PERCEPTUAL IMAGE DISTORTION

Patrick C. Teo and David J. Heeger

Department of Computer Science and Department of Psychology
Stanford University, Stanford, CA 94305

ABSTRACT

In this paper, we present a perceptual distortion measure that predicts image integrity far better than mean-squared error. This perceptual distortion measure is based on a model of human visual processing that fits empirical measurements of the psychophysics of spatial pattern detection. The model of human visual processing proposed involves two major components: a steerable pyramid transform and contrast normalization. We also illustrate the usefulness of the model in predicting perceptual distortion in real images.

1. INTRODUCTION

A variety of imaging and image processing methods are based on measures of image integrity (distortion measures). Examples include: image data compression, dithering algorithms, flat-panel display and printer design. In each of these cases, the goal is to reproduce an image that looks as much as possible like the original.

It has long been accepted that distortion measures like mean-squared error (MSE) are inaccurate in predicting perceptual distortion. For example, in the context of image data compression, a number of methods have exploited the human visual system's insensitivity to higher spatial frequencies to achieve higher compression rates than schemes that simply used MSE as their distortion measure. Recently, some researchers have found that they were able to tolerate coarser quantization in areas of higher "texture energy" and still achieve the same perceptual results [12, 11]. However, these techniques have often been rather ad hoc. A notable exception is recent work by Watson [15] that is based on psychophysical masking data.

In this paper, we present a perceptual distortion measure that predicts image integrity far better than MSE. Our distortion measure is a generalization of the model used by Watson [15], and it encompasses the "texture energy" masking phenomenon mentioned above [11, 12].

Our perceptual distortion measure is based on fitting empirical measurements of: (1) the response properties of neurons in the primary visual cortex (also called visual area V1), and (2) the psychophysics of spatial pattern detection, that is, peoples' ability to detect a low contrast visual stimulus. We will present only the latter in this paper.

It is important to recognize the relevance of these empirical spatial pattern detection results to developing measures of image integrity. In a typical spatial pattern detection experiment, the contrast of a visual stimulus (called the target) is adjusted until it is just barely detectable. In some experiments (called masking experiments), the target is also superimposed on a background (called the masker). Again, the contrast of the target is adjusted (while the masker contrast is held fixed) until the target is just barely detectable. Typically, a target is harder to detect (i.e., a higher contrast is required) in the presence of a high contrast masker. A model that predicts spatial pattern detection is obviously useful in image processing applications. In the context of image compression, for example, the target takes the place of quantization error and the masker takes the place of the original image.

In recent years, we and others have developed a nonlinear model of early vision, hereafter referred to as the *normalization model*, to explain a large body of data [2, 3, 6, 7, 8]. The normalization model explains three general classes of spatial pattern detection results. First, it explains baseline contrast sensitivity (detection of a target when there is no masker). Second, the model explains the usual phenomena of contrast masking (when the target and masker have the same orientation). Third, and unlike previous models of spatial masking (like that used by Watson [15]), the normalization model explains the masking effect that occurs when the target and masker have very different orientations.

2. THE MODEL

The model consists of four stages: (1) front-end linear filtering, (2) squaring, (3) normalization, and lastly (4) detection. The first stage of the model decomposes the image locally into its spatial frequency and orientation components. The coefficients of the linear filtering are then squared to yield local energy measures. Because the human visual system is differentially sensitive to local image frequency composition, the third stage normalizes the squared coefficients accordingly. Both the reference and distorted images are subjected to the first three stages; the final detection stage then determines the amount of distortion visible in the distorted image.

2.1. Linear Transform

Many researchers have suggested a variety of linear transforms which resemble the orientation and spatial frequency tuning of cortical receptor fields [10, 9, 16] or psychophysically determined visual sensors [5]. These linear transforms

This research was supported by NIMH grant 1-R29-MH50228-01 and by NASA grant NCC2-307.

0-8186-6950-0/94 $4.00 © 1994 IEEE

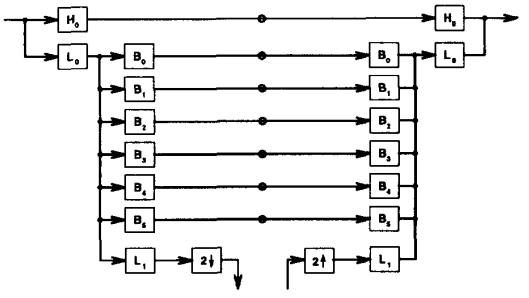

Figure 1: Analysis/synthesis representation of the steerable pyramid transform implemented. H_0 is a high-pass filter; the L_i's represent low-pass filters and the B_i's represent orientation selective filters.

often have the following characteristics: (1) octave spacing and frequency bandwidths and (2) narrow orientation selectivity. In addition to these considerations, we are also concerned about the computational efficiency of the transform (and its inverse).

