Visible differences predictor: an algorithm for the assessment of image fidelity

Daly, Scott

Scott J. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," Proc. SPIE 1666, Human Vision, Visual Processing, and Digital Display III, (27 August 1992); doi: 10.1117/12.135952

SPIE.

Event: SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology, 1992, San Jose, CA, United States

The visible differences predictor: an algorithm for the assessment of image fidelity

Scott Daly
Imaging Group Research Laboratories, Eastman Kodak Company
Rochester, NY 14650–1816
daly@kodak.com

ABSTRACT

Image fidelity is the subset of overall image quality that specifically addresses the visual equivalence of two images. This paper describes an algorithm for determining whether the goal of image fidelity is met as a function of display parameters and viewing conditions. Using a digital image processing approach, this algorithm is intended for the design and analysis of image processing algorithms, imaging systems, and imaging media. The visual model, which is the central component of the algorithm, is comprised of three parts: an amplitude nonlinearity, a contrast sensitivity function, and a hierarchy of detection mechanisms.

1.0 INTRODUCTION

The visible differences predictor (VDP) is motivated by the need to quantitatively describe the visual consequences of decisions regarding the design and quality control of imaging products. Intended to be used in the development of image processing algorithms, imaging system hardware, and imaging media, it is a design tool that can find wide areas of application. The differences due to imaging systems may begin as *mathematical differences* (i.e., incorrect code values), but ultimately end up as *physical differences* (i.e., incorrect luminances) once the image is displayed. The goal of the VDP is to determine the degree to which these *physical differences* become *visible differences*. Commonly used techniques[1–3] analyze parameters such as the system's MTF and noise power spectra, often with a one-dimensional integration, and calculate a single number describing image quality. While these techniques perform well for many aspects of analog media, they have not been as successful for describing digital image quality, the effects of adaptive algorithms, or the nonlinear aspects of analog media. The problems with these techniques lie in their lack of phase information in the analysis, their inability to deal effectively with nonlinearities, and their simplicity relative to the complexity of the visual system.

To solve these problems, the VDP uses a digital image processing approach. The use of two-dimensional images, rather than just parameters of the imaging system, enables the preservation of phase information. This information is necessary to predict visual distortion because of the masking properties of the visual system, in which the location of the image error is as important as the magnitude. Further, nonlinearities in the media or algorithms pose problems for the current approaches that use power spectra and MTF, because of their implications of linearity. An image processing approach can easily incorporate such systems parameters as MTF and noise power spectra through simulation, yet it also allows for more exact simulation of the nonlinearities of both the media and the visual system.

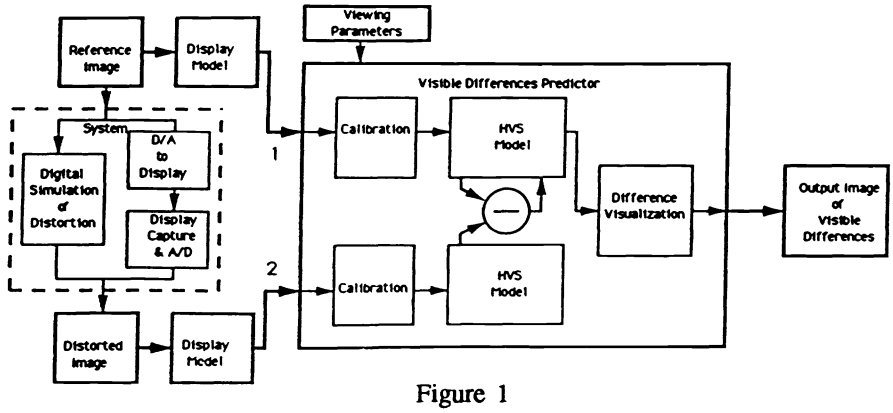

Figure 1

The VDP is a relative metric since it does not describe an absolute metric of image quality, but instead addresses the problem of describing the visibility of differences between two images. The usage of the VDP is shown in Figure 1, and the

 0-8194-0820-4/92/$4.00

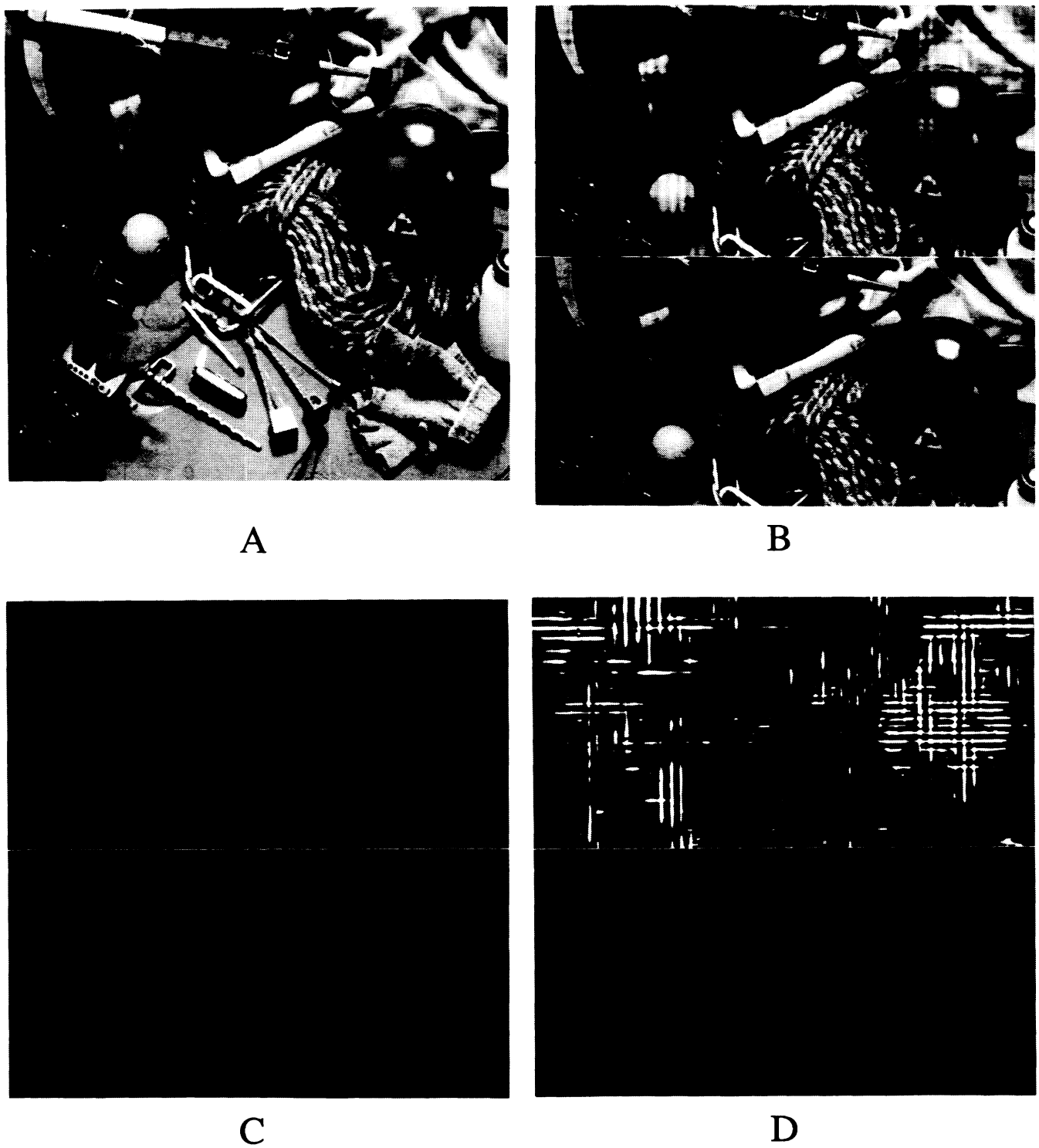

Figure 2

(A) Reference image. (B) Two distorted versions of the reference image, both having a mse of 30 (33dB SNR). The upper image has a banding distortion while the lower has a tone-scale distortion. (C) Physical error images for the banding (upper) and the tone-scale (lower) distortions. (D) The VDP results for the banding (upper) and tone-scale (lower) distortions. Full white or black indicate where the distortion has a probability of detection of 1.00.

