AN ANALYSIS AND COMPARISON OF TWO VISUAL DISCRIMINATION MODELS

by

BEI LI

A THESIS

Presented to the Department of Computer
and Information Science
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Master of Science

June 1997

"An Analysis and Comparison of Two Visual Discrimination Models," a thesis prepared by Bei Li in partial fulfillment of the requirements for the Master of Science degree in the Department of Computer and Information Science. This thesis has been approved and accepted by:

Prof. Gary W. Meyer

Date

Accepted by:

Vice Provost and Dean of the Graduate School

An Abstract of the Thesis of

Bei Li for the degree of Master of Science

in the Department of Computer and Information Science

to be taken June 1997

Title: AN ANALYSIS AND COMPARISON OF TWO VISUAL DISCRIMINATION MODELS

Approved: _____
Prof. Gary W. Meyer

Visual models are often used to analyze the performance of image processing systems. Two of the leading models are the Daly and the Sarnoff model which have been designed to predict the visibility of luminance differences between static input images. They accomplish this by attempting to reproduce the functional responses of every physiological mechanism in the visual pathway of the brain.

These two models are based on the same set of psychophysical facts about human vision. Therefore, they have a similar basic architecture and some similar mechanistic features. However, the Daly and the Sarnoff models take totally different approaches to modeling visual perception: the frequency domain approach and the spatial domain approach respectively.

A comparison of these two models is made based on a detailed description of

their structures and on detection test results. Similarities and differences of both models are discussed along with their strengths and weaknesses.

CURRICULUM VITA

NAME OF THE AUTHOR: Bei Li

PLACE OF BIRTH: Chengdu, China

DATE OF BIRTH: October 22, 1970

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon
Fudan University, China

DEGREES AWARDED:

Master of Science in Computer and Information Science, 1997,
University of Oregon
Bachelor of Science in Electronic Engineering, 1993,
Fudan University

AREAS OF SPECIAL INTEREST:

Image Processing, Computer Graphics, Databases, Electronic Engineering

PROFESSIONAL EXPERIENCE:

Intern, Digital Image Technique Center, Xerox Corporation, Webster, New York, 1996

Teaching and Research Assistant, Department of Computer Science, University of Oregon, Eugene, 1995-97

Assistant Manager, Shenzhen Communication Industry Company, Shenzhen, China, 1993-94

AWARDS AND HONORS:

Prize Winner, National Youth Computer Programming/Designing Competition, 1984, 1985
Prize Winner, National Youth Physics Competition, 1987, 1988
Honor Student of City of Chengdu, 1985, 1988

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to Professor Gary Meyer for his guidance during this work and for his assistance in the preparation of this manuscript. Without his insightful understanding of the subject and his strong sense of the direction vision modeling is heading to, this work would not have been possible.

Special thanks to Andreas Philipp for his industrious proof-reading, constructive suggestions, and technical support. I am also grateful to Xiaofan Fang, Victor Klassen and Mark Bolin for the valuable discussions we had about vision, image processing, and psychophysical experiments. My deep appreciation goes to Weidong Yang for sharing his knowledge in mathematics and optics, for his help in retouching some of the final pictures.

This work is supported by a grant from the Digital Image Technique Center, Xerox Cooperation to Professor Gary Meyer at the University of Oregon.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Human visual models have been developed for a wide variety of applications. These include image compression, distortion detection, halftoning, image synthesis, and image quality measurement. The ultimate goal is to render imagery with optimal visual quality given constraints on computational speed, memory, bit rate, and display. For instance, visual metrics can be used to help assess simulated print quality for different printing specifications. These metrics can be exploited to synthesize more pleasant images to the human eye with limits placed on resolution. A psychologically meaningful quantitative quality metric is very useful in evaluating images. Visual models provide such objective and reproducible metrics.

There has been a wide range of different approaches taken to the measurement of image quality. The most straight-forward but also the most expensive way to measure the fidelity or quality of an image is to collect data from the responses of human assessors. It is a time consuming and subject dependent process. Moreover, the results are hard to reproduce.

There is a need for so-called objective measures, that is, mathematical formulas or algorithms which will predict the visual quality of an imaging system. The most common objective measure of image fidelity is the root mean squared error

(RMSE). Objective metrics performed on machines are less expensive, repeatable and independent of subjective assessment. They are more suitable for the increasing demand on image quality assessment.

Simple objective metrics such as RMSE do not include elements of human vision. The RMSE is not an appropriate measure for many image quality testings, because it incorrectly assumes that errors of equal magnitude are equally visible. It also assumes that the final combined perceptual error is simply the addition of all errors. RMSE is widely used only because of its simplicity.

Many early visual discrimination models fail in many situations where local luminance adaptation or spatial masking is involved. A visual model has to model many features of the human visual system (HVS) before it can serve as an accurate measure of image quality or fidelity. These features must also be based on psychophysical measurements. In general, they consist of various sorts of channeling and non-linear processing of the images.

Psychophysical measurements of perception provide guidance for understanding and building visual models. For example, we can determine the threshold of perception of various components of an image. All information below this threshold can be removed from the image without human detection of the loss. Based on this observation, the contrast sensitivity function has played a critical role in the first vision systems. Some important psychophysical discoveries are discussed in Chapter II.

A new class of visual models have been proposed to accommodate the underly-

ing psychophysical facts about human vision: mechanistic visual models. They have a similar basic architecture and similar mechanical features. These models can be classified into two categories: frequency domain models and spatial domain models. The first class mainly operates in the frequency domain and the second one only in the spatial domain. They are two totally different approaches to capturing the central features of human visual perception.

The Daly model and the Sarnoff model are typical examples of the frequency and spatial domain approaches respectively. They are considered the leading visual models nowadays. Both models share the same structure in their mechanistic visual system (Lubin, 1993), particularly when it comes to modeling early vision.

In the Daly and the Sarnoff model, the following characteristics of human vision are simulated: 1) It is the contrast, and not the linear difference, that determines the visibility of luminance variations. 2) Brightness (or the perception of lightness) is a nonlinear function of luminance. Brightness contrast is processed in a nonlinear fashion. 3) Human contrast sensitivity is a function of spatial frequency. 4) Spatial frequency processing in the human cortex has the properties of radial selectivity and orientation selectivity. 5) Spatial masking reduces the detectability of a given stimulus by the (simultaneous) presence of an additional stimulus.

Both models embody the same known fundamental psychovisual observations and their detection results are psychophysically plausible. Although these two models have evolved and been tested for years, there is relatively little literature devoted to comparing the two. In the interest of comparison we give a thorough analysis of

the models as a whole and of each of their elements. Like most other visual models, these two were originally developed for some specific tasks. For example, the Sarnoff model was specifically tailored to model vision in a cockpit environment. They were designed with different resource priority considerations. Both models have their own advantages and disadvantages.

In this thesis first the background of visual perception and human visual modeling is established. Some important facts drawn from well-known psychophysical measurements are introduced. For each visual model, a detailed description of the model structure is given. The functionalities of each stage are discussed with physiological justification. Detection results and model performance are presented as well. In comparing the two models, the similarities and differences of the two are analyzed along with their strengths and shortcomings. Finally, conclusions are given and future work is described.

CHAPTER II

PSYCHOPHYSICAL FOUNDATIONS

Both knowledge about human vision and an awareness of visual modeling techniques are crucial to understanding and designing a good human visual model. The development of human visual models is closely coupled with advances in psychophysical studies of visual perception. More and more aspects of human vision are being incorporated into visual models to make them more powerful.

In this chapter, some time-tested facts from human visual perception are presented. They are the foundation of all visual models. Human visual perception is so complicated that it is more plausible to design a visual model on the basis of psychophysical facts than on the basis of speculation about physiological structure (Schreiber, 1993). Discussions are limited to the luminance (intensity) domain. Color vision is not covered in this thesis.

Contrast Sensitivity

The absolute amplitude of luminance is not the major factor contributing to the visibility of luminance variations. This is because the human eye adapts over a large dynamic range. Instead it is the contrast that matters. The definition of contrast C is

$$C = \frac{\Delta L}{L}$$

where L is luminance.

The human contrast sensitivity curve is shown in Figure 1 (Schreiber, 1993). When the luminance levels are anywhere between 0.0001 millilamberts and 0.01 millilamberts, the slope of the curve is around 0.42 (or simply 0.5) on a log-log scale.

$$\frac{\Delta L_{min}}{L} = K L^{-0.5} = \frac{K}{\sqrt{L}} = C_{min}$$

where K is a constant.

When luminance rises to the photopic range, the contrast sensitivity curve begins to level off on a log-log plot. This can be seen in the first transition stage in the plot above. From 0.1 millilambert to 800 millilamberts, $\frac{\Delta L}{L}$ is within a factor of 2 of its minimum value. The slope of the contrast sensitivity curve, known as the Weber-Fechner fraction, remains constant over a large brightness range.

$$\frac{\Delta S}{\Delta L} = \frac{dS}{dL}$$

$$\Delta S = \frac{dS}{dL} \cdot \Delta L = constant$$

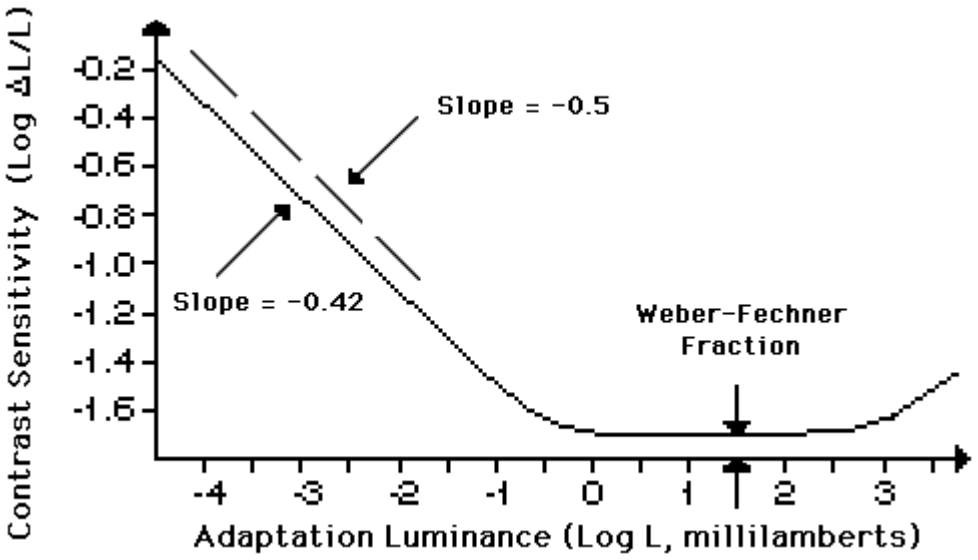

FIGURE 1: Contrast Sensitivity vs. Luminance (after Schreiber 1993)

where

S: Sensation.

ΔS: Minimum detectable change in S.

If brightness rises, saturation sets in and $\frac{\Delta L}{L}$ falls again. This is shown in the second transition stage in the plot.

Brightness vs. Intensity: Nonlinearity

It is widely agreed that there is a nonlinear relationship between brightness and luminance, or in other words, between the magnitude of the sensation and the intensity. There are a number of empirical scales that have been proposed to approximate the relationship of these two quantities: linear, logarithmic, modified

log, power law, bilinear, and Munsell. This will be covered in depth in the Daly model description. We will see that there is a close relationship between the brightness curve and contrast sensitivity. Nonlinearity takes different forms at different intensity levels. There are nonlinearities within and across intensity levels.

Point Spread Function of the Optics of the Eye

The perception of graphic primitives is affected by the pupil aperture which causes them to be blurred. As the light passes from the cornea through the pupil, diffraction occurs due to fringing at the pupil's edge. As a result, when a point of light enters the eye the image projected on the retina has a bell-shaped distribution. This is known as the point spread function (PSF). For line primitives, i.e. stimuli restricted totally along one orientation, a line spread function (LSF) is applicable. A typical PSF and LSF are shown in Figure 2.

The LSF can be used to describe the retinal output image in one dimension. In contrast to that, the PSF can be used to describe the quality of the optical system in two dimensions.

