A Metric of Perceived Image Degradation Based on Foveal and Peripheral Visual Performance

Jian Yang and Michael E. Miller
Electronic Imaging Products, R&D, Corporate Design & Usability
Eastman Kodak Company
Rochester, New York

Abstract

The human visual system has a lower spatial resolution in the periphery than in the fovea. This property may be useful to reduce system bandwidth in applications where the observers' fovea is allowed to scrutinize a very small portion of a picture. For example, an image frame in a video or movie is presented for less than 100 milliseconds. When viewed with such a short duration, most parts of the frame is seen by the peripheral retina. Since the resolution of the spatial resolution is low, one could reduce information in image areas that will be viewed by the periphery without causing perceptible image degradation. The purpose here is to present a quantitative metric for evaluating the image quality for such non-uniform degradations, by considering visual performance in the periphery. According to this metric, the image difference between the degraded image and the original is decomposed into five levels of pyramid error images based on a measured or assumed fixation position. The contrast of each error image is scaled by a contrast threshold map that is a function of spatial frequency, eccentricity, and image content. Finally, the perceived image degradation is calculated as the square root of the sum of the mean squared contrast over levels and color channels. Experiments were conducted to obtain subjective image quality with non-uniformly degraded images using two different algorithms. Five observers participated in the experiment and were instructed to rate the image quality on a ratio scale. The resulting image quality metric accounts for 92% of the variance in the image quality ratings. As a benchmark, RMS difference accounts for only 67% of the variance.

Introduction

Human vision is spatially inhomogeneous. When operating under photopic luminance levels, human vision has its greatest spatial and chromatic resolution at the fovea and lower spatial and chromatic resolution in the periphery. This property may be useful in applications where image communication demands very high system bandwidth. For example, the required bandwidth can become exceedingly high when the field of view of the video screen needs to be large, such as in an immersive entertainment venue. In these environments, the time interval between video frames is generally very short and precluding saccadic eye movements within a frame. In this environment, the human fovea is not able to scrutinize more than one location of a single frame and therefore, most parts of the image frame are seen by the peripheral retina. In these environments, there is an opportunity to reduce the required communication bandwidth by using eccentricity-dependent filtering methods to eliminate some information from the image. The question is how to evaluate the perceptible image degradation when the information reduction is not uniformly allocated in space.

Usually, vision models consider only foveal visual performance in estimating the perceptible image degradations (e.g., Refs 1 to 3), under an assumption that there will be enough viewing time for the fovea to scrutinize the entire image. In motion pictures, however, it is most likely that image degradations in the peripheral retina are often not detected, although one would predict they are perceptible degradations based on a foveal vision model. Therefore, if one is to model the image quality of imaging systems that take advantage of the non-homogeneity of the human visual system, one must incorporate peripheral visual performance into the estimation of the perceived image degradation. We are not aware of any models in the existing literature, especially for estimating the magnitude of visible degradation. The work of Peli[4] and Geisler and Perry,[5] however, are closely related to this topic and they are useful for the development of such a model.

Peli[4] suggests that a local contrast metric can correlate to the perceived contrast in an image. In calculating this metric, the image is first decomposed into several frequency bands, i.e., pyramid error images. The amplitude of each band is scaled by the luminance value obtained from all lower frequency bands. Finally, the perceived contrast is a combination of the scaled contrast over all the bands. This particular model does not include peripheral properties, but it provides a method to estimate suprathreshold magnitudes.

The work of Geisler and Perry[5] concerns the filtering of image content at different eccentricities. In their method, an image is also decomposed into a pyramid of error images. The algorithm keeps only the image content at each band

that is expected to be visible based on a contrast threshold map, which varies with eccentricity and spatial frequency. Although this filtering method matches human visual performance at different eccentricities it does not provide a metric to evaluate the final image quality.

The purpose here is to develop a quantitative metric of image degradation that matches subjectively perceived image quality, using some model components that have been discussed in Refs. 4 and 5. For a comparison, we will also compute the root-mean-square error, as well as CIELab ΔE metric, between a foveated image and its original.

