A Comparison of Two Image Quality Models

Bei Li[a], Gary W. Meyer[b] and R. Victor Klassen[c]

[a,b] Computer Science Department, University of Oregon, Eugene, OR 97403

[c] Color and Digital Imaging Systems, Xerox Corporation, Webster NY 14580

ABSTRACT

In recent years a number of different vision models have been proposed to assist in the evaluation of image quality. However, there have been few attempts to independently evaluate these models and to make comparisons between them. In this paper we first summarize the work that has been done in image quality modeling. We then select two of the leading image quality models, the Daly Visible Differences Predictor and the Sarnoff Visual Discrimination Model, for further study. We begin by describing our implementation, which was done from the published papers, of each of the models. We next discuss the similarities and the differences between the two models. The paper ends with a summary of the important advantages of each approach. The comparison of these two models is presented in the context of our research interests which are image quality evaluation for both computer imaging and computer graphics tasks. The paper includes illustrations drawn from these two areas.

1. INTRODUCTION

Visual difference metrics are useful as tools for evaluating image processing algorithms, particularly algorithms designed to leave the image visually unchanged (such as compression algorithms) and algorithms designed to improve the image (such as by halftoning to enhance the print-ability). They also give us a new tool for optimizing the time/quality tradeoff in image synthesis.

Working only from the references cited, we have implemented two metrics: the Daly Visual Differences Predictor,[1] and the Sarnoff Visual Discrimination Model.[2] Our purpose was twofold: firstly, to provide an independent validation of the models as described in the literature, and secondly to compare the two. We found the two models to behave similarly in many respects (as might be expected of two models of the human visual system), however we found strengths and weaknesses in both.

The remainder of the paper is structured as follows. We begin with a brief overview of visual difference models, followed by short descriptions of each of the models. We describe the implementations at the functional level, providing details only where they are not easily available from the literature. We show examples of the predictions of the two models, and characterize their performance. After discussing the two models separately, we compare them, pointing out particular advantages of each. Finally we conclude and make suggestions for further work.

2. BACKGROUND

The two models that are compared in this paper are the result of over twenty five years of research in image quality evaluation. During this period of time, the effect of separate elements of the human visual system on image quality has been explored by researchers working in the fields of image processing, image science, and vision science. These individual visual model components have been continuously refined and integrated, and the two image quality models described in this article are the result.

The initial work in this area was performed by image processors seeking a better way to identify whether the error in a picture was visible to an observer. The nonlinear response of the human visual system to light formed the basis of the first vision model employed for image processing.[3] This model made it possible to map image intensities to more effectively utilize the dynamic range of a display device. The next addition to the image processing models was spatial frequency filtering.[4,5] Better image coding was made possible by transforming an image into the frequency domain and by taking into account the contrast sensitivity function of the human visual system.

The most recent advances in image quality modeling have been accomplished by vision scientists attempting to build better models of human vision. Spatial frequency channels are one example of this.[6] As a result of this discovery, computational techniques were developed so that a spatial frequency hierarchy could be determined and spatial frequency selectivity could be included in the vision models.[7,8] This lead to the incorporation of masking

effects as part of the image quality models.[9] Summing the outputs of the frequency channels[10,11] resulted in a map that could be used to visualize the perceived differences between two images. This set the stage for the two image quality models that are considered in this paper.

3. THE DALY VISIBLE DIFFERENCES PREDICTOR

The Daly Visible Differences Predictor (VDP) receives as input two images and produces as output a difference map, which predicts the probability of detection for dissimilarities throughout the two images. If the images vary substantially, the probability of prediction will be one, and as the differences increase further, the probability will not increase further. The predictor operates solely in the realm of differences below and near threshold.

The basic blocks of the predictor are shown in Figure 1. Key features are an initial non-linearity, frequency domain weighting with the human contrast sensitivity function (CSF),[12,13] and a series of detection mechanisms.

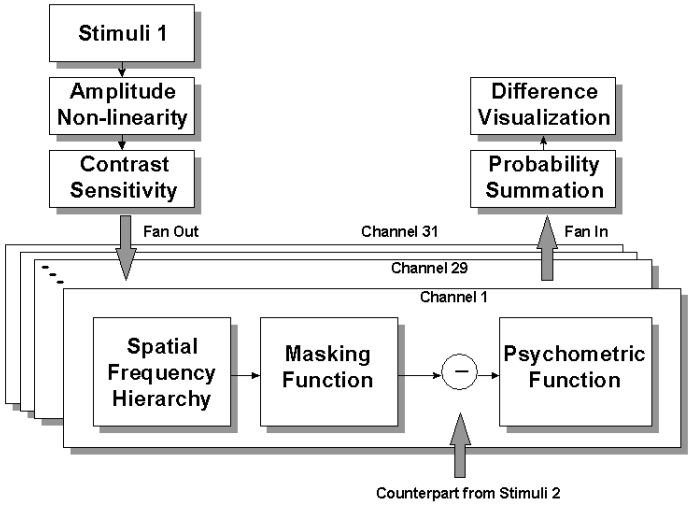

Figure 1. Daly VDP Model

Each image is initially passed through a non-linear response function to account for both adaptation and the non-linear response of retinal neurons. In the Daly VDP model, it is assumed that the observer's visual system adapts separately to each pixel. The adaptation model used approximates the relationship between brightness sensation and luminance. At low luminance levels, it fits the cube-root power law, while at higher luminance levels it approximates the logarithmic dependence. Both of these relationships are in common use, but neither on its own accommodates the entire effect of lightness adaptation for human eyes.

After the initial non-linearity, the image is converted to the frequency domain. The transformed data is weighted with the CSF. That is, the scaled amplitude for each frequency is multiplied by the CSF for that spatial frequency. The peak sensitivity in the CSF varies with environmental luminance. We found that the analytic Barten modulation transfer function,[13] although isotropic, gave reasonable detection results. The weighted data is then converted to local contrast information by dividing each point (amplitude) by the original image mean.

