A PERCEPTUAL DISTORTION METRIC FOR DIGITAL COLOR IMAGES

Stefan Winkler

Signal Processing Laboratory
Swiss Federal Institute of Technology
1015 Lausanne, Switzerland
http://ltswww.epfl.ch/~winkler/
Stefan.Winkler@epfl.ch

ABSTRACT

This paper presents a comprehensive distortion metric for digital color images. It is based on a normalization model of the human visual system that incorporates color perception. The model is shown to accurately fit psychophysical contrast sensitivity data as well as intra- and inter-channel contrast masking data from several different psychophysical experiments. The output of the metric is compared with subjective data for natural images.

1. INTRODUCTION

Compression of digital images has been studied for many years. However, algorithms have been (and still are) designed and optimized based on simple error measures like peak signal-to-noise ratio (PSNR) despite the fact that such measures correlate poorly with perceived quality. The problem is that PSNR does not take into account viewing conditions and the behavior of the human visual system. Subjective tests on the other hand, undoubtedly the "ultimate truth" about image quality, are complex and time-consuming and hence do not lend themselves to applications where automation is desired.

These problems necessitate methods of objective quality assessment. Ideally, such a system would perceive and measure impairments just like a human being. Considering the variety of compression algorithms available and the rapid change of technology in this field, a distortion metric that is independent of the particular algorithm is preferable. Therefore metrics based on models of the human visual system like the one introduced here are the most general and potentially the most accurate ones.

The vision model presented here is based on experimental findings about the following aspects of visual information processing in the brain:

- color perception and the theory of opponent colors, which defines a color space whose principal components (black-white, red-green and blue-yellow) correspond to the pathways in the brain [8, 9];
- the response properties of neurons in the primary visual cortex [1, 3];
- the sensitivity to spatial patterns, i.e. contrast sensitivity and contrast masking experiments [2, 6, 12].

The latter two are the basis for a gain control or normalization model of early vision, which is described below.

Pattern sensitivity is of paramount importance for perceptual distortion measurements. In a typical experiment, the contrast threshold for the detection of a particular stimulus (usually a sinusoid or a Gabor patch) is measured. To investigate visual masking, such a stimulus (also referred to as target) is superimposed on a background stimulus (the masker). With some exceptions, the target is harder to see as the contrast of the masker increases. In the context of image processing and compression, the coding noise takes the place of the target, and the original image becomes the masker. Hence, a model that is able to predict pattern sensitivity is of obvious value for these applications.

Based on analyses of the response properties of neurons in the primary visual cortex [1, 3], models of early vision have been proposed that explain both contrast sensitivity and contrast masking, even when masker and target have different orientations [2, 5, 13, 17], which is a case of inter-channel masking. The vision model presented here combines these aspects with a theory of color perception.

2. VISION MODEL

The basic structure of the model is shown in Figure 1. After conversion to opponent-colors space, each of the resulting three components is subjected to a perceptual decomposition, yielding a number of perceptual channels, which subsequently undergo normalization. Both the reference image and the processed image are used as input to the model and go through the same processing steps. The final detection stage performs pooling over sensor outputs to compute a distortion measure. Each of these steps is explained in more detail below.

2.1. Conversion to Opponent-Colors Space

Digital images are usually coded in RGB color space, which cannot be used directly within a vision model, because it is neither perceptually linear nor device-independent. Therefore, a conversion to a perceptually more meaningful color space is necessary. Device independence can be achieved by determining the absorption rates of the three types of cones on the retina. These cone absorption rates can be calculated from the frame-buffer values via a linear transform

