Image Preference Scaling for HDR Image Rendering

Jiangtao Kuang, Garrett M. Johnson, and Mark D. Fairchild, Munsell Color Science Laboratory, Rochester Institute of Technology, Rochester, New York, USA

Abstract

Two psychophysical experiments were performed to evaluate image preference of 6 high-dynamic-range (HDR) image rendering algorithms. The experiments were split into a paired comparison experiment examining overall preference, and a rating scale experiment judging individual preference for 6 image attributes: highlight details, shadow details, overall contrast, sharpness, colorfulness and artifacts. The paired comparison experiment was analyzed using Thurstone's law to generate interval scales. In addition, dual scaling analysis indicates a single perceptual dimension accounting for the variance of overall preference. The overall preference shows high correlations with shadow details, overall contrast, sharpness and colorfulness, which represent the most important factors in observers' preference judgment. Stepwise regression of various image attributes to the overall preference results showed that for many images the preference scales of a single attribute can predict the overall image preference.

Introduction

High-dynamic-range (HDR) imaging has been an active research area in the imaging community in recent years. HDR images are typically images containing a large range of luminance information and are represented by more than 8-bits per channel. Imaging technology has advanced such that the capture and storage of this broad dynamic range is now possible. However, due to the limitation of the luminance range of common desktop display as well as hardcopy output, displaying these images is still a complicated problem. 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.1 For digital photographic applications, many tone-mapping algorithms are designed to produce pleasing or preferred images to observers. Image preference becomes an important benchmark for the success of rendering for a tone-mapping algorithm. Psychophysical evaluations with human observers have been applied for testing algorithms' performance on image preference. Kuang et al.2 used a paired-comparison paradigm to scale the image preference for 8 tone-mapping algorithms using 10 different image scenes. More concerns on HDR rendering algorithms evaluation have been described by Johnson3 recently.

There are many factors that can influence image preference, including tone reproduction, sharpness, colorfulness and the visibility of unnatural artifacts in the images. These factors, called image appearance attributes, can use the same definitions of those in image quality models, referred as "nesses" by Engeldrum.[4] High-Dynamic range tone-mapping can be thought of as an extreme form of gamut mapping. Previous gamut mapping research[5] has shown that the performance of gamut mapping algorithms have strong correlation with image characteristics, and an automatic approach[6] was proposed to select appropriate algorithms based on image analysis. Keelan.[7] proposed a multivariate method for predicting overall image quality from individual image attributes by altering a single attribute and then determining the influence of this attribute on the overall image quality. This begs the question, when observers judge their impression of the merit or excellence in the HDR image rendering results, do they examine all the individual image attributes? If not, what are the most important attributes that determine the overall preference? In this particular experiment, the overall image preference and preference of each individual attribute were evaluated separately. Several psychological scaling techniques were performed in an attempt to reveal what attributes observers are using to judge the HDR rendering preference. The experiments also illustrates the performance of the tone-mapping algorithms performance in specific image attributes, showing areas for their improvement and designing more robust algorithms in the future.

Overall Image Preference

The psychophysical experiment described is a continuation of research first discussed by Kuang et al.[2] The goal of this experiment is to evaluate HDR image rendering algorithms based on overall image preference.

Experimental Design

A paired comparison psychophysical experiments was performed to scale the image preference of 6 HDR rendering algorithms using 12 different pictorial image scenes. Based on the results of the original experiment.[2] two of the least preferred algorithms were eliminated from testing, and two new image scenes with human subjects were added for evaluation in this experiment. These additional images were created to test the rendering algorithms performance on human portraits, a very important category in photography. Six algorithms were selected to represent different tone mapping and spatial processing approaches. Sigmoid transformation (S)[8] and histogram adjustment technique (H)[9] are global operators; Retinex(R),[10] iCAM (I),[11] bilateral filter (B)[12] and photographic reproduction (P)[13] are local operators. The selection of test scenes is an important consideration for this investigation. Twelve scenes from a variety of categories including indoor, outdoor, day, night, natural scenery, portrait and computer rendered, and covering different overall dynamic ranges and mean luminances. The scenes used are shown in Figure 1.