Following the non-linearity and CSF weighting is the series of detection channels. At this point the data is split into 31 independent streams or channels. The visual system is known to have specific selectivities based on orientation (60 degrees per orientation division) and spatial frequency (approximately one octave per channel). In the Daly VDP, each of five overlapping spatial frequency bands is combined with each of six overlapping orientation bands to split the image into thirty channels. These, with the orientation-independent base band, give a total of 31 channels. At this point the individual channels are back-transformed into the spatial domain.

Each channel has associated with it a mask contrast which is a function of location in the image. The presence of masking information at a specific location, spatial frequency and orientation increases the threshold of detectability for a signal with those characteristics. For each channel, a threshold elevation map is computed as a function of

the mask contrast.[14] Finally, mutual masking is applied between the two sets of threshold elevation maps from both input images, to give a single threshold elevation map per channel.

Now we are ready to compute the detection probability. The contrasts of corresponding channels in one image are subtracted from those of the other image, and the difference is scaled down by the threshold elevation. The scaled contrast differences are used as the argument to a psychometric function to compute a detection probability. The psychometric function yields a probability of detection of a difference for each location in the image, for each of the 31 channels. The detection probabilities for all of the channels are combined using the assumption of independent probabilities, giving an overall signed detection probability for each location in the image.

4. SARNOFF VISUAL DISCRIMINATION MODEL

The Sarnoff Visual Discrimination Model (VDM) has been designed for physiological plausibility as well as speed and simplicity. While the Daly VDP is an example of a frequency domain visual model, the Sarnoff VDM operates in the spatial domain. The key elements of the VDM include spatial resampling, wavelet-like pyramid channeling, a transducer for just noticeable difference (JND) calculations and a final refinement step (CSF normalization and dipper effect simulation). Given two input images and a set of parameters for viewing conditions, the output of this model is a JND map. In this section, the influence and function of each stage of the Sarnoff VDM are addressed. The general structure of the model is shown in Figure 2.

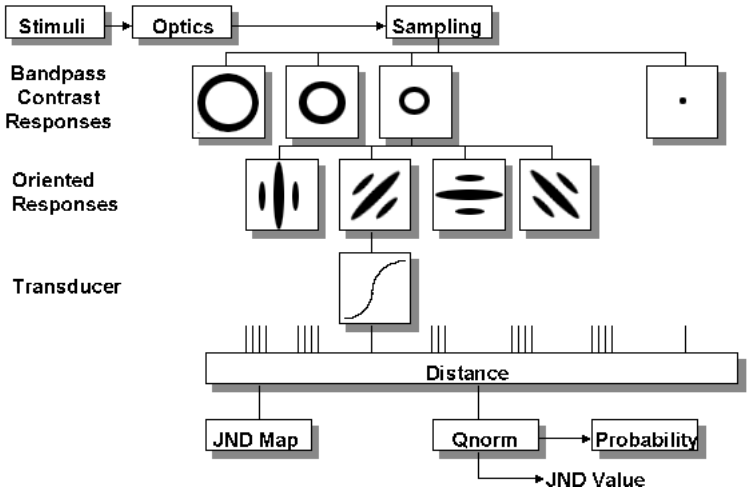

Figure 2. Sarnoff VDM (after Lubin[2])

The first stage of the Sarnoff VDM takes into account the optics of the eye and the retinal mosaic. A single point spread function (PSF) is used to predict the foveal performance of the two-dimensional optics of the eye, under the assumption that the PSF is circularly symmetric. The effect of the PSF convolution is a blurring of the input images. A spatial resampling (120 pixels per degree) is then performed to account for the fixed density of cones in the fovea. The resampling is essential in a spatial domain approach since the extraction of the different frequency bands is totally dependent on the resampling kernels and resampling rates. Special steps must be taken if the original image is too big and the local image quality cannot be assessed in a single glance. This leads to a block dividing process in which a big image is divided into N smaller blocks.

In the Sarnoff VDM a Laplacian pyramid[15] is used to store the wavelet representation of the resampled input image and a quadrature mirrored pair of convolution kernels is used to record information along each of four orientations. After this stage, the raw luminance signal has been converted to units of local contrast. A Laplacian pyramid is used to record decomposed information for all seven band-pass levels. Due to the use of a spatial domain convolution approach, the peak frequency of each level has to be a power of two. The seven bandpass levels have peak frequencies from 32 to 0.5 cycles/degree, where each level is separated from its neighbors by one octave. For reasons of simplicity and performance, a steerable pyramid was actually used to perform the decomposition in both the Sarnoff VDM

and our implementation. The steerable pyramid is a multi-scale, multi-orientation, image transform with both frequency and orientation components.[8,16] The final step in the decomposition process is the computation of a phase-independent energy response by squaring and summing odd-phase and even-phase coefficients. They are obtained by convolving the quadrature mirrored pair filters (oriented operators and their Hilbert transforms) with a certain frequency band.

The normalization stage, as a preprocess to the transducer stage, is the counterpart to the contrast sensitivity function normalization in the Daly VDP. The energy measure is normalized by the square of the reciprocal of the contrast sensitivity function. A transducer is then used to refine the JND map by taking the spatial masking dipper effect into consideration. The dipper shape reflects one characteristic of the contrast discrimination function. This stage involves the transformation by a sigmoid non-linearity. Finally, the model includes a pooling stage in which transducer outputs are averaged over a small region by convolving with a disc-shaped kernel.

After getting the JND difference map for each channel, the last stage is devoted to putting together the contributions from all the channels. This leads to the concept of a space of multiple dimensions. There are 28 channels involved in the summation: seven pyramid levels times four different orientations. For each spatial position, the final JND distance can be regarded as the distance between two 28-dimensional vectors.

