

Sharpness Rules

Garrett M. Johnson and Mark D. Fairchild
Munsell Color Science Laboratory, Chester F. Carlson Center for Imaging Science
Rochester Institute of Technology, Rochester, New York

Abstract

A large-scale psychophysical experiment was performed examining the effects of various simultaneous variations of image parameters on perceived image sharpness. The goal of this experiment was to unlock some of the rules of image sharpness perception. A paired comparison paradigm was used to compare images of different resolution, contrast, noise, and sharpening. In total, 50 people performed over 140,000 observations. The results indicate that there are several very interesting trade-offs between the various parameters of contrast, noise, resolution, and spatial sharpening. An interval scale of image sharpness was created. This scale was then used to test the results of several existing models of color and spatial vision. The ultimate goal of this experiment, along with the visual modeling is to obtain a mathematical model of perceived image quality.

Introduction

The ongoing quest of modeling perceived image quality is one rich in both past and present research. Recent work has illustrated how close, yet how far we might actually be from obtaining the elusive goal of mathematically predicting image quality.[1,2,3,4] The approach proposed in this research can be summarized with one simple hypothesis:

> *An image quality metric can be derived as a measure of the perceived difference from an ideal image.*

This hypothesis assumes that any change in image quality results from a perceived color difference. This difference might be the effect of color and tonal reproduction, or other spatial aspects of color appearance such as spatial resolution, sharpness, noise, or half-toning algorithms.

To examine this hypothesis, we envision a four-step process of deriving an image quality metric:
(1) psychophysics to create interval scales of image quality,
(2) formulating a vision model to build a difference metric,
(3) deriving the relationship between the difference metric and image quality scales, and
(4) establishing an anchor image for the interval scale.

While this research is concentrating on the first step, using psychophysics to create interval scales of image quality, a brief discussion of the other steps is in order.

Vision Model and Difference Metrics

The current standard for color difference specification is the CIE94 color difference equation, based on an extension of the CIELAB color space.[5] The CIE94 equation, however, was created using simple color patch stimuli, in well-defined viewing conditions. If the viewing conditions of stimuli are more dynamic, a color appearance model must be used instead. The CIECAM97s model represents the current standard in color appearance modeling.[6,7] While better able to handle complicated color appearance changes such as changes in white point, and viewing luminances, this model largely was developed based on simple colorimetric stimuli. These models tend to neglect some of the spatial aspects of images, such as sharpness and noise, which tend to have a great effect on perceived image quality.[8] In order to do that, a model of spatial vision is necessary. Much work has been done in spatial vision modeling, as can be witnessed by Lubin's Sarnoff Model,[9] and Daly's Visible Differences Predictor.[10]

Whereas these models provide impressive predictions of image difference between spatially complex stimuli, these stimuli tend to be monochromatic. The treatment of color, and specifically color appearance is not emphasized. Other models have attempted to bridge this gap between color difference modeling and spatial vision modeling. One such model is the S-CIELAB color difference metric, from Zhang and Wandell.[11] This model combines spatial filtering of color stimuli, with the CIELAB color difference equation. This model has been extended with the multi-channel approach of Daly's model by Jin *et al.*, to create the color visual difference model (CVDM).[12] Pattanaik *et al.* formulated a multi-scale model (MOM) of spatial and color vision that is capable of predicting a wide variety of spatial threshold and color appearance data and incorporates an intrinsic model of light and chromatic adaptation.[13] These three models will be examined, and perhaps refined, in future research, to correlate with perceived image quality scales.

The fundamental assumption for this approach to image quality modeling is that there is a perceived difference between any given image and an ideal image. The psychophysical and mathematical modeling of image quality relies on the concept of an "original" image, in order to make the comparisons. Many times these original images have already been subject to several image quality degradations from the imaging systems used to obtain them. Previous research has been done to synthesize high photometric resolution images, of arbitrary spatial and

spectral resolution.[14] The synthesis process used to render these images can create images that are not subject to any degradation caused by an imaging system. One such scene was rendered, and used as an original image for this research.