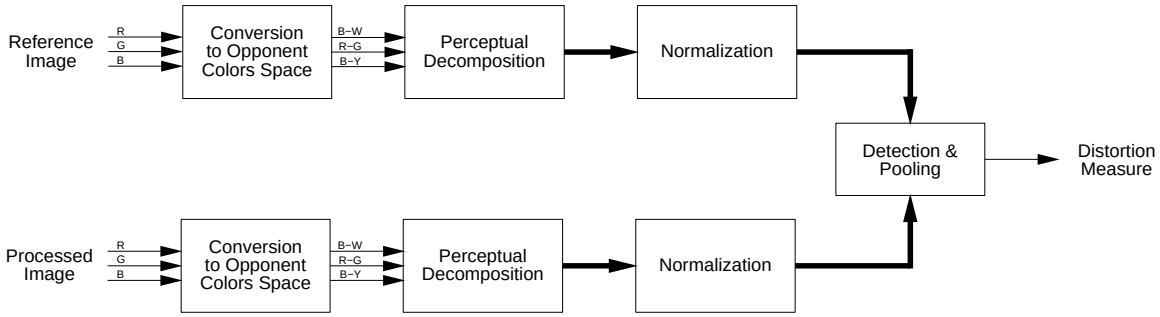

Figure 1: Block diagram of the metric. After conversion to opponent-colors space, each of the resulting three components is subjected to a perceptual decomposition, yielding several perceptual channels, which subsequently undergo normalization. The final detection stage performs pooling over sensor outputs to compute a distortion measure.

taking into consideration the spectral power distribution of light emitted from the display phosphors and the spectral sensitivities of the cones [15].

Via another linear transform, these cone absorption rates are then converted to opponent-colors space, whose principal components black-white (B-W), red-green (R-G) and blue-yellow (B-Y) are perceptually orthogonal. In accordance with recent psychophysical experiments, which suggest that the human visual representation of simple colored patterns is also pattern-color separable, the specific opponent-colors space employed here is based on the model by Poirson and Wandell [8, 9]. The spectral sensitivity of each component of this color space is shown in Figure 2.

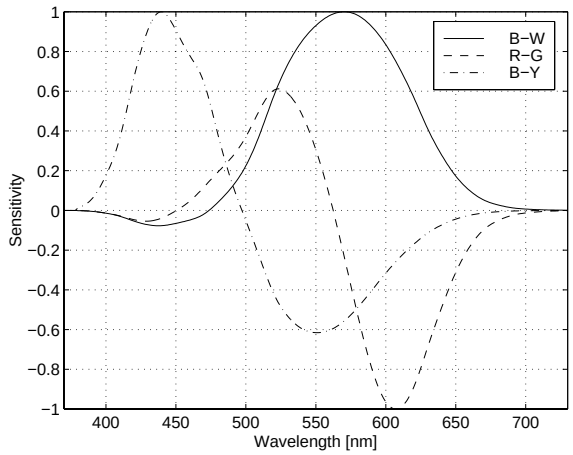

Figure 2: Spectral sensitivities of the three components black-white (solid), red-green (dashed) and blue-yellow (dot-dashed) of the pattern-color separable opponent-colors space derived by Poirson and Wandell [8, 9].

2.2. Perceptual Decomposition

The steerable pyramid transform [10,11] is used for the perceptual decomposition. This transform decomposes the image into a number of spatial frequency levels, each of which is further divided into a set of orientation bands. The basis functions are directional derivative operators.[1] The orientation decomposition is steerable, i.e. the response of a filter tuned to an arbitrary orientation can be obtained through a linear combination of the responses of the basis filters at that location. This property is important as it implies that the decomposition is rotation-invariant and self-inverting and minimizes the amount of aliasing in the subbands [10].

In the present implementation, the basis functions have octave bandwidth and octave spacing; four orientation bands are used (see Figure 3). For the B-W channel, four levels (with four orientations each) plus one (isotropic) lowpass are used. For the R-G and B-Y channels, two additional levels are needed in order to be able to capture the empirical fact that the sensitivity of the chromatic channels does not decrease significantly for lower spatial frequencies, as opposed to the luminance channel [6].

