Perceptual preference for noise and color saturation tradeoff in digital camera images

Xuemei Zhang, Richard L. Baer; Agilent Technologies Laboratories; Palo Alto, CA, USA

Abstract

An experiment was done to explore the noise and color saturation trade-offs in color processing of digital camera images.Usually pixel noise during capture is amplified when the color saturation is dialed up in the image processing. The data collected in the present study allow us to understand how perceptual preference changes along the noise and color saturation trade-off continuum. This makes it possible to pick the most preferred point on the noise-saturation trade-off line for individual cameras.

Introduction

Engineers and scientists working in the color processing of digitally captured images are well aware of the trade-off between noise amplification and color saturation. Saturation enhancement can result in higher visual noise in the image. On the other hand, image noise can be reduced by minimizing the negative off-diagonal terms of the color correction matrix, at the cost of desaturated colors.

Noise and color saturation preference is highly subjective. Some people may like bright colors and tolerate graininess, some may like smooth pictures and tolerate muted colors. Given the noise properties of a particular camera system and the light level during capture, we are interested in finding an optimal level of color saturation in color correction, so that the pixel noise is not amplified to a level that is unpleasant to the observer. The experiments reported here were designed to give us a relative preference map in the noise and color saturation space, so that we can pick out preferred noise-saturation trade-off points at different SNR levels for any sensor.

Experiment design

Due to the high-variance nature of preference experiments, we chose to use the relatively stable paired-comparison method for our study, and ran the experiments using a large number of subjects. To reach a large number of willing participants, we conducted the experiment on Agilent Technologies internal websites. However, it is difficult to control viewing condition and display calibration when subjects could be sitting anywhere using any computer within the company. To test sensitivity of the results to display calibration and ambient lighting, a pilot experiment was done first with a small number of subjects both in a controlled imaging laboratory and on desk computers. The pilot experiment confirmed that the preference result was not sensitive to changes in ambient lighting and typical display variation.

Image preparation

For test images shown to subjects, we used 12 raw images captured on a few different cameras, including a consumer digital camera made by Hewlett Packard, a consumer digital camera made by Kodak, and CMOS image sensors made by Agilent Technologies. The images were all captured under high ambient light level, and thus were low noise to begin with. They were further downsampled through binning to smaller sizes, further reducing noise. These images were assumed to be noise-free. The contents of the images included landscape, people, animal, and still objects, and included both indoor and outdoor scenes.

Spatially independent pseudo-random additive gaussian noises were added to the raw images to generate 5 different levels of noise per original image. The noise standard deviation varied among different images and different color saturation levels, with the constraint that the final processed image had average S-CIELAB ΔE_{94} [1] values of 0, 1.2, 2.4, 3.6, and 4.8. The use of S-CIELAB ΔE_{94} values instead of standard deviation or SNR for noise specification is to ensure the generalization of experimental data to other sensors and image pipelines which likely will employ different demosaic and denoising algorithms.

Aside from the noise levels, each image was also rendered at 5 different color saturation levels through adjustment of the color correction matrix. The color matrices were generated from spectral sensitivity measurements of each camera using a Bayesian method [2]. For each camera and scene illuminant, a full saturation color matrix was generated assuming RLab adapation [3] at $1000 cd/m^2$ scene light level. Then a monochrome matrix, which gave luminance (Y) estimates, were generated using the same Bayesian method. Images of different saturation levels were generated by color correcting with a weighted combination of the full color matrix and the monochrome matrix. Figure 1 shows the relationship between color matrix saturation (expressed as percentage of full color matrix used) and a perceptual saturation measure (CIECAM97s [4] saturation value of the Macbeth red patch). Saturation levels 0, 0.25, 0.5, 0.75, and 1.0 were used for the pilot experiment, and levels 0.1, 0.25, 0.5, 0.75, and 1.0 were used for the web-based full experiment.

After noise injection, the images were processed with a basic camera image pipeline, with demosaicking, white balance and color correction (at 5 different color saturation levels), and display gamma correction. Steps that would affect noise and color saturation properties, such as denoising, non-linear tone mapping, and highlight desaturation were not used, to ensure the effect of color correction on noise amplification is reflected in the final image. For each near noise-free raw image, a total of 25 versions were rendered at 5 noise levels and 5 color saturation levels.