This research presents the first of a series of psychophysical experiments designed specifically for the derivation, and testing of image quality metrics. While only one of the many perceived appearances that make up image quality, it has been noted that sharpness plays a very important role.[8] Therefore, the study of sharpness presents an ideal starting point towards bridging the gap between spatial and color image quality modeling.

Experimental

This experiment examines the simultaneous variations of four image parameters: spatial resolution, additive noise, contrast adjustment, and spatial sharpening filters.

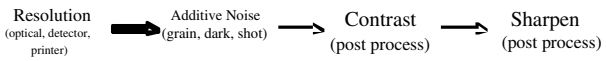

Figure 1. Image Processing Path Used to Create Samples.

Spatial Resolution

Previous research has indicated that for pictorial images, 300 pixels-per-inch at 8 bits-per-pixel is adequate for printed color image quality.[15] Thus, we focused on three levels of spatial resolution: 300 ppi, 150 ppi, and 100 ppi. These images were created by sub-sampling a higher resolution image, and then using nearest-neighbor interpolation to expand the lower resolution image back to the original size, effectively creating the appearance of larger pixels, for the lower resolution images.

Noise

To examine the influence of additive noise on perceived image quality, four levels of uniform, channel independent RGB noise were created: no noise, 10 digital count, 20 digital count, and 30 digital count noise. Each of the noise levels was uniformly distributed around a mean of 0.

Contrast Enhancement

Three levels of contrast enhancement were used in the experiment. This includes the standard "non-enhanced" level, and two levels of contrast enhancement. The enhancement was performed using sigmoidal exponential shaping functions.

The three levels of contrast (none, exponent 1.1, exponent 1.2) were performed on the independent image RGB values, indicative of a typical image-processing situation.

Sharpening

There exists many image editing tools which allow an end-user the ability to enhance the sharpness of an image, through the use of spatial or frequency filters. One common tool is *Adobe Photoshop®.* In this experiment there are two levels of image sharpening: none, and the Photoshop sharpen filter. This is similar to post processing one might do on pre-existing images.

Experimental Design

The four different image parameters described above combine to form 72 images, when simultaneous variations are included (3 resolution * 4 noise * 3 contrast * 2 sharpening). The order that the simultaneous variations occur can have a great impact on the resulting images. For this research a real imaging system, such as a digital camera, was simulated. Figure 1 illustrates the flow-chart followed to process the experimental images used.

Figure 2 shows an image matrix representing the 4 image variations, in the order the samples were prepared.

The 72 images were then used in a paired-comparison experiment. In the paired-comparison paradigm, the 72 different images result in 2556 pairs for evaluation (72*71/2). Combined with 4 distinct scenes, as shown in Figure 4, this requires a staggering 10224 observations.

The pairs of images were displayed on an Apple Cinema digital LCD display, driven by a Power Macintosh G4/450. The 22-inch diagonal display allowed two 4x6 inch images to be displayed simultaneously.

The images were presented on a white back-ground, with a maximum luminance of 154 cd/m^2. Previous work by Gibson has shown that LCD monitors are capable of performing as well as, if not better than, high quality CRT displays.[16] To simulate 300-ppi resolution, the display was placed at a viewing distance of 5ft, which is approximately 3.5 times a normal print viewing distance of 18 inches. The images presented were 630 by 420 pixels, which subtended roughly 7 degrees of visual angle when viewed at this distance. To facilitate the speed at which pairs could be viewed all 288 different images (72 images x 4 scenes) were loaded into memory. All possible pairs were then randomized and were presented to the observer with random selection between right and left side of the display. The observer was given a left hand and right hand mouse, which they clicked to select their chosen image. With this set up, it was easily possible to present a new image pair in less than .5 seconds.

Observers were then presented with the rather simple task of choosing which of the two images “appears sharper.” A single session presented 500 pairs of images to an observer. On average, an observer was able to finish a session in 20 minutes. Observers could then choose to continue on for multiple sessions, if they desired, or quit after a single session. Since no person could perform all 10224 observations in a single setting, the experiment was designed to allow an observer to finish a session and resume where they left off at a later date.