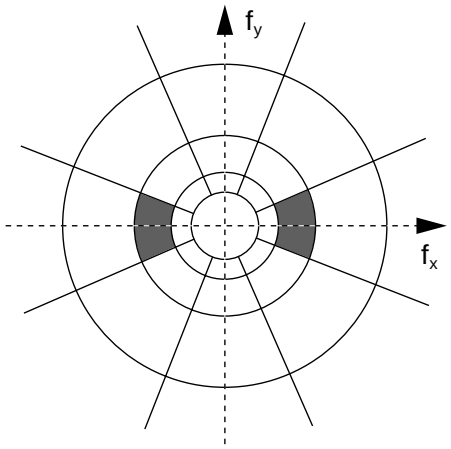

Figure 3: Idealized illustration of the partition of the spatial frequency plane by the steerable pyramid. Three levels plus one (isotropic) lowpass filter are shown. The bands at each level are tuned to orientations of 0, 45, 90 and 135 degrees. The shaded region indicates the spectral support of a single subband.

[1] Source code and filter kernels for the implementation of the steerable pyramid transform are available at http://www.cis.upenn.edu/~eero/steerpyr.html.

2.3. Normalization

The perceptual decomposition yields a large number of coefficients for the image. First a power-law nonlinearity is applied to these coefficients, then they are normalized in order to account for the limited dynamic range of visual sensors as well as intra- and inter-channel masking effects [13].

Let $a = a(c, f, \theta, x, y)$ be a coefficient of the perceptual decomposition for color channel c at location x, y in frequency band f and orientation band θ. Then the corresponding normalized sensor output $s = s(c, f, \theta, x, y)$ is computed as

$$s = k\frac{a^p}{\sigma^2 + a^q * h}. \tag{1}$$

The excitatory path in the numerator consists of a simple power-law nonlinearity with exponent p. The inhibitory path in the denominator controls the gain of the excitatory path. It also includes a nonlinearity with a possibly different exponent q. Additionally, filter responses are pooled over different channels in the inhibitory path by virtue of a convolution with the pooling function h. In its most general form, this pooling operation may combine coefficients from the dimensions of color, frequency, orientation, space (x and y) and phase. For this paper, complete summation over phase is assumed, and spatial pooling is omitted as there is evidence that it is in fact very localized [17]. For further simplification, pooling over frequency and color channels is disregarded. A Gaussian pooling kernel is used for pooling over orientation channels. The saturation constant σ is added to prevent division by zero; k is used to adjust the overall gain of the mechanism.

In Teo and Heeger's implementation [13], which is based on a direct model of neural cell responses [3], the exponents of both the excitatory and inhibitory nonlinearity are fixed at $p = q = 2$ so as to be able to work with local energy measures. However, this procedure leads to a rapid saturation of sensor outputs, which is why they have to use multiple contrast bands (i.e. several different k's and σ's) for all coefficients in order to cover the full range of contrasts. This leads to an undesirable increase in the number of model parameters and complicates the fitting process.

Watson and Solomon [17] showed that the response of such a combination of several rapidly saturating mechanisms can be approximated satisfactorily by a single non-saturating mechanism with an excitatory exponent $p > q$. This approach is chosen here as well because it reduces the number of parameters significantly and results in a more efficient implementation.

2.4. Detection and Pooling

The final stage computes a Minkowski sum of the difference between the sensor outputs from the reference image $s_0 = s_0(c, f, \theta, x, y)$ and the sensor outputs from the processed image $s_1 = s_1(c, f, \theta, x, y)$:

$$\Delta s = \sqrt[\beta]{\sum |s_0 - s_1|^\beta}. \tag{2}$$

Δs is a measure of the amount of perceptual distortion in the image. If desired, a transformation such as

$$Q = \frac{5}{1 + K\Delta s} \tag{3}$$

can be used to obtain a measure of perceptual quality on the common scale from 0 to 5 (low to high quality). K has to be chosen on the basis of the vision model to ensure the proper mapping [5].

