# A Groundtruth Database for Testing Objective Metrics for Image Difference

*Elaine Jin and Sharon Field*
*Eastman Kodak Company, Rochester, New York*

## Abstract

An objective metric that can predict the perceived difference of a processed image from an original would be very useful in optimizing image-processing algorithms (e.g., JPEG compression). Ideally, such a metric must be validated against human data before it is used in real imaging applications. There have been numerous efforts to develop such metrics and a few attempts to validate these metrics, but none of the previous validation work intended to cover a wide range of image differences. As a result, the metric developed may not be useful for general purposes. In the present study, we developed a comprehensive database of images and psychophysical data and used it as a tool to test various models of image difference. Several image manipulations were performed to introduce image differences of different types (such as density shift, JPEG compression, and image blur). A psychophysical study was performed to obtain subjective evaluations of image differences from ten observers. As a first step, we tested CIE 2000 and S-CIELAB models against the database. Our results indicate that simple models, such as the CIE 2000 color difference model, can predict density shift and image blur well, but models that incorporate spatial components (such as S-CIELAB) are better in predicting the results of JPEG compression.

## Introduction

An objective metric that can predict the perceived difference of a processed image from an original would be very useful in developing image-processing algorithms. For example, it can be used to predict the effect of an image-processing algorithm (e.g., JPEG compression). Ideally, such a metric must be validated against human data before it is used in real imaging applications. However, because of the tremendous amount of work needed in collecting human data and validating a model, the state-of-the-art in validation of image-difference models appears to lag the development of image-difference models. Numerous image-difference models have been developed in recent years.[1-6] In contrast, only a few studies have been published that applied the psychophysical approach for model validation.[7-10]

Model validation can take different approaches. One approach is to measure the responses of human observers to image differences at the pixel level.[7] Because the output of an image-difference model is a pixel-by-pixel map, the model output can be directly compared with the measurement from the human observers. The drawbacks of this approach are that the process of acquiring data is extremely tedious, and a model validated this way cannot be applied directly in a real imaging situation. For example, such a model cannot answer the question: “If two image pairs have different image-difference maps, are they perceived to be different?” A second approach would be to have human observers judge the overall image difference and use the data to validate a model.[8-9] To apply this approach, integration rules must be developed to link the pixel map output by the model to the single number given by the human observers. Simple statistical quantities, such as mean, median, and standard deviation, have typically been used to perform this integration. More complicated methods include using Minkowski and probability summation. The integration rules continue to be a subject of study because there is not yet any robust psychological evidence for how human observers integrate local differences into one global impression. A third approach[10] is a compromise of the previous two. In this approach, an image is divided into small blocks, and a human observer rates each block for image difference. A statistical predictor is derived from the pixel-by-pixel map for each individual block, and linked with the human rating data. This approach is almost as demanding as the first one, in terms of human data collection. Unfortunately, it also has the same disadvantage of the second approach because the local differences must be integrated to allow the model to be assessed.