To find a PSF that correctly reflects the anatomy of the eye and the characteristics of the optics system, a shift invariant optical system has to be understood first. Shift-invariance describes one of the properties of the eye's optics. When the position of the input light is shifted, the output pattern on the retina will be shifted by a corresponding amount, but the shape of the output pattern will remain invariant. For example, when the input is a harmonic (sinusoid) of a certain frequency,

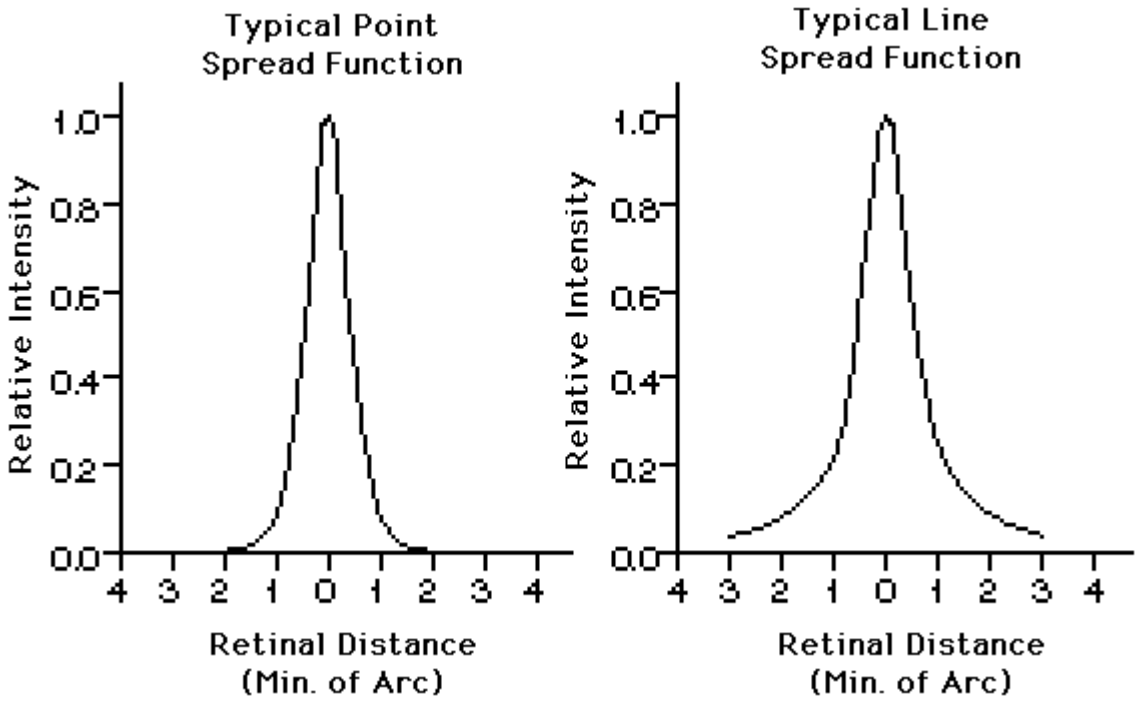

FIGURE 2. Typical PSF and LSF

the output will be a harmonic (sinusoid) of the same frequency although it might have a different scale and phase. The shift-invariance is due to the uniformity of the foveal optics (foveal vision). Shift-invariant systems are a special form of linear systems (Wandell, 1995).

We can take advantage of shift-invariant properties of the eye. Under the assumption that the PSF is circularly symmetric, a single PSF can be used to predict the performance of the two dimensional optics of the eye in foveal vision. But actually the PSF is not circularly symmetric. This will be discussed more in Section IV.1.

Modulation Transfer Function

The modulation transfer function (MTF) is one of the major specifications of an optical system. It is used to describe the performance of the optical system, e.g. the imaging quality of an optical lens, or how well the lens can transfer information. For most optical systems, the MTF takes the form of a low-pass filter if the system is in focus. But for the optical system of the human eye, the spatial MTF is a band-pass filter.

From MTF to CSF: Plausibility and Restrictions

Humans are complex biological systems rather than simple mechanical systems. When it comes to human subjects, a psychophysical measurement of sensitivity is adopted rather than an objective measurement of the human optical system. Correspondingly, the concept of the contrast sensitivity function (CSF) is used instead of the MTF.

The MTF can only be successfully used to describe homogeneous systems, because the MTF contains no spatial description. A system is called spatially homogeneous if its characteristics are constant across space. The optics of the eye are only homogeneous within the fovea or near the optic axis. In the periphery of the retina, the densities of the rods and cones vary. In addition, the visual pathway as a whole is inhomogeneous (Cornsweet, 1970). Fourier techniques only have a limited application in such a case. In the Sarnoff model (Chapter IV), where both

the foveal and peripheral vision are considered, it is inappropriate to make use of Fourier techniques.

A single one dimensional MTF is not sufficient to describe an anisotropic system like the eye. However, two one dimensional MTF's (or a two dimensional MTF) are sufficient since all other orientations can be interpolated from them (Cornsweet, 1970). A system is called isotropic if its characteristics are the same in all orientations.

Due to the astigmatism of human eyes, their optics are anisotropic, which discourages the use of a single MTF. But since effects of anisotropy are relatively small, a second MTF is not necessary (Cornsweet, 1970).

Contrast Sensitivity Function

As mentioned earlier, the human CSF is a band-pass function, instead of a low-pass function. The low spatial frequency drop-off is due to lateral spatial antagonism within the retina (Schwartz, 1994). A typical receptive field for photoreceptors will vigorously respond when different luminances are observed between the center of the receptive field and its surroundings. If spatial frequency decreases (which means, not much brightness variation across space), there will hardly be any luminance variation within receptive fields. Thus sensitivity will drop.

The high spatial frequency cutoff is due to the limited photoreceptor density within the retina (Schwartz, 1994). Each photoreceptor can be seen as a sample taker and all photoreceptors within the retina can be seen as a sampling matrix. It

is obvious that the finer the sampling matrix, the better the resolution will be. Due to biological limitations, the density of the photoreceptor matrix is fixed, which leads to a fixed upper bound of human eye acuity. In addition, due to optical aberrations any optical system shows high-frequency limitations even if it is perfectly in focus.

There is a difference between the detection and the discrimination of contrast. It is easy to find a stimulus threshold in psychophysical experiments, but it is very difficult to deal with perturbations above the threshold (Schreiber, 1993).

Selectivities in Cortex

The cortex has both frequency and orientation selectivity. It has been shown that (under certain conditions) the visual system has the ability to separate signals into different frequency ranges. This channeling concept is directly embraced in many image processing applications, such as orientation analysis (edge detection), subband coding, and multiresolution image splines (Burt and Adelson, 1983B). It was used in HVS-based coding (see the Daly model in Chapter III) and wavelet coding (see the Sarnoff model in Chapter IV) as we shall see later.

It is safe to point out that the independent detection capability of each channel does not exclude interactions between channels in human vision. There is no reason to assume an independent quantization of frequency-specific coefficients.

Spatial Masking

The term masking is commonly used to refer to any destructive interaction or interference among transient stimuli that are closely coupled in space or time (Legge, 1980). Masking designates the reduction of detectability of a given stimulus by the simultaneous presence of an additional, in general, suprathreshold stimulus. Masking occurs between periodic patterns (e. g. sinusoidal gratings) with similar orientation and radial frequency, and between aperiodic patterns (e. g. at luminance borders).

Masking effects can be measured and described in terms of luminance threshold elevation T_e. It is defined as the difference between two thresholds: 1) the threshold needed to distinguish the signal in the presence of the mask, and 2) the threshold needed to distinguish the signal in the presence of the uniform background. The basic masking effect can be generalized in a generic masking plot, shown in Figure 3. In this plot, zero slope indicates no variation in T_e. In other words, there is no masking effect when the masking contrast is low. The threshold elevation rises when normalized mask contrast increases indicating a rising masking effect.

The dipper effect, also known as negative masking, is not shown in Figure 3. Several studies have shown that a low contrast masker increases the detectability of a signal (Legge, 1980). Accordingly the threshold variation becomes negative. Correspondingly there should be a dipper segment in the T_e curve when the mask contrast is low.

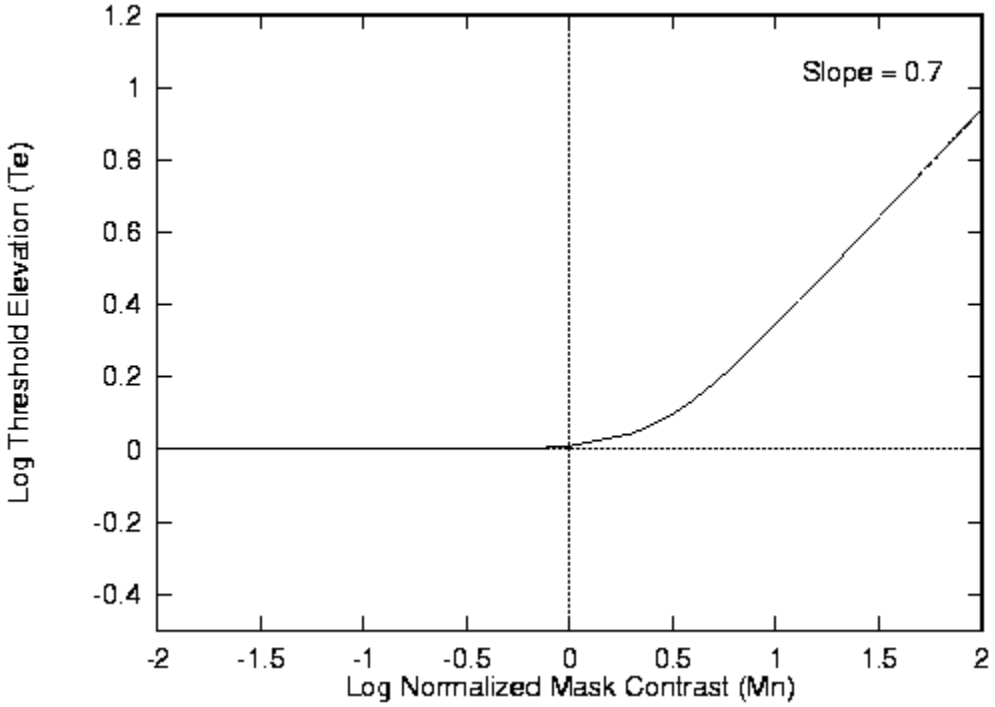

FIGURE 3: Threshold Elevation (T_e) vs. Normalized Masking Contrast (M_n)

CHAPTER III

DALY MODEL

The Daly model, also called the visible differences predictor (VDP), interprets early vision behavior, from retinal contrast sensitivity to spatial masking. It predicts the visible difference between images and mainly operates in the frequency domain.

The VDP models many perception elements. It takes into consideration a number of exclusive issues like mutual masking, phase-coherent masking, and phase-incoherent masking. The Daly model is a threshold model that can handle threshold detection well. However it is incapable of discriminating among different suprathreshold errors and of quantifying them. Although better formulas or modeling structures could be used, it has been shown that the VDP is a working model that produces reasonable results for a number of different data sets.

Description of the Model

The Daly model computes the visible difference map for two images. Given a pair of input images and a set of parameters for viewing conditions and calibration, the output of the model is a probability map for detecting the difference between the two input images. The general structure of the model is shown in the block diagram in Figure 4.