Design and procedures

Subjective preference data were obtained using a paired comparison method. For each trial, subjects were shown two different renderings of the same original image, and asked to select

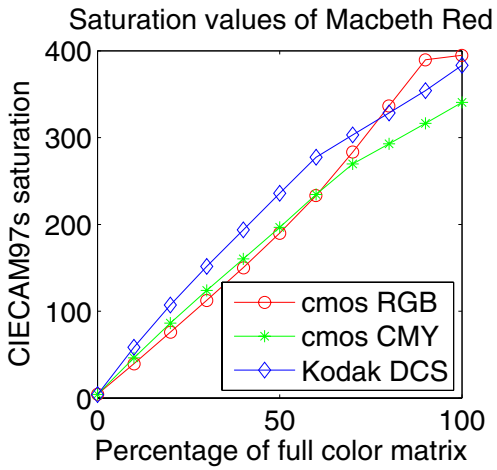

Figure 1. *Perceptual color saturation level as a function of percentage of full color matrix used in processing an image. For 3 sensors with very different spectral properties, the perceptual saturation levels correspond well with the percentage of full color matrix used.*

the version they prefer (the version they would choose to keep in a digital album).

Due to the large number (25) of renderings per image, not all possible pairings were presented to the subjects to avoid tediousness. Pairings that would have an obvious preference (e.g. same color saturation, higher vs lower noise, or same noise level, higher vs lower color saturation, or high saturation low noise vs high noise low saturation) were not included in the comparisons. After eliminating the "obvious" pairings, 25 paired versions were left per image. This is still a lot of paired comparisons to do. To reduce tediousness for subjects in the web experiments, we further reduced the number of pairings per image to 13, and distributed the complete list of pairings across different subjects. For the web experiment, each subject completed 156 paired comparisons, which included 12 different images, each with 13 pairs of comparisons. For the pilot experiments (which had more motivated subjects), each subject completed 300 paired comparisons, including the same 12 images, each with 25 pairs of comparisons.

The main experiment was conducted using a web interface. A JAVA paired-comparison program was written to control the presentation of images and recording of data over the web. To participate in the experiment, subjects went to an internally advertised web site. They first read a brief introduction and instructions to the experiment, which asked them to maximize the browser window, and make sure the display was set at 24 or 32 bit depth. Afterwards, they completed 156 paired comparisons by comparing two side-by-side images and clicking on a button below the image they prefer. After the experiment, they were asked to estimate the gamma value of their displays using a visual tool [5]. The paired comparison data and the estimated gamma values were both recorded.

For part of the pilot experiment, the paired comparisons were done on a calibrated computer in a controlled lab environment, with the experiment program run in Matlab. The set up was otherwise very similar to the web experiment.

Data analysis

The experimental design of the present study required paired comparisons among 25 different renderings (300 possible pairings) per image. To keep the length of the study tolerable for subjects, at most one trial per possible pairing was done by each subject. Existing methods to convert binary data to interval data typically require non-unanimous responses per comparison to give reliable scale estimates [6, 7]. To use these scaling techniques on our data, we could lump together different subjects' responses for the preference scale estimates, and then use a resampling procedure to get an estimate of between-subject variation. However, one of the things we wanted to check was whether there were qualitative differences among different subjects' results (e.g. whether there are sub-groups of subjects who just looked at image noise, and others who only looked at color when doing the comparisons), for which preference scale estimates for each individual subject would be useful. Therefore, we developed a method to convert individual subject's paired comparison responses into an interval scale which works well even when the number of trials per comparison is low, and when not all possible pairings were tested for a particular set of samples.

The details of our scaling method is outside the scope of this presentation, and will be the subject of another paper. Very briefly, we used a Bayesian method, and specified prior distributions on the relative perceptual distance between the samples. Without any data we assume all samples are indistinguishable from each other, but it is possible for the distances to take on any value according to a normal distribution centered at zero. Paired comparison data, even when incomplete, can be used to derive a posterior distribution of the relative perceptual distances, which then put the samples on an interval scale. Testing with simulated data showed the method to be very stable with respect to unanimous responses and incomplete pairings. All paired comparison data from our experiments were converted to interval data using this technique.

Pilot experiment and results

Before the large scale web experiment, we conducted a pilot experiment with a small number of subjects to investigate the effect of viewing environment (light surround vs dark surround), which was not controlled in the web study; and to check consistency of data across subjects, across images, between different demosaic methods, and between web-based and lab-based experiments.