3. RESULTS

3.1. Model Fitting

The model contains several parameters that have to be adjusted in order to accurately represent the human visual system. k, the gain factor, p, the excitatory exponent, and σ, the saturation constant, are used as parameters in the fitting process. q, the inhibitory exponent, is fixed at $q = 2$ in this implementation, as it is mainly the difference $p - q$ which is relevant [17].

An exponent of $\beta = 2$, i.e. a simple squared-error norm is used in the final detection stage of the metric. This choice can be justified in terms of an ideal observer model. The sensor outputs can be considered as the mean values of noisy sensors. Assuming an additive, independent, identically distributed Gaussian noise with zero mean and a standard deviation independent of the sensor outputs, a squared-error norm detection stage gives the probability that the ideal observer would detect the distortion. If a standard deviation of 1 is assumed for the noise, the ideal observer is able to detect the distortion with a probability of 76% when the squared-error norm is equal to one [13]. Assuming this detection efficiency for the model, $\Delta s = 1$ at threshold.

Contrast sensitivity is modeled by setting the gains of the filters in the steerable pyramid in such a way that the ensemble produces the empirical contrast sensitivity curve. For the B-W channels, the gains are chosen so as to match contrast sensitivity threshold data for sinusoidal gratings from Kelly [4]. For the R-G and B-Y channels we rely on similar data from Mullen [6], as demonstrated in Figure 4.

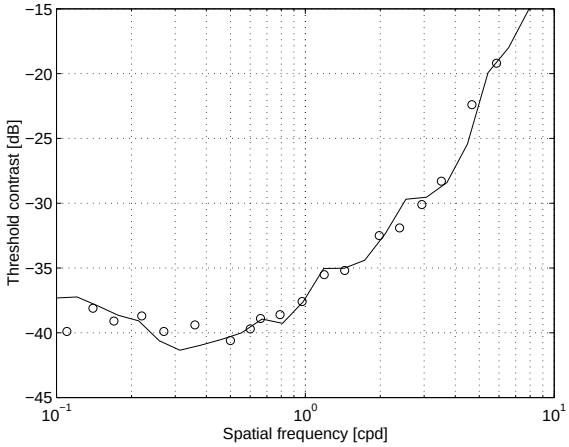

Figure 4: Contrast sensitivity data for blue-yellow gratings [6] as threshold contrast vs. spatial frequency in cycles per degree; the solid curve shows the model fit.

The parameters k, p and σ are determined by fitting the model's responses to masked gratings. For the B-W (luminance) channel, empirical data from several contrast masking experiments conducted by Foley [2] is used, where the contrast thresholds of Gabor targets masked by sinusoidal gratings of different orientations were measured. The parameters k, p and σ of the R-G and B-Y channels are adjusted to fit similar data presented by Switkes et al. [12], as shown in Figure 5 for the R-G channel.

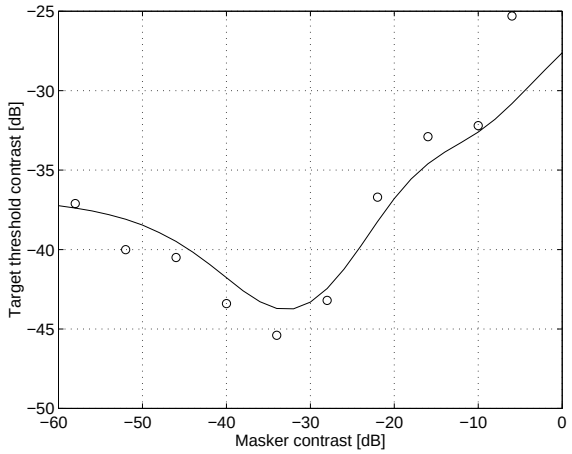

Figure 5: Contrast masking data for red-green gratings [12] as target threshold contrast vs. masker contrast; the solid curve shows the model fit.