Calibration is used to avoid selecting the model parameters on a case by case basis. The procedure is divided into two steps. The first step makes sure the CSF fits the psychophysical data.[17] The second step adjusts the variables in the transducer function so that its outputs match those from human vision. Calibration of the transducer was found to have a large impact on the accuracy of the final detection results. An analysis was done to show the impact that each parameter of the transducer function has on the dipper effect. From these studies the range for reasonable values of the parameters was determined. Informal studies with human subjects further refined the choice of the parameters and showed that optimal detection results are obtained when the parameters are within the predicted theoretical range.

5. DETECTION RESULTS AND PERFORMANCES

5.1. Daly VDP Detection Results

In this section, both the input images and the output detection images of the Daly VDP are discussed and compared. To facilitate comparison with the results from the Sarnoff VDM, the visualization of detection maps is slightly different from that used in the original Daly VDP. Instead of using signed probabilities, we show absolute values. The brightness of each pixel in the detection map is proportional to the probability that distortion can be seen at this pixel. The brighter a pixel, the more likely the distortion will be noticed.

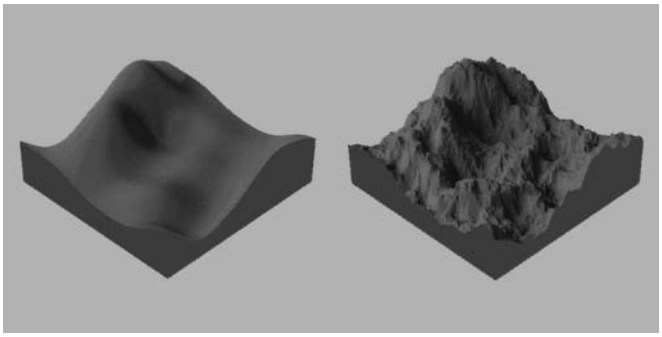

Figure 3. Mountains with Different Levels of Detail

The input images tested include computer generated patterns, synthesized images (the mountain image from Bolin and Meyer[18] in Figure 3), and natural pictures (the chapel image in Figure 6). The distortions introduced into the original images and to be detected by the Daly VDP include blurring (Figure 7a), patterned noise (Figure 4a, 8a), and quantizations (Figure 5a). A standard computer monitor with a resolution of $100dpi$ was used as a display device. The maximum luminance of the monitor was 50 cd/m^2 and gamma correction was done. The results shown below were obtained at a viewing distance of 0.5 meter.

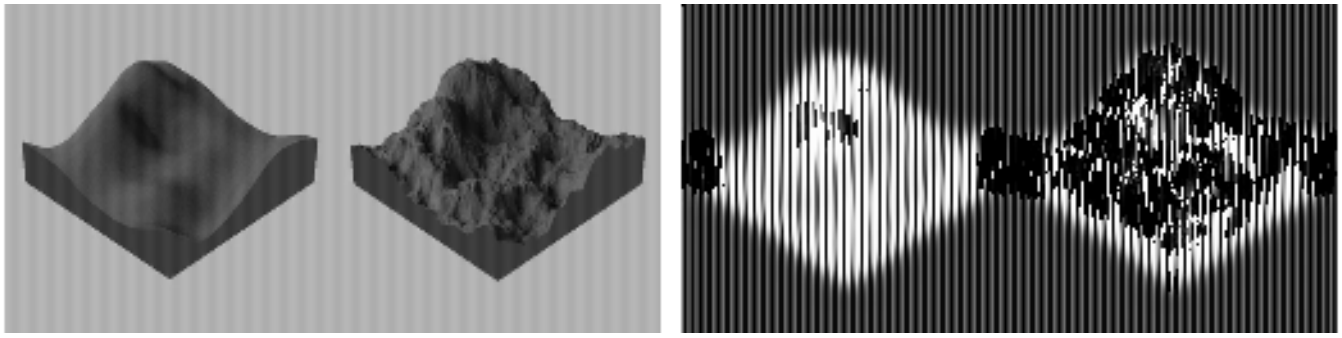

Figure 4. a. Mountains with Sine Waves (9 Cyc/Deg); b. Daly VDP Detection Map

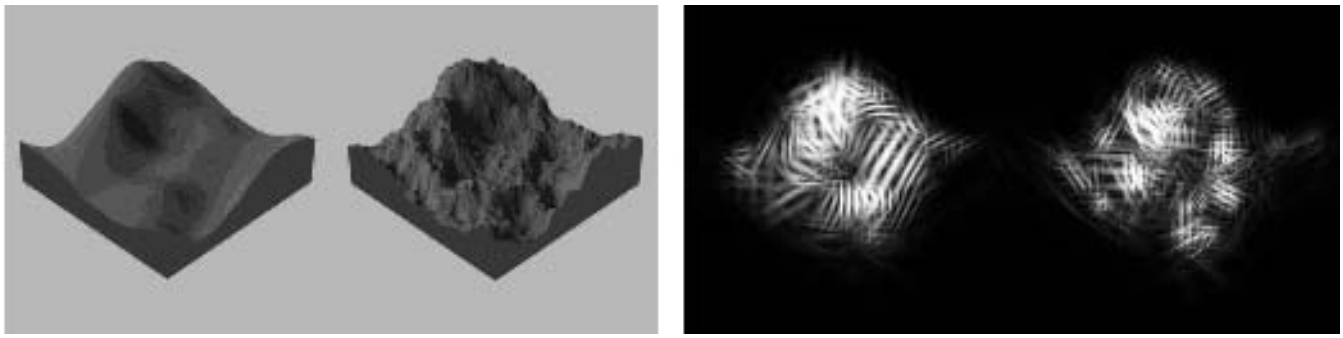

Figure 5. a. Quantized Mountains (4 Bits/Pixel); b. Daly VDP Detection Map

The image in Figure 3 illustrates two mountains with different levels of detail. The gray scale depth of the image is 8 bits/pixel. It is a good test image because it has two distinguishable regions with different frequency ranges. When a sine wave noise pattern (14 cycles/degree) is added onto the original mountain image (Figure 4a) the noise is visible everywhere except in the part of the image containing the rough mountain. The corresponding detection map is shown in Figure 4b. When the image is quantized to 4 bits/pixel (Figure 5a), the banding effect is more visible in the smooth surface of the left mountain than in the rough surface of the right one. The prediction of the model is shown in Figure 5b.