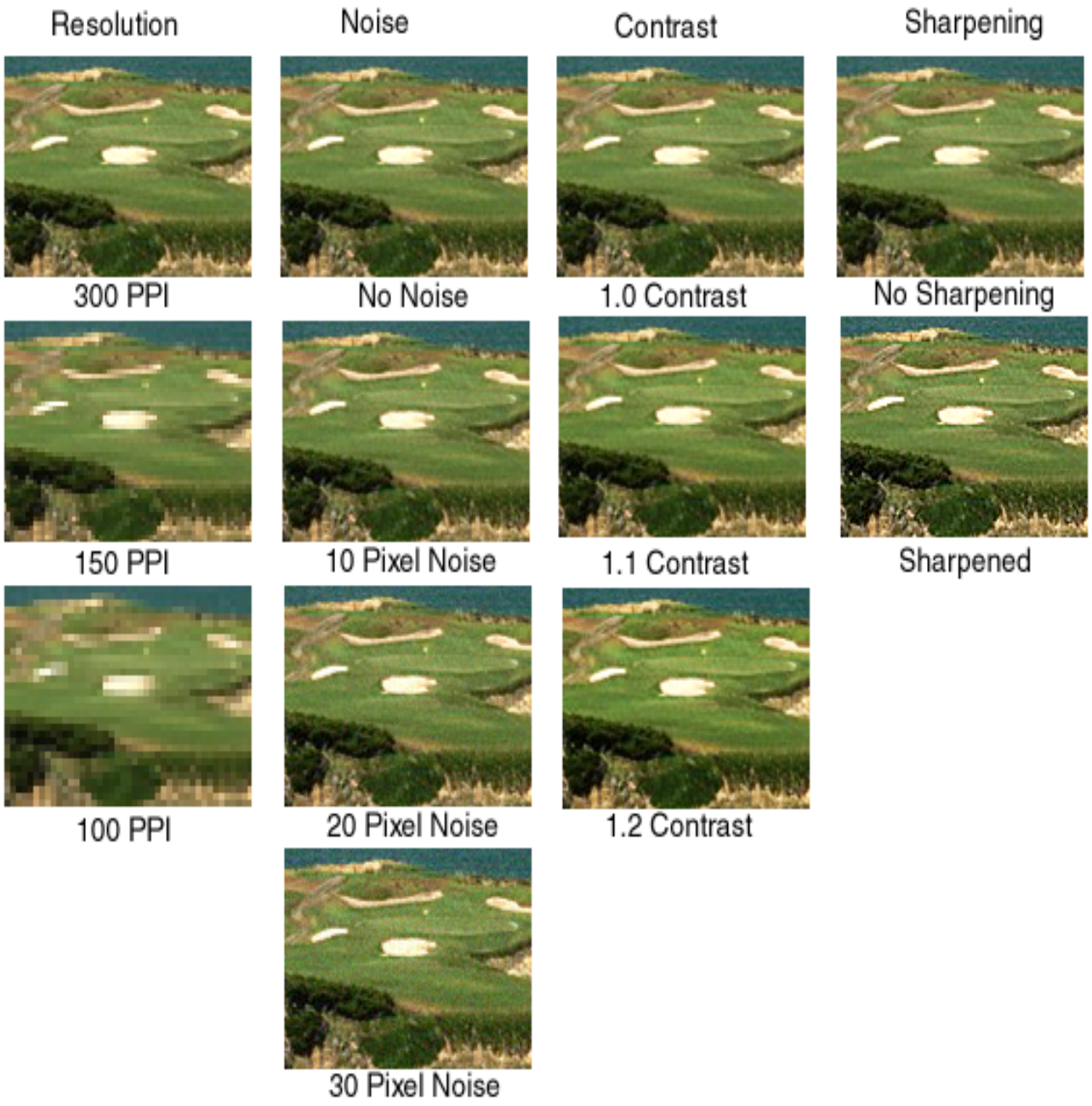

Figure 2. Image Matrix Representing Four Simultaneous Variations

Results

A total of 50 observers completed over 140,000 observations. Five observers completed all 10224 observations, while the average observer completed roughly 2500 image pairs.