A total of 8 subjects, 7 males and 1 female, participated in the first pilot experiment, which was conducted in an imaging lab with a calibrated (very close to sRGB) display, and controlled lighting. Half of the subjects did the experiment in complete darkness, half did the experiment with overhead fluorescent illumination measured at 326 lux at the display location. All subjects completed 300 paired comparions between different renderings of 12 different images as described before. The paired comparisons were converted to interval data. Figure 2 shows the comparison of relative preference scale of different renderings under the light and the dark viewing conditions. The preference results seem to be relatively insensitive to moderation variation in ambient light level.

Figure 3 shows the preference scales compiled for different source images. Again, aside from a general variation of the data, no systematic deviation as a functions of image type emerged

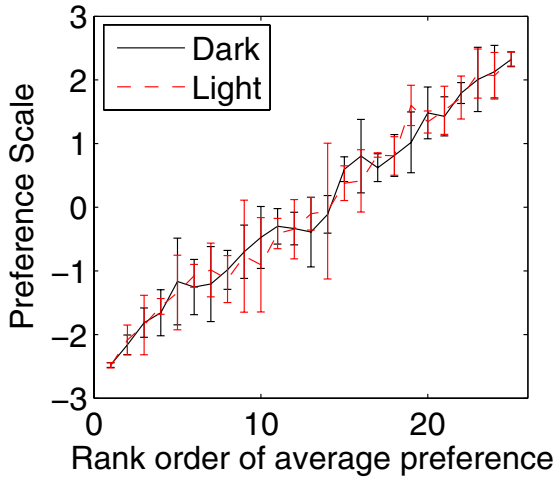

Figure 2. *Comparison of preference results when images were viewed under light vs dark surround. The error bars were calculated among the 4 subjects' results in each lighting condition. No significant difference was found.*

from the data, showing a general insensitivity of the preference data to scene type.

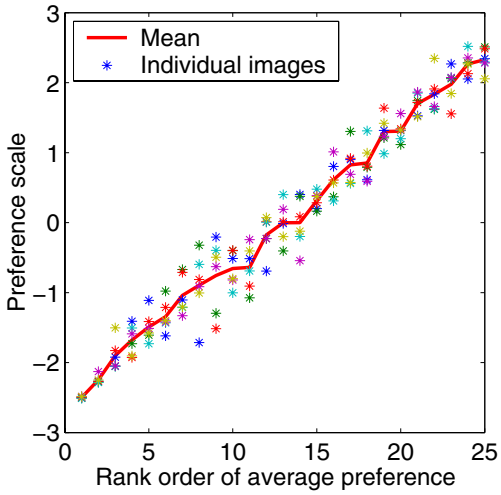

Figure 3. *Comparison of preference results among different image contents.*

Figure 4 shows the variation of preference scales for individual subjects in the pilot experiment. The 8 subjects included 4 people in the imaging field (expert viewers), and 4 naive subjects. Again, aside from general variability of the data, subjects' results were largely consistent, with no clear diverging patterns.

For our second pilot experiment, 6 subjects from the first experiment did the experiment again using the same images processed with a different demosaic method. The demosaic method used in the first pilot experiment was a Bayesian convolution method [8], and the method used for the second experiment was a proprietary directional linear interpolation method. The first method used a large kernel, and resulted in much higher local correlation of noise than the second method. Our goal is to check whether characterization of noise using the S-CIELAB ΔE metric

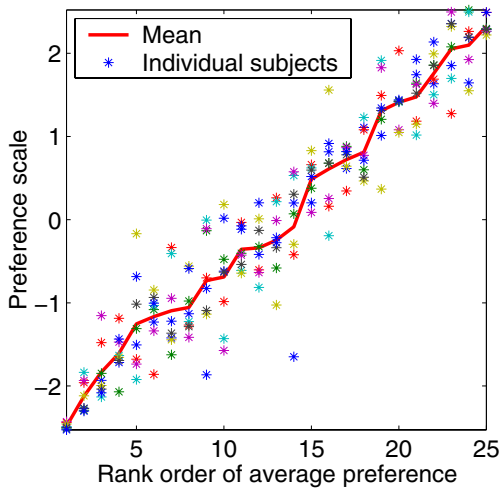

Figure 4. *Comparison of preference results among different subjects in the pilot experiment.*

for this purpose is sufficient when the type of noise is different. Figure 5 shows the preference scales for different renderings when comparing the two different demosaic methods. Again, no significant difference was found if the final noise levels were similar according to the S-CIELAB metric.

