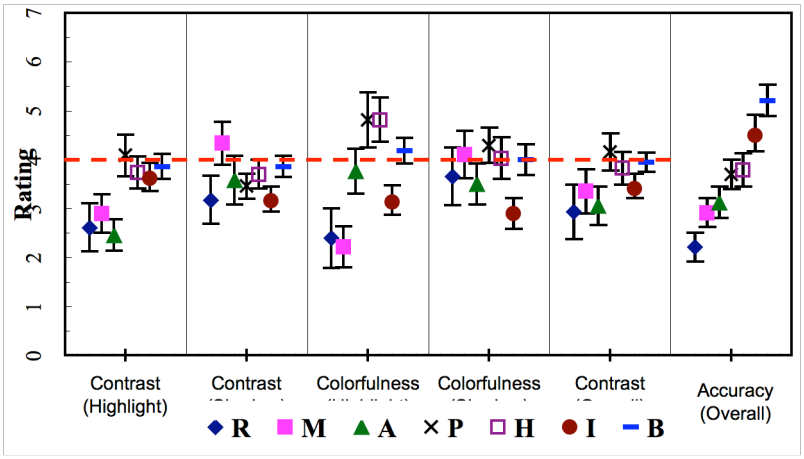

Fig. 18 Overall accuracy scores for HDR rendering algorithms using rating-scale method

The overall accuracy rates were very similar to the results from the paired comparison experiment (Fig. 14), with exactly the same ranks. This suggests that the simpler rating method can produce identical results to the paired comparison. The average accuracy ratings of bilateral filter are very close to the "good match" line, corresponding to the best algorithm for overall accuracy. For most of the test algorithms, their accuracy ratings are below the lines indicating that the rendered images are "too low" along those image attributes, which explains why they tend to perform worse than the bilateral filter. The rates for colorfulness in the highlight and shadow for the photographic reproduction and histogram adjustment were above the accurate match lines, indicating that these algorithms boost too much colorfulness in the rendered images, perhaps resulting in an artificial appearance.

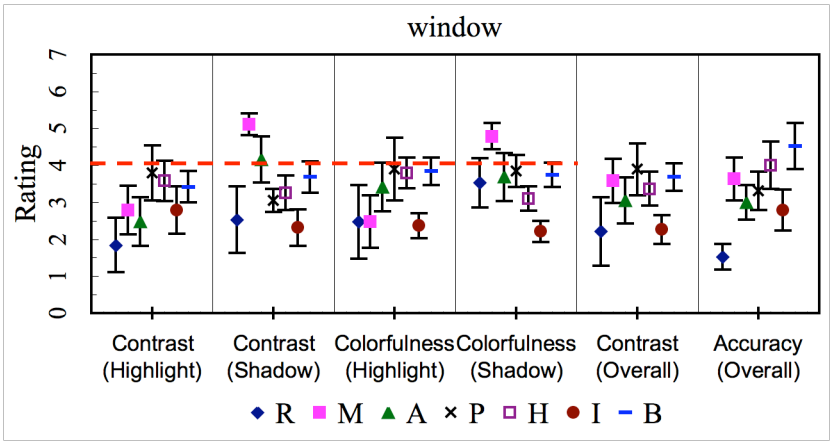

Fig. 19 Accuracy scores for window by image attribute using rating-scale method

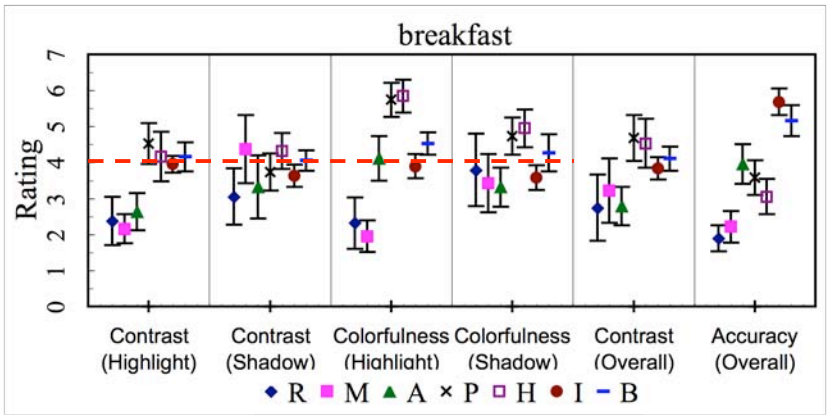

Fig. 20 Accuracy scores for breakfast by image attribute using rating-scale method