Thurston's Law of Comparative Judgement, Case V, was used to analyze the results of the paired comparison experiment, and convert the data into an interval z-score scale. Due to vast difference between some of the image pairs, there were several zero-one proportion matrix problems. This was solved using Morrisey's incomplete matrix solution, which uses a linear regression technique to fill in the missing z-scale values.[17]

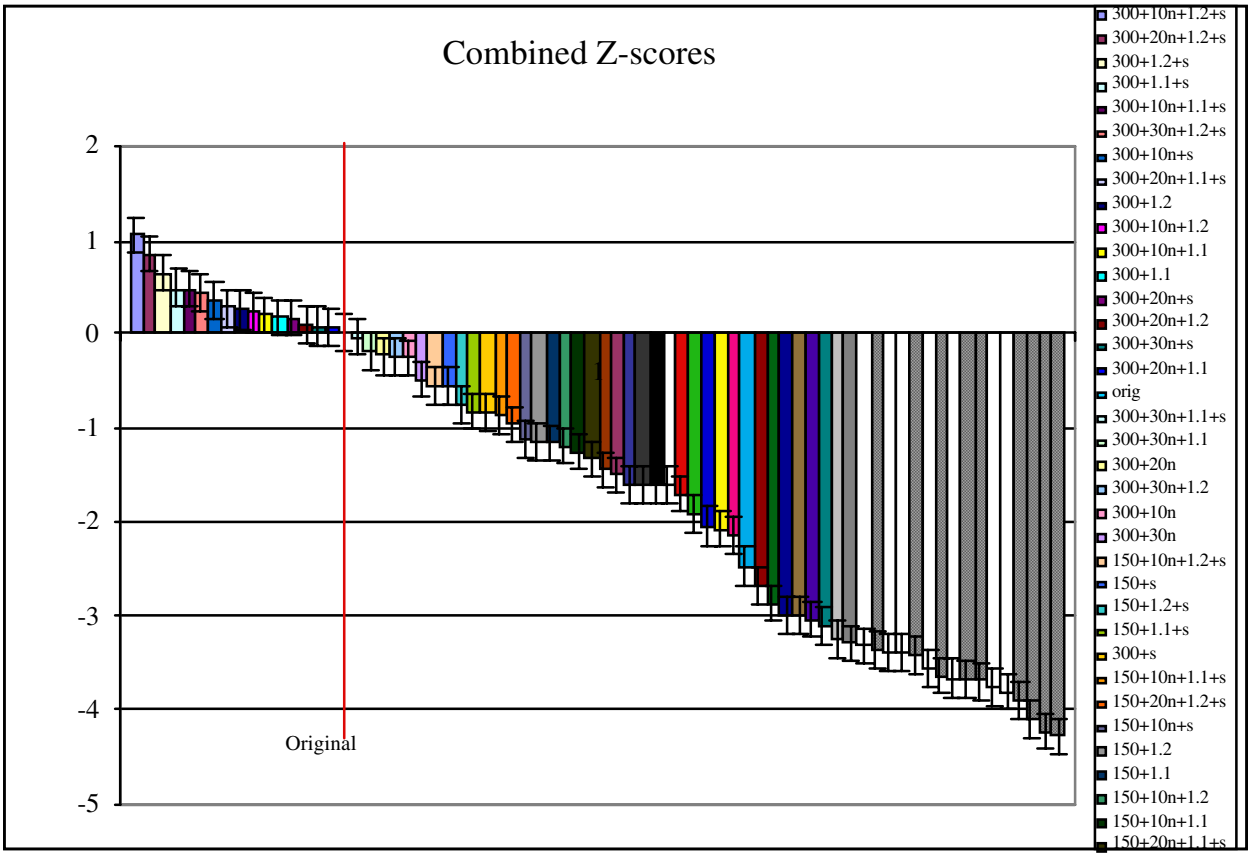

Figure 3. Normalized Z-scores of Combined Scenes. The legend on the right provides the rank order of the image variations, from best to worst.