Peripheral Visual Performance

Based on the study of Peli et al.,[6] the contrast threshold for detecting a patched grating of spatial frequency f against a uniform field at an eccentricity r can be described as

$$C_t(r, f) = C_t(0, f) \exp(k f r), \quad (1)$$

where $\mathbf{C_t(0, f)}$ is the contrast threshold at the fovea, and **k** is a parameter. In their formulation of this equation, Peli et al.[6] fit this equation to data from six reports in the literature that showed how the contrast threshold of monochromatic gratings varied with spatial frequency and eccentricity. These fits demonstrated that the k value ranged from 0.030 to 0.057. Peli and Gary[7] further showed that this equation is useful for estimating image degradation in real complex images.

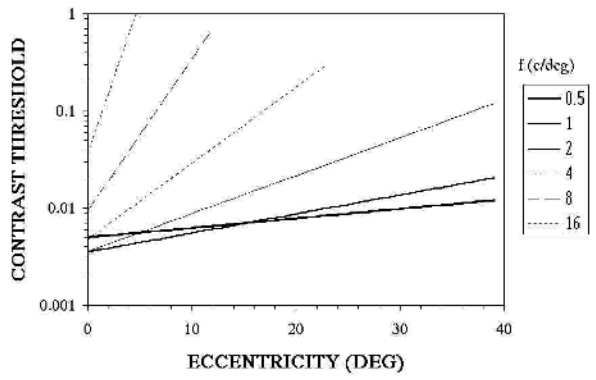

Figure 1. Contrast threshold versus eccentricity. The visual stimuli were grating patches with spatial frequency ranging from 0.5 to 16 cpd. The thresholds were calculated based on Eqs. 1 and 2, with the ***k, N, h, s,*** *and* ***a*** *being 0.045, 0.024, 0.058, 0.1 cpd, and 0.17 degree, respectively.*

Equation 1 gives the relative human contrast threshold at different eccentricities. To describe the contrast threshold completely, one also needs to determine the contrast threshold at the fovea. Yang et al.,[8] developed a model to capture foveal performance, a simplified version of this equation can be expressed as:

$$C_t(0, f) = [N + h\, s^2/(f^2 + s^2)] \exp(a\, f), \quad (2)$$

where, N, η, σ, and α are parameters. The parameter values based on the study of Yang and Stevenson[9] are 0.024, 0.058, 0.1 cpd, and 0.17 degree, respectively. The calculated contrast threshold versus eccentricity at six different spatial frequencies according to Eqs. 1 and 2, and a nominal set of parameter values are shown in Fig. 1.

Image Preparation

The purpose of creating foveated images is to allow the transmission of large, high-resolution images with the least possible information from the original image while allowing minimal perceptual degradation. We consider two foveated algorithms here.

Kortum and Geisler[10] developed an algorithm to sample the original image based on the required sampling interval at different eccentricities. The sampling intervals can also be calculated from Eq. 1. The cut-off frequency $\mathbf{f_c}$ at an eccentricity **r** can be defined by setting the contrast threshold $\mathbf{C_t}$ to 1, which gives $f_c = -\{\ln[N + \eta\sigma^2/(f_c^2 + \sigma^2)]\}/(\alpha + k\, r)$, and it can be further approximated to

$$fc = -\ln(N)/(a + k\, r), \quad (3)$$

when the cut-off frequency $\mathbf{f_c}$ is much higher than **s**, which is often the case. Any frequency components that are higher than the cut-off frequency would not contribute to visual perception, and thus can be discarded. On the other hand, one needs to make sure that lower frequency components up to the cut-off frequency $\mathbf{f_c}$, i.e., the Nyquist frequency, will be retained. To satisfy this condition, the sampling interval should not be larger than

$$\Delta x = 1/(2f_c) = -0.5\,(a + k\, r)/\ln(N). \quad (4)$$

In the Kortum and Geisler approach, the original pixels that are within the sampling widow are assigned with their mean value. This sampling was called a 'SuperPixel'.