From the previous detection maps we can see that the masking effect is captured by the model. Overall the detection results match what we see when we look at the pictures. However, for these images, the model over-predicts noise in the lower luminance regions. For example, the masking effect is actually stronger in the dark rough mountain surface than predicted by the model. In contrast to that, the model is not sensitive enough to detect the minor distortion in the high luminance background.

For the chapel image in Figure 6, two kinds of distortion were also introduced: blurring and sine waves. The image in Figure 7a is obtained by convolving the original chapel image with a 3 by 3 blurring window. In Figure 7a the blurring effect is very obvious in the area of the window pane, along the edges of the walls, and at the borders of the shadows. The detection results from running the Daly VDP are consistent with these observations (Figure 7b). Sine waves at a frequency of 8 cycles/degree are added as phase-coherent noise in Figure 8a. In this image, the sine wave noise is less noticeable in the window pane area, especially in the right hand roof where the lighting is brighter. The detection map is presented in Figure 8b.

Spatial masking is most effective when the signal frequency equals the noise frequency. By running the Daly VDP on a star image with continuously changing frequency and orientation (taken from the IEEE Facsimile Chart), we found that the VDP correctly detects this effect. When one sine wave is superimposed on top of another, the interference pattern becomes strongest when the two waves are orthogonal and weakest when they are parallel. This effect is captured by the model as well.

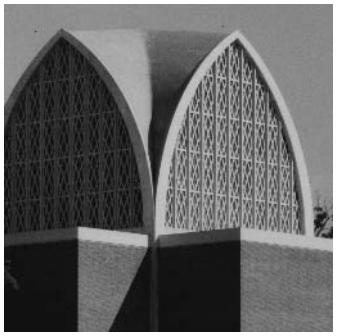

Figure 6. Original Chapel

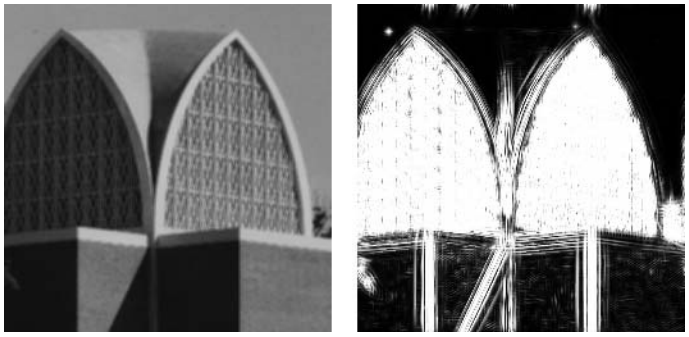

Figure 7. a. Blurred Chapel; b. Daly VDP Detection Map

5.2. Daly VDP Performance

The most time-consuming operations in the Daly VDP are the Fourier transformations. The complexity of the Fourier transformations is $O(N^2)$ where N is the number of entries in the two dimensional matrix. If the FFT and the FFT^{-1} are used before the CSF normalization stage and after the spatial masking stage respectively, the complexity for transformations between spatial and frequency domains can be reduced to $O(N \log N)$. Our analysis shows that up to 40% of the time is used in the FFT and FFT^{-1} stages. The complexity of the FFT determines the overall complexity. The complexity of the model is therefore $O(N \log N)$ with an upper bound of $O(N^2)$.

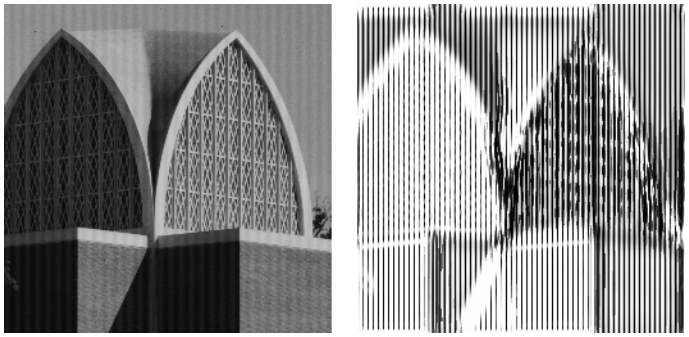

Figure 8. a. Chapel with Sine Waves (8 Cyc/Deg); b. Daly VDP Detection Map

5.3. Sarnoff VDM Detection Results

For comparison, the same input images that were used to test the Daly VDP were also used to evaluate the Sarnoff VDM. The tests were done in the same lighting environment with a standard computer monitor that has resolution of $100dpi$. The maximum luminance of the monitor was $50cd/m^2$ and the viewing distance was 0.8 meter. The reason for choosing 0.8 meter and not 0.5 meter as in the Daly VDP test was that at that distance and with the above display resolution the resampling rate of the retina is roughly 60 cycles/degree which leads to an integer expansion rate in the resampling stage. Convolution interpolation in resampling is easier with an integer expansion rate.