Figure 1.Thumbnails of experimental images

The HDR rendered images were displayed on a colorimetric characterized 23-inch Apple Cinema HD LCD monitor with a maximum luminance of 180 cd/m^2. The experiment was conducted in a darkened room. For the first experiment, pairs of images rendered from different algorithms were displayed simultaneously. Observers were asked to choose the image they preferred. A total of 33 color normal observers participated in the first 10 scenes evaluation, and 19 observers in the last two human portrait scenes evaluation. More details of this experimental design can be found in Kuang's paper.[2]

Experimental Analysis

The paired comparison data were first analyzed using Thurstone's Law of Comparative Judgments, Case V. This analysis results in an interval scale of image preference. Thrustone's law relies on the assumption of a one-dimensional scale. The uni-dimensional preference scale constructed from paired comparison data should avoid intransitive judgments (e.g., A is preferred to B, B to C, and C to A). The interval scale results for 12 scenes are summarized in Table I, as also shown in Figure 2, with the error bars representing the 95% confidence interval. From the results, bilateral filter and photographic reproduction have the best average rendering performance among the algorithms; however, it is clear that there are distinct scene dependencies from the rendering image preference. No single algorithm consistently performs well for all images, indicating that like traditional gamut mapping there may be a strong image dependency.

Table 1: Interval Scales of Image Preference

	S	R	H	I	P	B
belgium	-0.62	-1.28	0.62	0.42	0.91	1.58
bristolb	-0.36	-0.02	-0.14	0.32	0.69	1.13
church	0.28	0.31	0.36	0.16	0.57	0.54
colorcube	-0.81	-1.28	0.87	0.94	1.05	1.62
garage	0.48	-2.39	0.82	1.01	1.17	0.83
lamp_pete	0.09	-0.21	-0.40	0.06	1.18	-0.02
lamp_up	-0.73	1.32	0.49	0.76	0.50	0.68
tahoe1	-0.44	0.06	0.67	-0.16	0.79	1.12
clockbui	0.24	-0.84	0.50	0.61	0.51	0.86
split_cube2	0.55	-0.02	0.36	-0.86	0.41	1.07
ashi01	0.39	0.10	-1.25	-0.21	0.29	0.69
ashi05	-1.00	0.26	-0.84	1.29	-0.46	0.75
Average	-0.39	-0.37	-0.05	-0.01	0.31	0.51

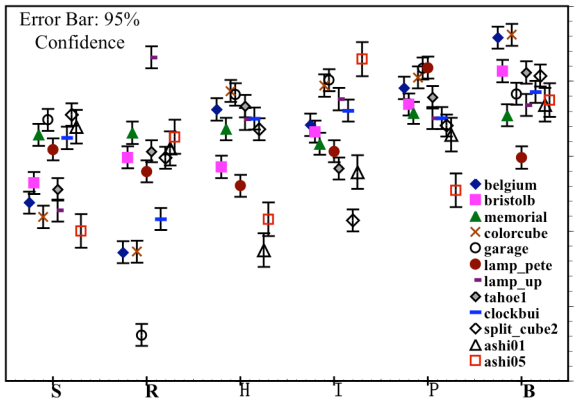

Figure 2. Interval scales of HDR rendering image preference (List the algorithm names along with the letter abbreviations here)

The paired comparison data were also analyzed using dual scaling,[13] a multidimensional technique that can delineate relations among variables, linear or nonlinear, from multivariate categorical data. The dual scaling analysis determines the number of independent dimensions that characterizes observers' preference,

and the percent of the variance that each dimension accounts for in the resulting scale. The results of the dual scaling analysis for all testing scenes are shown in Figure 3. From the percentage of the variance shown in Figure 3, we can see the first dimensions in all scenes are dominant, accounting for over 90 percent of the variance in the preference judgments, and the remaining dimensions are by comparison marginal. This singular dimensionality supports the assumption used when constructed interval scale using Thurston's law.

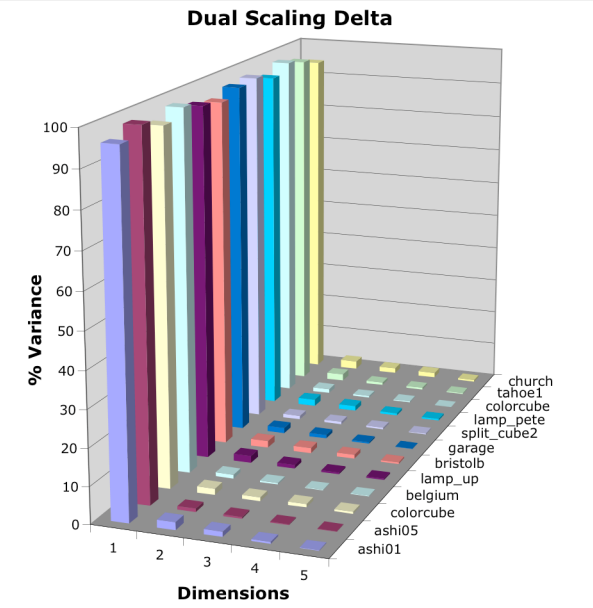

Figure 3. The percentage of variance accounted for each dimension of dual scaling analysis