In our previous work, we used a quadrature mirror filter suite on a hexagonally-sampled image [14]. The transform is orthogonal and thus is compact in its representation and efficiently computed. Unfortunately, being orthogonal, the basis functions describing a local region of an image severely alias one another.[1] Moreover, as noted previously, the orientation bandwidth of the hex-QMF's are a little too broad.

In this paper, we adopt the steerable pyramid transform introduced by Simoncelli *et al* [13]. The transform decomposes the image locally into several spatial frequency levels within which each level is further divided into a set of orientation bands. Figure 1 shows an analysis/synthesis representation of the transform. The basis functions for each level of the pyramid have octave bandwidths and are separated from those of neighboring levels by an octave as well. In our implementation, we divide every level into six orientation bands with bandwidths of approximately thirty degrees. The orientation decomposition at each level is steerable [4], i.e. the response of a filter tuned to any orientation can be obtained through a linear combination of the responses of the six basis filters computed at the same location. This property is important as it implies that the orientation decomposition is locally rotationally-invariant. The pyramid is also designed to minimize the amount of aliasing within each subband. Thus, the steerable pyramid, unlike the hex-QMF transform, is overcomplete and non-orthogonal. Even so, the transform is self-inverting which allows the inverse to be efficiently computed despite its non-orthogonality.

2.2. Squaring and Normalization

The front-end linear transform yields a set of coefficient values for every region in the image. These coefficients are next squared to obtain energy measures of the local orientation and spatial frequency components.

[1] In quadrature mirror filters, the aliasing introduced during subsampling is cancelled only during reconstruction.

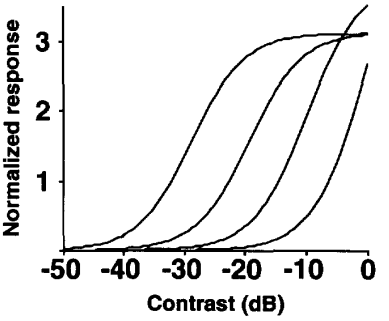

Figure 2: Response of the four normalized sensors as a function of log contrast for a sinusoidal grating image. The responses saturate for high contrasts. The dynamic range (the range of contrasts for which this sensor can discriminate contrast effectively) of each sensor is limited to the rising portion of each curve.

Since the front-end transform is linear, a coefficient's magnitude increases linearly with the contrast of the input image. Furthermore, these linear coefficients are equally sensitive (or insensitive) to perturbations of the input regardless of image contrast. Squaring introduces a simple contrast-dependence on sensitivity. However, squaring alone does not account for masking effects. Furthermore, the magnitude of the response of each sensor can potentially be very large. On the other hand, the dynamic range of the mechanisms in the visual system is limited. Normalization is required to predict masking effects and to restrict the range of response magnitudes of our hypothetical visual sensors.

The normalization scheme is divisive and is determined by two parameters: an overall scaling constant, k, and a saturation constant, σ^2. Let A^θ be a coefficient of the front-end linear transform. The squared and normalized output, R^θ, is computed as follows:

$$R^\theta = k \frac{(A^\theta)^2}{\sum_\phi (A^\phi)^2 + \sigma^2} \quad (1)$$

where ϕ ranges over all sensors tuned to different orientations. In our implementation, $\phi \in \{0, 30, 60, 90, 120, 150\}$.

We treat each spatial frequency level of the pyramid separately and conduct this pooling only over sensors tuned to different orientations. Hence, the normalized output of a sensor tuned to orientation θ is computed by dividing its original squared response, $(A^\theta)^2$, by the sum of the squared responses of a pool of sensors over all orientations in the same region of the image. Since this summation, $\sum_\phi (A^\phi)^2$, *includes* the term, A^θ, that appears in the numerator (i.e., each sensor suppresses itself), as long as σ is nonzero, the normalized sensor response will always be a value between 0 and k, saturating at high contrasts.