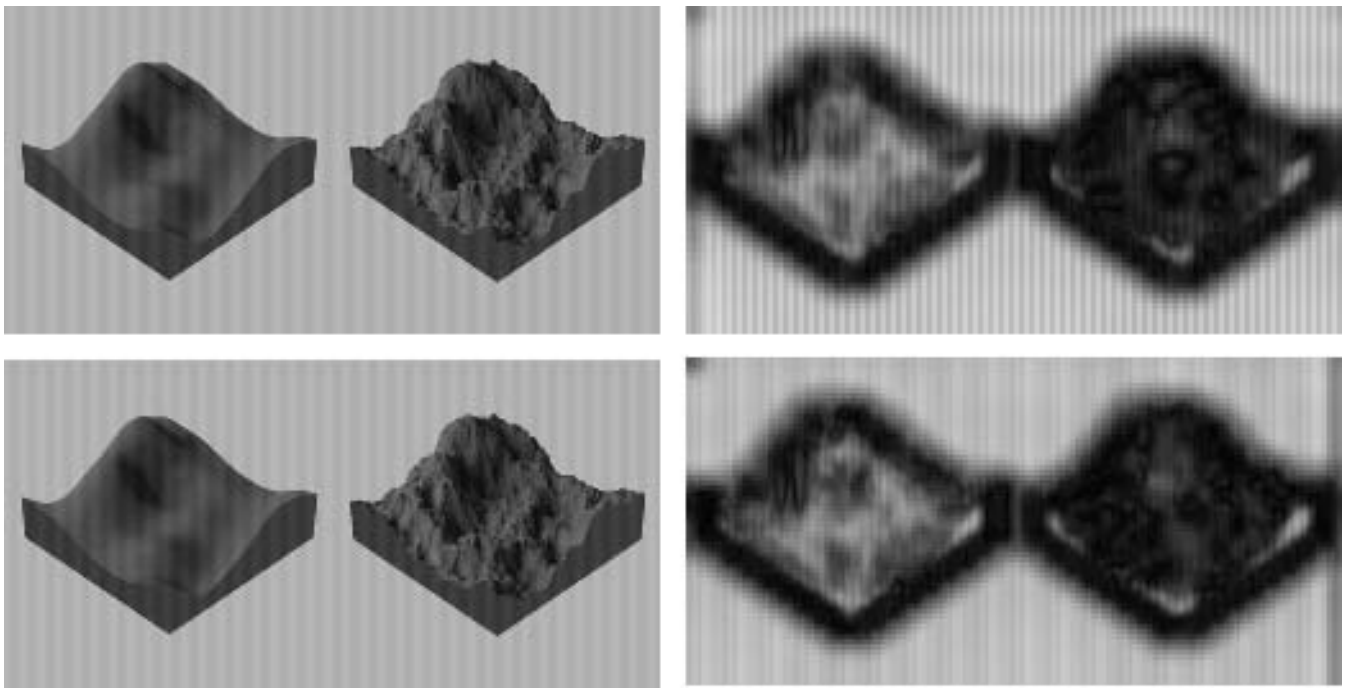

Figure 9. Mountains with Sine Waves (a. 8 Cyc/Deg, c. 9 Cyc/Deg) and Sarnoff VDM Detection Maps (b. 8 Cyc/Deg, d. 9 Cyc/Deg)

The reconstructed image in Figure 9a has a sine wave mask of 8 cycles/degree (one of the seven peak frequencies in the steerable pyramid representation). Due to spatial masking a large distortion difference exists between the two mountain areas in Figure 9a. This distorted image is fed into the Sarnoff VDM along with the original mountain image in Figure 3. As shown in Figure 9b, for this picture the masking effect is accurately predicted. The noise pattern in the background is also properly detected. The maximum JND of this detection map is 4.33 and the mean JND is 2.36.

The quantized mountain image (Figure 10a) at 4 bits/pixel and the original mountain image (Figure 3) at 8 bits/pixel were used as another test pair. The severe quantization aliasing shown in the smooth mountain surface and the strong masking effect in the rough surface of the mountain are both correctly predicted by the Sarnoff VDM. The detection map is shown in Figure 10b with a maximum JND of 2.86 and an average JND of 0.31.

The detection map of the blurred chapel (Figure 11a) is shown in Figure 11b. As in the Daly VDP, the most distorted part of the image, the panes and the edges, is correctly detected. But the detection results indicate a somewhat stronger distortion across the wall than can be observed by the human eye. The maximum and average JND of this detection map are 3.92 and 1.04 respectively. The detection map in Figure 12b for the input image pair of the original chapel (Figure 6) and the chapel with sine waves (Figure 12a) shows a similar prediction, for this image, of the masking effect to that from the Daly VDP. However, the model over-predicts the distortion in the dark area (e. g. the walls in the shadow). The maximum and average JND's for this picture are 7.46 and 1.41.