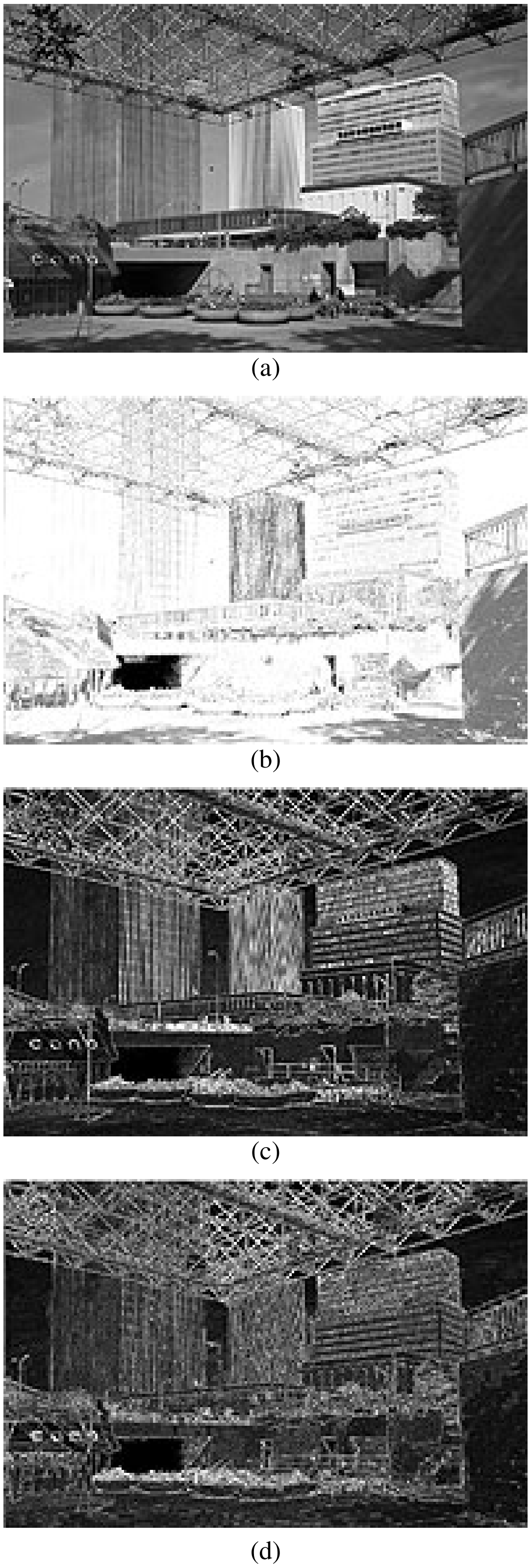
An ideal image-difference model would be able to predict image differences for any application. To validate such a model, the database for human responses must contain a variety of image differences. Images can be different along many dimensions such as in color or in the spatial domain. Image differences can also differ based on how the energy is distributed across an image. For example, some differences are local (e.g., image blur), while others are global in nature (e.g., density shift). Some local differences are edge related (e.g., image sharpening), while others are object related (e.g., saturation boost). Different conditions, as mentioned above, can be used to test different aspects of an image-difference model.

In the present study, we developed a comprehensive database of images and psychophysical data and used it as a tool to test various models of image differences. Several image manipulations were performed to introduce image differences of different types (such as density shift, image blur, and JPEG compression). The manipulations were performed in a parametric way so that the psychophysical data would allow an estimation of both within and between scene effects. Magnitude estimation was performed to obtain subjective evaluations of image differences from ten observers. As a first step, we tested CIE 2000 and S-CIELAB models against the database. Pixel-by-pixel image-difference maps were generated by each of the two models. Next, numerous statistical predictors were derived from the image-difference maps. The relationship between the model predictors and the subjective ratings were evaluated. It has been shown that simple models, such as the CIE 2000 color difference model, can predict density shift and image blur well, but models that incorporate spatial components (such as S-CIELAB) are better in predicting the results of JPEG compression.

## Methods

Eight test scenes were used as originals in the present study. They represented a variety of subject matter, such as people, animal, and landscape. The images also had different spatial frequency content, colorfulness, and were with or without the presence of memory colors. Three manipulations were applied to the original images. For density shift, an equal amount of code value shift was applied to R, G, and B channels of all pixels in the image in the ERIMM color space.[11] The amount of the code value shift was used as a control parameter. For image blur, unsharp masking was applied to the original images in the ERIMM color space. The control parameter for image blur was the gain of the unsharp masking algorithm. JPEG-DCT baseline compression scheme was applied for JPEG compression in sRGB space. A Q-table was designed for the viewing conditions used in the present study. The scale factor of JPEG-DCT was used as a control parameter. For each scene, nine parameter levels were selected. The visible image differences between the original and the processed images ranged from small to large. The test stimuli were AgX color images printed on a calibrated 4-inch CRT printer. The CRT printer was characterized, and had a resolution of 250 DPI. All images were converted from their original color spaces to the printer code value space prior to printing.

The three manipulations resulted in different types of image differences. Figure 1 shows the code value differences between the original and the processed images for one of the scenes. In the difference images, the bright areas have bigger differences compared to the dark areas. Images with density shift manipulation had global and color-related differences (Fig. 1b). Images with blur manipulation had local and edge-related differences (Fig. 1c). Images with JPEG compression had local and edge-related differences when the compression level was low, and had global and blocking type of differences when the compression level was high (Fig. 1d).


(a)

(b)

(c)

(d)

*Figure 1. One original image (a) and image differences introduced by (b) density shift, (c) image blur, and (d) JPEG compression.*

The experiment was conducted in a gray room. The walls were painted with a neutral color (Munsell color N/5). The illumination was CIE Standard Daylight D5000. A headrest restrained the forehead of the observers, and the viewing distance was fixed at 16 inches. The test stimuli were presented against a large gray wood panel (21" x 23"). There were two rails on the panel. On the top rail, there were three images. The one in the center was always the original image. One image on the side was also the original image, and the other was a processed image. On the bottom rail, there was a reference print that had two neutral patches against a neutral background. The background had an L* of 50. The two neutral patches served as a reference pair for image difference and had a luminance difference of 10 ΔE Units.