VDP can be seen to consist of components for calibration of the input images, a human visual system (HVS) model, and a method for displaying the HVS predictions of the detectable differences. The input to the algorithm includes two images and parameters for viewing conditions and calibration, while the output is a third image describing the visible differences between them. Typically, one of the input images is a reference image, representing the image quality goal, while the other is a distorted image, representing the system's actual quality. The block components outside of the VDP generically describe the simulation of the distortion under study. The VDP is used to assess the image fidelity of the distorted image as compared to the reference. Its output image is a map of the probability of detecting the differences between the two images as a function of their location in the images. This metric, *probability of detection*, provides an accurate description of the threshold behavior of vision, but does not discriminate between different suprathreshold visual errors. The VDP can therefore be summarized as a threshold model for suprathreshold imagery, capable of quantifying the important interactions between threshold differences and suprathreshold image content and structure.

Unlike most image quality metrics, including other recent image processing based approaches[4,5], the VDP does not collapse its output into a single number, which has advantages as well as disadvantages. An advantage is that an imaging system designer can see the nature of the difference, and use this information to further improve the design. A disadvantage is that it cannot be used for suprathreshold ranking. Nevertheless, we feel this model is an initial step toward the aim of a complete suprathreshold model. As a result of the approach taken, the VDP can be used for all types of image distortions including blur, noise, banding, blocking, pixellation, algorithm artifacts, and tone-scale changes.

2.0 OVERVIEW OF THE HVS MODEL

The HVS model concentrates on the lower order processing of the visual system, such as the optics, retina, lateral geniculate nucleus, and striate cortex. The overall approach taken is to model the visual system as a number of processes that limit visual sensitivity. Without these limitations, only a physical model of the displayed image would be needed. The necessity for an HVS model can be provided by demonstrating the severe failure of a common physical error metric: mean-squared error (mse). In Figure 2A, we show a reference image and two distorted versions of this image having identical mse appear in Figure 2B. Though these images have the same physical distortion as measured by mse, the degree of visibility of the distortion is vastly different. The error images corresponding to these distortions appear in Figure 2C.

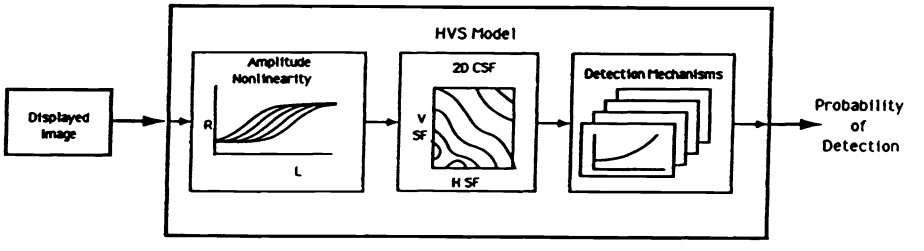

Figure 3

The HVS model addresses three main sensitivity variations. These are the variations as a function of light level, spatial frequency, and signal content. Sensitivity, S, is defined as the inverse of the contrast required to produce a threshold response,

$$S = \frac{1}{C_T} \quad , \tag{1}$$

where C_T is generally referred to as simply the threshold. The Michelson definition of contrast,

$$C = (L_{max} - L_{mean})/L_{mean} \quad , \tag{2}$$

is used, where L_{max} and L_{mean} refers to the maximum and mean luminances of the waveform, respectively. Sensitivity can be thought of as a gain, though the various nonlinearities of the visual system require caution in this analogy. The variations in sensitivity as a function of light level are primarily due to the light adaptive properties of the retina, and we shall refer to this overall effect as the *amplitude nonlinearity* of the HVS. The variations as a function of spatial frequency are due to the optics of the eye combined with the neural circuitry and these combined effects are referred to as the *contrast sensitivity function (CSF)*. Finally, the variations in sensitivity as a function of signal content are due to the post-receptoral neural circuitry, and these effects are referred to as *masking*. The HVS model consists of three main components that essentially model each of these sensitivity variations. In the current state of development of the VDP, these three components are sequentially

cascaded in the straightforward manner shown in Figure 3. The first component is the amplitude nonlinearity, implemented as a point process, while the second is the CSF, implemented as a filtering process. The final component in the cascade is the detection process, which models the masking effects. It is implemented as a combination of filters and nonlinearities.

3.0 HVS MODEL: AMPLITUDE NONLINEARITY

It is well known that visual sensitivity and perception of lightness are nonlinear functions of luminance. The amplitude nonlinearity of the VDP describes the sensitivity variations as a function of the grey scale, as well as the changes that occur due to different illumination levels, and is based on a model of the early retinal network. The original model[6] was a shift-variant nonlinearity, which had the problem of non-invertability. As a result, we have developed a simplified version that is shift-*in*variant, invertible, and implemented as simple point nonlinearities[7]. It assumes a state of adaptation resulting from an observer fixating a small image area, accomplished by a reduction in eye movements. It is referred to as the local amplitude nonlinearity and is implemented in the VDP as a function of pixel location (i,j),

$$R(i,j)/R_{max} = \frac{L(i,j)}{\left(L(i,j) + (c_1 L(i,j))^b\right)} \quad , \tag{3}$$

where R/R_{max} is the normalized response, L is the luminance as a function of location, b is 0.63 and c_1 is 12.6 for the units of cd/m^2. For this model, the adaptation level for a pixel in the image is solely determined from that pixel. Although physiological data indicates that visual system cannot adapt to indefinitely small areas of the image, we assume that the observer may view the image at any arbitrarily close distance. This removes any spatial attributes in the amplitude nonlinearity component and allows us to model all spatial frequency aspects separately in the CSF component. Though shift-invariant, the local amplitude nonlinearity model *is* an adaptive function, and the change in the model is shown in Figure 4 for a series of illumination levels. Though this component of the VDP gives the shape of the nonlinearity, the actual sensitivity is determined by the CSF.

Most of the other models used in the literature to describe this aspect of the visual system's nonlinearity with respect to gray level are also shift-invariant. The early image processing models of the HVS typically used a logarithmic front-end[8,9], and are often referred to as homomorphic models. This is a reasonable approximation, but will overestimate the detectability of differences in the dark areas for all but the brightest light adaptation levels. Conversely, linear front end models will overestimate the detectability of differences in the light areas. A second wave of image processing models began using power functions close to the cube-root[10,11], which perform very closely to the nonlinearity used in the current VDP. Figure 5 shows a comparison of these models with our local amplitude nonlinearity model. Even though it is similar to the cube root function for light levels near the practical range of 100 cd/m^2, the cube root function does not change shape with illumination level. The actual change in sensitivity as a function of the grey level would be described by the slope of the curves in Figure 5. Two very recent image processing models take unique approaches to modeling the front-end adaptation properties. One of these[4], combines the front end nonlinearity with the generation of the multiple channel filters by using the ratio-of-Gaussians (ROG) rather than the more typical difference-of-Gaussians (DOG). Though this appears to be a better approach with regard to visual experiments studying harmonic distortion products, the overall model failed to predict the difference between quantization in the linear and log domains. The other model[5] uses a normalization process to center the signal on a "retinal transducer function" that precedes the generation of the multiple frequency channels.