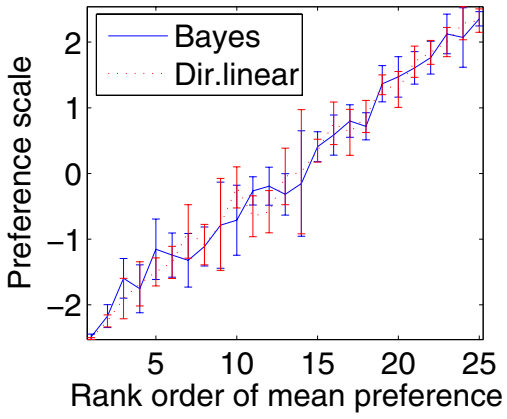

Figure 5. *Comparison of preference results when images were processed with different demosaic algorithms. Error bars were calculated among 6 subjects who participated in pilot experiments one and two.*

For the third pilot experiment, three subjects completed the same experiment as the first pilot experiment, this time using a web interface, and on un-calibrated displays on their desk, with ambient office lighting. Figure 6 shows the consistency of the preference results between the web-based version and the lab-based version. The general pattern of preference is consistent between the two data collection modes.

The pilot experiments confirmed that the noise-saturation trade-off preferences for digital images are relatively insenstive (to the first order) to ambient lighting condition, small variations in display calibration, image content, and individual subject variation. We felt comfortable in proceeding to the large-scale web-

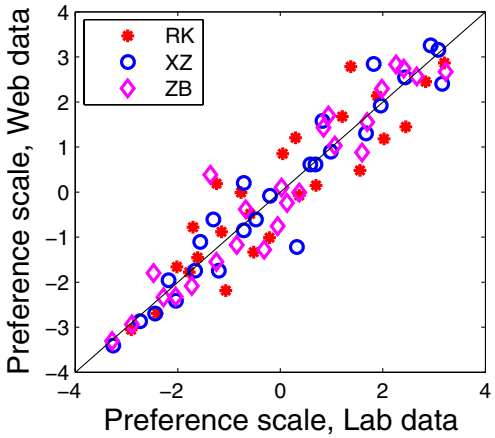

Figure 6. *Comparison of preference results between web-collected data and lab-collected data in third pilot experiment.*

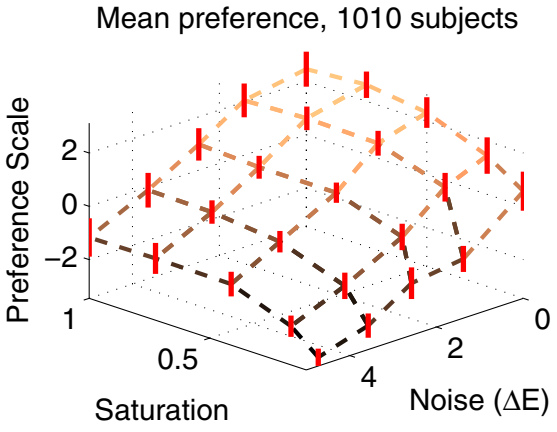

Figure 7. *Preference surface for images with different color saturation and noise levels, averaged over 1010 subjects.*

based experiment.

Web experiment and results

In order to recruit a large number of naive subjects in the preferece experiment, we advertised the study in Agilent's internal newsletter to all US employees, including research facilities, sales organizations, and factories. All participants who completed the experiment were entered into a prize drawing for 2 digital cameras.

In the web experiment, each subject viewed 156 pairs of images (two different renderings of the same scene each time, total of 12 scenes) side-by-side in the web browser, controlled by a java program. They were allowed as much time as possible for each trial. They clicked on a button below the image version they prefer to indicate their preference, and clicked on a "continue" button to move to the next paired comparison. The "continue" button was activated only after a preference choice was made, to discourage subjects from clicking through the experiment without actually making a preference choice.

At the end of the experiment, subjects were asked to estimate the gamma value of their displays using a visual tool [5]. They were then asked to provide an email address volunteerily if they wanted to enter into the prize drawing.

A total of 1010 subjects completed the experiment during the 20 day period the experiment site was open. For each subject, paired comparison data for different source images were lumped together, and converted to an interval preference scale as a function of color saturation (in percentage of full color matrix) and noise level (in S-CIELAB ΔE). Each of the 1010 individual preference surfaces (preference scaled plotted agains color saturation and noise level) were inspected, and they were all substantially similar in shape to each other. This provided some confidence that the subjects were indeed cooperating and making serious comparisons.

The avearage preference scales for all subjects, along with the error bars, are plotted in figure 7. The mean preference scales shown in the plot are also listed in Table .