Each of the normalized sensors has a limited dynamic range as shown in Figure 2. In other words, each sensor is able to discriminate contrast differences only over a narrow range of contrasts. This range is determined by the scaling and saturation constants, k and σ^2, respectively. Hence,

several contrast normalization mechanisms, each having different k_i's and σ_i^2's, are required to discriminate contrast changes over the full range of contrasts. In the current implementation of the model, we have four different contrast discrimination bands (that is, four different choices of σ_i^2 and k_i).

In summary, the front-end linear transform yields a set of coefficients which measure the different orientation and spatial frequency components in each local region of the image. With squaring and multiple normalizations, the number of measurements for each local region is increased fourfold. However, these local image measurements now analyze the image into its orientation, spatial frequency and contrast components. Furthermore, masking effects over orientation are captured by the pooling step in the normalization.

2.3. Detection

The detection mechanism determines locally if a distortion is visible. Let $\mathcal{R}_{ref}$ be a vector of normalized sensor responses from a local region in the reference image. Let $\mathcal{R}_{dist}$ be the vector of normalized responses from the corresponding region in the distorted image. The detection mechanism adopted by the model is the simple squared-error norm (i.e., the vector distance between $\mathcal{R}_{ref}$ and $\mathcal{R}_{dist}$):

$$\triangle\mathcal{R} = ||\mathcal{R}_{ref} - \mathcal{R}_{dist}||^2 \qquad (2)$$

One might include all of the normalized sensor responses (all spatial positions, spatial frequencies, orientations, and contrast discrimination bands) in the vectors, $\mathcal{R}_0$ and $\mathcal{R}_1$, and compute a single number representing the overall detectability of differences between the two images. We find it more informative, however, to implement the detection mechanism independently for each local patch (or block) of the images.

The vector distance detection mechanism can be justified in terms of an ideal observer model.[2] The vector of normalized sensor responses from each image correspond to the mean responses of noisy sensors. For the purposes of the model, we assume the noise to be additive, independent, identically-distributed, zero-mean Gaussian noise. Furthermore, the standard deviation of the noise is independent of the mean response. With these assumptions and an ideal observer model, the vector distance detection mechanism gives the likelihood that the ideal observer would detect the distortion. For example, assuming a standard deviation of one for the noise, the observer is able to detect the distortion 76% of the time when the squared difference is exactly one. In signal detection theory, this corresponds to a d' of one. In our model, we assume detection at this efficiency. Hence, $\triangle\mathcal{R}$ in equation (2) is equal to one at threshold.

[2] An ideal observer is assumed to have knowledge of the joint probability distribution of all its mechanisms in response to each stimulus. It then makes its decision so as to minimize its probability of error.

3. RESULTS

3.1. Model Fitting

The parameters of the model were fit to psychophysical data from spatial pattern detection experiments [3, 1]. In particular, data on contrast and orientation masking were measured by Foley and Boynton [3]. The task in their experiments was to detect a target Gabor pattern superimposed on a sinusoidal masker pattern. Target threshold contrast versus masker contrast (TvC) curves for several masker orientations were obtained from these experiments. When the masker and target have the same orientation, the TvC curve characterizes the sensitivity of the visual system to the target as a function of contrast, a phenomenon known as contrast masking. With different masker and target orientations, the set of TvC curves together record the effect of orientation masking. Baseline sensitivity to the target is captured by all the TvC curves when masker contrast is zero.

These TvC curves were used to tune the performance of the model. The only free parameters of the model are the pairs of scaling and saturation constants (k_i and σ_i). We found experimentally that four was the minimum number of pairs required to fit the data. Since each level of the steerable pyramid has six orientation bands, there are a total of twenty-four normalized sensor responses at each spatial position. The task, therefore, was to pick the k_i's and σ_i's such that the model accurately predicts the empirical TvC curves at various masker orientations.

Figure 3 shows the result of fitting the model to empirical TvC data at different masker orientations. The fit to 0-degree masker orientation TvC data (0-degree TvC data) is extremely good. The overall goodness of fit indicates that four contrast tuning mechanisms are sufficient to reproduce the properties described by the TvC data.

One important characteristic of the TvC data is the presence (or absence) of a "dipper". The presence of a dipper indicates that within that range of masker contrasts, the masker facilitates the detection of the target. A pronounced dipper can be observed in the 0-degree and 11.25-degree TvC data indicating that facilitation occurs at low contrasts for similarly oriented stimuli. The dipper is almost absent in the 22.5-degree TvC data (not shown) and completely absent in TvC data involving greater orientation differences.