Figure 3 presents a graph of z-scores obtained from the combined results of all 4 scenes. The z-scores have been normalized, so that the original image has a scale value of 0. Therefore, all images that have a positive value appear sharper than the original, while those with negative values appear less sharp. The legend shown in Figure 4 shows the ranking of all the image variations, from best to worst. The images are labeled as follows: first, the resolution of the image is listed, followed by the amount of noise, followed by the contrast level, and a sharpness key. For example, image 300+20n+1.2+s is a 300dpi image, with 20 pixel noise, a contrast enhancement of 1.2, and sharpened in Photoshop.

A test of the Average Probability Deviation on the resulting z-scores resulted in an average error of 0.026. This suggests that the Case V model fits the data well.

These results indicate that 21 images appeared as sharp or sharper than their respective original images. At least 6 images were judged significantly sharper than the original. All of these images had a resolution of 300 dpi. This indicates that spatial resolution is of the highest priority. The 300-dpi image, with a noise level of 10, a contrast increase of 1.2, and with spatial sharpening was determined to be statistically sharper than all other images. The 300-dpi image, with noise level 20, contrast increase of 1.2, and spatial sharpening was also judged significantly sharper.

The data for all the images individually were then examined to see if any scene dependencies were present.

For the Cow scene, the Average Probability Deviation calculated was 0.043, indicating less than 5% error. This indicates that the model used was a good fit for the data. It is important to also note that for the Cow scene, adding noise and increasing contrast to an image was at times able to mask some of the resolution differences between the 300dpi and the 150dpi images. Several enhanced 100dpi images were also judged to appear as sharp as some 150dpi images. Another interesting artifact for the cow scene, was the effect of spatial sharpening. For most images, the highest ranking images tended to have spatial sharpening, while for the cow this was not the case. Instead, there were many cases where lower resolution images were selected over the spatially sharpened higher resolution image. This suggests that perhaps the edges of the computer rendered cow were already too crisp, since they had suffered none of the degradation that usually occurs in an imaging system.

For the remaining scenes the Average Probability Deviations were determined to be 0.044, 0.046, and 0.043 for the Bear, Cypress, and Man images respectively. All of these errors were less than 5 percent. This indicates that the Case V model was a good fit for all of the image scenes. For the bear scene in particular, there were several different occasions where a lower resolution image was selected to be sharper than several higher resolution images. This was particularly the case for the 150-dpi vs 300-dpi images. This occurrence was also found in the Cypress images, and less so in the Man images. For all scenes, the sharpest images had some form of contrast enhancement.

To determine whether the combined data analysis masked any particular features evident in the individual scenes, the individual scene Z-scores were plotted against the combined Z-scores. Figure 5 illustrates these plots for two of the scenes, the Cow and Cypress images.

The cow scene fits with the combined data reasonably well with a correlation coefficient of 0.81, though there are some interesting outlying points. All of the data that do not match up well with the combined results involved images that were spatially sharpened. The most noticeable outlying point is the sharpened 300dpi image. While consistently one of the highest ranked images for the other scenes, it was ranked very low for the cow scene.

Figure 4. Four Different Scenes Used in Experiment (Cow, Bear, Cypress, Man)

The other scenes match the combined data rather well, with correlation coefficients of 0.90, 0.96, and 0.96 for the Bear, Man, and Cypress scenes respectively. This analysis seems to indicate that the data for all scenes can be combined. It is important to note that the slope of the lines fitting the data in the above figures is not important, but rather that the data can be fit well with a simple linear equation.

The individual image variations were then examined to try and gain an understanding of the rules of sharpness perception. All of the z-scores for a particular attribute were averaged, across the combined results, as well as individually for each scene. This created an average weight, for any given variation. Figure 6 provides a plot of this analysis.