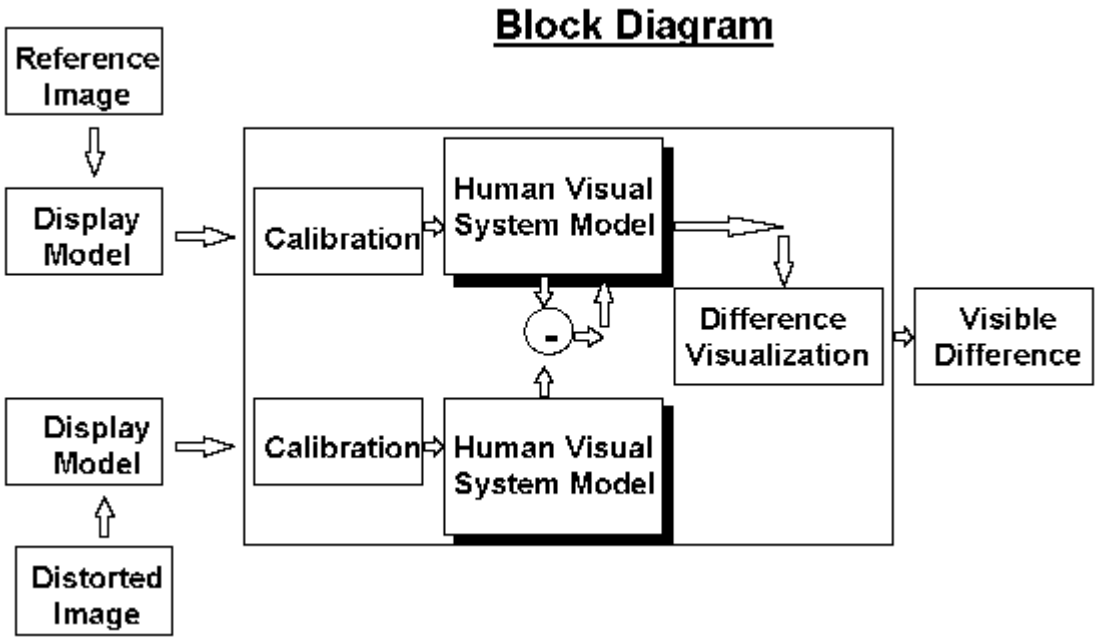

FIGURE 4. Daly Model

The key element of the VDP is a model of the human visual system (HVS). This model consists of three stages: amplitude non-linearity, the contrast sensitivity function (CSF), and the detection mechanisms. A simplified HVS is shown in Figure 5. While the first stage, amplitude nonlinearity, describes the relationship between visual sensitivity and intensity in the spatial domain, most of the other two stages operate in the frequency domain. The CSF characterizes the human response to contrast. The most complicated element of the VDP, the detection mechanism, incorporates visual selectivities and masking into the model. After spatial masking, the model returns to the spatial domain to sum up the detection results in the pooling stage. The final output of the model is a third image showing the predictability of visible difference as a function of pixel locations.

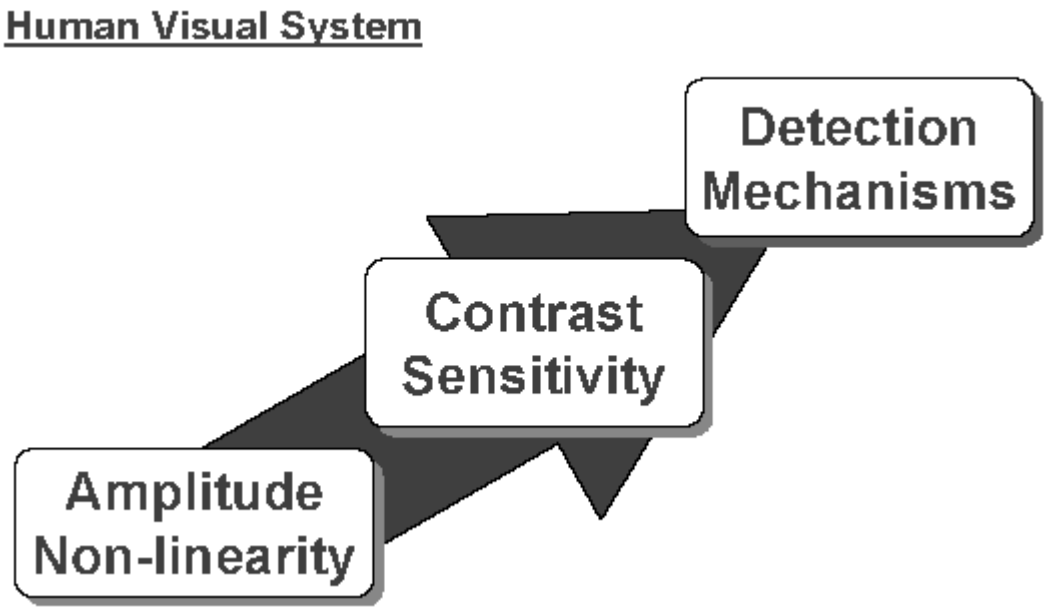

FIGURE 5. Human Visual System

By using actual input images and pointwise operations, the VDP preserves relative spatial information (i.e. phase relationships). For data transformation from the spatial domain to the frequency domain and back, usually the fast Fourier transformation (FFT or FFT^{-1}) is used. The frequency domain is suitable for applying the CSF and detecting spatial masking. In addition, frequency domain techniques are mathematically tractable and well understood. Theoretically the FFT followed by the modulation transfer function, or MTF (CSF is the inverse of MTF), is only suitable for linear systems. Nonlinearities of the algorithm and of the visual system itself thus pose problems to the FFT and the MTF. However the influence of nonlinearity on the mathematical computation can be shown to be limited.

Local Non-linearity

Nonlinear amplitude response is regarded as one of the most important perceptual behaviors. As shown in Figure 6, it describes sensitivity variations as a nonlinear function of intensity.

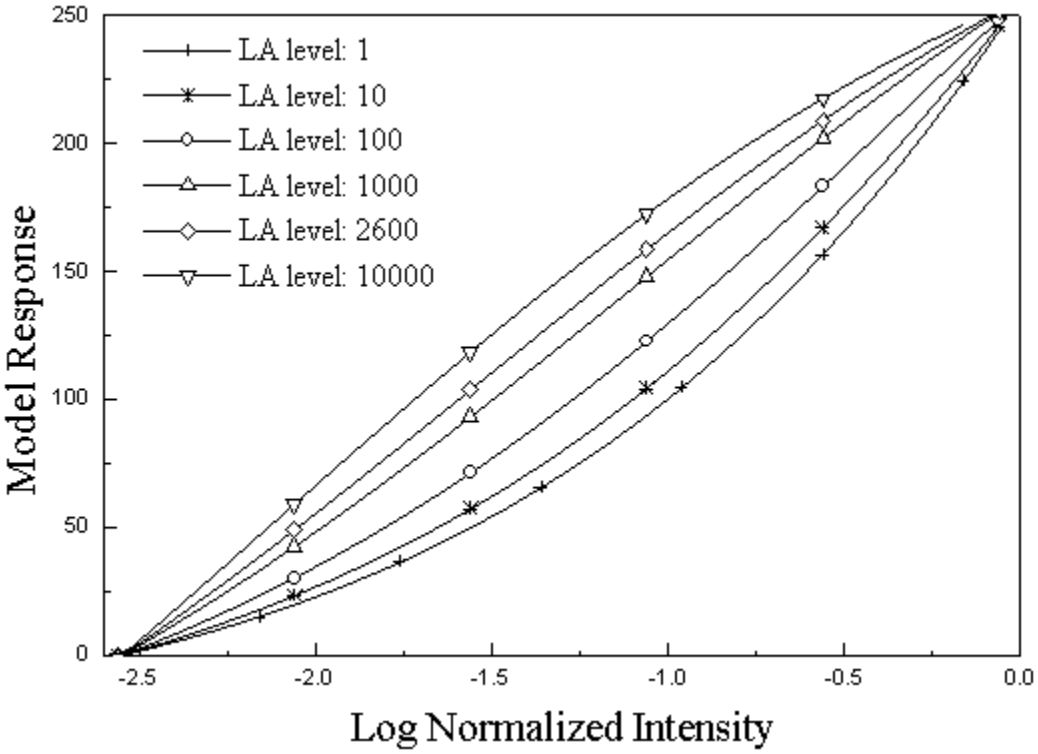

FIGURE 6. Perceptual Sensitivity vs. Intensity.

In Figure 6, different curves represent different illuminance levels. The slope of each curve varies with intensity. The way in which the slope varies is different from one curve to the other. Nonlinearities exist both within and across intensity levels. In the Daly model, a major assumption is made concerning the intensity adaptation of the human eye. It is assumed that the observer may view the image at an

arbitrarily close distance and that their visual system can adapt to the intensity level of a pixel. With this in mind, the amplitude nonlinearity is expressed as a function of pixel location (i, j):

$$\frac{R(i,j)}{R_{max}} = \frac{L(i,j)}{L(i,j) + (cL(i,j))^b}$$

where

L: Luminance.

i, j: Pixel location.

b: Constant 0.63.

c: Constant 12.6.

R/R_{max}: Normalized response.

The series of curves in Figure 6 shows the effect of long-term adaptation on the instantaneous response curves. Each response curve is normalized by its maximum response so that it lies on a 2.55 log range (Daly, 1993).

The local nonlinearity scale can also be justified by the relationship between lightness curves and contrast sensitivity functions (Schreiber, 1993). From the earlier description of Weber's law, we have

$$\Delta S = \frac{dS}{dL} \cdot \Delta L = K$$

$$\Longleftrightarrow \Delta L = \frac{K}{\frac{dS}{dL}}$$

$$\Longleftrightarrow \frac{\Delta L}{L} = \frac{K}{L \cdot \frac{dS}{dL}} = C$$

So given the nonlinear function of brightness (or the lightness curve) the contrast sensitivity C, defined as $\frac{\Delta L}{L}$, can readily be calculated. In other words, given luminance (or intensity) the contrast response is known as well.

There are a number of approximations for the nonlinear relationship between brightness sensation S and luminance L. Two of the most important ones are discussed below.

Case 1

If S and L follow the logarithm law, then

$$S = K_1 \cdot \log L$$

$$\Longrightarrow \frac{dS}{dL} = \frac{K_1}{L}$$

$$\Longrightarrow \frac{\Delta L}{L} = \frac{K}{L \cdot \frac{K_1}{L}} = C = K_2$$

The relationship between constant C and luminance L fits the Weber-Fechner Fraction portion of the curve in Figure 1 (Schreiber, 1993).

Case 2

If S and L follow the power law (e.g. cube root), then

$$S = K_1 \cdot L^{\frac{1}{3}}$$

$$\Longrightarrow \frac{dS}{dL} = K_2 \cdot L^{-\frac{2}{3}}$$

$$\Longrightarrow \frac{\Delta L}{L} = \frac{K}{L \cdot K2 \cdot L^{-\frac{2}{3}}}$$

$$C = K_3 \cdot L^{-\frac{1}{3}}$$

The relationship between C and L fits another segment of the contrast sensitivity curve in Figure 1 where the visual acuity, $\log(\frac{\Delta L}{L})$ or $\log C$, is proportional to the reciprocal of $L^{0.42}$ (or sometimes for simplicity $L^{0.5}$). The corresponding luminance level of this segment is lower than the normal case where Weber's Law holds.

Discussion

These two cases show that when the luminance level is low, the power law of the lightness curve is more suitable for approximating the nonlinear relationship between S and L, while the logarithm law is a better fit when the luminance level is higher. In both cases, however, the same curve is used across all luminance levels. Neither the logarithm law alone nor the power law alone can accommodate the entire effect of lightness adaptation for human eyes.

The local non-linearity law adopted in the Daly model tries to capture the nonlinearities both within and across different luminance levels. In the plot of the

lightness curves used in the Daly model (Figure 6), the sensitivity curve is more like a power function (case 2 above) when luminance is low and more like a log function (case 1 above) when luminance gets higher.

Contrast Sensitivity Function

The CSF is defined as the inverse of the modulation transfer function which in turn is defined as the contrast required to produce a threshold response at a particular spatial frequency. Therefore the band-pass shaped CSF describes how visual sensitivity varies with spatial frequency.

Before discussing the CSF further we give the definition of contrast used in the Daly model. This definition is called local contrast. It uses the baseband image as the mean of the image.

$$C = \frac{B(i,j)}{B_K(i,j)}$$

where

i, j: Pixel location.

$B(i,j)$: Value of the image at pixel (i,j).

$B_K(i,j)$: Value of the baseband at pixel (i,j).