Image Preference Modeling

While the paired comparison psychophysical experiment and analysis described above show the observers' preference about the HDR rendering images, they do not provide total insight into the actual criteria observers use to scale image preference. This information is important for further research in order to understand what specific areas are most important for image preference of HDR rendering algorithms. By measuring the preference of various image attributes from the rendering results, we can compare the performance of a specific attribute in 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. 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 research are: highlight details, shadow details, overall contrast, sharpness, colorfulness and the appearance of artifacts. Many tone mapping algorithms have a tendency of reproducing the maximum details in highlight and shadow areas, often sacrificing the overall contrast at times. However, observers may base preferences on the overall contrast, as well as sharpness or colorfulness when they judge images. In addition, certain unnatural artifacts in the images, such as halos around light sources, can have also influence the overall preference. The image attributes as defined in the observer instructions are defined below.

Shadow Details (SD) – ability to see information in just the shadow (dark) areas
Highlight Details (HD) – ability to see information in just the highlight (bright) areas
Overall Contrast (OC) – the overall variation of image lightness between the highlight and shadow regions.
Sharpness (S) – the overall amount of detail or clarity in the image
Colorfulness (C) – the overall amount of color present in the image
Artifacts (A) – color or spatial errors resulting from the image processing that impact the image negatively

Experimental Design

In order to evaluate the preference of the rendered images for each image attribute and also determine the overall contribution towards overall preference, a rating experiment was designed using the same rendered images that were used in the previous paired-comparison experiments. When considering methods for scaling preference, the paired-comparison technique is a good candidate due to the nature of the task. Most people are able to easily judge the "better" image from a pair without any training or background knowledge of the experiment, and by comparing each image to all the others this can result in an accurate and robust preference scale. However, as the comparison samples increase, the number of judgments climbs rapidly, and the experiment can become tiresome for the observers. For example, in this research, 6 image attributes were evaluated individually from 6 HDR rendering algorithms across 12 scenes, totaling 1080 pairs for comparison. For future research scaling perceptual accuracy, tone mapped images on the display might need to be compared directly against their corresponding real-world scenes, the huge luminance differences can lead to increased difficulties. Observers might need 30 seconds or longer for light adaptation. This could make the paired comparison paradigm even more overwhelming.

As a pilot experiment for future accuracy evaluation experiments, a rating-scale method was used in this research. The subjects were asked to evaluate their preference for each of the 6 image attributes at a single time, comparing the rendering to the "perfect" image in their mind. A rating scale number from 0 to 10 was used to express the preference. The attribute "artifacts" now was substituted with "lack of artifacts (LA)", which made a rating of 0 always mean the least preferred and 10 mean the most preferred for all attributes. For consistency the other experimental settings and viewing conditions were the same as those in the overall preference evaluation experiment. A total of 19 color normal observers participated. The entire procedure consisting of 72 images (6 algorithms, 12 scenes) took approximately 40 minutes to complete.

Experimental Analysis

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 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.[14] The normalized rating scales along with 95% confidence intervals for each image attribute over the 12 scenes are shown in Figure 4.

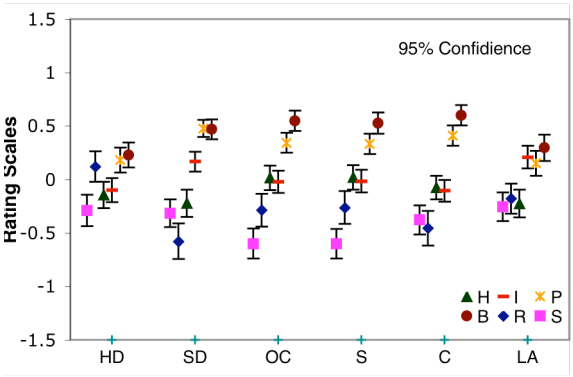

Figure 4. Rating scales of 6 image attributes over 12 scenes

The results show that the bilateral filter and photographic reproduction (represented by B and P in Figure 4) have good rendering preference in all image attributes, with significant higher rating scales than other algorithms in shadow details, overall contrast, sharpness and colorfulness. The strengths and weaknesses of all the algorithms can be ascertained from this data. For instance 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, are 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 rating scales to show the similarity of the tone mapping algorithms in each image attribute. This was performed for the average of all the scenes. The Mahalanobis distances among tone mapping algorithms are visualized in Figure 5 as dendrogram plots of the hierarchical binary cluster trees. The results show that the similarity patterns are very close to the overall preference ranks. Bilateral filter and photographic reproduction have very similar preference in almost all image attributes. iCAM and histogram adjustment also have high similarity except in shadow details and artifacts, where iCAM has better performance as shown in Figure 4.