3.2. Demonstration

In order to assess the model, we added bandpass distortion to a reference image and then computed the perceptual distortion between the original and distorted images. In particular, we distorted the second level of the steerable pyramid in two ways: first to maximize perceptual distortion, and second to minimize it. Figure 4 shows the original Einstein image along with the two distorted images. The perceived distortion is very different, yet the distortion was added so that standard distortion measures (mean squared error and peak signal-to-noise ratio) were very nearly the same. These standard distortion measures are, therefore, poor predictors of the perceived distortion.

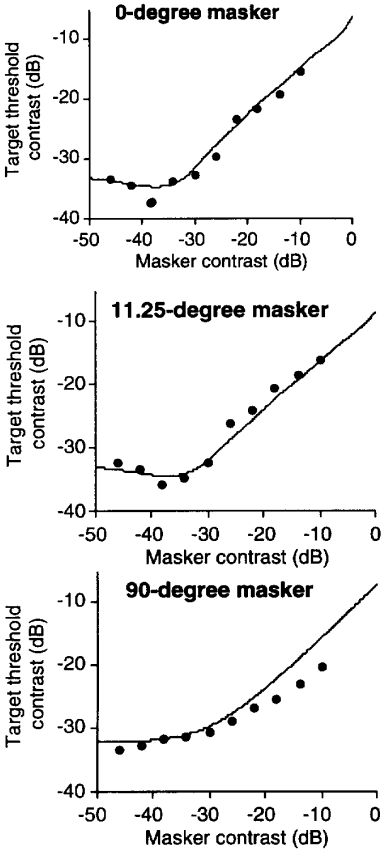

Figure 3: Result of fitting the model to empirical data. Empirical data are denoted by filled circles. Solid curves denote predicted target threshold contrasts.

Figure 4 also shows the perceptual distortion images for the minimally and maximally distorted Einstein images. Darker regions correspond to areas of lower perceptual distortion while brighter regions indicate areas of greater perceptual distortion.

4. CONCLUSION

We have described a model of perceptual distortion that is consistent with spatial pattern psychophysics. In particular, we have shown that the model explains both contrast and orientation masking. However, the accuracy of the model with regard to spatial frequency masking has not been determined. Also, the effects of spatial summation and different mean luminances have not been explored. While the detection mechanism can be explained in terms of an ideal observer model, its accuracy in predicting empirical psychometric curves needs to be verified as well. These modeling issues and practical concerns such as demonstrating the usefulness of the model in applications like image data compression we defer to future work.

5. REFERENCES

[1] Ahumada and Peterson. Luminance-model-based dct quantization for color image compression. In *Human Vision, Visual Processing and Digital Display III, SPIE*, pages 365–374, 1992.

[2] D G Albrecht and W S Geisler. Motion sensitivity and the contrast-response funtion of simple cells in the visual cortex. *Visual Neuroscience*, 7:531–546, 1991.

[3] J M Foley and G M Boynton. A new model of human luminance pattern vision mechanisms: Analysis of the effects of pattern orientation, spatial phase, and temporal frequency. In T A Lawton, editor, *Computational Vision Based on Neurobiology, SPIE Proceedings, volume 2054*, 1994.

[4] W T Freeman and E H Adelson. The design and use of steerable filters. *IEEE Pattern Analysis and Machine Intelligence*, 13:891–906, 1991.

[5] N Graham. *Visual Pattern Analyzers*. Oxford University Press, 1989.

[6] D J Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9:181–198, 1992a.

[7] D J Heeger. Half-squaring in responses of cat simple cells. *Visual Neuroscience*, 9:427–443, 1992b.

[8] D J Heeger. Modeling simple cell direction selectivity with normalized, half-squared, linear operators. *Journal of Neurophysiology*, 70:1885–1898, 1993.

[9] J P Jones and L A Palmer. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58:1233–1258, 1987c.

[10] S Marcelja. Mathematical description of the response of simple cortical cells. *Journal of the Optical Society of America A*, 70:1297–1300, 1980.

[11] T R Reed, V R Algazi, G E Ford, and I Hussain. Perceptually based coding of monochrome and color still images. In *Data Compression Conference*, pages 142–151, 1992.

[12] R J Safranek, J D Johnston, N S Jayant, and C Podilchuk. Perceptual coding of image signals. In *Twenty-fourth Asilomar Conference on Signals, Systems and Computers*, pages 346–350, 1990.

[13] E P Simoncelli, W T Freeman, E H Adelson, and D J Heeger. Shiftable multi-scale transforms. *IEEE Transactions on Information Theory, Special Issue on Wavelets*, 38:587–607, 1992.