The CSF employed in the Daly model generalizes all global spatial frequency effects, no matter where they actually occur along the visual perception path. This approximation is a function of viewing parameters, adaptation parameters, and

eccentricity. Its expression is given as follows:

$$S(\rho, \theta, i^2, d, e) = P \cdot \min(S_1(\frac{\rho}{r_a, r_e, r_\theta}, l, i^2), S_1(\rho, l, i^2)),$$

in which S_1 is

$$S_1(\rho, l, i^2) = ((3.23(\rho^2 i^2)^{-0.3})^5 + 1)^{\frac{1}{5}} \cdot A_l \varepsilon \rho e^{-B_l \varepsilon \rho} \sqrt{1 + 0.06 e^{B_l \varepsilon \rho}}$$

$$A_l = 0.801(1 + \frac{0.7}{l})^{-0.2}$$

$$B_l = 0.3(1 + \frac{100}{l})^{0.15}$$

where

S: Visual sensitivity.

P: Absolute peak sensitivity of CSF.

ρ: Spatial frequency in $cycles/deg$.

θ: Orientation in degrees.

l: Light adaptation level in cd/m^2.

i^2: Image size in visual degrees.

d: Viewing distance in meters.

ε: A frequency scaling constant. 0.9 for the luminance CSF.

A plot of the CSF described above is shown in Figure 7. The peak of the CSF is at about 4 cycles/degree. The CSF is not perfectly bandpass shaped. While

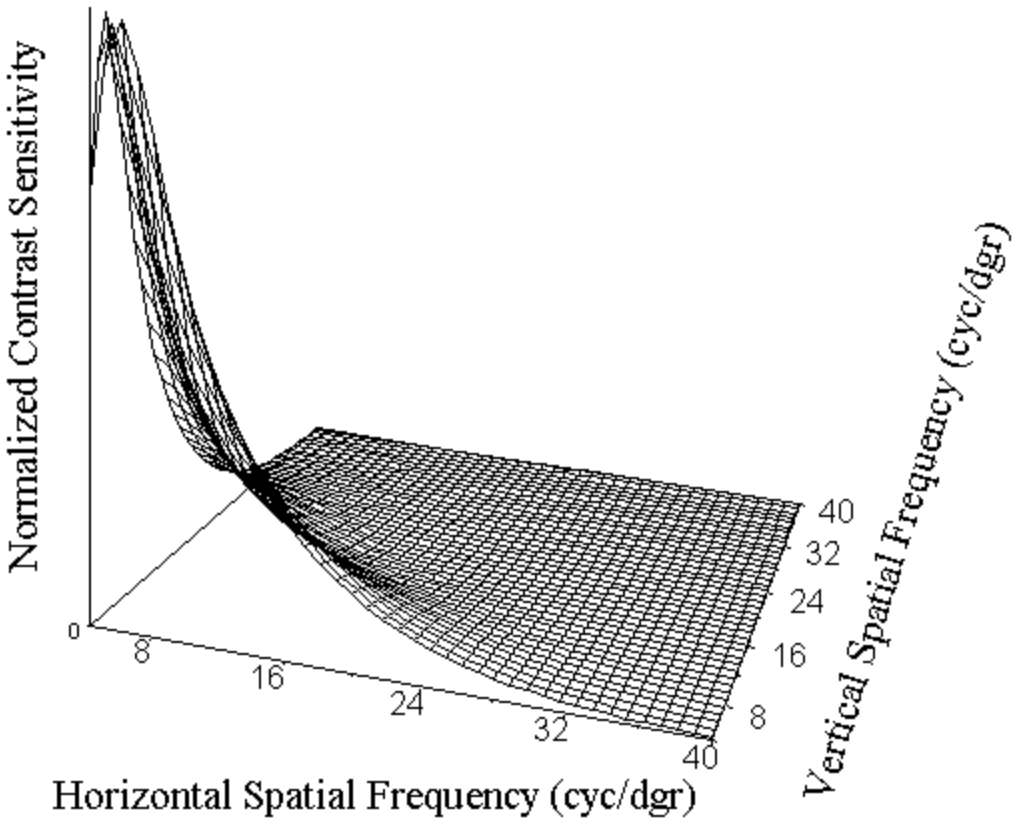

FIGURE 7. Contrast Sensitivity Function

the sensitivity to higher frequencies dies out dramatically, the CSF only drops off slowly in the lower frequency bands. The high frequency cutoff is due to the fixed photoreceptor array and optical aberrations. The low frequency drop-off is caused by lateral inhibition within the retina (Section II.6). This reflects one of the early vision properties that the CSF embodys: a low-pass-like function for suprathreshold signals. In the context of compression, this means that higher frequencies can be more coarsely quantized in frequency based encoding systems without serious visible degradation.

An additional property of the CSF is orientation anisotropy. The CSF shows a preference for vertical and horizontal orientations. This characteristic is also known as the oblique effect.

In our implementation a fixed viewing distance has been adopted. The eccentricity is assumed to be irrelevant for our applications and is therefore set to zero. We have found that using the original Barten MTF (Barten, 1989) produces better detection results when the peak sensitivity P is not a fixed number but a function of adaptation luminance.

Detection Mechanisms

The third stage, detection mechanisms (Figure 8), consists of a number of substages: image channeling, spatial masking, psychometric function, and probability summation.

Cortex Filtering

First we will discuss the cortex filters that channel the input image into a set of band-filtered images. From the diagram shown in Figure 8, we see that after this stage, the input image is fanned out from one channel to 31 channels or bands. Each channel is associated with one cortex filter which in turn consists of a radial filter (*dom* filter) and an orientational filter (*fan* filter). The radial filter performs radial frequency selectivity and the orientational filter performs the orientation selectivity. The baseband is an exception where no *fan* filter is needed.

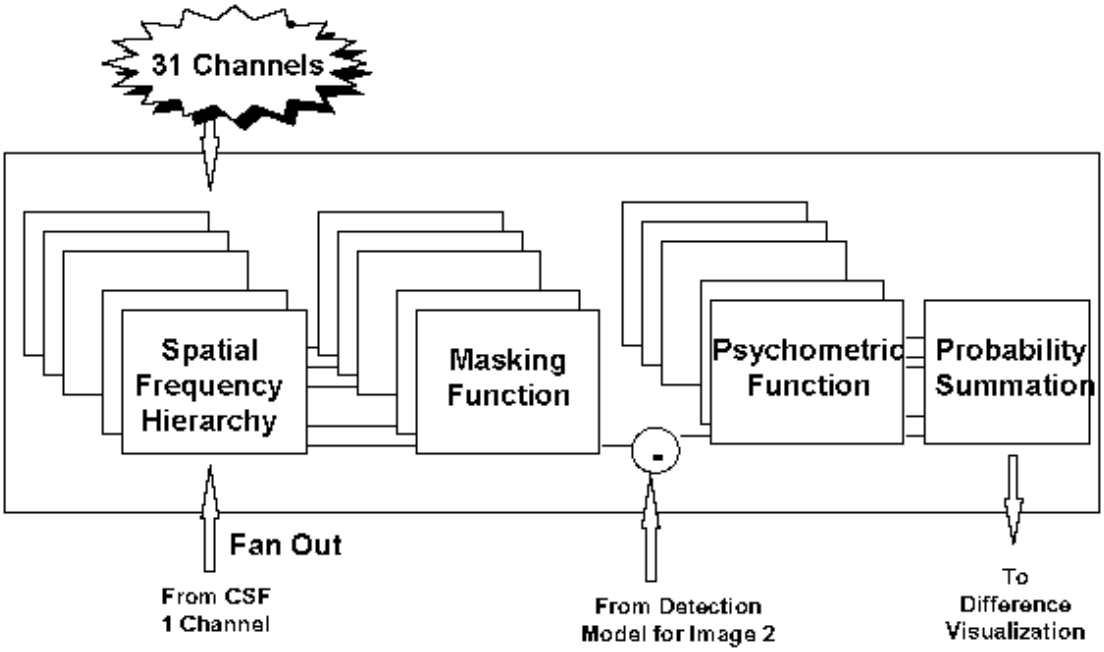

FIGURE 8. Detection Mechanisms

Dom Filters

Dom (difference of mesa) filters are defined as the differences between two consecutive *mesa* filters in the frequency domain. They are a set of overlapping band-pass filters, which extract information of different frequency ranges. The *mesa* functions, basically low-pass filters, are given below.

$$mesa(\rho) = \begin{cases} 1.0 & \text{for } \rho < \rho_{1/2} - \frac{tw}{2} \\ \frac{1}{2}(1 + \cos(\frac{\pi(\rho - \rho_{1/2} + \frac{tw}{2})}{tw})) & \text{for } \rho_{1/2} - \frac{tw}{2} < \rho < \rho_{1/2} + \frac{tw}{2} \\ 0.0 & \text{for } \rho > \rho_{1/2} + \frac{tw}{2} \end{cases}$$

The half-amplitude $\rho_{1/2}$ and the transition width tw of each filter are respectively defined as:

$$\rho_{1/2} = 2^{-k}$$

$$tw = \frac{2}{3}\rho_{1/2}$$

where k is the radial filter level. To compute the *dom* filters, the values of k in the *mesa* function range from 1 to $K-1$. K is the total number of radial filters which is set to six in this model.

For the lowest-frequency band (i.e. the base band), a truncated Gaussian function is used instead of a *mesa* function to weaken the ringing effect which is usually associated with sharp edges of the filter window.

$$base(\rho) = \begin{cases} e^{-(\rho^2/2\delta^2)} & \text{for } \rho < \rho_{1/2} - \frac{tw}{2} \\ 0 & \text{for } \rho \geq \rho_{1/2} + \frac{tw}{2} \end{cases}$$

where

$$\delta = \frac{1}{3}(\rho_{1/2} + \frac{tw}{2})$$

To compute the *dom* filters the possible values for variable k in this function are $K-1$ and K.

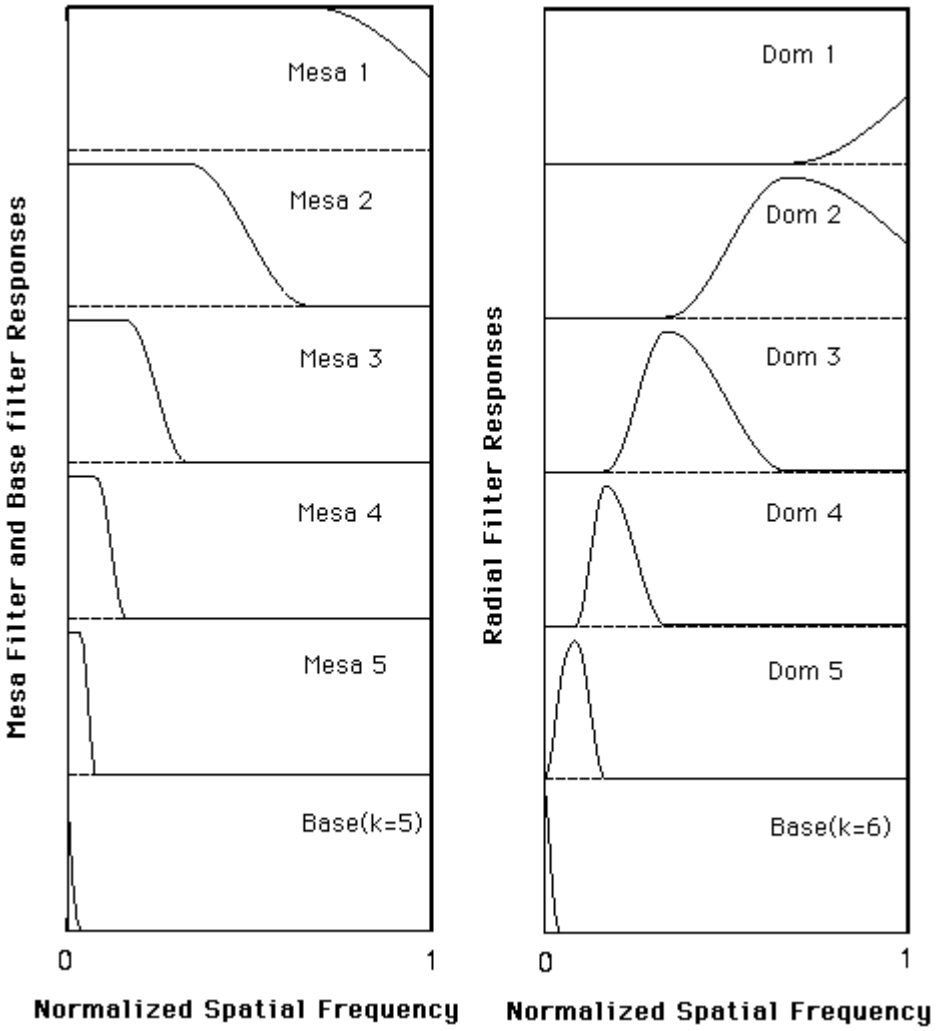

FIGURE 9. Mesa, Base and Dom Filters

In Figure 9, the responses of five *mesa* filters and one base filter with variable k equal to $K-1$ are shown. The response is anywhere between zero and one. With *mesa* functions and base function well defined, radial filters can be derived. There are two types of radial filters: *dom* filters and the base filter. *Dom* functions can be expressed as the formula below. They are basically band-pass filtering windows (Figure 9). As a radial filter the base filter is a base function with variable k equal to K, whose response is shown in the lower-right box in Figure 9.