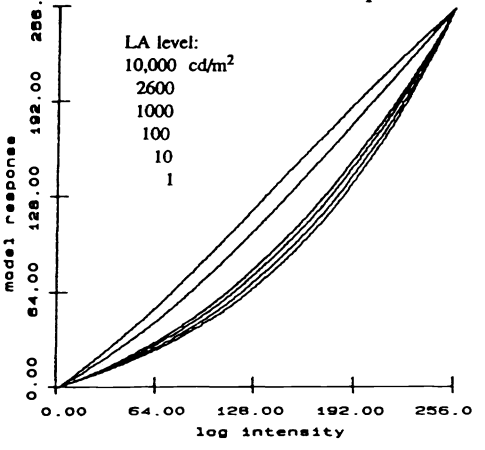

Figure 4

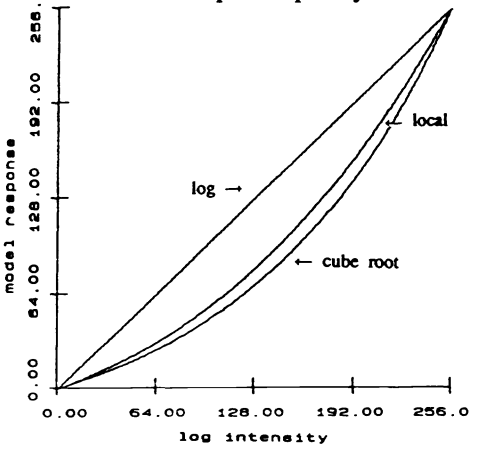

Figure 5

4.0 HVS MODEL: CONTRAST SENSITIVITY FUNCTION

The contrast sensitivity function (CSF) describes the variations in visual sensitivity as a function of spatial frequency. These variations are primarily due to the optics of the eye, the sampling aperture of the cone photoreceptor, and both passive and active neural connections. All of these contributing biological components are highly adaptive, with the result that the CSF changes as a function of light adaptation, noise, color, accommodation, eccentricity, and image size. All global spatial frequency effects are modeled in this component, whether they actually occur due to the optics, cones, cortical neurons or elsewhere in the visual system.

The following equations model the sensitivity, S, as a function of radial spatial frequency ρ in cy/deg, orientation θ in degrees, light adaptation level l in cd/m^2, image size i^2 in visual degrees, lens accommodation due to distance d in meters, and eccentricity e in degrees,

$$S(\rho,\theta,l,i^2,d,e) = P \cdot min\left[S\left(\frac{\rho}{bw_a \cdot bw_e \cdot bw_\theta}, l, i^2\right), S(\rho,l,i^2)\right] \tag{4}$$

The parameter, P, is the absolute peak sensitivity of the CSF, which unfortunately depends on the individual, but we can refer to studies such as Virsu[12] that measured the CSF for a large population of observers. Based on these, the value of 250 will be used in the current implementation of CSF. The parameters bw_a, bw_e, and bw_θ, model the changes in bandwidth due to the accommodation level, eccentricity and orientation, respectively, via the following equations,

$$\begin{aligned} bw_a &= 0.856 \cdot d^{0.14} \qquad \text{where } d \text{ is the distance in meters} \\ bw_e &= \frac{1}{1+ke} \qquad \text{where } e \text{ is the eccentricity in visual degrees, } k = 0.24 \\ bw_\theta &= ((1-ob)/2)cos(4\theta) + (1+ob)/2 \qquad \text{where } ob = 0.7 \end{aligned} \tag{5}$$

where the bandwidth shift due to eccentricity is from a model for spatial scale[13]. The remaining effects to be modeled include the effect of the image size and the light adaptation level,

$$S(\rho,l,i^2) = \left(\left(3.23(\rho^2 i^2)^{-0.3}\right)^5 + 1\right)^{-1/5} \cdot A_l \varepsilon \rho e^{-(B_l \varepsilon \rho)} \sqrt{1 + 0.06 e^{B_l \varepsilon \rho}} \tag{6}$$

$$\begin{aligned} A_l &= 0.801(1 + 0.7/l)^{-0.2} \\ B_l &= 0.3(1 + 100./l)^{0.15} \end{aligned}$$

where l is the light adaptation level in cd/m^2, i^2 is the image area in deg^2, and ε is a frequency scaling constant that equals 0.9 for the luminance CSF. The first half of the equation models the sensitivity as a function of image size, while the second models the changes in sensitivity and bandwidth as a function of light adaptation level and is based on an earlier model[3]. Shown in Figure 6 is a typical CSF for the 2–D frequency plane, showing the bandpass nature and oblique effect.

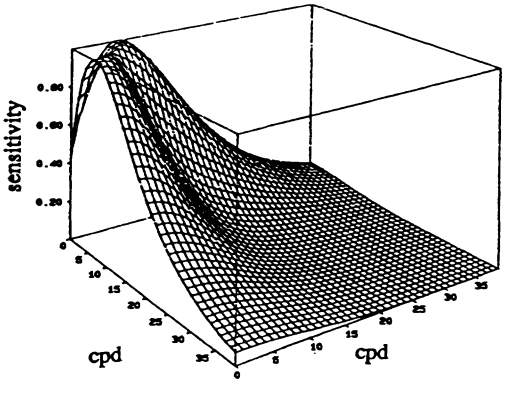

Figure 6

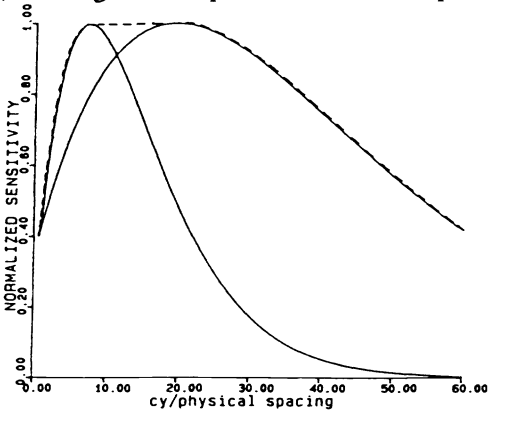

Figure 7

In the VDP algorithm, the CSF is modeled in the units of cy/deg and then mapped to the digital frequency domain by using the calibration parameters of viewing distance, horizontal pixel spacing, and vertical pixel spacing. The other

calibration parameters (maximum light level, image size, etc.) also affect the CSF, according to the adaptive model. Noise is not input as a parameter since it is modeled explicitly in the external image simulation steps, and its effects on the VDP model appear in the detection mechanisms. Although it is unrealistic to use a single distance in most applications, it is not unrealistic to assume the observer will remain within a range of viewing distances. For the VDP we have developed an extension of the CSF model, which is appropriate for a range of viewing distances, such that the input parameter to the VDP will include a minimum and maximum viewing distance. This extended CSF model is shown in Figure 7. The CSF model for a range of distances is shown as the dashed line and is formed as the envelope of all CSFs mapped to the digital frequency domain for the range specified by the minimum and maximum distances. This method is a shortcut since the range CSF is used with the rest of the algorithm only once, whereas the proper way would be to run the algorithm multiple times with viewing distances spanning the range and then combining the results. An observer would need to view the image for the entire range of distances to see as well as the VDP model's prediction, but this approach can ensure no visible distortions are perceived within the applications range.