[14] P Teo and D J Heeger. Perceptual image distortion. In *Proceedings of SPIE, volume 2179*, pages 127–141, San Jose, CA, Feb 1994.

[15] A B Watson. Visually optimal DCT quantization matrices for individual images. In *Data Compression Conference*, pages 178–187, 1993.

[16] R A Young. The Gaussian derivative model for spatial vision: I. Retinal mechanisms. *Spatial Vision*, 2:273–293, 1987.

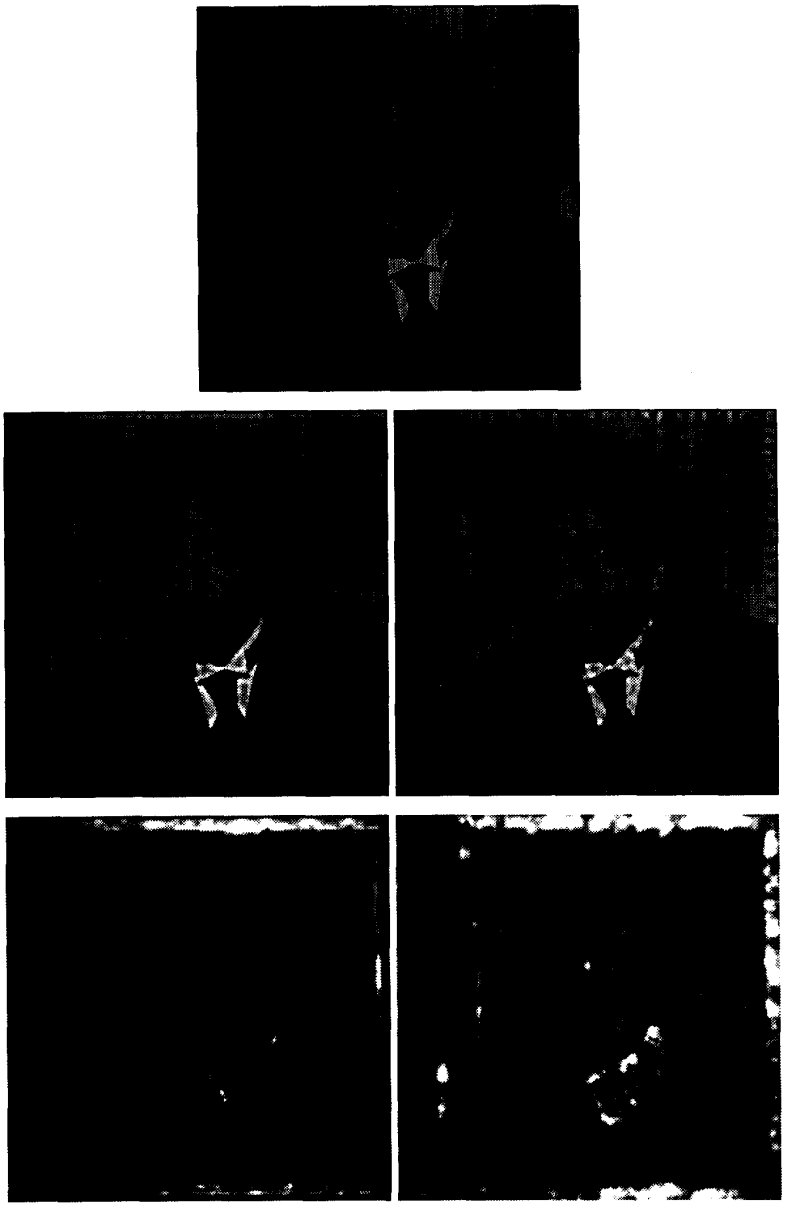

Figure 4: (Top) Original Einstein image. (Middle-left) Image was distorted so as to minimize perceptual distortion (RMSE = 9.01, peak-SNR = 29.04 dB). (Middle-right) Image was distorted so as to maximize perceptual distortion (RMSE = 8.50, peak-SNR = 29.54 dB). Both distorted images have nearly identical mean-squared error and peak-SNR. The overall perceptual-distortion-measures for the left and right images are 3.59 and 4.64 respectively. (Bottom-left) Perceptual distortion measured from the minimally distorted image. Darker regions correspond to areas of lower perceptual distortion while brighter regions indicate areas of greater perceptual distortion. (Bottom-right) Perceptual distortion measured from the maximally distorted image.