The simulation results shown in the figures demonstrate that the overall quality of the fit is very good and close to the difference between measurements from different observers; all of the effects found in the above experiments are captured by the model.

3.2. Validation

The database used for validation consists of distorted versions of a 320×400 color image showing the face of a child surrounded by colorful balls. The original was compressed with the JPEG algorithm, and the coding noise was determined in YUV space by computing the difference between the original and the compressed image. Subsequently, the coding noise was scaled by some factor ranging from -1 to 1 in the Y, U and V channels, respectively, and was then added to the original in order to obtain the distorted images. Several test series were created by varying the noise intensity along specific directions in YUV space [14].

Psychophysical data was collected for two subjects using a QUEST procedure [16]. The subjects were shown the original image together with two test images, one of which was the distorted image, and the other one the original. Subjects had to identify the distorted image, and the percentage of correct answers was recorded. The responses for one test series are shown in Figure 6.

Such data can be modeled by the psychometric function

$$P(C) = 1 - 0.5e^{-(x/a)^b}, \qquad (4)$$

where $P(C)$ is the probability of a correct answer and x is the stimulus strength; a and b determine the midpoint and the slope of the function [7]. These two parameters are estimated from the psychophysical data; the variable x is equivalent to the distortion measure Δs in this procedure. The resulting function is used to map the distortion measure onto the "% correct"-scale. Figure 6 shows the results obtained in such a manner for one test series. A comparison of the estimated detection thresholds of the two subjects for each test series with the predictions of the metric is shown in Figure 7. It can be seen that the predictions of the proposed distortion metric are quite accurate for most of the test series. The remaining discrepancies can be explained by the fact that the subjective data for some test series were relatively noisy, making it harder to compute a reliable estimate of the detection threshold.

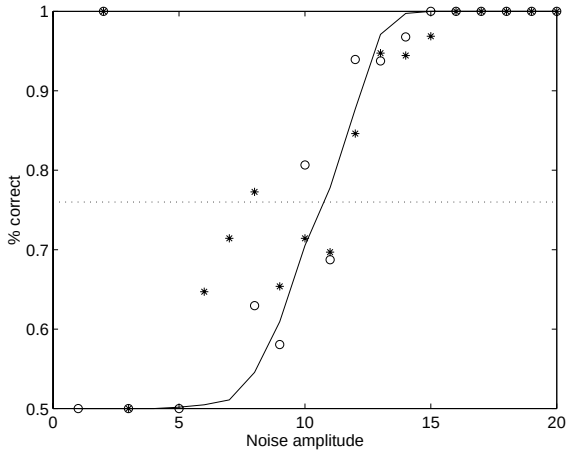

Figure 6: Percentage of correct answers vs. noise amplitude for the two subjects (circles and stars) and the transformed distortion measure (solid curve).

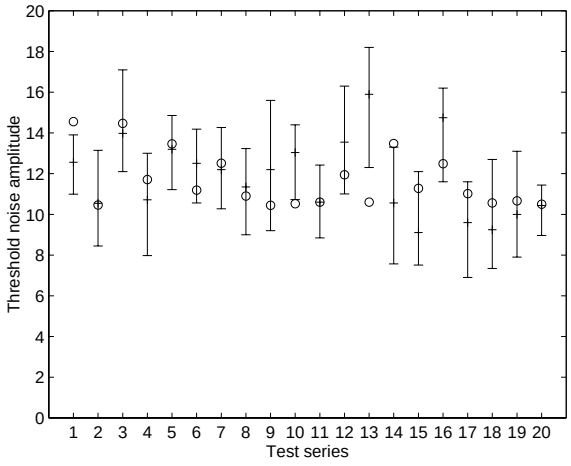

Figure 7: Estimated detection thresholds of the two subjects for each test series (error bars) compared to the metric's predictions (circles).