The theoretical feature of frequency analysis in human vision has not been used in the SuperPixel. Geisler and Perry[7] further developed a foveated multiresolution pyramid. In this approach, different levels of the pyramid were circularly truncated based on the estimated cut-off spatial frequency of the visual system at different eccentricities. The reconstructed image from the zone-limited pyramid contains fine structure at the center of the fixation, and it gets more and more blurred towards the peripheral retina. This is the so-called foveated multiresolution pyramid (FMP).

One drawback of this particular approach is that it could produce visible hard boundaries between zones. To eliminate this nuisance, Geisler and Perry[7] smoothed the transition boundaries in a somewhat arbitrary way.

In the current study, we do not truncate each pyramid level into circular zones. Instead, we employ a filtering technique that was proposed by Peli.[4] In this method, one computes a contrast threshold map based on Eqs. 1 and 2 and the peak frequency of the pyramid level, and uses the contrast-threshold-map to threshold the image content. Whenever the image contrast of a specific frequency band is

below the corresponding visual contrast threshold, the image contrast is set to zero, that is, thresholded. All other image content is kept unchanged. In this way, the foveated zones are not circular, but have irregular boundaries. The modified approach does not produce hard boundaries (see Fig. 2).

Figure 2. Examples of the degraded images of a degradation level using the FMP algorithm. The crosses at the upper-left corner of each of the images indicate the fixation position.

Subjective Image Quality Evaluation

The purpose of this experiment is to obtain visual performance data on suprathreshold image degradation, thus, for each of the above-mentioned algorithms (SuperPixel and FMP), we created eight degraded images for each original image by scaling the contrast threshold Eqs. 1 and 2.

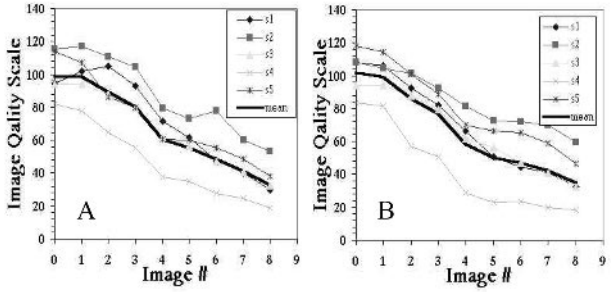

Figure 3. Subjective ratio scales versus degradation level (images #0 to #8) for the five observers, with (A) the scene DISNEY, and (B) the scene TAXI, processed with the FMP algorithm.

The perceivable image degradation is reflected in the subjective image quality scale in reference to the original non-degraded image, (i.e., image #0). Figure 3 shows the subjective image quality scale at different image degradation levels with the FMP algorithm. The data points are geometric means over 20 repetitions for each observer. The dark heavy lines in each panel show the geometric means over the five observers.

The results of the SuperPixel algorithm are shown in Fig. 4. These curves are similar to those shown in Fig. 3. The main difference is the image-processing algorithm.

It is helpful to clarify here that the image displayed prior to the start of a run was the original scene, i.e., image #0, and the observers were instructed to assign a number of 100 to this image. It is interesting to see that this scale held quite well through all the subsequent trials, as the mean scores for each observer were not far from 100 when evaluating the image #0. Furthermore, as the four heavy curves show, the average scores across observers for image #0 are very close to 100.

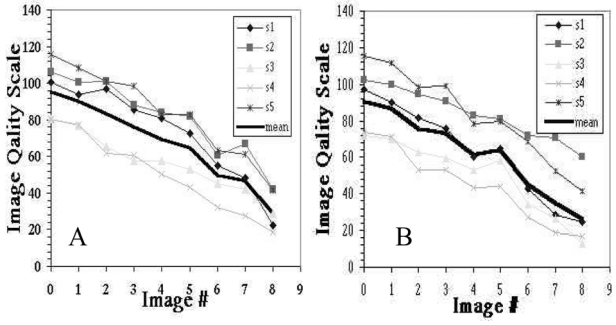

Figure 4. Subjective ratio scales versus degradation level (images #0 to #8) for the five observers, with (A) the scene DISNEY, and (B) the scene TAXI, processed with the SuperPixel algorithm.