In other visual models based on spatial frequency mechanisms, the most common approach is to weight the sensitivities of the discrete channels such that the sum is close to an aim CSF[14,15]. A disadvantage of this approach is that it is more difficult to include the adaptive behavior of the CSF if it is not decoupled from the multiple channels. In comparison, our technique of separating the CSF model from the multiple channel generation allows for more flexibility in modeling the CSF adaptivities, distance changes, and range effects. However, the VDP's technique of preceding the cortex filtering with a CSF causes a slight distortion in the shape of the frequency channels from that described by the cortex filters.

5.0 HVS MODEL: DETECTION MECHANISMS

The final HVS component is comprised of the multiple detection mechanisms, which are modeled with four sub-components as shown in Figure 8. The sub-components include the *spatial frequency hierarchy* that models the frequency selectivity of the visual system and creates the framework for the multiple detection mechanisms, the *masking function* that models the magnitude of the masking effect, the *psychometric function* that describes the threshold in a detailed manner, and finally, *probability summation* that combines the responses of all the detection mechanisms into a unified perceptual response.

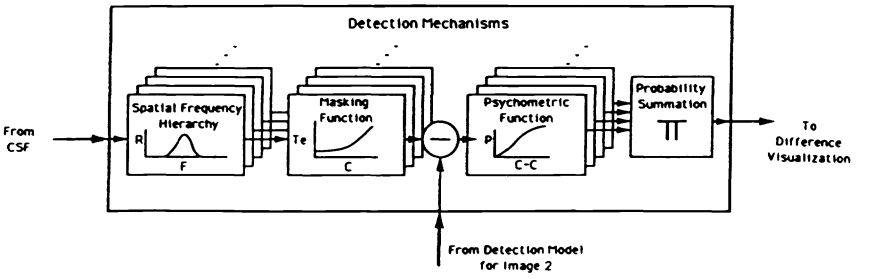

Figure 8

5.1 Spatial frequency hierarchy

The spatial frequency selectivity of the visual system refers to the existence of specialized mechanisms, which are selective for, or tuned to, narrow ranges of spatial frequency. This property has been known for a long time from both neurophysiological recordings[16,17] and psychophysical studies in adaptation and masking[18,19]. Although the masking studies were originated as a tool to elucidate properties of the visual cortex, they are directly relevant to image fidelity assessment. These studies have found a radial frequency selectivity that is essentially symmetric on a log frequency axis with bandwidths nearly constant at one octave. In addition, these studies have revealed an orientation selectivity with symmetry about a center peak angle and tuning bandwidths varying as a function of radial frequency, ranging from 30 degrees for high frequencies to 60 deg for low frequencies[20]. The current understanding of these important mechanisms is that they are limited in both frequency and space, thus giving rise to the popularity of Gabor and Wavelet modeling approaches.

The frequency selectivity of the visual system is modeled here by using a hierarchy of filters. This assumption of discrete spatial frequency channels is merely an approximation to the visual system. By operating on the spatial images generated from such a filter bank, we can model the space-frequency localization aspects of the visual system. The spatial frequency hierarchy is the base algorithm necessary to properly model the signal dependency of masking. Currently, there is a vast array of choices in implementing the spatial frequency hierarchy. These all have the properties of multiresolution decomposition with localization properties in both the space and frequency domains based on a set of dilated and translated scaling functions, which are consequently self-similar from one resolution to the next. Such multiresolution schemes can

be generally referred to as wavelets, though there is still definitional debate regarding wavelets, such as the necessity of orthogonality.

Of these, the **cortex transform** was selected as the basic structure to implement the spatial frequency selectivity in the VDP. Its log-polar frequency structure was based on the combination of the spatial frequency selectivity measured psychophysically and the measurements of 2–D cortical receptive fields. It was introduced by Watson[21], who later incorporated it into a more complete visual model[15]. Some of the advantages of the cortex transform include its reversibility, its ease of implementation, and its flexibility. The reversibility allows the CSF to be modeled independently. Since the purpose of the VDP is to analyze other imaging systems and algorithms (which may themselves be visually based to various degrees), the emphasis was on accuracy and not on efficiency, and the flexibility of the cortex transform aids in this goal. It is not without its disadvantages, however, which include its non-orthogonality and computational complexity.

Before describing the cortex transform in more detail, it is worthwhile to briefly discuss the other approaches. The **Gabor** function is ideal for space-frequency localization. As a result, a number of visual models based on the Gabor function in a log polar structure have been proposed[5,22]. One approach[5] is to filter the image by the bank of Gabor filters resulting in an expanded set of images, which essentially over–represent the image, yet are not complete. This method limits the visual model as a research tool and poses problems with regard to the CSF. The other approach is the Gabor transform[22], which creates a set of coefficients for the basis set of Gabor functions. The Gabor functions sample the image with multiple rates such that the total number of degrees of freedom are preserved. The main problem with this approach is the computational complexity. As an example, the Gabor function is truncated at one σ for computational reasons[23], which causes the algorithm to lose its property of reversibility and can give rise to tiling effects in images. Another problem with all Gabor techniques is that the filters are not symmetric on a log frequency axis, an issue that relates to the fit of the gabor to the receptive fields of cortical neurons as well as the appropriateness of the type of space-frequency localization[24]. The **QMF subband transform** has successfully been used as a structure for perceptually based image compression[25,26] and it is the most computationally efficient approach for generating the spatial frequency bands. It can be orthogonal and reversible and can be set up to have the desired log spacing of frequency bands, by implementation in a pyramid fashion. Its main problem is what is referred to as a mixed orientation band, in which the diagonal band contains frequencies near 45° and 135°. This leads to significant failures in its prediction of visual masking effects, since this mixed band assumes contrast energy from these distant orientations mask each other as well as similar orientations. A version of pyramidal subband coding on a hexagonal grid[26] removes the mixed orientation band, but then the orientation channels are either limited to three (60° apart) or bear no relationship to the shape of the frequency selectivity found psychophysically or physiologically. Another recently proposed **wavelet** decomposition[27] has the same problems as this approach, since it is also Cartesian separable. The **HOP transform**[28] begins with a hexagonally sampled image, physiologically motivated by the hexagonal structure of the cones in the fovea. Difference-of-Gaussian (DOG) functions are applied to the samples in the hexagonal array. This model has proved useful in understanding how the cortical receptive fields are formed from the retinal inputs, but there are only three orientation channels, and as a result, the tuning is too broad for use in a visual model. The visual system may indeed do processing very similar to this model, but the hexagonal structure of the fovea undergoes apparently random shifts in local orientation about every 20 cones[29], which could allow more orientation channels to be formed in the cortex.

In the Cortex transform, the radial frequency selectivity and the orientational selectivity are modeled with separate classes of filters that are cascaded to describe the combined radial and orientational selectivity of cortical neurons. This combined filter has been termed the *cortex* filter. Four major modifications to the originally proposed Cortex transform appear in our model. The first modification is that the filter transitions use Hanning functions, whereas the original version used a Gaussian amplitude distribution function. We have found the Hanning transition results in less ringing, especially if the Gaussian is truncated as in the original description. Another modification is that the orientation transitions are defined in polar degrees, whereas the original transitions were Euclidean. A third modification is our use of a Gaussian baseband to reduce ringing in the baseband. A final change was to abandon the use of a high frequency residual, where all the radial frequencies greater than 0.5 cy/pix were either lumped in a single filter or merely discarded. In the model presented here, these high frequencies are split into a series of orientation bands as with all the radial frequency filters, which was necessary for blur prediction and other high frequency distortions.