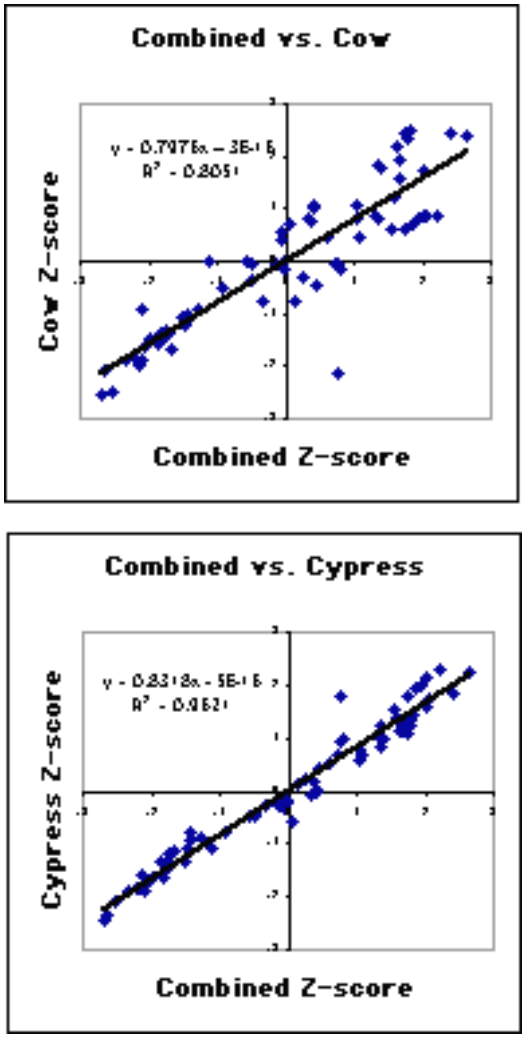

Figure 5. Individual scene Z-scores against the combined scene Z-scores. Top of figure shows Cow image, Bottom shows Cypress.

It is clear from the analysis that spatial resolution, which can be thought of pixel size or addressability, is by far the most important influence on perceived image sharpness. Other interesting "rules" can be interpreted from the results. Enhancing contrast increases the perception of sharpness for all scenes, except for the bear. Additive noise increased perceived sharpness, up to a certain amount of Pixel noise, and then decreased sharpness. Spatial filtering had a significant effect of sharpness for all scenes, except the Cow scene where it decreased perceived sharpness. These effects were most noticeable in the 300 and 150 dpi images. At 100 dpi, the effects were similar, though less distinct.

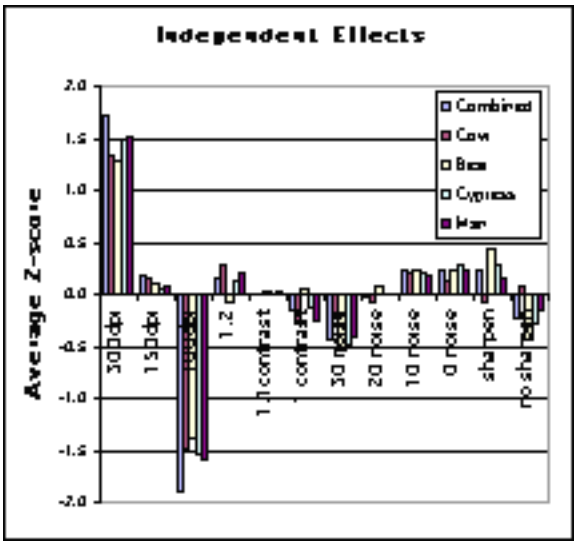

Figure 6. Average Z-scores of individual image parameters, indicating relative importance towards perceived sharpness

Model Analysis

We then ran all of the images through four different "vision" models. These were the CIE ΔE_{94} Color Difference Equation on a pixel-by-pixel basis, S-CIELAB and the CVDM model, combined with the ΔE_{94} equation, and the Multiscale (MOM) Model. The CVDM model was modified slightly from the published version to utilize the same Contrast Sensitivity Functions as S-CIELAB. This was done to determine the effects of the visual masking functions on the different spatial and orientation sub-bands. The Multiscale model has not yet developed a standard method for measuring color differences, at this point in time. For our purposes, we did a simple Euclidean difference on the lightness, and opponent color channels.