$$dom_k(\rho) = \begin{cases} mesa(\rho)|_{\rho_{1/2}=2^{-(k-1)}} - mesa(\rho)|_{\rho_{1/2}=2^{-k}} & \text{for } k = 1, \cdots, K-2 \\ mesa(\rho)|_{\rho_{1/2}=2^{-(k-1)}} - base(\rho)|_{\rho_{1/2}=2^{-k}} & \text{for } k = K-1 \end{cases}$$

Fan Filters

Fan filters model the orientation attributes of spatial frequency selectivity. Information concerning different orientations is filtered out by a series of overlapping Hanning windows (Figure 10). Since there is less orientation sensitivity in the lower frequency bands, it is overkill in the Daly model to use six *fan* filters when frequency decreases. In the Daly model the *fan* filter l is expressed as below.

$$fan_l(\theta) = \begin{cases} \frac{1}{2}[1 + \cos(\frac{\pi|\theta - \theta_c(l)|}{\theta_{tw}})] & \text{for } |\theta - \theta_c(l)| < \theta_{tw} \\ 0.0 & \text{for } |\theta - \theta_c(l)| \geq \theta_{tw} \end{cases}$$

where

θ: Any orientation in degrees.

θ_{tw}: Angular transition width in degrees.

$\theta_c(l)$: Orientation of the *fan* filter in degrees.

θ_{tw} and $\theta_c(l)$ are further defined as:

$$\theta_c(l) = (l-1) \cdot \theta_{tw} - 90$$

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

where

L: Total number of *fan* filters, i. e. 6.

Cortex Filters

As mentioned earlier in this section, cortex filters are simply the product of a *dom* filter and a *fan* filter in the frequency domain.

$$cortex_{k,l}(\rho,\theta) = \begin{cases} dom_k(\rho) \cdot fan_l(\theta) & \text{for } k = 1 \cdots K-1, l = 1 \cdots L \\ base(\rho) & \text{for } k = K \end{cases}$$

Figure 11 shows the layout of all *dom* and *fan* filters in the spatial frequency plane. The series of arabic numbers from 1 to 5 specifies the *dom* filters at different frequency levels. *Dom* No.1 covers the highest range of frequencies. The highest

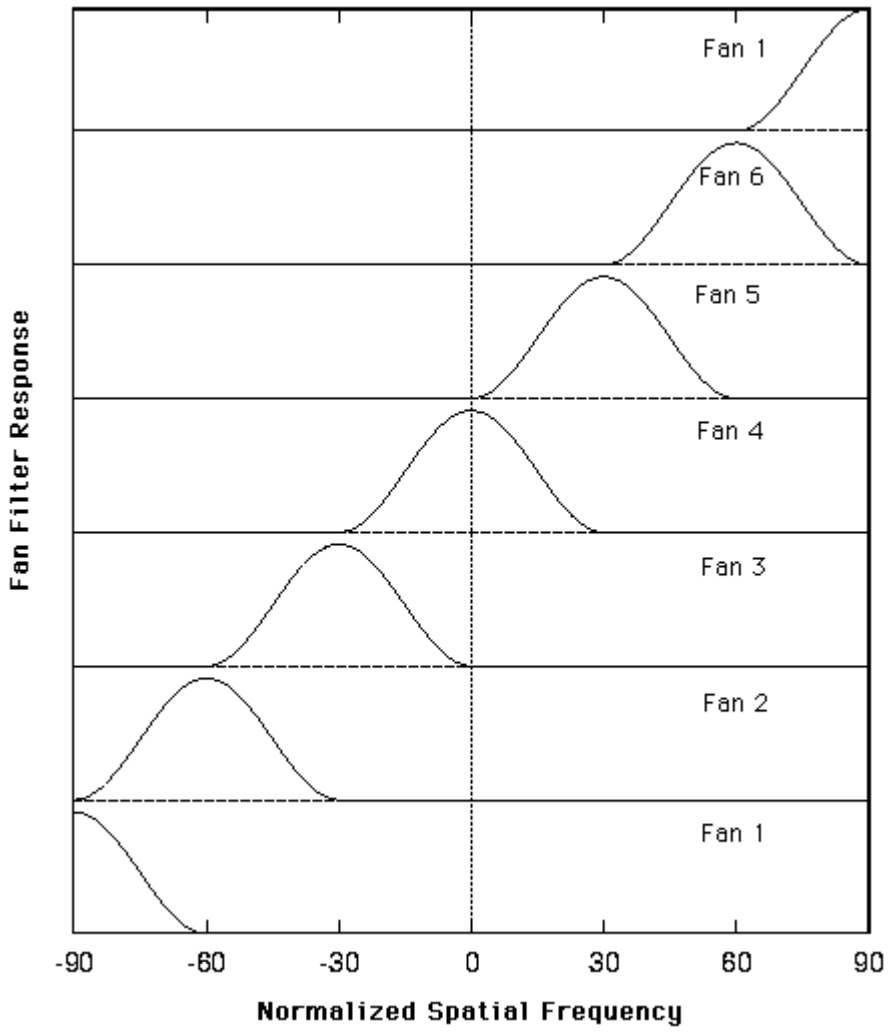

FIGURE 10. Fan Filters

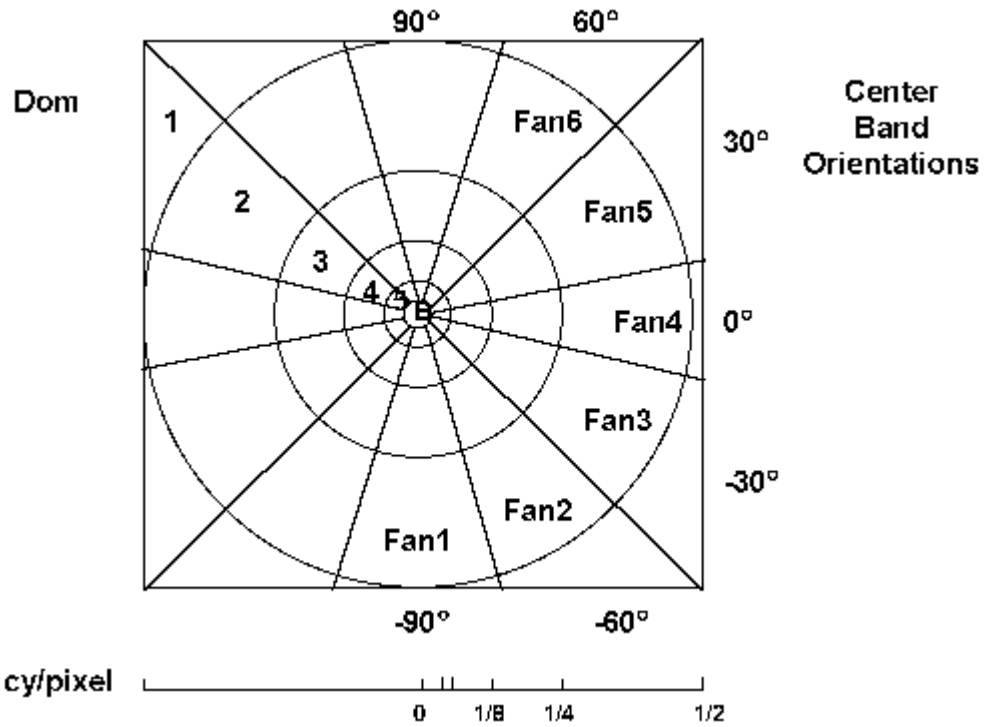

FIGURE 11. Cortex Filter Layout in Frequency Domain

frequency of any image is half a cycle per pixel. The bigger the serial number k of the *dom* filter is, the lower the frequency range it resides in. From Figure 11, we see that the outside radius of successive *dom* filters decreases by a factor of 2. Each frequency band covers a range of one octave in the frequency domain. This way the frequency localization aspects of the visual system can be modeled.

The center band orientation of the *Fan1* filter is 90 degrees or -90 degrees. The *fan* filters are ordered in a counter-clockwise fashion. Each *fan* filters covers a range of 30 degrees. After six sequential *fan* filters, the same order repeats. The orientation discussed here is not directional.

Selectivities

The selectivities are demonstrated in Figure 12. Each filtered image in Figure 12 is obtained by applying either a single *dom* filter or a single *fan* filter to the original star image. The reason we chose the star image as a test image is that it has a wide range of frequencies and all orientation information. Each frequency band except for the base band is further fanned out into six channels of different orientation. For the base band, no orientation selectivity is observed. Thus 5 (frequency bands) times 6 (orientation per frequency bands) plus 1 (baseband) is 31. In this manner a 31 channeled representation of the original input image is obtained.

When a *dom* and a *fan* filter are applied together, information of a certain frequency range and a certain orientation can be filtered out from the source picture. Figure 13 shows two filtered images obtained by using different cortex filters, i.e. different combinations of the *dom* and *fan* filters.

Spatial Masking

As discussed in Section II.8, spatial masking reduces the detectability of a given stimulus through the simultaneous presence of an additional suprathreshold stimulus. A masking function quantifies such a generic effect in terms of the variation of the detection threshold (T_e) as a function of the normalized mask contrast. This relationship is shown in Figure 3 where the signal frequency and mask frequency are the same. When the normalized mask contrast is low, there is no variation in the threshold. When the normalized mask contrast increases, the threshold elevation