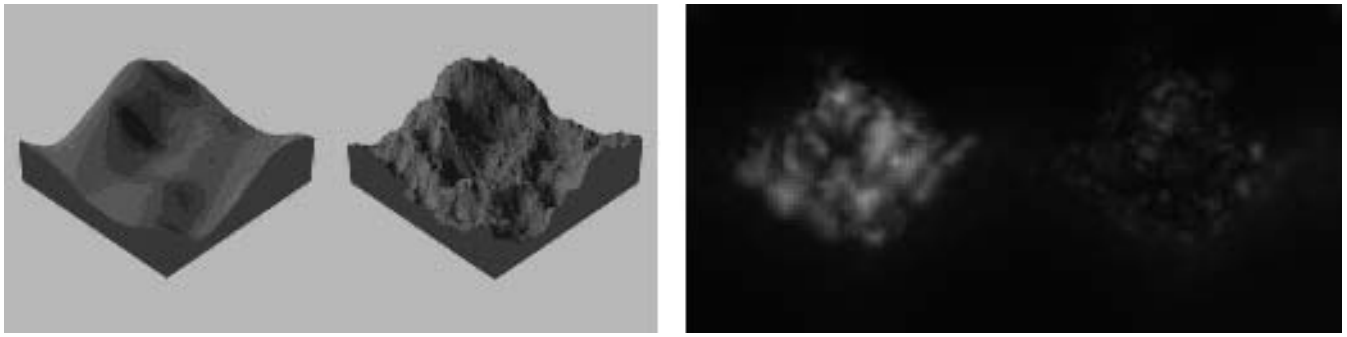

Figure 10. a. Quantized Mountain (4 Bits/Pixel); b. Sarnoff VDM Detection Map

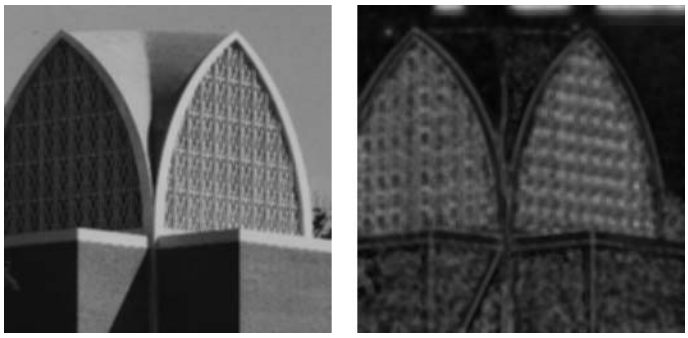

Figure 11. a. Blurred Chapel; b. Sarnoff VDM Detection Map

5.4. Sarnoff VDM Performance

The Sarnoff VDM functions purely in the spatial domain with simple operations. The modeling of each perceptual stage is interpreted either as one-pass filtering (e. g. the PSF blurring, pooling stage), two-pass filtering (e. g. cortex channeling), or as straightforward pixel-by-pixel calculations (e. g. , the CSF normalization). Theoretically, the complexity of the model is linear to the number of pixels in the resampled input images. The upper bound of the complexity is $O(N)$, where N is the number of the pixels. This linear relationship between the execution time and the size of the detected images has been verified by making performance measurements.

However, the Sarnoff VDM gains its speed at the cost of memory. Its image decomposition must record data at all frequency levels and all orientations. The generation and maintenance of the wavelet pyramids, local mean

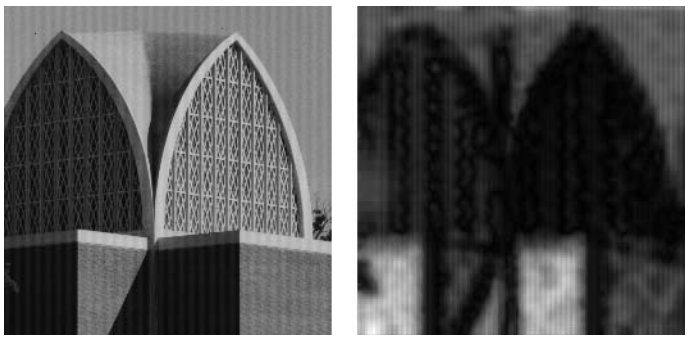

Figure 12. a. Chapel with Sine Waves (8 Cyc/Deg); b. Sarnoff VDM Detection Map

pyramids, and contrast pyramids takes a considerable amount of memory. By comparison, in the Daly VDP only one representation of the image in the frequency domain is needed.

6. COMPARISON OF THE DALY VDP AND THE SARNOFF VDM

The Daly VDP and the Sarnoff VDM each have their respective advantages and disadvantages. The differences between the two models come from 1) the different approaches they represent (i.e. the frequency domain approach vs. the spatial domain approach, accurate threshold modeling vs. good overall supra-threshold results), 2) emphasis on different aspects of human visual perception, and 3) different implementation techniques.

6.1. Similarities Between the Models

While mathematical metrics such as the root mean squared error (RMSE) measure tend to treat the entire human visual system as a "black box," both the Daly VDP and the Sarnoff VDM use physiological and psychophysical data to open the black box. As a result, input images and parameters are needed not only for the system as a whole but also for a number of component mechanisms within.

The threshold concept is used in both models. They both use JND as the metric to quantify the differences between the input images. To generate a JND map as a function of pixel location, the luminance contrast at each pixel must first be calculated. At the next stage it is necessary to apply the CSF normalization to convert the contrast into the JND metric. Spatial masking based on spatial tuning is the final modification of the JND values.

Both models have a decomposition and a summation mechanism. Decomposition based on frequency channeling and orientation tuning makes spatial masking an easier task. The output of the filters which are tuned to different frequencies, orientations, and spatial positions are passed through the summation mechanism to convert the output of those channels into a single map as a function of pixel location.

A pipeline structure is shared by both the Daly VDP and the Sarnoff VDM. Each stage of both models can be modified without interfering with its neighboring stages. Since there are various alternative theories and models to explain each element of the human visual system as a whole, we can always select the most appropriate model for a given application. If there is any advancement in psychophysical study of the human visual system, refinements of the mechanistic models can be easily done without major changes to their basic architecture.

6.2. Differences Between the Models

In the Daly VDP the optics point spread function (PSF) is not explicitly modeled as an element of the human visual system to avoid a shift-variant nonlinearity and the accompanying problem of noninvertibility. If the PSF were used, the blurring effect from convolving the PSF with the input images could lead to a better approximation of the adapted luminance in the retina. This is a coarse approximation, although the process is invertible, which is what is usually preferred for signal processing. In the Sarnoff VDM there is a stage devoted to the optical PSF. However in this model it is assumed that the PSF is circular symmetric, which it is not.

The Daly VDP includes a separate stage to handle the non-linear relationship between brightness and intensity: amplitude nonlinearity. A lightness curve is used to convert the raw luminances into sensitivities. The Sarnoff VDM does not explicitly include brightness nonlinearity.

Although eccentricity can be used as an input parameter in the Daly VDP, the model is mainly dedicated to foveal vision. The original application of the model is the assessment of image fidelity which primarily uses foveal vision. The Sarnoff VDM can be applied to more general situations like aircraft cockpit vision simulations. When an application is limited to image quality measurement, these two models can be regarded as the same as far as foveal vision is concerned.

The Daly VDP is a typical example of the frequency approach. It employs FFT and filtering mechanisms to construct a spatial frequency hierarchy. The Sarnoff VDM only operates in the spatial domain. It builds a steerable pyramid instead of a frequency hierarchy.