The output of each model was a "difference" image between the original image and the 71 variations of that image. Simple image statistics were then performed on the resulting error images to determine the mean, variance, and median of each error image. These statistics were then plotted against the interval scale developed from the psychophysical experiment. The hope was to find a relationship between the output statistics, and the experimental results. Figure 7 illustrates a plot of the average mean error across the four scenes determined by the vision models against the absolute Z-scores found experimentally. The Figure shows the models in order of complexity (ΔE_{94}, S-CIELAB, CVDM, MOM).

As expected, doing a pixel-by-pixel color difference on the images did not correlate well with the results of the experiment. The wide scatter in the plot and the poor correlation coefficient help illustrate this. S-CIELAB actually fared worse at predicting the data, with almost no correlation. The CVDM model also had a poor overall fit to the data but does show some interesting artifacts. It appears as if there are two distinct linear series, rather than a single. This appears to be a result the models prediction on the Photoshop sharpened images. This could be an attributed to the visual masking functions or the spatial and orientation filters, since those are the only differences between CVDM and S-CIELAB. Future analysis of these trends might lead to a simple alteration of the CVDM model that might better predict the data. The MOM model was able to fit the experimental results rather well, as illustrated in the bottom of Figure 6, despite not having a standard method of computing color differences. The relationships between the median and variance of the error image, rather than the mean illustrate similar results.

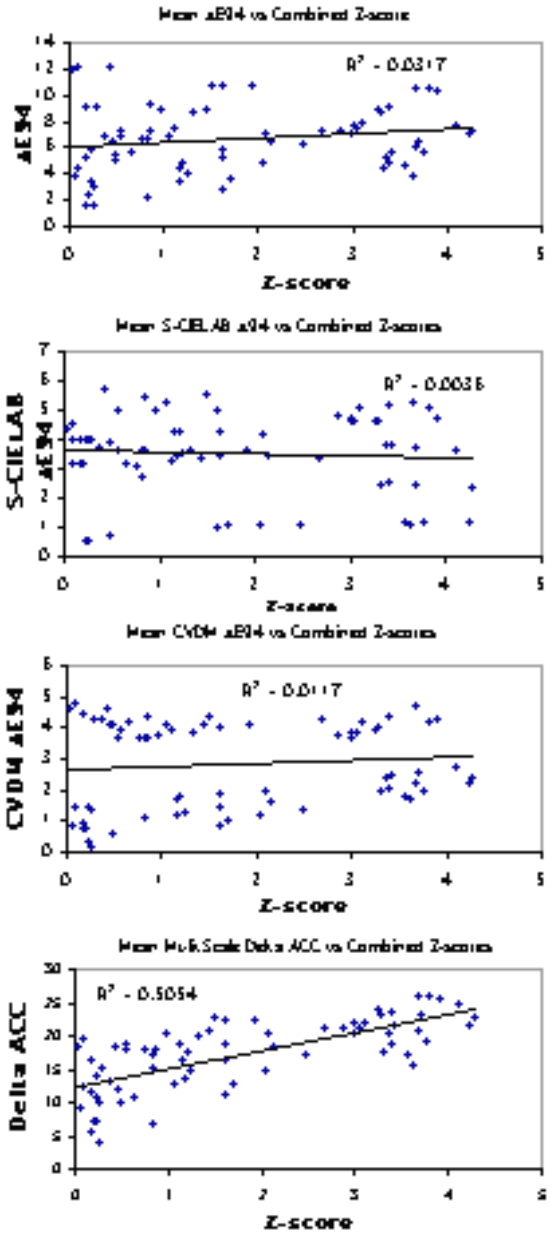

Figure 7. Model predictions plotted against experimental results.

Conclusions

A large-scale paired comparison experiment was developed to test the simultaneous effect of contrast, noise, resolution, and spatial filtering on perceived sharpness. In all, 50 observers performed a total of 140,000 observations. Psychophysical analysis was used to create an interval scale of sharpness. For every scene there were many images that were judged to be sharper than the original. From this analysis we determined several "rules" of sharpness. Resolution is by far the most important sharpness factor when dealing with 300, 150, and 100 dpi images. Increasing contrast will, in general, increase the appearance of sharpness. Additive uniform noise also increases sharpness up to a certain level of noise, and then decreases sharpness. Finally, spatial sharpening generally causes an increase in perceived sharpness. Of course, like all "rules," these are sometimes broken.