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 data of 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 strong correlation with 4 attributes: shadow details, overall contrast, sharpness and colorfulness. It seems that for these scenes the highlight details and 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 in image preference, which is also shown in Calabria's experiment.[15]

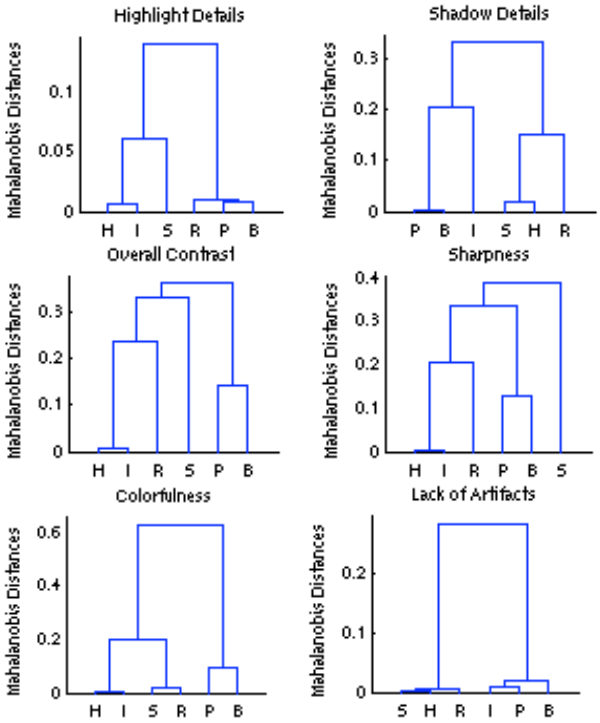

Figure 5. Hierarchical cluster trees of Mahalanobis distances among tone mapping algorithms for each image attribute

Table 2: Correlation Values Among Overall Preference and Image Attributes

	HD	SD	OC	S	C	LA
Preference	0.45	0.75	0.82	0.81	0.77	0.47
HD		0.38	0.55	0.53	0.51	0.61
SD			0.79	0.74	0.82	0.48
OC				0.93	0.91	0.57
S					0.90	0.59
C						0.58

The correlations between the overall preference and the individual image attributes for each scene are summarized in Table III. It is clear that there are distinct scene dependencies and different image attributes correlate better to the overall image preference. For instance, ashi01 and ashi05 are similar scenes with the same human subject, but the individual attributes correlate differently. The perceived artifacts have different impairments to the overall

preference. In ashi01, most people tend to ignore the blurred background and focus on the human subject, while they are very sensitive to the color balance change in the highlight parts in ashi05, as shown in Figure 6. Thus the perceived artifacts are more important to the overall preference judgments for ashi05. It confirms a common image quality phenomenon that serious degradations can dominate minor ones in overall quality.[5]

Table 3: Correlation Values Between Overall Preference and Image Attributes for 12 Scenes

	HD	SD	OC	S	C	LA
belgium	0.05	0.91	0.93	0.90	0.91	0.88
bristolb	0.29	0.73	0.93	0.91	0.97	0.45
church	0.01	0.45	0.81	0.76	0.83	0.43
colorcube	0.93	0.95	0.92	0.89	0.96	0.97
garage	0.04	0.96	0.89	0.92	0.98	0.30
lamp_pete	0.44	0.68	0.70	0.71	0.89	0.87
lamp_up	0.69	0.83	0.99	0.99	0.88	0.83
tahoe1	0.18	0.38	0.99	0.99	0.92	-0.17
clockbui	0.49	0.99	0.92	0.82	0.98	-0.17
split_cube2	0.73	0.88	0.97	0.94	0.96	0.92
ashi01	0.90	0.77	0.88	0.95	0.93	0.30
ashi05	0.85	0.35	0.95	0.97	0.82	0.93

Figure 6. Artifacts in the rendering images (Left image is rendered by sigmoid transform, the right image is rendered by photographic reproduction)

A stepwise regression[16] 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 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 fitting R-square of 0.98. This is illustrated in Figure 7.

While a single attribute of colorfulness might be capable of predicting the preference of average scenes, it is of interest to determine whether that is the case for any individual scene. 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 IV. Other than the church scene, all scenes were fit with one single attribute, although that attribute differs from scene to scene. Compared to the correlation values in Table III, it is obvious that the image attributes needed are the ones with the strongest correlation to the preference for that scene. As the correlations among image attributes are also very strong, such as contrast, sharpness and colorfulness, many of these attributes are able to fit the preference results almost as well. These results seem to agree with the conclusion from the dual scaling analysis, which suggested that the overall image preference scale could be explained in a single dimension.