In one session, the observers would see stimuli with only one manipulation. They were shown examples of the type of image difference that would occur in the session but were not given any verbal description of the image difference. In each trial, the observers would see a triplet of the test images and were asked to indicate which image on the side was different from the center image, and to rate the image difference in relation to the reference pair. The reference pair was given a difference rating of 100. Another trial would start after the experimenter recorded the response. There were 72 image pairs in each session (8 scenes x 9 manipulation levels). The order of presenting the stimuli was randomized. Difference manipulations were presented in different sessions. Ten observers participated in the study. They all had normal or corrected-to-normal visual acuity and normal color vision.

Two results came out of the study. The 2 AFC constant stimuli method gave results that could be used to derive a detection threshold. The magnitude-estimation method gave a rating result in relation to the magnitude of image difference shown for the particular image pair. In this paper, we will only discuss the rating results. A data-cleaning step was performed on the rating data based on results of a “null trial,” i.e., the response of an observer when all three images are the same. After the data cleaning, the arithmetic mean of responses from all observers was calculated. This was used as the overall image-difference rating for an image pair.

For model testing purposes, the test images were converted from the printer code value space back to the PCS space via an inverse ICC profile and, finally, to the CIE XYZ space. The XYZ images were taken as the input to the image-difference models. Two models were tested in the present study: the CIE 2000 color difference model and the Spatial-CIELAB model (S-CIELAB).[3] The CIE 2000 color difference model is the state-of-the-art in predicting color differences for large, uniform color patches. It is yet to be proven that this model can be used to predict color differences for digital images. The original S-CIELAB codes were modified so that the CIE 2000 color difference model was used as the back end for this model. Compared to the CIE 2000 color difference model, S-CIELAB model is one step closer to simulating the human visual system in that it incorporates contrast sensitivity functions for the three opponent color channels.

## Results

### Rating Results

As mentioned in the Methods section, the manipulations were performed in a parametric manner by way of changing the control parameter. Figure 2 (a, b, and c) shows the rating results in each of the three manipulations. Each line in the figures represents a single scene with varying parameter level. As can been seen from the figures, the rating results increased monotonically with the increase in parameter level for almost all scenes and all manipulations. The mean ratings for difference scenes, however, can be very different for a given parameter level, indicating scene dependency in the manipulations. Scene dependency is the greatest for JPEG compression, and is the smallest for image blur. A good image-difference model should be able to predict perceived image difference across scenes and, hence, largely reduce scene dependency.

Figure 3 shows the standard error of the mean (SEM) as a function of the mean rating for each of the manipulations. There is considerable amount of variation between observers in their responses to the image differences. In general the SEM increases with the increase of the mean rating. JPEG compression seems to have smaller SEMs compared to the other two manipulations.

### Model Prediction by CIE 2000 Color Difference Model

The CIE 2000 color-difference model was applied to each image pair to generate a pixel-by-pixel image-difference map. The statistical means were extracted from the difference map as model predictors including regular means, such as minimum, maximum, mean, standard deviation, and the 25, 50, 75, and 95 percentiles. The mean, median, and standard deviation were also calculated for the non-zero pixels in the image-difference map. The correlation coefficients between each of the statistical predictors and the human rating results were calculated for all 72 image pairs in each manipulation. These correlation coefficients are referred to as the correlation coefficients for the pooled data. The statistical predictors that showed high correlations were identified, as shown in Table 1.

The far-right column of Table 1 shows the highest correlation between the rating responses of any two observers in the psychophysical study for that particular manipulation. This correlation sets an upper limit for the model fitting to the experimental data. If the model correlation were higher than the upper limit, the model would be fitting noise instead of meaningful variation. The results show that the CIE 2000 predictions are performing well for density shift and image blur. The correlation between model predictors and the mean ratings are equal or beyond the upper limit for the two manipulations. The prediction for JPEG compression, however, is very poor, i.e., much lower than the upper limit. Another interesting finding is that different statistical predictors provided the best fit to the human data for different manipulations. For example, if the mean difference were used to predict the image blur data rather than the maximum, the correlation would be reduced from 0.928 to 0.774. Conversely, if the maximum difference were used to predict the density results, the correlation would also be much lower than if the mean value were used (0.646 vs. 0.954).