In order to ensure the filter set sums to 1.0, the radial frequency filters are formed as differences of a series of 2–D low-pass *mesa* filters, characterized by a flat passband, a transition region, and a flat stop-band region. The transition region is modeled with a Hanning window such that the mesa filter can be completely described by its half-amplitude frequency,

$\rho_{1/2}$, and transition width, *tw*, as follows,

$$\begin{aligned} mesa(\rho) &= 1.0 && \text{for} \quad \rho < \rho_{\frac{1}{2}} - \frac{tw}{2} \\ &= \frac{1}{2}\left(1 + cos\left(\frac{\pi\left(\rho - \rho_{\frac{1}{2}} + \frac{tw}{2}\right)}{tw}\right)\right) && \text{for} \quad \rho_{\frac{1}{2}} - \frac{tw}{2} < \rho < \rho_{\frac{1}{2}} + \frac{tw}{2} \\ &= 0.0 && \text{for} \quad \rho > \rho_{\frac{1}{2}} + \frac{tw}{2} \quad . \end{aligned} \tag{7}$$

The radial frequency selectivity is modeled by *dom* filters (differences of mesas) formed from two mesa filters (or filtered images) evaluated with different half-amplitude frequencies. The *k*th dom filter is given by,

$$dom_k(\rho) = mesa(\rho)|_{\rho_{\frac{1}{2}}=2^{-(k-1)}} - mesa(\rho)|_{\rho_{\frac{1}{2}}=2^{-k}} \quad , \tag{8}$$

where the | symbol means the mesa filter is to be calculated with the indicated half-amplitude frequency, $\rho_{1/2}$. This frequency is expressed in the units of cy/pixel, which are the units used in the detection mechanisms model. It does not use cy/deg because of the separate modeling of the detection mechanisms and the CSF. Increasing values of *k* correspond to higher levels of the hierarchical pyramid, which are consecutively lower frequency bands. In the current implementation, the transition width of each filter is a function of its defined half-amplitude frequency as follows,

$$tw = \frac{2}{3}\rho_{\frac{1}{2}} \quad , \tag{9}$$

and the resulting radial frequency set is shown in Figure 9. This transition width configuration gives approximately constant behavior on a log frequency axis with a bandwidth of 1.0 octave and symmetric responses.

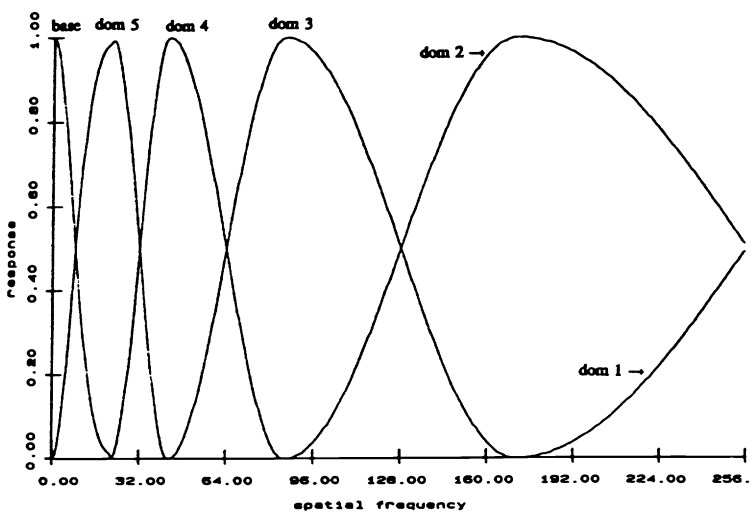

Figure 9

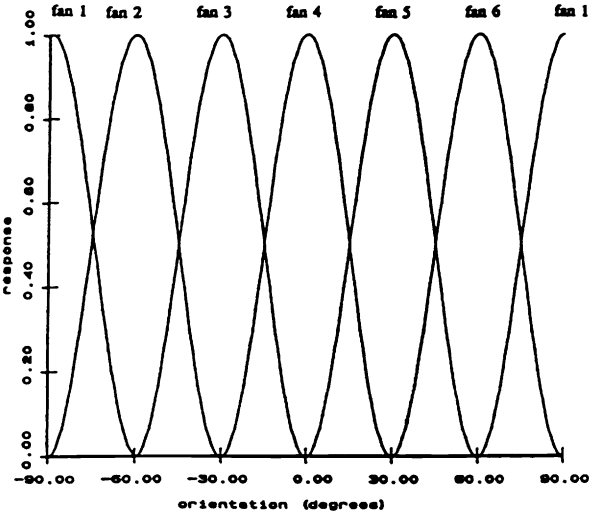

Figure 10

The orientation attributes of the spatial frequency selectivity are modeled with *fan* filters. A Hanning window will also be used for these filters, which is determined in angular degrees, θ, in the Fourier plane. The equation for fan *l* as a function of orientation is,

$$\begin{aligned} fan_l(\theta) &= \frac{1}{2}\left(1 + cos\left(\frac{\pi|\theta - \theta_c(l)|}{\theta_{tw}}\right)\right) && \text{for} \quad |\theta - \theta_c(l)| \le \theta_{tw} \\ &= 0.0 && \text{for} \quad |\theta - \theta_c(l)| > \theta_{tw} \quad , \end{aligned} \tag{10}$$

where θ_{tw} is the angular transition width and $\theta_c(l)$ is the orientation of the center, or peak, of fan filter *l* given by,

$$\theta_c(l) = (l-1)\cdot\theta_{tw} - 90 \quad . \tag{11}$$

If we set the transition width equal to the angular spacing $\theta_{\Delta c}$ between adjacent fan filters,

$$\theta_{tw} = \theta_{\Delta c} = \frac{180}{L}, \tag{12}$$

where L is the total number of fan filters, then the filter set sums to 1.0 and the resulting orientation bandwidth also equals the angular spacing. In the current implementation L is 6, which gives an orientation bandwidth of 30 degrees and this fan filter set is shown in Figure 10.

The cortex filters are formed as the separable product of the dom and fan filters as,

$$\begin{aligned} cortex_{k,l}(\rho,\theta) &= dom_k(\rho)\cdot fan_l(\theta) \qquad \text{for} \quad k = 1, K-1 \; ; \; l = 1, L \\ &= base(\rho) \qquad\qquad\qquad\quad \text{for} \quad k = K \qquad\qquad , \end{aligned} \tag{13}$$

where the particular cortex filter can be denoted by the dom and fan filter indices, *k,l* respectively. Since no orientation selectivity is attributed to the baseband (k = K), its index l is a dummy parameter. The total number of cortex filters in the set is ({K-1}*L+1), which in the current implementation is 31 (K = 6, L = 6). The set of cortex filters is reversible, i.e.,

$$\sum_{k=1,K}\sum_{l=1,L} cortex_{k,l}(\rho,\theta) = 1 \qquad \text{for all } \rho,\theta \quad . \tag{14}$$

5.2 Contrast units in the cortex filtered images

The two input images are filtered by the cortex filter set after they have been modified by the amplitude nonlinearity and CSF models described previously. Their pixel values must be converted to the units of contrast in order to calibrate to psychophysical results. For arbitrary waveforms, contrast should be defined as a function of location, since the waveform may change its entire character from one region to the next. It is also useful to allow the contrast to be both positive and negative, to delineate points on the waveform greater or less than the mean. The contrast for the cortex bands is modified from equation 2 into a function of pixel location (i,j) given by,

$$C_{k,l}(i,j) = \left(B_{k,l}(i,j) - \overline{B_{k,l}}\right)/\overline{B_{k,l}} \qquad , \tag{15}$$

where $C_{k,l}(i,j)$ is the contrast in band k,l, $B_{k,l}(i,j)$ is the value of the filtered image, and $\overline{B_{k,l}}$ is the mean of band k,l. However, for all but the baseband image, the mean of the cortex bands, $\overline{B_{k,l}}$, is zero, making the equation indeterminate. To avoid this problem, we use a constant value for the denominator based on the input image mean (which can conveniently be found from the baseband mean $\overline{B_K}$) and leave the numerator mean as zero,

$$C_{k,l}(i,j) = B_{k,l}(i,j)/\overline{B_K} \qquad . \tag{16}$$

5.3 Masking function

Masking refers to the decreased visibility of a signal due to the presence of a suprathreshold background (i.e., the mask) and there are two main ways to model this effect, given the image is decomposed into the appropriate spatial frequency bands. One of these is to use a threshold elevation image, while the other is to use a contrast transducer function[14] to modify the pixel dependent contrast of a band. This latter technique has the advantage of being able to retain the degree of suprathreshold differences, and is easier to incorporate in image compression algorithms. However, most of the models using this approach use a contrast transducer function calculated from the masking curve of contrast discrimination experiments. We feel the contrast discrimination experiment is not appropriate for the majority of practical distortions in real world images. The threshold elevation approach, which uses a masking function directly, seemed the best approach for threshold prediction.