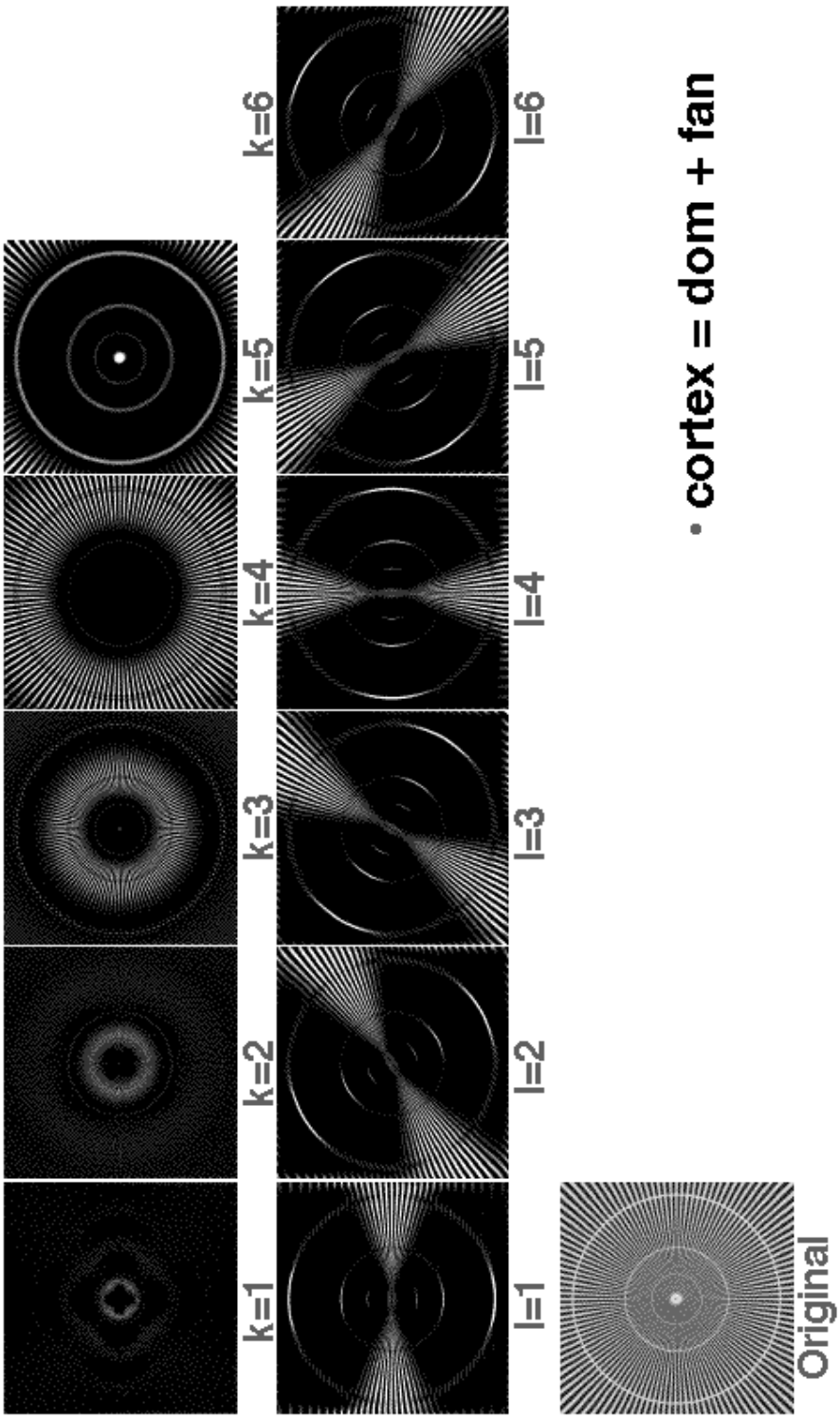

FIGURE 12. Selectivities of the Individual Dom and Fan Filters

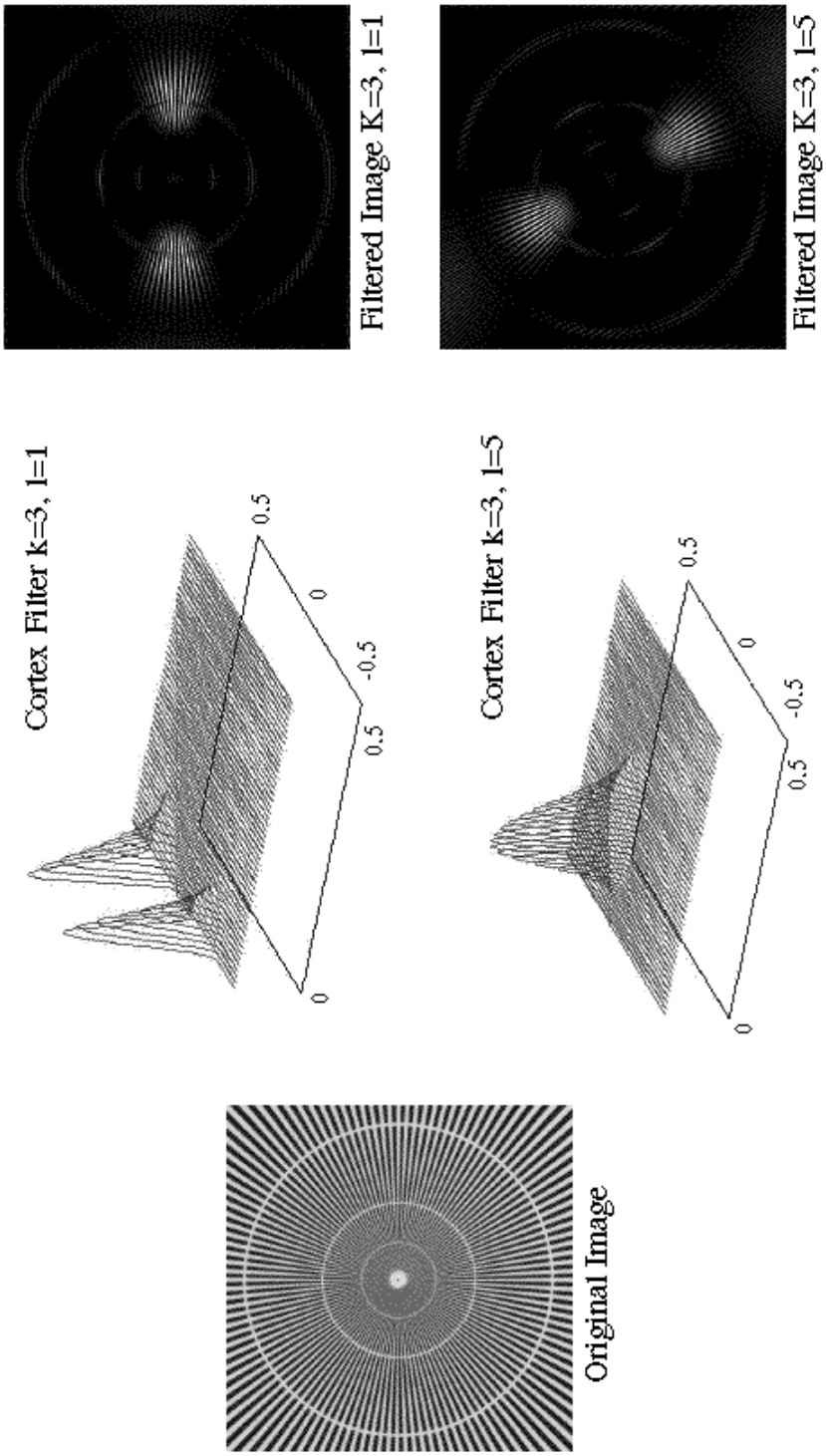

FIGURE 13. Selectivities of Two Cortex Filters

rises monotonically at a fixed rate.

CSF Normalization

In Figure 3, both the horizontal axis and the vertical axis are normalized by the detection threshold at the test frequency f in a uniform field (i.e. , the inverse of the CSF at frequency f). The normalization by the CSF shown in the following expressions makes sure the high mask contrast asymptotes are of the same slope.

$$M_n(\rho, m) = \frac{m}{T(\rho, 0)} = m \cdot csf(\rho)$$

$$T_e(\rho, m) = \frac{T(\rho, m)}{T(\rho, 0)} = T(\rho, m) \cdot csf(\rho),$$

where

M_n: Normalized mask contrast.

T_e: Threshold elevation.

m: Mask contrast.

ρ: Test frequency.

Cortex Filter Normalization

So far, we have only discussed the situations in which signal frequency and mask frequency are the same. When the frequencies of the two signals are different the degree of masking decreases. Accordingly the mask function in Figure 3 will shift

horizontally to the right. Only after the cortex filter normalizations are introduced is it possible to use a single mask function for all frequencies.

By normalizing the horizontal axis with the cortex filter, the masking functions for all signal and mask frequency combinations can be shifted horizontally to ensure that this single curve shown in Figure 3 is suitable for all signal and mask frequency combinations. Now the CSF-normalized mask contrast on the horizontal axis has been further modified as follows:

$$M_n(\rho_t, \theta_t) = m(\rho_m, \theta_m) \cdot csf(\rho_m, \theta_m) \cdot cortex_{k,l}(\rho_m, \theta_m)$$

In this expression, $M_n(\rho_t, \theta_t)$ is the normalized mask contrast at frequency (ρ_m, θ_m) seen by the mechanism that detects the test signal at frequency (ρ_t, θ_t).

Phase-Coherent and Phase-Incoherent Masks

The usage of the masking function also depends on some other factors, such as mutual masking, learning effects, and the nature of the masking noise (Daly, 1993). The masks can be categorized into two kinds: the phase-coherent masks and the phase-incoherent masks. When masks are phase-coherent, e. g. sine waves, the dipper effect emerges. The dipper effect makes T_e drop below the uniform field threshold (i. e. $1/csf$) when the normalized mask contrast is around 1 (Section II.8). The dipper effect is actually not incorporated into the Daly model under the assumption that this effect diminishes very fast when differences in the signal frequency and

mask frequency increase. A simplified version of the threshold elevation is used.

$$T_e(\rho, m) = \begin{cases} \frac{T(M_n)}{T(0)} = [\frac{M_n}{T(0)}]^{0.7} & \text{for } M_n > T(0) \\ 1.0 & \text{for } M_n \geq T(0) \end{cases}$$

where M_n is the normalized mask contrast and $T(0)$ is the uniform field threshold.

When masks are phase-incoherent (e.g. white noise) the slope of the high-contrast asymptote increases to 1, indicating the behavior of Weber's Law. Moreover, the dipper effect does not occur.

Learning Effect and Mutual Masking

The learning effect states that when the mask is easier to learn, the masking effect is less obvious. For example, either longer exposure to the test images or the repeated usage of a certain mask will make detection easier. In general when the predictability of the mask and its familiarity to the observer increases, the spatial masking decreases. In the diagram shown in Figure 14, stronger masking effects are reflected as the steeper slope of T_e in the higher contrast range.

As the last element of spatial masking, mutual masking is used. Since the masking image cannot be robustly derived from the original image or from the distorted one alone, masking images are computed from both input images. With two masking images, a point-by-point comparison between these two images is made. For each pixel the smaller value of the two is chosen as the final T_e.

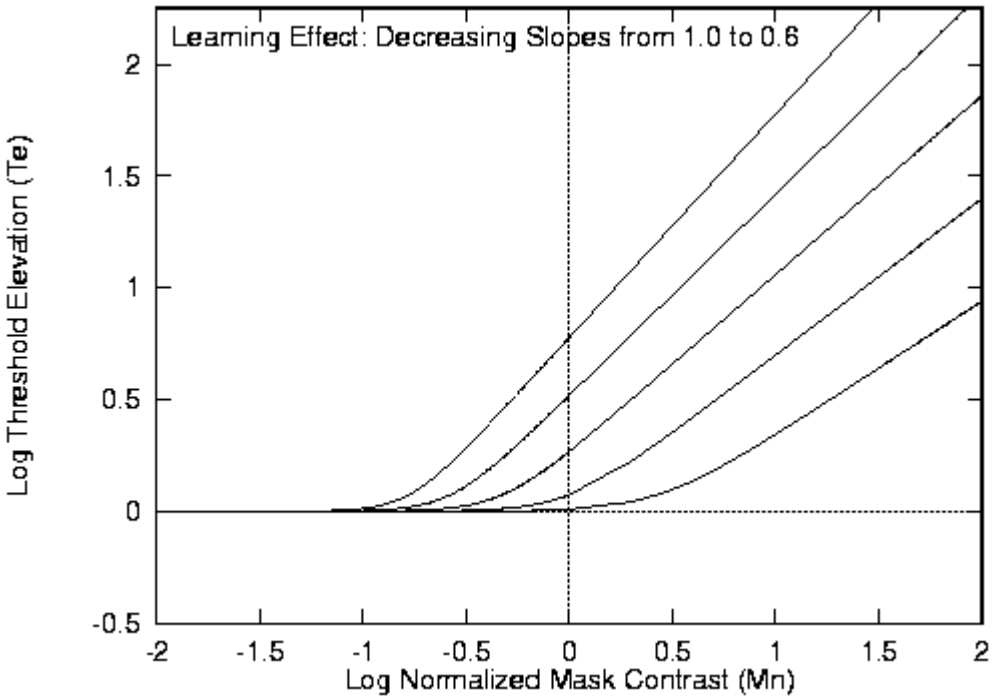

FIGURE 14. Learning Effect in Spatial Masking

$$T_{em}^{k,l}[i,j] = min(T_{e1}^{k,l}[i,j], T_{e2}^{k,l}[i,j])$$

where

$T_{em}^{k,l}$: Threshold elevation map of band k,l from mutual masking.

i,j: Pixel position.