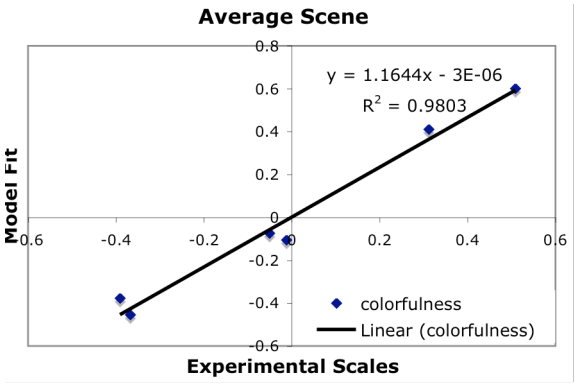

Figure 7. Average preference scale estimation using colorfulness rating scales

Table 4: 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

Conclusion

Two psychophysical experiments were performed to scale image preference of HDR rendering images. A total of 6 tone-mapping algorithms were tested over 12 HDR scenes. The rendering results were displayed on a desktop LCD monitor. The overall preference was first evaluated using a paired comparison experiment. Thurstone's law was used to generate interval scales of preference. Dual scaling analysis was performed on the scales, indicating 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.

The overall image preference was then predicted using preference scales of six image attributes. The scales were generated from a

rating scale experiment. The Bilateral filter and photographic reproduction showed consistently good performance for each attribute, as well as the overall preference. Correlations between preference and each of the image attributes were tested. The result shows that shadow details, overall contrast, sharpness and colorfulness have high correlations with preference. The degradations of artifacts are dependent on scenes and artifact types. While the initial results indicate that highlight details are not as important as other attributes in a HDR image, it is dangerous to draw this conclusion. More HDR scenes with important highlight details, such as human subjects in the bright regions, need to be tested in further experiments. A stepwise regression analysis showed that the rating scale of one image appearance attribute is capable of predicting the overall preference. As the image attributes are highly correlated with each other, more than one attribute is able to predict the preference as well. These modeling results will provide some hints for the design of new tone-mapping algorithms. The image preference scaling techniques presented in this paper may also be valuable for the evaluation of other image processing or color rendering operations.

References

1. K. Devlin, A Review of Tone Reproduction Techniques, Technical Report CSTR-02-005, Department of Computer Science, University of Bristol (2002).
2. J. Kuang, H. Yamaguchi, G. Johnson, M. Fairchild, Testing HDR image rendering algorithms, IS&T/SID 12th Color Imaging conference (2004).
3. G.M. Johnson, Cares and Concerns of CIE TC8-08: Spatial Appearance Modeling & HDR imaging, SPIE/IS&T Electronic Imaging Conference, San Jose, (2005).
4. P.G. Engeldrum, a Framework for Image Quality Models, Jour. Imag. Sci. & Tech. 39:312 (1995).
5. J. Morovic, Y. Wang, Influence of Test Image Choice on Experimental Results, 12th Color Imaging Conference, pg. 143-148, (2003)
6. P. Sun, Z. Zheng, S. Hsin, Selecting Appropriate gamut mapping algorithms based on a combination of image statistics, SPIE Vol 5667, pg. 211-219 (2005)
7. B.W. Keelan, Handbook of Image Quality: characterization and Prediction, Marcel Dekker, New York, NY (2002).
8. G.J. Braun and M.D. Fairchild, 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 (1999).
9. 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).
10. B. Funt, F. Ciurea, and J. McCann, Retinex in Matlab, Proceedings of the IS&T/SID Eighth Color Imaging Conference: Color Science, Systems and Applications, pg. 112-121 (2000).
11. G.M. Johnson and M.D. Fairchild, Rendering HDR images, *IS&T/SID 11th Color Imaging Conference*, Scottsdale, pg. 36-41 (2003).
12. 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).
13. S. Nishisato, Elements of Dual Scaling: An Introduction to Practical Data Analysis, Lawrence Erlbaum Associates, New Jersey (1994)
14. P. Engeldrum, Psychometric Scaling: a Toolkit for Imaging Systems Development, Imcotek Press, Winchester (2000)
15. A.J. Calabria, M.D. Fairchild, compare and Contrast: Perceived Contrast of Color Images, 10th Color Imaging Conference (2002)
16. Draper, N., and H. Smith, Applied Regression Analysis, Second Edition, John Wiley and Sons, Inc., 1981, pg.307-312

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.