The masking function describes the threshold of a test signal as a function of the mask contrast. If plotted on log-log axes, it can be characterized by two asymptotic regions. One of these has a slope of zero (indicating no threshold elevation) and occurs for low mask contrasts, while the other has a positive slope, indicating the threshold increases with increases in mask contrast. The actual high contrast slope depends on whether the mask is phase coherent (i.e., a sine wave), giving slopes near 0.7, or phase incoherent (i.e., a noise field), which gives a slope of 1.0. One assumption behind our model is that if the threshold and mask contrast are normalized by the uniform field threshold (i.e., $\frac{1}{csf}$), then a single curve can describe the masking function for all frequencies, providing the test and mask frequencies are the same. Another is that if the test and

mask contrast are further normalized by the inverse of the sensitivity of the detecting mechanism's receptive field (modeled by the cortex filter), then a single curve can describe the threshold elevation even for test and mask frequencies that differ. As a result, the normalized contrast for a mask of frequency (ρ_m,θ_m) is given by,

$$m_n(\rho_t, \theta_t) = m(\rho_m, \theta_m) \cdot csf(\rho_m, \theta_m) \cdot cortex_{k,l}(\rho_m, \theta_m) \quad , \tag{17}$$

which describes the mask contrast seen by the mechanism that detects the test signal with frequency (ρ_t,θ_t). The indices *k,l* on the cortex filter correspond to the filter that is centered on the test signal frequency. These csf and cortex normalization assumptions are not supported by all data, but are indicative of the first order effects.

5.3.1 Learning Effect Model: The dependence of the high contrast asymptote slope on the phase coherence of the mask poses problems for those trying to use the data in applied visual models. The question arises as to which type of synthetic structure (sine waves or noise fields) is more appropriate to describe the masking effect in actual images. A study that seems to unify the disparate results between the noise and sine wave experiments has been performed by Smith and Swift[30]. By clever design of experiments, they were able to obtain slopes of 1.0 for sine masks and slopes of 0.7 for noise fields, which is reversed from usual. Thus the difference in slopes was not inherent in the noise or sinusoidal waveforms. The source of the difference in slopes between the two masks seems to lie in the differing degrees of uncertainty caused by the masking pattern; an uncertainty that can be reduced through learning. The sine masking pattern is fairly easy to learn, and therefore to detect the signal added to it one merely looks for differences in the expected appearance of the mask. On the other hand, the noise mask is difficult to learn, and this difficulty results in a higher slope, indicating more masking and a lower detector efficiency.

In the Smith and Swift study, the same noise field was used repeatedly and the slope of the high contrast asymptote was measured over a large number of trials. As the observer's familiarity with the noise mask increased, the slope dropped from 1.0 to about 0.65. Further proof that uncertainty was limiting the detection for the noise masks was done by using a 3AFC technique. In this technique, three identical noise fields were presented simultaneously, and the signal was added to one of them. The task is to simply pick the field that looks different, and for this experiment they obtained slopes between 0.65 and 0.80, even though they changed the noise field for each presentation. No learning was necessary in this case. To obtain slopes of 1.0 for the sine masking case, they used naive observers and collected data during the customary training sessions. They could only measure two points on the asymptote, but it does underscore the learning effect, since the measured slopes dropped from 1.0 to about 0.65. This was similar to that found for the noise masks, but the effect happened much faster, since the sine masks were easy to learn. The results of this learning experiment for sine masks are shown as the data points in Figure 11.

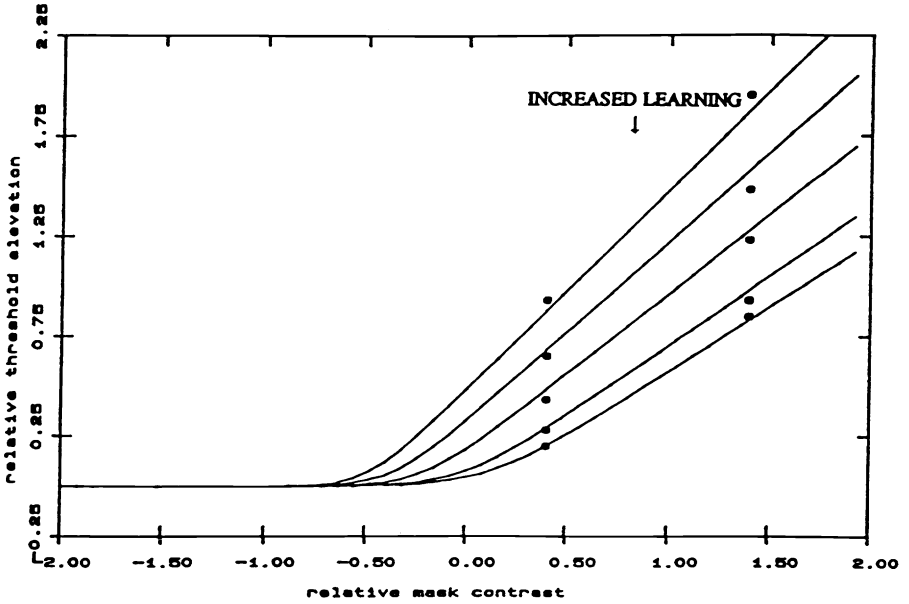

Figure 11

We have modeled the masking function incorporating the uncertainty/learning effect with the following equation

$$T_e(m_n) = \left(1 + (k_1(k_2 \cdot m_n)^s)^b\right)^{1/b} \quad , \tag{18}$$

where T_e is the threshold elevation and s corresponds to the slope of the high masking contrast asymptote, which ranges between 0.65 and 1.0. For a high uncertainty (low learning level), the slope is 1.0 and as learning increases, the slope (and

uncertainty) reduces to 0.65. The parameters k_1 and k_2 determine the pivot point of all the learning slopes shown in Figure 11. This equation is shown fitting the data by using values of s ranging from 1.0 to 0.65, and it simplifies to a signal-to-noise ratio detection model with internal noise and an SNR of 6.0 when s = 1.0. The parameter b determines how closely the curve follows the asymptotes in the transition region and we are currently using values between 2 and 4 for different bands. Although we have unified the different experiments, we are still left with the problem of choosing a slope, or consequently, the expected degree of learning. After many experiments with natural and synthetic images, we have chosen a value for the learning slope that depends on the cortex band.