$T_{e1}^{k,l}$: Threshold elevation map for band k,l of image No.1.

$T_{e2}^{k,l}$: Threshold elevation map for band k,l of image No.2.

Operating Domains

In the spatial masking stage, the model first operates in the frequency domain until the CSF and cortex filter normalizations are done. The input images are first transformed into the Fourier domain, followed by CSF normalization and cortex filter normalization. Then each channel k, l is converted back to the spatial domain where the normalized mask contrast m_n can be obtained as a function of location. The whole process is reflected in the following formula of $m_n^{k,l}[i,j]$.

$$m_n^{k,l}[i,j] = \mathcal{F}^{-1}(\mathcal{L}[u,v] \cdot csf[u,v] \cdot cortex^{k,l}[u,v])$$

where

u, v: Cartesian frequency components.

$\mathcal{L}[u,v]$: Fourier transform of the input image processed by the amplitude nonlinearity.

$\mathcal{F}$: Reverse Fourier transformation.

$cortex^{k,l}$: Cortex filter for frequency level k and orientation l.

After 31 normalized mask contrast maps are obtained, a set of 31 threshold elevation maps can be computed with the following formula for each input image.

$$T_e^{k,l}[i,j] = (1 + (k_1(k_2|m_n^{k,l}[i,j]|^s)^b)^{1/b}.$$

Finally, mutual masking is applied between the two sets of threshold elevation maps from both input images. One set of $T_e^{k,l}$ is calculated and piped into the next stage: the psychometric function.

Psychometric Function

Psychometric functions are usually obtained by repeating psychophysical experiments on a large number of individuals. The psychometric function given below describes the increase of the probability of detection as the signal contrast increases.

$$P(C) = 1 - e^{-(\frac{C}{\alpha})^{\beta}}$$

where

C: Contrast.

P: Probability of detecting a signal of contrast c.

α: Threshold for normalization.

β: Slope controller.

When contrast C increases, P increases monotonically at a rate controlled by the parameter β. The change in the value of threshold α leads to the shifting of the curve along the abscissa. Let C_n be the normalized contrast:

$$C_n = \frac{C}{\alpha}$$

Thus the P can be rewritten as:

$$P(C_n) = 1 - e^{-C_n^{\beta}}.$$

$P(C_n)$ is plotted in Figure 15 with linear coordinates. The abscissa C_n is the contrast normalized by the threshold α.

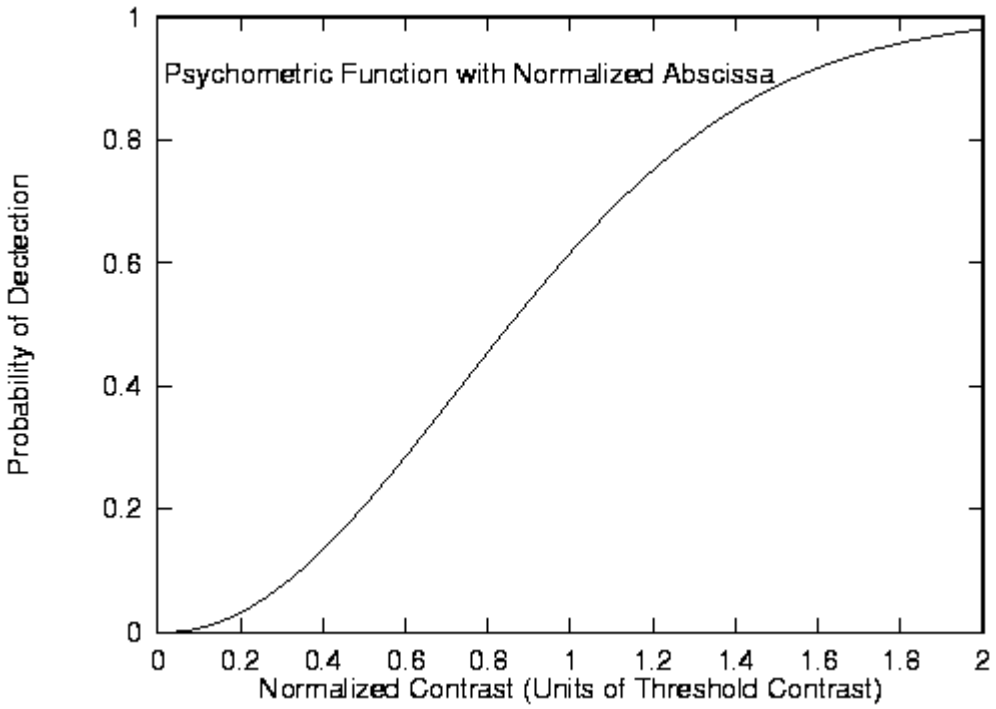

FIGURE 15. Psychometric Function

Probability Summation

Once the detection probabilities have been computed for each band of the spatial filter hierarchy, the probability images are combined into a single image

that, for every pixel, describes the overall probability of detecting an error in the image.

With the psychometric function, we can compute the probability of detection in a single band k, l for each pixel.

$$P_{k,l}[i,j] = 1 - e^{-(\frac{\Delta C_{k,l}[i,j]}{Tem[i,j] \cdot T(0)})^{\beta}}$$

The overall probability is obtained by pooling together probability contributions from all k, l bands as a function of position.

$$P_{total}[i,j] = 1 - \prod(1 - P_{k,l}[i,j])$$

Performance Analysis

Detection Results

In this section, both the input images and the output detection images of the Daly model are discussed and compared. In our implementation, 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.

The input images tested include computer generated patterns (the quarter star in Figure 26), synthesized images, e. g. the mountain image from (Bolin and

Meyer, 1995) in Figure 16, and natural pictures (the chapel image in Figure 21). The distortions introduced into the original images and to be detected by the Daly model include blurring (Figure 22), patterned noise (Figure 17, 24, and 27), and quantizations (Figure 19).

A standard computer monitor with a resolution of $100dpi$ was used as a display device. The maximum luminance of the monitor is about 50 cd/m^2. The results shown below are obtained at a viewing distance of about 0.5 meter.

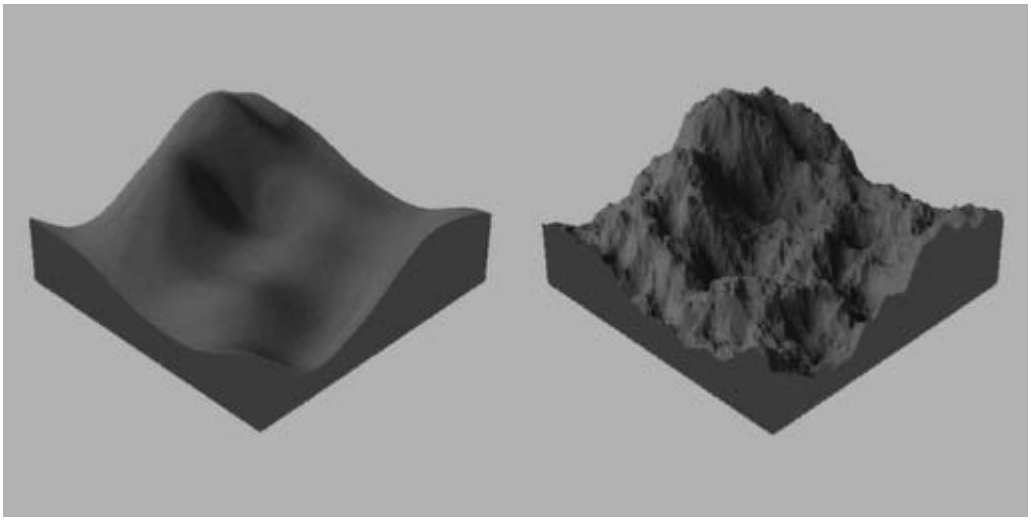

FIGURE 16. Mountains with Different Levels of Detail

The image in Figure 16 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. The impact of the distortion can be shown and compared because these two regions are side by side. When a sine wave noise pattern (14 cycles/degree) is added onto the original mountain image (Figure 17), it is visible everywhere except in the part of the image

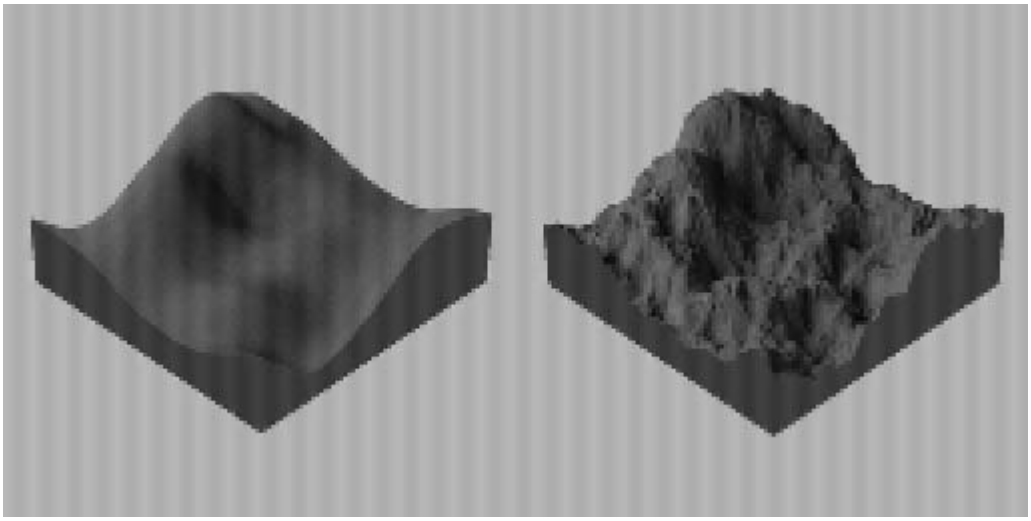

FIGURE 17. Mountains with Sine Waves (9 Cyc/Deg)

containing the rough mountain. The corresponding detection map is shown in Figure 18.

When the image is quantized to 4 bits/pixel, the banding effect is more visible in the smooth surface of the left mountain than in the rough surface of the right one (Figure 19). The quantization affects the appearance of the uniform background only slightly so the distortion is basically invisible there. The prediction of the model is shown in Figure 20.

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 with our eyes. However, 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 distortion in the higher luminance background.

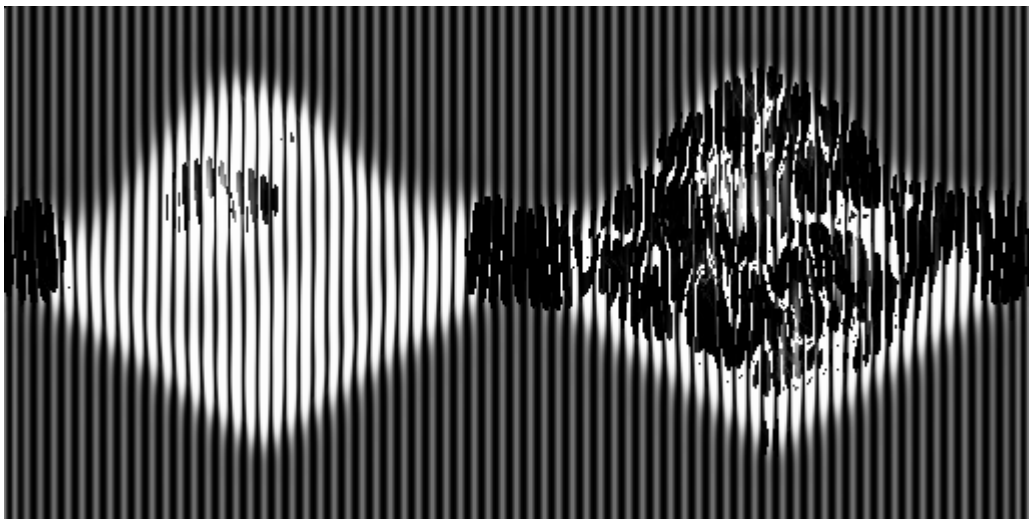

FIGURE 18: Daly: Detection Map of Mountains with Sine Waves (9 Cyc/Deg)

For the chapel image in Figure 21, two kinds of distortion were also introduced: blurring and sine waves. The image in Figure 22 is obtained by convolving the original chapel image with a 3 by 3 blurring window. In Figure 22 the blurring effect is very obvious in the area of window pane, along the edges of the walls, and at the borders of the shadows. The detection results from running the Daly model are consistent with these observations (Figure 23).