The preference surface shows nothing surprising. Subjects preferred rendering with lower noise, and with higher color saturation to an extent (after which the preference level asymptotes, and possibly will drop if saturation goes up further). If we are free to move about the noise and saturation space, the zero-noise, full-saturation corner is obviously where we want to be.

Average preference scales for different saturation and noise levels.

noise ΔE	saturation				
	0.1	**0.25**	**0.5**	**0.75**	**1.0**
0	0.56	1.45	2.32	2.62	2.46
1.2	-1.24	0.79	1.77	2.10	2.05
2.4	-1.50	-0.55	0.43	0.85	0.97
3.6	-2.56	-1.67	-0.70	-0.25	-0.12
4.8	-3.13	-2.52	-1.69	-1.28	-1.17

However, for most cameras, it is not possible to achieve both optimal noise level and color saturation at the same time. For any given lighting and exposure level on a digital camera, which gives a particular signal-to-noise-ratio (SNR), the color processing results can vary along a line in the saturation and noise space, and thus this preference surface data becomes useful in finding the most preferred point along those noise-saturation tradeoff lines for different SNR levels of a particular camera.

Using preference data to find optimal noise-saturation trade-off for a sensor

Here we show an example of using the preference data to guide the choice of optimal noise-saturation trade-off points for a particular sensor and exposure setting. The preference data were given as a function of color saturation specified as percentage of full color matrix calculated assuming 1000 cd/m^2 scene light level in the RLab adapatation model [3], and noise level calculated as mean S-CIELAB ΔE [1] values on the final processed image. In general, for a fixed raw image SNR level, increased color saturation is related to increased perceptual noise. For different sensors, the perceptual noise level is related to the color saturation level by a different function, and need to be calculated specifically for each sensor.

Figure 8 shows the saturation-noise trade-off lines plotted on

top of the preference surface for a particular sensor at raw image SNR levels from 5 to a little over 100. The noise levels were calculated by simulating a noise-injected medium (20at the specified color saturation level for this sensor, and then calculating the S-CIELAB ΔE values from a uniform grey patch. When SNR level is high, increasing color saturation does not increase noise level by much (solid black line along upper right edge of plot); when SNR level is low, increasing color saturation has a large impact in perceived noise (solid black lines diagonally across the middle of the surface).

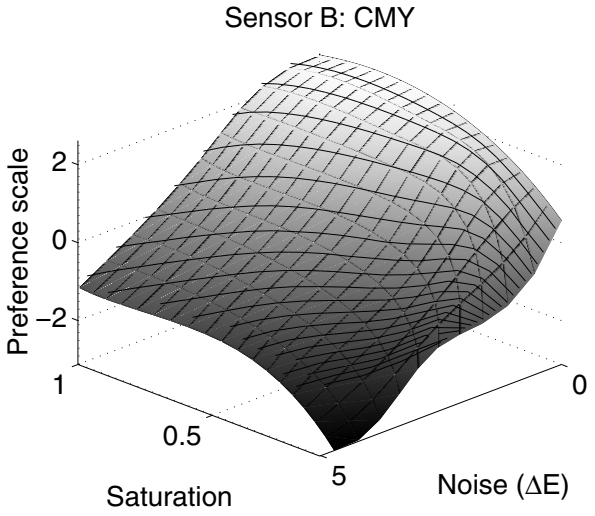

Figure 8. *Interpolated preference surface with saturation-noise trade-off lines plotted on top for a particular sensor at different raw image SNR levels.*

If we plot the preference scales in Figure 8 as a function of color saturation for each of these trade-off lines at different raw SNR levels, as shown in Figure 9, we can easily find out the most-preferred color saturation levels at different raw SNR levels for this sensor. It is easy to see that at high SNR levels (top lines), the optimal color saturation levels are uni-modal and on the high saturation end, i.e. when noise is low, users prefer high color saturation in the image. When raw SNR is very low (bottom lines), the optimal color saturation levels are also uni-modal, but on the low end of the saturation scale. This means when raw noise level is high, user prefer lower color saturation so that noise in the final image is not amplified too much. When raw SNR level is medium (middle lines), then the preference values are sometimes bi-modal, with two peaks at medium saturation level and low saturation level. This might mean that when raw noise level is moderate, there are two "sweet spots" on the saturation-noise trade-off line: one is a "noise-dominating" mode, where lower color saturation is preferred to reduce final perceptual noise; another is a "color-dominating" mode, where higher color saturation is preferred at the cost of higher perceptual noise. It is interesting to note that subjects sometimes pick either noise or color saturation as the dominant concern, and do not always pick a "compromise" when faced with the saturation-noise trade-off.