The averaging effect in the pooling stage of the human visual system (HVS) is simulated in the Sarnoff VDM when the output of the transducer is convolved with a disc-shaped kernel. The same disc-shaped convolving kernel is used for each transducer output resolution. Therefore, the contributions from the lower frequency signals are more extensively blurred. The Daly VDP does not consider this property of the HVS.

As opposed to the Sarnoff VDM, the Daly VDP uses a simplified spatial masking function without taking the dipper effect into account. When the mask contrast is below the CSF value, no masking is considered. For higher mask contrasts, a constant slope is used to simulate the asymptotic behavior of the spatial masking function. In the Sarnoff VDM, a careful calibration of the transducer is needed to achieve the right dipper effect.

The two vision models have different ways of visualizing the detection results. In the Daly VDP, a psychometric function is used to convert the normalized threshold contrasts into detection probabilities. As a result, the final output visualization is a map of the detection probabilities as a function of location. Omitting the psychometric function, the Sarnoff VDM uses the JND map directly as the final result.

As mentioned in the last paragraph, a psychometric function describing the relationship between the threshold contrasts and detection probabilities is used in the Daly VDP. The mechanical summation in the Daly VDP is the summation of the probabilities, whereas in the Sarnoff VDM it is the computation of the distance between two multi-dimensional JND vectors.

Since the Sarnoff VDM operates solely in the spatial domain, its ability to select signals of an arbitrary frequency is limited. As shown in Section 6.2, the VDM performs best when the dominant frequencies (e. g. phase-coherent sine wave noise) in the input images primarily fall into one of the seven bands. For example, when the frequency of the sine wave noise is 8 cycles/degree, the detection result is correct and clear. If the frequency of the sine wave falls between two neighboring frequency bands (e. g. 9 cycles/degree), the detection result does not produce a distinguishable pattern. To illustrate this, sine wave noise of different frequencies has been introduced into the original mountain image (Figure 3). Two distorted input images are shown in Figures 9a and 9c. The sine wave frequencies in these two input images are respectively 8 and 9 cycles/degree. The detection results are shown in Figures 9b and 9d. In the same order, the maximum JND's are 4.33 and 4.18. The mean JND's are 2.36 and 2.27.

6.3. Common Problems Shared by the Models

Although the mechanism used to handle the local luminance mean in the Sarnoff VDM is more appropriate than the one in the Daly VDP (Section 6.5), it is still not robust. Consider the following implementation problem: if there is a big patch of uniformly black pixels in the input image, the local luminance mean for many pixels in this area will still be zero even though some averaging has been done. If the local luminance mean of a pixel is zero, its contrast computation will be undefined. In our current implementation, a non-zero local luminance mean is found by increasing the number of neighboring pixels for averaging.

Both models face difficulties in finding a correct general CSF representation. In the Daly VDP, the peak sensitivity is picked for different environments. This parameter adjustment has to be done before each application of the VDP. In the Sarnoff VDM, calibration is needed for CSF normalization. However, in different luminance environments, CSF's change and so do the CSF normalizations. Therefore, the question boils down to the following: at which environment/adaptation luminance level should the CSF test and calibration be done to get optimal results?

The number of orientation filters used in these two models is either more than sufficient or just barely enough (Section 6.4 and 6.5). A hybrid of the two could be adopted: four different orientation filters could be used for lower frequency bands where orientation selectivities are relatively weak, and six different orientation filters (or more) could be used for higher frequency bands where orientation selectivities are stronger.

In both models, spatial masking contributions from all channels are treated independently and equally. Cross channel masking is not considered.

6.4. Advantages of the Daly VDP Model

The Daly VDP, like several other psychophysically based approaches, performs in the frequency domain. Frequency domain analysis has given rise to the concept of frequency tuning or channeling which is quite prevalent in psychophysical models. Frequency channeling assumes that there are pathways in the HVS specifically tuned to detect certain spatial frequency stimuli. Moreover, frequency domain analysis (e. g. the CSF) can be easily performed using some well-understood mathematical computations (e. g. FFT and FFT^{-1}).

Recall that the CSF describes the variations in visual contrast sensitivity as a function of spatial frequency. It is more natural to make use of this function in the frequency domain. The advantage of frequency domain models, such as the Daly VDP, is to have a precise and continuous CSF normalization. In the Sarnoff VDM, CSF normalization

is approximated by performing it in only seven discrete frequency bands (levels). For each band, a single peak frequency is used to get the CSF values.

The Daly VDP has a fine simulation of the orientation selectivities. Six orientation filters are used for each frequency band. Although this might be slightly over-complete, six filters do produce more accurate results. In the Sarnoff VDM, only four orientation channels are used. This is acceptable but it introduces some degradation.

When two images are compared and assessed, the mask cannot be derived solely from any one of them. Otherwise, it could not correctly predict bandwidth changes between the two images. The change in frequency content leads to changes in spatial masking and thus the generation of masking maps. Mutual masking is adopted in the Daly VDP so a minimum elevation threshold is used. This produces more plausible threshold elevation maps for all bands.

For the Daly VDP, there is no power-of-two limitation to the size of the image. However the FFT performs best when the base of the image size is a prime number. On the other hand, in the Sarnoff VDM the size of the input image (actually the image size after resampling) needs to be a power of two.

6.5. Advantages of the Sarnoff VDM

The Sarnoff VDM attempts to simulate the functionality of each element along the visual perception pathway. This includes optics, re-sampling, channeling, and cortex spatial masking. Since there is no physiological evidence that the HVS performs Fourier domain processing, the spatial domain model more closely parallels the underlying neural process. The Sarnoff VDM tries to reproduce the same functions that happen along the visual pathway.

Since the Sarnoff VDM performs solely in spatial domain, it is possible to represent the CSF normalization as a function of location. The CSF used in the Sarnoff VDM is a function of the local mean of each pyramid level. Theoretically, a CSF with phase information (i.e. as a function of pixel position) should simulate local luminance adaptation better. However, according to our tests, this refined CSF does not show a remarkable improvement over the CSF obtained with a single adaptation luminance.