Objective Image Quality Evaluation

As benchmarks, we first computed the root-mean-square error, as well as CIELab ΔE metric, between a foveated image and its original. No spatial inhomogeneities of the visual system are considered in the benchmark metrics. We will develop a metric based on human vision performance that varies with eccentricity and spatial frequency.

Physical RMS Difference Metric

RMS error is an often-used metric for measuring the difference between an altered image and its original. It has been reported that the RMS error metric does not match the perceptual difference of the images,[11] nevertheless, we use this metric to serve as a benchmark. The computation of the RMS difference is straightforward. The digital file of the original scene is represented by I_0, and a degraded image by I_d. The difference image in terms of luminance is

$$D(i, j, k) = [\, I_d(i, j, k) - I_0(i, j, k)] \, \mathrm{lum}(k)/255, \qquad (5)$$

where, **j and i** index the coordinates in **x** and **y** directions, respectively, and **k** indexes one of the RGB channels. The array **lum** contains the luminance values of the R, G, and B channels at the code value 255. The variance of the image difference for each channel is

$$V(k) = \sum_{i=1}^{N} \sum_{j=1}^{M} D(i, j, k)^2 / NM, \qquad (6)$$

where, N and M are the vertical and horizontal image sizes, respectively. The RMS difference metric is defined here as:

$$RMS = \sqrt{\sum_{k=1}^{3} V(k)}. \quad (7)$$

For each degraded image, we calculated the RMS difference and plotted it against the subjective rating score of the same image in Fig. 5. The rating scores are the geometric means shown in Figs. 3 and 4, with each of the curves normalized to a maximum value of 100. Therefore, all of the original images should have a rating score of 100 and an RMS difference of 0. The dashed line is a linear fit of the data that cuts across the coordinate (100, 0). Sixty seven percent of the variance in the data points is accounted for by the linear regression line.

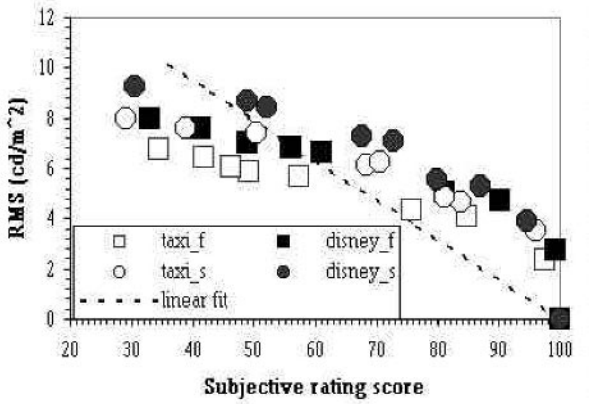

Figure 5. The relationships between the metric of RMS luminance difference and the subjective rating score, with the FMP processed TAXI (open squares) and DISNEY (filled squares) and with the sub-sampled TAXI (open circles) and DISNEY (filled circles).

It is obvious that the RMS difference does not share a strong linear relationship with the subjective rating score. Looking carefully, we can see that for degraded images that have a similar rating score, the DISNEY (filled symbols) scene tends to have a larger RMS difference than the TAXI (open symbols) scene does. Furthermore, the SuperPixel (circles) images tend to have a larger RMS difference than FMP (squares) images do.

Mean CIELab ΔE

It is well understood that human visual response is a nonlinear function of the input luminance. The CIELab color space intends to provide a space that correlates more or less linearly to the visual perception of uniform colored patches (see Ref. 12).