Summary

We have conducted a large scale web-based study, validated by a series of smaller scale pilot studies conducted in the lab, to find out subjective preference of images in the noise and color saturation space. This data can be used to guide the choice of optimal noise-saturation trade-off in color processing of digitally captured images.

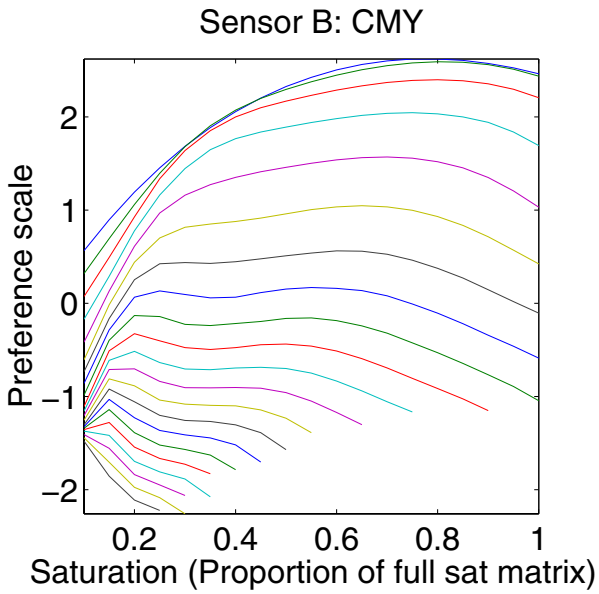

Figure 9. *Preference scales along saturation-noise trade-off lines plotted as a function of color saturation level at different raw image SNR levels.*

Acknowledgements

We wish to thank Russell Iimura, Ram Kakarala, Zachi Baharav, Mark Butterworth, Dwight Poplin, Dieter Vook, and Bob Taber for help with java coding, logistics arrangements for the experiments, acting as subjects, and for helpful discussions during the course of this study. This study was conducted inside, and sponsored by, Agilent Technologies.

References

[1] Xuemei Zhang and Brian A. Wandell, A spatial extension to CIELAB for digital color image reproduction, Journal of the SID, 5(1), 61 (1997).

[2] Xuemei Zhang and David H. Brainard, Bayesian color correction method for non-colorimetric digital image sensors, Proceedings of the 12th IS&T/SID Color Imaging Conference (2004).

[3] Mark D. Fairchild and Roy S. Berns, Image color-appearance specification through extension of cielab, Color Research and Application, 18(3), 178 (1993).

[4] M. R. Luo and R. W. G. Hunt, The structure of the CIE 1997 colour appearance model (CIECAM97s), Color Research and Application, 23, 138 (1998).

[5] Xuemei Zhang, Yingmei Lavin and David Amnon Silverstein, Display gamma is an important factor in web image viewing, in Proceedings of the IS&T/SPIE 9th Annual Symposium on Electronic Imaging, vol. 4299, pp. 455–462 (2000).

[6] L. L. Thurstone, A law of comparative judgment, Psychological Review, 34, 273 (1927).

[7] David Amnon Silverstein and Joyce E. Farrell, Efficient method for paired comparison, Electronic Imaging, 10(2), 394 (2001).

[8] David H. Brainard, Bayesian method for reconstructing color images from trichromatic samples, Proceedings of the IS&T 47th Annual Meeting, pp. 375–380 (1994).

Author Biography

Xuemei Zhang is currently a research scientist at Micron Technology, developing image processing algorithms and image quality metrics. She received her PhD in psychology and MS in statistics from Stanford University (1997), and BS in psychology from Beijing University (1988). Since then she has worked in HP Labs and Agilent Labs, on topics including digital camera image processing, image quality metrics, machine vision, and color reproduction. She is a member of IS&T and SID.

Rick Baer manages the Imaging Systems Group at Micron Technology, which develops image quality measurement methods, performs systems-level analysis of imaging systems, and investigates new imaging applications. He was employed at Agilent Laboratories from 2000-2005, where he worked on CMOS imaging and developed a camera for low-resolution image sensing and interpretation in wireless sensor networks. Previously he was employed at Hewlett-Packard Laboratories for 16 years where he worked on many different projects including digital photography, acoustic biosensing, wireless communications and mass spectrometry. He received his Ph.D. from Stanford University, bachelor's degree from MIT, both in electrical engineering.