In the Daly VDP, the luminance of the pixel itself is used as the local luminance mean under the assumption of an arbitrarily close viewing distance. In the Sarnoff VDM the local luminance mean of each pixel is the average of the luminance of neighboring pixels, which is a better approximation. A more appropriate local luminance leads to a better local contrast.

The complexity of Daly VDP is $O(N \log N)$, as opposed to $O(N)$ in the Sarnoff VDM. The Sarnoff VDM operates only in the spatial domain. It avoids the expensive FFT and FFT^{-1} transformations which take up to 40% of the execution time in the Daly VDP.

In the Sarnoff VDM the CSF normalization is done after the contrast pyramid is obtained. Therefore, distortion introduced by the CSF cannot interfere with the image decomposition. On the other hand, in the Daly VDP the CSF modulation is done before the cortex filtering. The signals in the frequency domain are therefore slightly distorted before spatial selectivities are applied. According to Legge and Foley[19] an important feature of the masking model is the ordering of its elements. It is better to place the linear stages before the cone nonlinearities.

7. CONCLUSION

Working from the published references, two image quality models, the Daly VDP and the Sarnoff VDM, were successfully implemented. Both models were tested on images similar to those in their original publications and comparable results were obtained. This paper thus serves as an independent verification of the algorithms presented in the articles where these two models were introduced.

Each of these methods takes a different approach to modeling the human visual system and computing image quality. The Daly VDP emphasizes threshold accuracy by duplicating psychophysical results concerning the visual system. This leads to a careful computation of the initial nonlinear response to light, the application of the contrast sensitivity function in a continuous frequency domain, and the decomposition of the original image into a relatively large number of orientation bands. The Sarnoff VDM focuses on modeling the physiology of the visual pathway. This produces a careful simulation of the optical PSF, the handling of extra-foveal vision, and the incorporation of an eccentricity dependent pooling stage.

The tests that were performed showed that both models are able to detect the major artifacts that they were designed to identify. On our limited set of images, the Sarnoff VDM was somewhat more robust giving better JND

maps and requiring less re-calibration. However, its limited number of orientation bands did make it susceptible to failure when the frequencies to be detected fell between the available bands. The Sarnoff VDM had better execution speed than the Daly VDP model but at the expense of using significantly more memory.

On the overall, the most important contribution of this paper is the verification of the major features of these models. Both models perform as their authors said they would. However, a complete evaluation would require a larger number of test images and a careful set of psychophysical tests. In this way a more detailed analysis could be performed under a greater variety of conditions, and the models' ability to detect artifacts could be completely characterized.

8. ACKNOWLEDGMENTS

Primary funding for this work was provided by Xerox Corporation with additional support from the National Science Foundation under grant number CCR-9619967.

REFERENCES

1. S. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, ed., pp. 179–206, MIT Press, 1993.
2. J. Lubin, "A visual discrimination model for imaging system design and evaluation," in *Vision Models for Target Detection and Recognition*, E. Peli, ed., pp. 245–283, World Scientific, 1995.
3. T. G. Stockham, "Image processing in the context of a visual model," *Proceedings of the IEEE* **60**, pp. 828–841, 1972.
4. C. F. Hall and E. L. Hall, "A nonlinear model for the spatial characteristics of the human visual system," *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-7**, pp. 161–170, 1977.
5. O. D. Faugeras, "Digital color image processing within the framework of a human visual model," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **ASSP-27**, pp. 380–393, 1979.
6. H. Wilson and J. Bergen, "A four mechanism model for threshold spatial vision," *Vision Research* **19**, pp. 19–31, 1979.
7. A. B. Watson, "Estimation of local spatial scale," *Journal of the Optical Society of America* **A 4**, pp. 1579–1582, 1987.
8. W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**, pp. 891–906, 1991.
9. A. B. Watson, "Efficiency of a model human image code," *Journal of the Optical Society of America* **A 4**, pp. 2401–2417, 1987.
10. C. Zetzsche and G. Hauske, "Multiple channel model for the prediction of subjective image quality," *Society of Photo Instrumentation Engineering Proceedings* **1077**, pp. 209–216, 1989.
11. C. Lloyd and R. Beaton, "Design of a spatio-chromatic human vision model for evaluating full-color display systems," *Society of Photo Instrmentation Engineering Proceedings* **1249**, pp. 23–27, 1990.
12. W. F. Schreiber, *Fundamentals of Electronic Imaging Systems - Some Aspects of Image Processing*, Springer-Verlag, 1993.
13. P. G. J. Barten, "The square root integral (sqri): A new metric to describe the effect of various display parameters on perceived image quality," in *Human Vision, Visual Processing, and Digital Display, Proc. SPIE* **1077**, pp. 73–82, 1989.
14. The use of phase uncertainty filtering by the Daly VDP is described in US patent #5,394,483.
15. P. J. Burt and E. H. Adelson, "The laplacian pyramid as a compact image code," *IEEE Transactions on Communications COM-31* , pp. 532–540, 1983.
16. A. Karasaridis and E. Simoncelli, "A filter design technique for steerable pyramid image transforms," in *Proceedings of ICASSP-96, Proceedings of ICASSP-96* , 1996.
17. O. Schade, "Electro-optical characteristics of television systems. I. characteristics of vision and visual systems," *RCA Review 9* , pp. 5–37, 1948.
18. M. R. Bolin and G. W. Meyer, "A frequency based ray tracer," in *Proceedings ACM SIGGRAPH'95, Proceedings ACM SIGGRAPH'95* , pp. 409–418, 1995.
19. G. E. Legge and J. M. Foley, "Contrast masking in human vision," *Journal of Optical Society of America* **70**, pp. 1458–1470, 1980.