Therefore, we want to check whether the CIELab ΔE metric correlates well to the subjective quality measurements. In the calculations, we first converted the image RGB code values to XYZ values based on the measured chromaticity and luminance values of the R, G, and B channels. The CIELab L*, a*, b* values were calculated following their definition formulas (see Ref. 12, page 220), with the white point chosen to be the color with the R, G, and B values of 255. The ΔE map is calculated as,

$$\Delta E(i,j) = \sqrt{\Delta L_d^*(i,j)^2 + \Delta a_d^*(i,j)^2 + \Delta b_d^*(i,j)^2}. \quad (8)$$

The mean **ΔE** is defined as

$$Mean\_\Delta E = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M} \Delta E(i,j)^2 / NM}. \quad (9)$$

For each degraded image, we calculated the **Mean_ΔE** value and plotted it against the subjective rating score of the same image in Fig. 6. For all the original images, again, have a rating score of 100 and a **Mean_ΔE** of 0. The subjective rating scores are the same as those in Fig. 5. The dashed line is a linear regression of the data points. This regression equation accounts for 56% of the variance in the data, which is lower than the variance accounted for by the RMS metric.

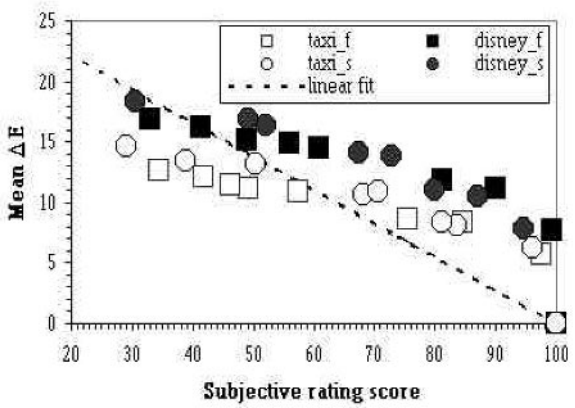

Figure 6. The relationships between the CIELab mean ΔE difference and the subjective rating scores, with the FMP processed TAXI (open squares) and DISNEY (filled squares) and with the sub-sampled TAXI (open circles) and DISNEY (filled circles).

Scaled RMS Difference with Contrast Threshold

From Fig. 1, we know that human vision weights image information differently, based on the frequency content and retinal location. Thus, it is reasonable that the simple physical models using either RMS difference or Mean_ΔE do not correlate well with the subjective scores of image degradation. One reason for this is that the difference in the images at a very high spatial frequency or at a larger eccentricity, should be less perceptible if the same difference is at a lower spatial frequency or near the fovea.

To implement the eccentricity and spatial frequency dependent visual behavior in the difference metric, we follow the basic approach proposed by Peli[4] and Geisler and Perry.[5] We first decompose the difference image as determined by Eq. 5 to five frequency-bands of error images DB(i, j, k, l). The contrast at each frequency band and location is scaled by the corresponding contrast

threshold to obtain the visibility value at each location, band, and channel:

$$VB(i, j, k, l) = DB(i, j, l, k)/\text{mean\_lum}(k)/C_t(r(i, j), f(l)), \quad (10)$$

where, l indexes one of the five frequency bands, and the array **mean_lum** contains the mean luminance values of the R, G, and B channels of the image. $\mathbf{C}_t$ is the contrast threshold determined at eccentricity **r** and peak frequency **f** (see Eqs. 1 and 2). The variance of the image difference, indicated by visibility, for a given band and channel (**l , k**) is then:

$$V(k,l) = \sum_{i=1}^{N} \sum_{j=1}^{M} VB(i,j,k,l)^2 / NM. \quad (11)$$

The square root is computed for the summed variance over different bands and channels to obtain the scaled RMS visibility:

$$RM = \sqrt{\sum_{l=1}^{5} \sum_{k=1}^{3} V(l,k)}. \quad (12)$$

For each degraded image, we calculate the scaled RMS visibility metric and plot it against the subjective rating score of the same image in Fig. 7. For all the original images, again, they have a rating score of 100 and a scaled RMS visibility of 0.