5.3.2 Masking in Cortex Bands: The previous discussion of the masking function was described in the context of the visual system's spatial frequency selectivity rather than the cortex filter's discrete channel approximation. The main difference in the discrete filter approach is that there is not a detector centered on every test frequency. We need to broaden our concept of the test frequency to include all the frequencies present in the cortex band. Likewise, our concept of the masking frequency should now encompass all the possible frequencies in the band. In the algorithm, many frequencies contribute to the normalized mask contrast m_n within a particular band k, l as a function of location (i,j). It is calculated as,

$$m_n^{k,l}(i,j) = \mathcal{F}^{-1}\{\mathcal{R}(u,v) \cdot csf(u,v) \cdot cortex^{k,l}(u,v)\} \quad , \tag{19}$$

where $\mathcal{R}$ is the Fourier transform of the input image as modified by the amplitude nonlinearity and u, v are the Cartesian frequency components. Though the normalized mask contrast may be positive or negative, it is its magnitude that affects the masking function, so that the threshold elevation in equation 18 implemented as a function of location becomes,

$$T_e^{k,l}(i,j) = \left(1 + \left(k_1\left(k_2\left|m_n^{k,l}(i,j)\right|\right)^s\right)^b\right)^{1/b} \quad . \tag{20}$$

The band specific threshold elevation, $T_e^{k,l}(i,j)$, is a function of pixel location and is referred to as the *threshold elevation image.*

5.4 Psychometric function and probability summation

The psychometric function describes the increase in the probability of detection as the signal contrast increases. It is given by the following equation, which can describe data for nearly all combinations of frequencies,

$$P(c) = 1 - e^{-(c/\alpha)^\beta} \tag{21}$$

where $P(c)$ is the probability of detecting a signal of contrast c. The threshold is given by the parameter α, which shifts the function along the contrast axis. The slope of the pyschometric function is described by the parameter, β, and is largely invariant across many experiments. This has enabled the single parameter, α, to retain the information describing the entire psychometric function region. In the context of the VDP algorithm, we calculate the probability of detection as a function of location for each of the cortex bands as,

$$P_{k,l}(i,j) = 1 - e^{-\left(\Delta C_{k,l}(i,j)/T_e^{k,l}(i,j)\right)^\beta} \quad , \tag{22}$$

The contrast difference as a function of location for band k,l, $\Delta C_{k,l}(i,j)$, is calculated as,

$$\Delta C_{k,l}(i,j) = \frac{B1_{k,l}(i,j)}{\overline{B_K}} - \frac{B2_{k,l}(i,j)}{\overline{B_K}} \quad , \tag{23}$$

where $B1_{k,l}$ and $B2_{k,l}$ are the two filtered input images for band k,l.

As mentioned earlier, the role of the CSF in the VDP algorithm is to normalize all frequencies so their uniform field threshold is equal as seen by the subsequent spatial filter hierarchy. By the calibration techniques chosen, a code value excursion of 1.0 corresponds to the uniform field threshold contrast in each band of the hierarchy and this allows the substitution of $T_e^{k,l}(i,j)$ for α in equation 22. This approach assumes the psychometric function does not change shape when the threshold has been elevated by masking, which is a simplification from actual data. The sign of the error is also calculated, which will be used in the visualization section of the algorithm. The sign indicates whether the error appears lighter or darker than the reference image and also gives shape to the predicted error. The signed probability of detection, *SP* is simply calculated from,

$$SP_{k,l}(i,j) = sign(\Delta C_{k,l}(i,j)) \cdot P_{k,l}(i,j) \quad . \tag{24}$$

Once the detection probabilities are computed for each band of the spatial filter hierarchy, these probability images are combined into a single image, which describes the overall probability of detecting an error for every pixel in the image. The technique of probability summation is used, with the equation given by the product series,

$$P_t(i,j) = 1 - \prod_{k=1,K;l=1,L} (1 - P_{k,l}(i,j)) \quad , \tag{25}$$

where P_t is the total probability of detection resulting from all bands as a function of location. Similar to the case for the individual bands, where the sign of the error was preserved, we preserve the sign of the total probability of detection of differences between the two input images.

Although we have used probability summation in the VDP, the most common method used to combine the outputs from the different frequency mechanisms is to use a vector summation of differences across bands and pixel locations[4,5], which gives a single number output. This approach works in conjunction with the contrast transducer function described previously and has the capability of describing suprathreshold differences. However, its verification is much more difficult in the context of images where the distortion may be visible in several parts of the image and does not lend itself well to being described as a single signal. The vector summation approach, like all single number output models, treats the entire physical difference image as a single signal and then calculates objectionability of the global distortion. It does not indicate what parts of the distorted image are above and what parts are below threshold, nor whether the local distortions are lighter or darker than the reference. This makes it much more difficult to develop the algorithm. Another problem with a single number output model is that we would eventually like a single number metric to include the relative importances of certain image areas (such as faces) and models that integrate their results into a single number would be more difficult to extend to include such higher order attributes. We feel the method chosen for the VDP (masking function in conjunction with probability summation) is more appropriate for threshold predictions, and we feel threshold predictions are more relevant to image fidelity assessment.

6.0 VISUALIZATION AND INTERPRETATION OF THE DETECTION MAPS

The output of the probability summation is a map of the probability of detecting visible differences and is a function of pixel location. Disregarding the sign of the detection map, which describes the polarity of the visual difference, the magnitude ranges from zero to one. All suprathreshold visual differences will map to the value of one, since the algorithm cannot distinguish between visual differences, which are easily detectable, yet have different degrees of perceived contrast error. Since the entire range of the probability map extends from −1 to 1, it must be remapped to view on most display devices, and there are currently two methods of displaying the detection map.

6.1 Free-field difference map

The first of these methods, the *free-field difference map*, displays the predictions on a uniform field of grey, free of image structure. This method merely uses a linear mapping of the code values of the detection map, with the equation given by,

$$FF(i,j) = SP_t(i,j) * \left(\frac{max - min}{2}\right) + \left(\frac{max + min}{2}\right) \tag{26}$$

where FF is the free-field difference prediction, SP is the signed probability of detection from equation 24, and max and min are the display device's maximum and minimum values, respectively. The free-field difference map is monochromatic, and an example of the VDP's output in this format is shown in Figure 2D for the two distorted images of Figure 2C. Pixels that are lighter than the pedestal of (max+min)/2 indicate where the distorted image appears lighter than the reference, and vice versa for the darker pixels. Pixels that are full white or full black are detected completely, that is, with a probability of 1.0 and these regions have a clipped appearance because all the suprathreshold differences are mapped to their values of −1.0 and 1.0. The pixels that are shades of grey between these two extremes indicate where the visible differences are solely within the threshold region. The VDP results match the visibility of the distortions in Figure 2C, in that the banding distortions (A) are easily visible in many areas of the image, while the tone scale distortions (B) are extremely difficult to see, since they all lie within the threshold region.

6.2 In-context difference map

With the method just described, it is sometimes difficult to judge the correspondence between the predicted differences and the differences actually observed between the two input images. The second method, the *in-context difference map*, was designed to overcome this problem. This is accomplished by displaying the predicted differences in the context of the

reference image. So that the suprathreshold image content of the reference image does not mask the predicted results, we expand the monochromatic output to color, such that the predicted differences appear in color on the original monochromatic reference image. The reference image is copied to all three RGB planes, and the detection image is scaled and added to one of the color planes as follows,

$$IC(i,j) = SP_t(i,j) \cdot \left(\frac{max - min}{2}\right) + ref(i,j) \tag{27}$$

where *ref* is the reference image and the other terms are as described in equation 26. For example, if the red layer is used, then red indicates where the distorted image appears lighter than the reference, and cyan indicates where it appears darker than the reference.