Several models of color and spatial vision were then used in an attempt to predict the results. As expected, the standard CIE94 Color Difference Equation when applied on a pixel by pixel basis was unable to predict the results. Neither S-CIELAB nor the CVDM were able to predict the results, by using error image statistics. This is not a surprising result as neither model was created to be a model of human vision. Rather, these models were developed as a method for calculating image differences. The resulting error images from both of these models were much closer to actual perceived differences than the error images found by simply using a pixel-by-pixel difference. The MOM model was able to predict the results of the experiment reasonably. This model, as compared to the others, is a more complete model of both spatial and color vision, resulting in much more complexity and computational expense.

The future goal of this research is now to take a closer look at each of the vision models, and to refine them to be better suited to this type of image quality research. This might include developing a more robust difference equation for the MOM model and perhaps adding some of its local adaptation and masking functionality into the other models. This particular experiment also emphasized image sharpness and not overall image quality. While it is generally thought that sharpness is an important aspect of overall image quality, more work needs to be done to better understand that relationship.

References

1. P. Engeldrum, A framework for image quality models, J. Imag. Sci. Tech. 39, 312-318 (1995).
2. P. Engeldrum, Image quality models: Where are we?, IS&T PICS Conference, 251-257 (1999).
3. J.E. Farrell, Image quality evaluation, Ch. 15 in Colour Imaging: Vision and Technology, L.W. MacDonald and M.R. Luo, Eds., Wiley, Chichester, 285-314 (1999).
4. A.M. Ford, Determination of compressed image quality, Ch. 16 in Colour Imaging: Vision and Technology, L.W. MacDonald and M.R. Luo, Eds., Wiley, Chichester, 315-338 (1999).
5. CIE, Industrial Colour-Difference Evaluation, CIE Tech. Rep. 116, Vienna (1995).
6. CIE, The CIE 1997 Interim Colour Appearance Model (Simple Version), CIECAM97s, CIE Pub. 131 (1998).
7. M.D. Fairchild, Color Appearance Models, Addison Wesley, Reading, (1998).
8. L. MacDonald, Framework for an image sharpness management system, IS&T/SID 7th Color Imaging Conference, Scottsdale, 75-79 (1999).
9. J. Lubin, The use of psychophysical data and models in the analysis of display system performance, Ch. 12 in Digital Images and Human Vision, A.B. Watson, Ed., MIT Press, Cambridge, (1993).
10. S. Daly, The Visible Differences Predictor: An algorithm for the assessment of image fidelity, Ch. 13 in Digital Images and Human Vision, A.B. Watson, Ed., MIT Press, Cambridge, (1993).
11. X. Zhang and B.A. Wandell, A spatial extension of CIELAB for digital color image reproduction, SID 96 Digest, (1996).
12. E.W. Jin, X.-F. Feng, and J. Newell, The development of a color visual difference model (CVDM), IS&T PICS Conference, 154-158 (1998).
13. S.N. Pattanaik, M.D. Fairchild, J.A. Ferwerda, and D.P. Greenberg, Multiscale model of adaptation, spatial vision, and color appearance, IS&T/SID 6th Color Imaging Conference, Scottsdale, 2-7 (1998).
14. G.M. Johnson and M.D. Fairchild, Computer synthesis of spectroradiometric images for color imaging systems analysis, IS&T/SID 6th Color Imaging Conference, Scottsdale, 150-153 (1998).
15. A. Vaysman and M.D. Fairchild, Degree of quantization and spatial addressability trade-offs in perceived quality of color images, Color Imaging: Device Independent Color, Color Hardcopy, and Graphic Arts III, Proc. SPIE 3300, 250-261 (1998).
16. J.E. Gibson, Color tolerances in pictorial images presented on various display technologies (tentative title), RIT M.S. Thesis, in progress (2000).
17. P. Engeldrum, Psychometric Scaling: A Toolkit For Imaging Systems Development, Imcotek Press, Winchester (2000).