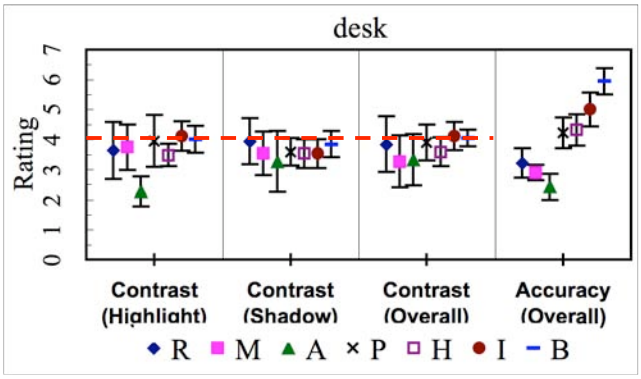

Fig. 21 Accuracy scores for desk by image attribute using rating-scale method

Image dependency can be seen from the results for the individual test scenes, shown in Fig. 19-21. Once again, iCAM performs poorly for *window*, while it performs well for *breakfast* and *desk*, which confirms the findings from the paired comparison accuracy experiment described above. Moreover, the overall low rates for the *window* scene would have significant influence on the average overall performance across all images, which could explain why the average scales for the individual image attributes of iCAM are farther from the ”accurate” lines but the algorithm still ranks second for overall accuracy as shown in Fig. 18. Comparing these results to those shown in Fig. 14 suggests that the analysis for the rating-scale method might not as reliable as that for the paired comparison method. The photographic reproduction’s rates for colorfulness for *breakfast* are significantly higher than the “accurate” line, indicating that this algorithm tends to over-boost the color for highly chromatic images. From the results in Fig. 19-21, it can be seen that the scales for most of the image attribute (with the exception of colorfulness) for most of algorithms fall below the “accurate” line, suggesting the trend of under-

estimating the image attributes for existing HDR rendering algorithms. Generally, the bilateral filter consistently does well for all three scenes, showing the robustness of this particular algorithm for rendering HDR scenes.

6.4 Paired comparison vs. Rating-scaling

To quantify how well the accuracy scales obtained using the two psychophysical methods described in this section agree with each other, the overall accuracy scales from the rating-scale experiment are plotted against the accuracy scales from the paired comparison experiment. A linear regression was performed and illustrates that the scales from these two methods correlate well with each other, with a correlation coefficient of 0.95. Hence, it can be said that the agreement between the results of the two psychophysical methods described above is very good. For most practical applications, either of these methods can be chosen when designing a psychophysical experiment to evaluate HDR rendering algorithms. Considering that the scales generated by the paired comparison method are considered more accurate and precise, we suggest that rating-scale method could be used for initial algorithms selection evaluation, to simplify experiment procedure before a more time consuming paired comparison experiment is utilized to create accurate experimental scales.

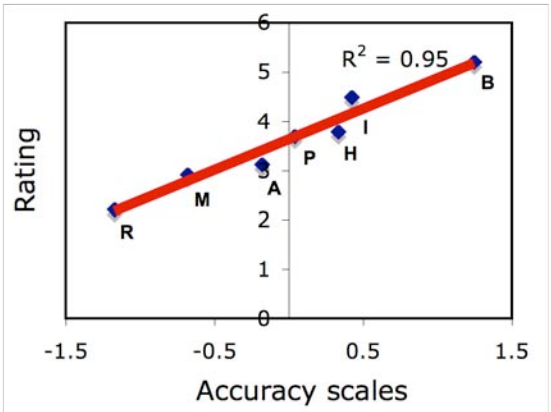

Fig. 22 Overall paired comparison results versus rating-scaling accuracy evaluation results for seven HDR rendering algorithms

6.5 Preference vs. Accuracy

In order to investigate the relationship between the accuracy of a rendering and its overall pleasantness, the preference scores were compared with the overall accuracy scores. Both of these scales were generated using the paired comparison method, using the three test scenes described above. A plot of the HDR rendering accuracy scales versus preference

scales is shown in Fig. 23. The R-square of the linear regression between these two scales is 0.94. This strong positive correlation between these results illustrates the general consistency of accuracy performance and overall preference, and reassures that rendering accuracy is highly correlated with image preference when the original scenes are present.