6.3 The use and interpretation of the maps

Although it is not an application, the initial use of the output detection maps is to evaluate the success of the algorithm. The maps indicate the shape and location of the predicted visual differences and can be verified by observation. One image with distortions equals many experiments, since there is essentially an experiment for each local artifact. Since it is a threshold model, the detection output maps can be analyzed and verified much easier than for a suprathreshold model where it is difficult for individuals to quickly and consistently rank magnitudes. As a result, we relied on informal psychophysical testing to determine the parameters of the algorithm in the developmental phases. This allows more time consuming formal psychophysical tests to be performed after the completion of the model.

Of course the primary function of the VDP is to aid in the design of imaging systems. The output detection maps can be used directly by imaging systems designers to see the nature of the distortion they are adding. This allows them to concentrate on the distortions that are visible, and the shapes and locations of the predicted differences may give the designer insight into ways to improve the performance/cost attributes of the system.

The number of applications for the VDP would be increased if the multi-dimensional detection map could be integrated into a single number metric. The VDP could then be used as a cost function in automated optimization routines and other related design approaches. Once confidence is established in the accuracy of the model for threshold results, it can be used as a framework to study potential metrics that reduce the prediction to a single number. With the establishment of such metrics, it would no longer be necessary to display the simulated image or visualize the detection maps, as the VDP output could be mathematically analyzed directly.

Currently, we have developed one simple method of reducing the detection map to a single number, and it can be used to determine if the two input images are visually identical for the specified viewing conditions. This is done by finding the maximum magnitude of all the pixels in the detection map. If this number, the peak probability of detection, is less than 1.0, we know any visual differences are entirely in the threshold region. If the number is less than 0.5, the images are *visually equivalent* for our purposes. We then perform the VDP for a number of viewing distances to determine the minimum distance where the two images are visually equivalent. We refer to this distance as the *critical distance* and use it to compare competing designs, where a lower number indicates a superior design. However, with the exception of this one approach, methods to combine the map of detectabilities into a single number are still under research, and we hope this model will be instrumental in such research.

7.0 CONCLUSION

An algorithm for the prediction of visible differences between two digital images has been developed. This paper provides a detailed description of the algorithm, its central component (the *human visual system* model), its intended goals, and its utility as a design aid for imaging systems. The algorithm has been tested for a wide variety of image distortions including synthetic images designed after psychophysical experiments and natural images with practical distortions. Some of these include blur, noise, data compression artifacts, banding, blocking, contouring, low frequency non-uniformities, hyperacuity, and tone-scale changes. The VDP's performance with these tests has been promising enough to warrant formal psychophysical tests, which will help to quantify its degree of success.

8.0 REFERENCES

1. G. Higgins "Image quality criteria" J. Appl. Photo. Eng. V. 3 #2, pp. 53–60 (1977).
2. C. R. Carlson and R. W. Cohen "A simple psychophysical model for predicting the visibility of displayed information" Proc. S.I.D. V. 21 #3. pp. 229–246 (1980).
3. P. G. J. Barten "The SQRI method: a new method for the evaluation of visible resolution on a display" Proc. S.I.D. V. 28 pp. 253–262 (1987).
4. C. Zetzche and G. Hauske "Multiple channel model for the prediction of subjective image quality" S.P.I.E. Proceedings V. 1077 pp. 209–216 (1989).
5. C. Lloyd and R. Beaton "Design of a spatio-chromatic human vision model for evaluating full-color display systems" S.P.I.E. Proceedings V. 1249 pp. 23–37 (1990).
6. R. A. Normann, B. S. Baxter, H. Ravindra, and P. J. Anderton "Photoreceptor contributions to contrast sensitivity: Applications in radiological diagnosis" IEEE Trans. Sys., Man and Cybernetics V. SMC-13 #5. pp. 946–953 (1983).
7. M. I. Sezan, K. L. Yip and S. Daly "Uniform perceptual quantization: Applications to digital radiography" IEEE Trans. Sys., Man and Cybernetics V. SMC-17 #4. pp. 622–634 (1987).
8. T. G. Stockham "Image processing in the context of a visual model" Proc. IEEE V. 60 #7. pp. 828–841 (1972).
9. C. F. Hall and E. L. Hall "A nonlinear model for the spatial characteristics of the human visual system" IEEE Trans. Sys. Man and Cybernetics V. SMC-7 #3 pp. 161–170 (1977).
10. J. L. Mannos and D. J. Sakrison "The effects of a visual fidelity criterion on the encoding of images" IEEE Trans. Inf. Theory. V. IT-20 pp. 525–536 (1974).
11. J. J. McCann, S. P. McKee and T. H. Taylor "Quantitative studies in retinex theory" Vis. Res. V. 16, pp. 445–458 (1976).
12. V. Virsu, P. Lehtio, and J. Rovamo "Contrast sensitivity in normal and pathological vision" Doc. in Oph. Proc. Series, Ed. by L. Maffei. V. 30 pp. 263–272 (1981).
13. A. B. Watson "Estimation of local spatial scale" J.O.S.A. A V. 4 pp. 1579-1582 (1987)
14. H. Wilson and J. Bergen "A four mechanism model for threshold spatial vision" Vis. Res.V.19 pp. 19–32 (1979).
15. A. Watson "Efficiency of a model human image code" J.O.S.A. A V. 4 pp. 2401–2417 (1987).
16. D. H. Hubel and T. N. Wiesel "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex" J. Phys. V. 160 pp. 106–154 (1962).
17. R. L. DeValois, D. G. Albrecht, and L. G. Thorell "Spatial frequency selectivity of cells in the macaque visual cortex" Vis. Res. V. 22 pp. 545–559 (1982).
18. C. Blakemore and F. W. Campbell "On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images" J. Phys. V. 203. pp. 237–260 (1969).
19. C. F. Stromeyer III and B. Julesz "Spatial-frequency masking in vision: Critical bands and spread of masking" J.O.S.A. V. 62 pp. 1221–1232 (1972).
20. G. Phillips and H. Wilson "Orientation bandwidths of spatial mechanisms measured by masking" J.O.S.A. A V. 1 pp. 226-232 (1984).
21. A. Watson "The Cortex Transform: rapid computation of simulated neural images" Comp. Vis. Graphics and Image Proc. V. 39 pp. 311–327 (1987).
22. J. Daugman "Complete discrete 2–D Gabor transforms by neural networks for image analysis and compression" IEEE Trans. Acou. Speech and Sig. Proc. V. 36 pp. 1169–1179 (1988).
23. J. Daugman "Relaxation neural network for non-orthogonal image transforms" Proc. ICNN-88 V. 1 pp. 547-560 (1988).
24. D. G. Stork and H. R. Wilson "Do Gabor functions provide appropriate descriptions of visual cortical receptive fields?" J.O.S.A. A V. 7 pp 1362-1373 (1990).
25. E. H. Adelson, E. Simoncelli, and R. Hingorami "Orthogonal pyramid transforms for image coding" S.P.I.E. Proc. V. 845 pp. 50–58 (1987).
26. E. Simoncelli and E. Adelson "Nonseparable QMF pyramids" S.P.I.E. Proc. V. 1199 pp. 1242-1246 (1989).
27. S. Mallat "A theory for multiresolution signal decomposition: the Wavelet representation" IEEE P.A.M.I. V. 11 pp. 674–693 (1989).
28. A. B. Watson and A. J. Ahumada "A hexagonal orthogonal-oriented pyramid as a model of image representation in visual cortex" IEEE Trans. Biomedical Eng. V. 36 pp. 97–106 (1989).
29. D. Pum, P. K. Ahnet, and M. Grasl "Iso-orientation areas in the foveal cone mosaic" Vis. Neuroscience V. 5 pp. 511–523 (1990).
30. R. A. Smith and D. J. Swift "Spatial-frequency masking and Birdsall's theorem" J.O.S.A. A V. 2 pp. 1593–1599 (1985).