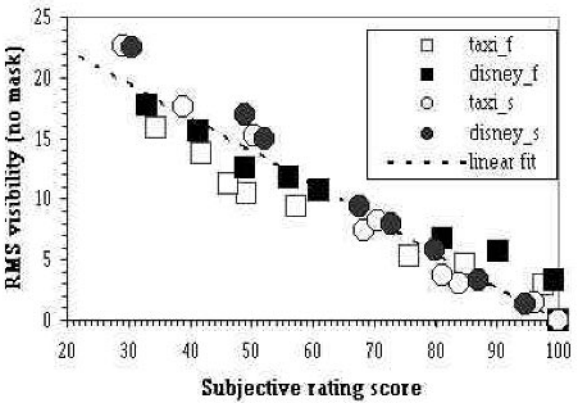

Figure 7. The relationships between the metric of the scaled RMS visibility and the subjective rating scores, with the FMP processed TAXI (open squares) and DISNEY (filled squares) and with the sub-sampled TAXI (open circles) and DISNEY (filled circles).

The dashed line is a linear regression of the data points. This metric accounts for 91% of the variance in the data. It is obvious that the scaled RMS visibility has a nice linear relationship with the subjective rating score, a great improvement from the benchmark metrics (i.e., the physical RMS difference and mean CIELab ΔE). However, the scaled RMS visibility still tends to over estimate the DISNEY (filled symbols) to the TAXI (open symbols) scene. Furthermore, the circle symbols (SuperPixel algorithm) show a steeper slope than the square symbols (FMP algorithm) do against the subjective rating score.

Scaled RMS Difference with Masking Threshold

As we mentioned earlier, the contrast threshold equations were based on the visual performance for detecting a grating against a uniform background field. The existence of a patterned background, in general, would result in an increase of the contrast threshold. The amount of threshold elevation increases as the contrast of the background field increases when the contrast is at a suprathreshold level.[13,14] As we look at the original DISNEY and TAXI scenes, one can see that the DISNEY scene is crowded and has a higher overall contrast than the TAXI scene. Based on a calculation of the physical RMS contrast, the DISNEY scene has an RMS contrast value of 20 cd/m^2 and the TAXI scene has a value of 16 cd/m^2. The difference in the RMS contrast and the related potential masking effect might explain why the RMS visibility for DISNEY is over estimated.

Although it is easy to describe the masking effect qualitatively, it is a difficult task to describe the effect quantitatively, especially when a complex image pattern is involved. Therefore, here we approximate the contrast threshold, Cm, in the presence of a pattern background as follows:

$$Cm(r, f) = Ct(r, f)\,(1 + 4\,pm(r, f)), \quad (13)$$

where **Ct** is the contrast threshold with a uniform background, and ***pm*** is the absolute value of the contrast map of the original image I_0. The value 4 is an empirical scaling factor for the masking strength that produced a good match between the DISNEY and TAXI scenes.

The contrast threshold Ct in Eq. 10 is replaced by the masked contrast threshold Cm to calculate the visibility values and the scaled RMS visibility. For each degraded image, we calculated the scaled RMS visibility metric that included the masking effect, and plotted it against the subjective rating score of that image in Fig. 8. The dashed line is a linear regression of the data points, and it accounts for 92% of the variance in the data. The masking consideration here improved the linear prediction slightly. Now the metric is about neutral to either the DISNEY or the TAXI scenes, as we can see that the filled and open symbols in the figure are nicely overlapped. However, the improvement is marginal. The circle symbols (SuperPixel algorithm) still show a steeper slope over the square symbols (FMP algorithm) against the subjective rating score.

Conclusions

As one would expect, the physical RMS difference metric does not correlate well with the subjective rating score of foveally processed images. The CIELab ΔE provides even worse predictions for this class of images.

The scaled RMS visibility metric, which is based on very simple visual performance model, provides a nice

linear prediction of the subjective rating scores. The inclusion of a model of the masking effect further improved the prediction, and handles well the difference in the original images. However, this metric still leaves some discrepancy for predicting the RMS visibility when the images are processed with different algorithms, namely, FMP and SuperPixel.