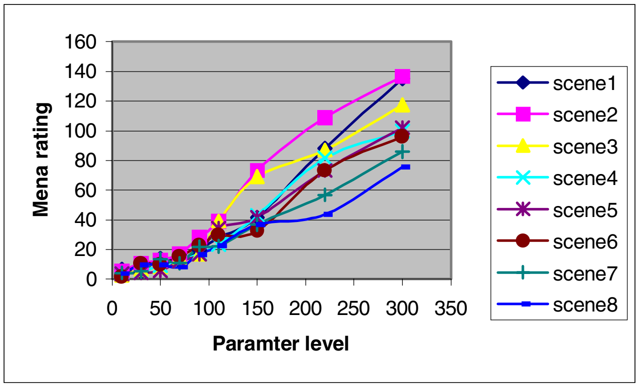


*Figure 2(a) Density shift rating results.*

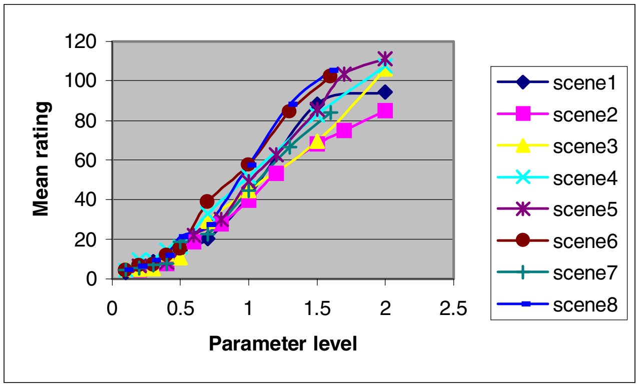


*Figure 2(b). Image blur rating results.*

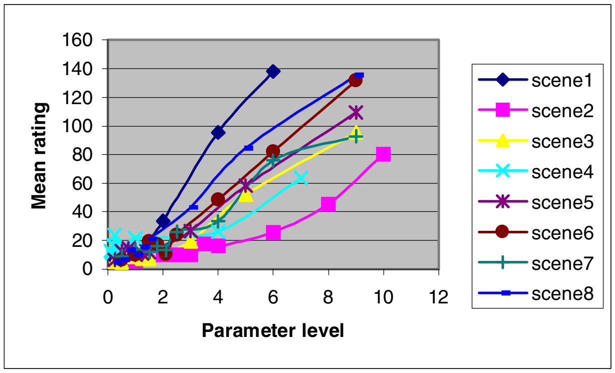


*Figure 2(c). JPEG compression rating results.*

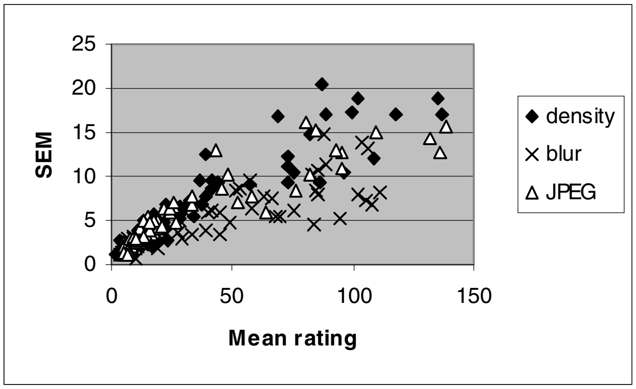


*Figure 3. Standard error of the mean (SEM) as a function of the mean rating for all three manipulations.*

**Table 1. The Correlation Between CIE 2000 Model Predictors and Human Ratings for the Pooled Data**

| Manipulation | Mean | Maximum | NZSTD | Upper Limit |
|---|---|---|---|---|
| Density shift | 0.954 | 0.646 | 0.852 | 0.907 |
| Image blur | 0.774 | 0.928 | 0.858 | 0.926 |
| JPEG compression | 0.311 | 0.395 | 0.368 | 0.896 |

**Table 2. The Correlation Between CIE 2000 Model Predictors and the Human Ratings for Individual Scenes**

| Scene | Densit y shift | Image blur | JPEG compression |
|---|---|---|---|
| | Mean | Maximum | Maximum |
| 1 | 0.970 | 0.917 | 0.937 |
| 2 | 0.983 | 0.996 | 0.792 |
| 3 | 0.980 | 0.929 | 0.850 |
| 4 | 0.966 | 0.984 | 0.860 |
| 5 | 0.988 | 0.982 | 0.684 |
| 6 | 0.981 | 0.994 | 0.869 |
| 7 | 0.978 | 0.990 | 0.955 |
| 8 | 0.977 | 0.991 | 0.935 |
| Average | 0.978 | 0.973 | 0.860 |