6.6 Summary of Experiment 3

A psychophysical experimental framework has been developed to evaluate the rendering accuracy of HDR rendering algorithms, along with image preference, using three real-world scenes. Based on this experiment, the bilateral filter consistently performed well for both preference and accuracy. A strong positive correlation between the results from the paired comparison and rating-scale experiments indicates that both methods may be utilized as an evaluation experiment paradigm. The consistency between the preference and accuracy results demonstrates the high correlation between rendering accuracy with its pleasantness.

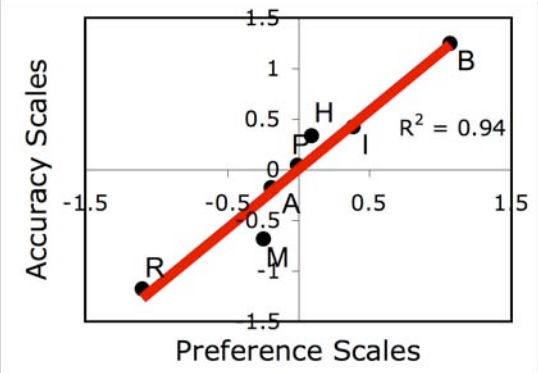

Fig. 23 Overall preference results versus accuracy results for seven HDR rendering algorithms

7. CONCLUSION

A thorough evaluation of image preference and rendering accuracy for many HDR rendering algorithms has been conducted through three psychophysical experiments in this study. The test results for overall preference and accuracy, as well as individual image attributes have illustrated areas for algorithm improvement and should help guide further development of more robust rendering algorithms. The results show that the bilateral filter significantly and consistently outperforms other test algorithms for both preference and accuracy, making it a good candidate for an obligatory or default algorithm that could be included in future algorithm evaluation 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 as described in Section 2 above. This could explain the performance differences for this particular algorithm shown in the experiments by [Ledda et al. 2005]. A strong correlation between the accuracy scales derived from the paired comparison and the rating-scale methods used in Experiment 3 indicates that both of these two methods are suitable for HDR rendering algorithm testing. The rating-scale method is recommend for pilot experiments of test algorithms selection for its simplicity, while the paired comparison method is suitable for further comparison and evaluation as it can provide more accurate and precise results. Finally, the consistency between the preference and accuracy results suggests that HDR rendering algorithms that perform well in terms of accuracy may also be used in a general photographic system to provide pleasant results, and vice-versa.

REFERENCE

Bartleson, C.J. AND Grum, F. 1984. Optical radiation measurements. Volume 5, Visual Measurements, Academic Press, Inc., pg. 467-471.

Biggs, W. 2004. Perceptual accuracy of tone mapping algorithms. MS. Thesis, Dalhousie University.

Braun, G.J. AND Fairchild, M.D. 1999. Image lightness rescaling using Sigmoidal contrast enhancement functions. IS&T/SPIE Electronic Imaging '99, Color Imaging: Device Independent Color, Color Hardcopy, and Graphic Arts IV, pg. 96-105.

Calabria, A.J. AND Fairchild, M.D. 2002. Compare and contrast: perceived contrast of color images, 10th Color Imaging Conference.

Day, E.A., Taplin, L.A. AND Berns, R.S. 2004. Colorimetric characterization of a computer-controlled liquid crystal display. Color Res. Appl. in press.

Debevec, P.E. AND Malik, J. 1997. Recovering high dynamic range radiance maps from photographs. Proc. SIGGRAPH '97, pg. 369-378.