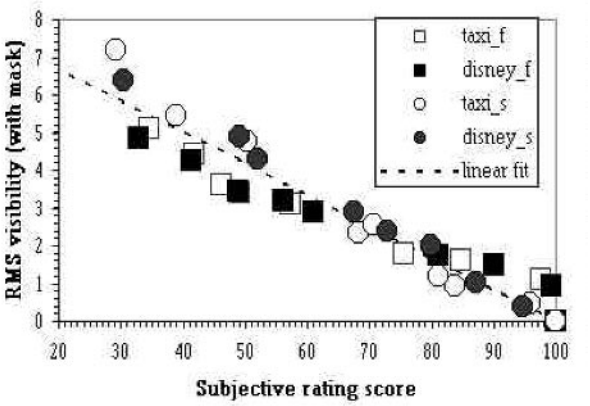

Figure 8. The relationships between the metric of RMS visibility with masking considered and the subjective rating scores, with the FMP processed TAXI (open squares) and DISNEY (filled squares) and with the sub-sampled TAXI (open circles) and DISNEY (filled circles). The dashed line is a linear fit of the data that cuts across the coordinate (100, 0). 92 percent of the variance in the data points is accounted for by the linear regression line.

The image degradations produced by the two foveated algorithms are distributed quite largely in space. At this stage, it is hard to predict how the metric correlates with the image degradations within a local area.

Acknowledgment

The authors thank James G. Stephens, John A. Agostinelli, and Edward Covannon for helpful discussions and Terry Coia for the assistance on running the experiments. We thank the observers who participated in the experiment.

References

1. S. Daly, Human Vision, Visual processing, and Digital Display III, *Proc. SPIE,* **1666,** pg. 2 (1992).
2. J. Lubin, In E. Peli (Ed.) Vision Models for Target Detection and Recognition, World Science, pg. 245 (1995).
3. X. M. Zhang and B. A.Wandell, *Proc. SID Symp., 27,* pg. 731 (1996).
4. E. Peli, *J. Opt. Soc. Am. A***7**, 2032 (1990).
5. W. S. Geisler and J. S. Perry, Human Vision and Electronic Imaging III, *Proc. SPIE*, **3299**, pg. 294 (1998).
6. E. Peli, J. Yang, and R. Goldstein, *J. Opt. Soc. Am. A,* **8,** pg. 1762 (1991).
7. E. Peli and G. A. Geri, *J. Opt. Soc. Am. A* **18,** pg. 294 (2000).
8. J. Yang, X. Qi, and W. Makous, *Vis. Res.*, **35,** pg. 1965 (1995).
9. J. Yang and S. B. Stevenson J. *Opt. Soc. Am. A* **15**, pg. 1027 (1998).
10. P. T. Kortum and W. S. Geisler, Human Vision and Electronic Imaging, *Proc. SPIE,* **2657,** pg. 350 (1996).
11. B. Girod, In A. B. Watson (Ed.), Digital Images and Human Vision, the MIT press, pg. 207 (1993).
12. M. D. Fairchild, Color Appearance Models. Addison Wesley Logman, Inc, (1997).
13. G. E. Legge *J. Opt. Soc. Am.*, **69,** pg. 838 (1979).
14. J. Nachmias and R. V. Sansbury, Vis. Res. **14,** pg. 1039 (1974).

Biography

Jian Yang received a BS degree in physics from Fudan University in 1982, an MS degree in optics from the Shanghai Institute of Optics and Fine Mechanics in 1984, and a PhD degree in experimental psychology from Northeastern University in 1991. He previously worked at the University of Rochester and the University of Houston. He joined Eastman Kodak Company in 1998. His current research is to apply human vision knowledge to image and information processing.

Michael E. Miller received BS and MS degrees in Industrial and Systems Engineering from Ohio University and a PhD in Industrial and Systems Engineering from Virginia Tech. He has been employed as a human factors engineer at IBM and joined Eastman Kodak Company as a human factors engineer in 1994. He currently leads a group of scientists interested in applying knowledge of human perception and cognition to imaging system design.