Table 2 shows how well the identified CIE 2000 model predictors could be used to linearly fit the rating results for individual scenes. The correlation coefficient between the model predictor and the human rating data was calculated for each individual scene and individual manipulation. For the density shift, the model predictor (mean difference) could linearly fit the rating results of all scenes well (correlation coefficient >0.96). For image blur, the model predictor (maximum difference) predicted two scenes poorer compared to the others (scenes 1 and 3). In general, a linear model could describe the rating data for individual scenes well. For JPEG compression, the model predictor (maximum difference) could linearly predict only three of the scenes fairly well (scenes 1, 7, and 8). This suggests that the model predictor from CIE 2000 is not a good predictor for the JPEG compression.

**Model Prediction by the S-CIELAB Model**

The S-CIELAB model was applied to each image pair to generate a pixel-by-pixel image-difference map. The same statistical predictors as used above were calculated from the image-difference map. The correlation coefficients between each of the statistical predictors and the human rating results were calculated for all 72 image pairs in each manipulation. The statistical predictors with high correlation coefficients were identified for each of the manipulations, as shown in Table 3.

The results in Table 3 show that density shift could be predicted equally well by the S-CIELAB model using the same predictor (mean difference). For image blur, the S-CIELAB prediction (maximum difference) was slightly worse than the prediction by the CIELAB model. For JPEG compression, however, the S-CIELAB model provided a much better prediction than did the CIE 2000 model (even though the correlation is still lower than the upper limit). The non-zero standard deviation provided the best prediction for JPEG compression. The difference numbers in a row show that if a sub-optimal model predictor were used in predicting ratings results, the correlation would be lower.

**Table 3. The Correlation Between S-CIELAB Model Predictors and the Human Ratings for the Pooled Data**

| Manipulation | Mean | Maximum | NZSTD | Upper Limit |
|---|---|---|---|---|
| Density shift | 0.951 | 0.771 | 0.821 | 0.907 |
| Image blur | 0.697 | 0.830 | 0.740 | 0.926 |
| JPEG compression | 0.739 | 0.571 | 0.763 | 0.896 |

Table 4 shows how individual scenes could be described in a linear fashion by the identified S-CIELAB model predictors. The correlation coefficient between the model predictor and the human rating was calculated for each individual scene. For density shift the model predictor (mean difference) could linearly fit the rating results of all scenes well (correlation coefficient > 0.97). For image blur the model predictor (maximum difference) showed poorer linearity in two scenes (scenes 1 and 4) compared to the others. For JPEG compression the model predictor (NZSTD) could linearly predict only two of the scenes well (scenes 1 & 8).

**Table 4. The Correlation Between CIE 2000 Model Predictors and the Human Ratings for Individual Scenes**

| Scene | Density shift | Image blur | JPEG compression |
|---|---|---|---|
| | Mean | Maximum | NZSTD |
| 1 | 0.970 | 0.710 | 0.905 |
| 2 | 0.987 | 0.982 | 0.788 |
| 3 | 0.982 | 0.925 | 0.780 |
| 4 | 0.975 | 0.844 | 0.898 |
| 5 | 0.990 | 0.945 | 0.688 |
| 6 | 0.983 | 0.953 | 0.673 |
| 7 | 0.984 | 0.983 | 0.838 |
| 8 | 0.978 | 0.931 | 0.960 |
| Average | 0.981 | 0.909 | 0.816 |

As mentioned before, the experimental design in the present study would allow an estimation of the model performance for both within and between scene effects. Table 5 shows the relationship between the average model prediction for individual scenes (the middle column) and for the pooled data (the second column to the right). If there were no scene dependency, i.e., a single model predictor would predict the rating results across scenes, the average correlation for individual scenes should be very close to the correlation for the pooled data. The results showed that the average correlation for individual scenes is always higher than that for the pooled data. The ratio of the two, as shown in the far-right column, can be used as a measure of scene dependency. The higher the ratio, the smaller the scene dependency.