4. CONCLUSIONS

A distortion metric for digital color images based on a model of the human visual system has been presented. The results show that the model can fit psychophysical contrast sensitivity and contrast masking data accurately and performs well with respect to subjective data for natural images. Future research will include extended inhibitory pooling in order to study the influence of interactions between chrominance and luminance channels. More extensive tests should reveal the gain in prediction accuracy of a color model over a simple luminance model.

5. REFERENCES

[1] Duane G. Albrecht, Wilson S. Geisler: "Motion selectivity and the contrast-response function of simple cells in the visual cortex." in *Visual Neuroscience*, vol. 7, pp. 531–546, 1991.

[2] John M. Foley: "Human luminance pattern-vision mechanisms: Masking experiments require a new model." in *Journal of the Optical Society of America A*, vol. 11, no. 6, pp. 1710–1719, June 1994.

[3] David J. Heeger: "Normalization of cell responses in cat striate cortex." in *Visual Neuroscience*, vol. 9, pp. 181–197, 1992.

[4] D[onald] H. Kelly: "Motion and vision. I. Stabilized images of stationary gratings." in *Journal of the Optical Society of America*, vol. 69, no. 9, pp. 1266–1274, September 1979.

[5] Pär [J.] Lindh, Christian J. van den Branden Lambrecht: "Efficient spatio-temporal decomposition for perceptual processing of video sequences." in *Proceedings of the International Conference on Image Processing*, vol. 3, pp. 331–334, Lausanne, Switzerland, September 16–19, 1996.

[6] Kathy T. Mullen: "The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings." in *Journal of Physiology*, vol. 359, pp. 381–400, 1985.

[7] Jacob Nachmias: "On the psychometric function for contrast detection." in *Vision Research*, vol. 21, pp. 215–223, 1981.

[8] Allen B. Poirson, Brian A. Wandell: "Appearance of colored patterns: Pattern-color separability." in *Journal of the Optical Society of America A*, vol. 10, no. 12, pp. 2458–2470, December 1993.

[9] Allen B. Poirson, Brian A. Wandell: "Pattern-color separable pathways predict sensitivity to simple colored patterns." in *Vision Research*, vol. 36, no. 4, pp. 515–526, 1996.

[10] Eero P. Simoncelli, William T. Freeman, Edward H. Adelson, David J. Heeger: "Shiftable multi-scale transforms." in *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 587–607, March 1992.

[11] Eero P. Simoncelli, William T. Freeman: "The steerable pyramid: A flexible architecture for multi-scale derivative computation." in *Proceedings of the International Conference on Image Processing*, pp. 444–447, Washington, DC, October 23–26, 1995.

[12] Eugene Switkes, Arthur Bradley, Karen K. De Valois: "Contrast dependence and mechanisms of masking interactions among chromatic and luminance gratings." in *Journal of the Optical Society of America A*, vol. 5, no. 7, pp. 1149–1162, July 1988.

[13] Patrick C. Teo, David J. Heeger: "Perceptual image distortion." in *Proceedings of SPIE Human Vision, Visual Processing and Digital Display*, vol. 2179, pp. 127–141, San Jose, CA, February 8–10, 1994, ISBN 0-8194-1474-3.

[14] Christian J. van den Branden Lambrecht, Joyce E. Farrell: "Perceptual quality metric for digitally coded color images." in *Proceedings of the European Signal Processing Conference*, pp. 1175–1178, Trieste, Italy, September 10–13, 1996.

[15] Brian A. Wandell: *Foundations of Vision*. Sinauer Associates, 1995, ISBN 0-87893-853-2.

[16] Andrew B. Watson, Denis G. Pelli: "QUEST: A Bayesian adaptive psychometric method." in *Perception & Psychophysics*, vol. 33, no. 2, pp. 113–120, 1983.

[17] Andrew B. Watson, Joshua A. Solomon: "Model of visual contrast gain control and pattern masking." in *Journal of the Optical Society of America A*, vol. 14, no. 9, pp. 2379–2391, September 1997.