Devlin, K. 2002. A review of tone reproduction techniques. Technical Report CSTR-02-005, Department of Computer Science, University of Bristol.

Draper, N., AND Smith, H. 1981. Applied regression analysis, second edition. John Wiley and Sons, Inc., pg.307-312

Durand, F. AND Dorsey, J. 2002. 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.

Engeldrum, P. 2000. Psychometric scaling: a toolkit for imaging systems development. Imcotek Press, Winchester.

Frankle, J. AND McCann, J. 1983. Method and apparatus for lightness imaging. US Patent #4,384,336.

Funt, B., Ciurea, F., AND McCann, J. 2000, Retinex in Matlab. Proceedings of the IS&T/SID Eighth Color Imaging Conference: Color Science, Systems and Applications, pg. 112-121.

Funt, B., Ciurea, F., AND McCann, J. 2002. Tuning Retinex parameters. Proceedings of the IS&T/SPIE Electronic Imaging Conference (2002).

Ikeda, E. 1998. Image data processing apparatus for processing combined image signals in order to extend dynamic range, U.S. Pattern 5801773.

Johnson, G.M. 2005. Cares and concerns of CIE TC8-08: spatial appearance modeling & HDR imaging. SPIE/IS&T Electronic Imaging Conference, San Jose.

Johnson, G.M. AND Fairchild, M.D. 2003. Rendering HDR images. *IS&T/SID 11th Color Imaging Conference*, Scottsdale, pg. 36-41.

Jones, L.A. AND Condit, H.R., 1941. The brightness of exterior scenes and the computation of correct photographic exposure. *Journal of the Optical Society of America A*, **31** 651-666.

Keelan, B. 2002. Handbook of image quality: characterization and prediction. CRC.

Kuang, J., Johnson, G.M., AND Fairchild, M.D. 2005. Image preference scaling for HDR image rendering. IS&T/SID 13th Color Imaging Conference.

Kuang, J., Liu, C., Johnson, G.M., AND Fairchild, M.D. 2006. Evaluation of HDR image rendering algorithms using real-world scenes. Conference of ICIS

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

Larson, G.W. 1998. LogLuv encoding for full-gamut, high-dynamic range images. Journal of Graphics Tools, vol. 3, 15-31.

Larson, G.W. Rushmeier, H. AND Piatko, C. 1997. A visibility matching tone reproduction operator for high dynamic range scenes, IEEE Transactions on Visualization and Computer Graphics, pg. 291-306.

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

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

McCann, J. 2004. Retinex at 40, Journal of Electronic Imaging, 13(1), pg. 139-145.

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

Montag, E. 2004. Louis Leon Thurstone in Monte Carlo: creating error bars for the method of paired comparison, Proc. of SPIE-IS&T Electronic Imaging, pg. 222-230

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

Naka, K.I. AND Rushton, W.A.H. 1966. S-potential from colour units in the retina of fish, Journal of Physiology, 185:536-555.

Nayar, S.K. AND Mitsunaga, T., 2000. High dynamic range imaging: spatially varying pixel exposures. Proc. IEEE CVPR, Vol. 1, pg. 472-479.

Nishisato, S. 1994. Elements of dual scaling: an introduction to practical data analysis. Lawrence Erlbaum Associates, New Jersey.

Reinhard, E., Stark, M., Shirley, P. AND Ferwerda, J. 2002. Photographic tone reproduction for digital images. In Proceedings of ACM SIGGRAPH 2002, Computer Graphics Proceedings, Annual Conference Proceedings, pg. 267-276.

Reinhard, E., Ward, G., Pattanaik, S., AND Debevec, P. 2006. High dynamic range imaging. Morgan Kaufmann Publisher, pg. 223-323.

Robertson, M.A., Borman, S. AND Stevenson, R.L. 1999, Dynamic range improvement through multiple exposures. IEEE International Conference on Image Processing.

Thurstone, L.L. 1927. A law of comparative judgment, Psychological Review, 34:273-286.

Ward, G. 2005. JPEG-HDR: A backwards-compatible, high dynamic range extension to JPEG. IS&T/SID's 13th Color Imaging Conference, pg. 283-290.