**Table 5. Measure of Scene Dependency**

| Model | Manipulation | Average correlation (A) | Pooled correlation (B) | Measure of scene dependency (B/A) |
|---|---|---|---|---|
| CIE 2000 | Density shift | 0.978 | 0.954 | 0.975 |
| | Image blur | 0.973 | 0.928 | 0.954 |
| | JPEG compression | 0.860 | 0.395 | 0.459 |
| S-CIELAB | Density shift | 0.981 | 0.951 | 0.969 |
| | Image blur | 0.909 | 0.830 | 0.913 |
| | JPEG compression | 0.816 | 0.763 | 0.935 |

For the density shift, the two models gave very similar results for both the individual scenes and the pooled data. As a result, the measures of scene dependency also were similar. For image blur, the CIE 2000 model fit both individual scenes and the pooled results better than the S-CIELAB model. The scene dependency was also slightly smaller for the CIE 2000 model. For JPEG compression, the CIE 2000 model performed better than the S-CIELAB in predicting individual scenes but did poorly for the pooled data. Scene dependency is much greater for CIE 2000 than for the S-CIELAB model. This suggests that S-CIELAB predicted fairly well for the pooled data, mainly because of the reduction in scene dependency.

## Conclusions

The rating results were obtained from ten observers for three image manipulations (density shift, image blur, and JPEG compression). The results showed monotonic increase with increase in the level of the control parameters. The variation among responses of observers, as measured by the standard error of the mean, increased with the increase in mean rating.

Both the CIE 2000 color-difference model and the S-CIELAB model were applied to predict the difference between the original and the processed images. For both models, the arithmetic mean of the pixel-by-pixel image differences was identified as the predictor for density shift, and the maximum image difference was identified as the predictor for image blur. For JPEG compression, none of the predictors out of the CIE 2000 model would predict the rating results well, and the non-zero standard deviation was identified as the best predictor for the S-CIELAB model. The improvement of the S-CIELAB prediction over the CIE 2000 prediction was mainly a result of the reduction in scene dependency.

In the present study, we developed a database of images and psychophysical data that covered a variety of image differences. We presented one approach of using the database to test two image-difference models. We demonstrated that there are a significant amount of errors in model prediction for some of the manipulations. We also showed that there are varying amounts of scene dependency in the model prediction. This suggests that none of the models tested can serve as a general-purpose, image-difference model, and there is a need to develop a better model for image difference.

## References

1. S. Daly, *Digital Images and Human Vision*, MIT Press, Cambridge, MA, 1993, pg. 179.
2. J. Lubin, *Vision Models for Target Detection and Recognition*, World Scientific, 1995, pg. 245.
3. X. Zhang and B. Wandell, A Spatial Extension of CIELAB for Digital Color Image Reproduction, *Proc. SID*, pg. 731 (1996).
4. E. Jin, X. Feng. and J. Newell, The Development of a Color Visual Difference Model (CVDM), *Proc. PICS*, pg. 154 (1998).
5. M. Lian, Image Evaluation Using a Color Visual Difference Predictor (CVDP), Human Vision and Electronic Imaging VI, *Proc. SPIE*, pg. 175 (2001).
6. M. Fairchild and G. Johnson, Meet iCAM: A Next-Generation Color Appearance Model, Tenth Color Imaging Conference, *Proc. IS&T/SID*, pg. 33 (2002).
7. X. Zhang, E. Setiawan, and B. Wandell, Image Distortion Map, Fifth Color Image Conference, *Proc. IS&T/SID*, pg. 120 (1997).
8. S. Bouzit and L. MacDonald, Colour Difference Metrics and Image Sharpness, Eighth Color Image Conference, *Proc. IS&T/SID*, pg. 262 (2000).
9. G. Johnson and M. Fairchild, Sharpness Rules, Eighth Color Image Conference, *Proc. IS&T/SID*, pg. 24 (2000).
10. C. Taylor, Z. Pizlo, and J. Allebach, Perceptually Relevant Image Fidelity, Human Vision and Electronic Imaging III, *Proc. SPIE*, pg. 110 (1998).
11. K. E. Spaulding, G. J. Woolfe, and E. J. Giorgianni, Reference Input/Output Medium Metric RGB Color Encodings (RIMM/ROMM RGB), *Proc. PICS*, pg. 155 (2000).

## Biography

**Elaine Jin** received her Ph.D. degree in Psychology from the University of Chicago in 1998. After graduation, she worked as an imaging scientist for Polaroid Corporation and as an image psychologist for Eastman Kodak Company. Her work has primarily focused on vision modeling, color preference, and image quality issues.

**Sharon Field** received her BA in Psychology with Honors from the University of Rochester in 1994. She currently works as a Research Technician in the Corporate Design and Usability lab at Eastman Kodak Company. Her interests are in human computer interaction and image quality evaluations.