Sine waves at a frequency of 8 cycles/degree are added as phase-coherent noise in Figure 24. In this image, the sine wave noise is less noticeable in the window pane area, especially in the right hand arch where the lighting is brighter. The detection map is presented in Figure 25.

A quarter star image (Figure 26) and vertical sine waves are also used as signal and noise respectively. They will be used to demonstrate the masking effect

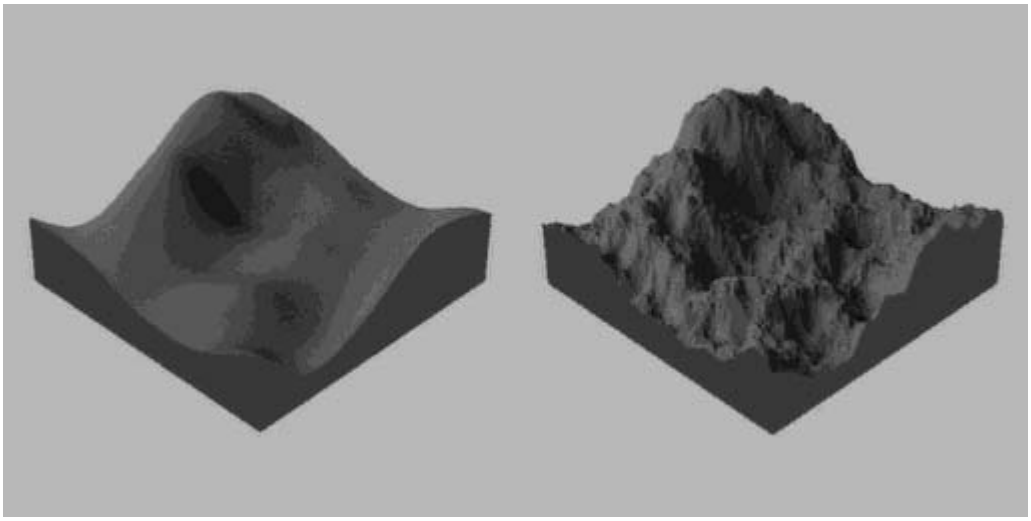

FIGURE 19. Quantized Mountains (4 Bits/Pixel)

and to test the predicting ability of the Daly model. The star pattern is a good test case. The quarter star image, like the full star image in Figure 12, consists of a pattern with a continuous range of frequencies and orientations. Since vertical sine waves will be used as the masking noise, the upper-left quarter of the full star image including the center area (where vertical patterns are located) is sufficient to illustrate the masking effect.

As discussed in Section III.1, spatial masking is most effective when the signal frequency equals the noise frequency. When the two frequencies are different the curve of the threshold elevation in Figure 3 shifts horizontally to the right, which indicates a decrease in the spatial masking effect. The addition of sine wave noise to a signal pattern with continuously changing frequency can clearly illustrate the frequency selectivity.

Likewise, signal patterns having a continuously changing orientation can show

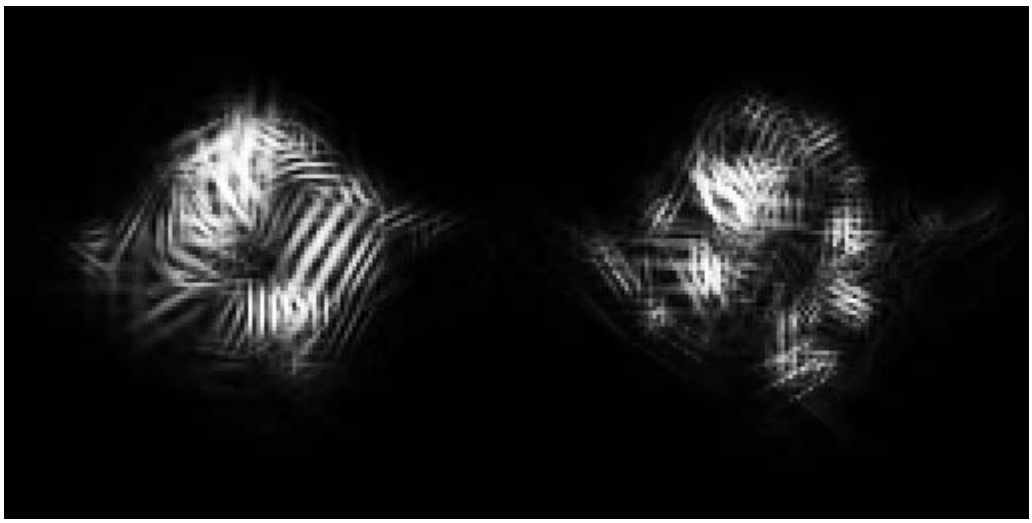

FIGURE 20. Daly: Detection Map of Quantized Mountains

the orientation selectivity of the spatial masking very well. When one sine wave is superimposed on top of another one, the direction and phase of wave patterns has an effect on their interference pattern. The interference pattern becomes strongest when the two waves are orthogonal and weakest when they are parallel.

The influence of frequency selectivity and of orientation selectivity is shown in Figure 27. The vertical sine pattern can be seen in most parts of the image, especially in the high frequency part that is within the second ring from the center. The detection results from the Daly model are shown in Figure 28. In this detection map, the area with the major masking effect is detected by the model. The black region shown in the map is where the frequency in the star pattern matches the frequency of the vertical sine wave noise. This is also where these two patterns are parallel.

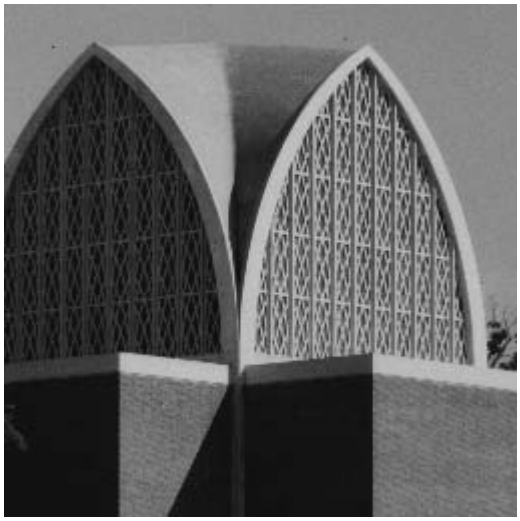

FIGURE 21. Original Chapel Image

Execution Time Profile

The most time-consuming operations in the Daly model are the Fourier transformations. The complexity of the Fourier transformations is $O(n^2)$ where n is the number of entries in the $2d$ 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 $O(n \log n)$. Figure 29 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. So the complexity of the model is $O(n \log n)$ with the upper bound of $O(n^2)$.

The profiling test of the Daly model has been conducted on a Sun Sparc 10. One can see that the results shown in Figure 30 verify the theoretical analysis of the complexity.

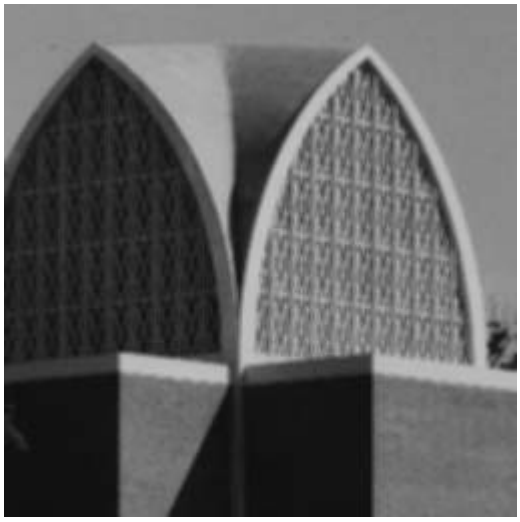

FIGURE 22. Blurred Chapel

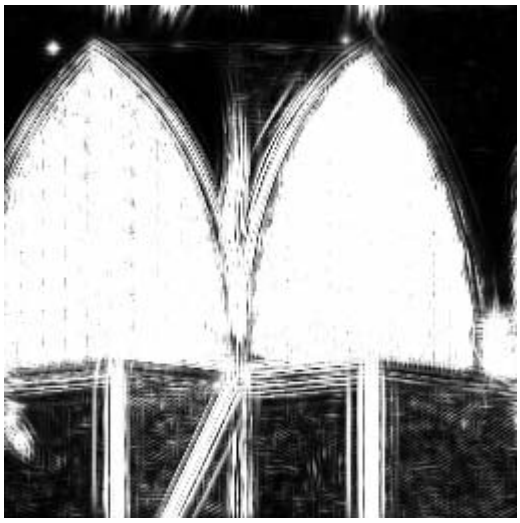

FIGURE 23. Daly: Detection Map of Blurred Chapel

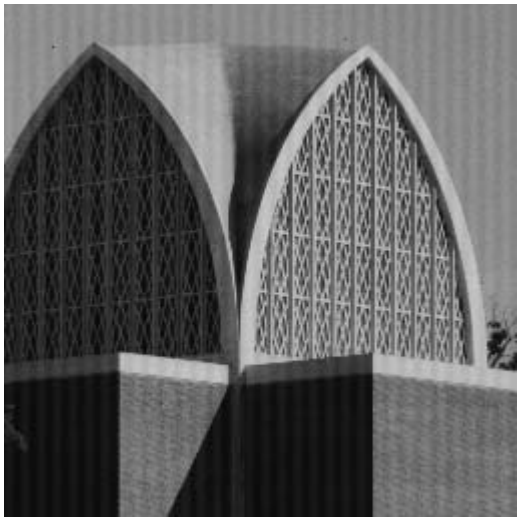

FIGURE 24. Chapel with Sine Waves (8 Cyc/Deg)

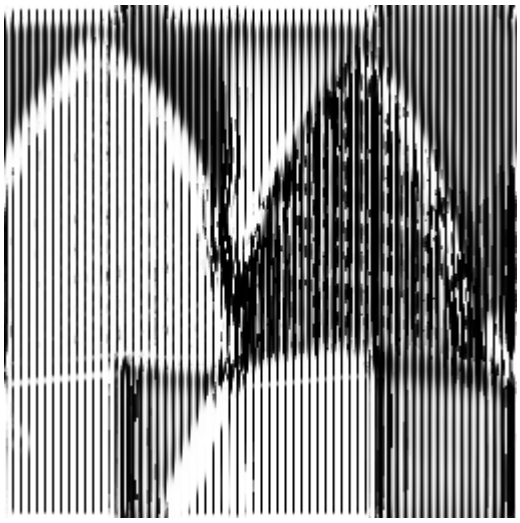

FIGURE 25. Daly: Detection Map of Chapel with Sine Waves

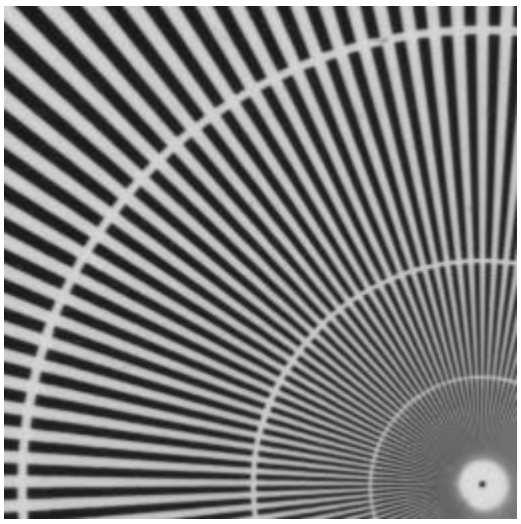

FIGURE 26. Original Quarter Star Image

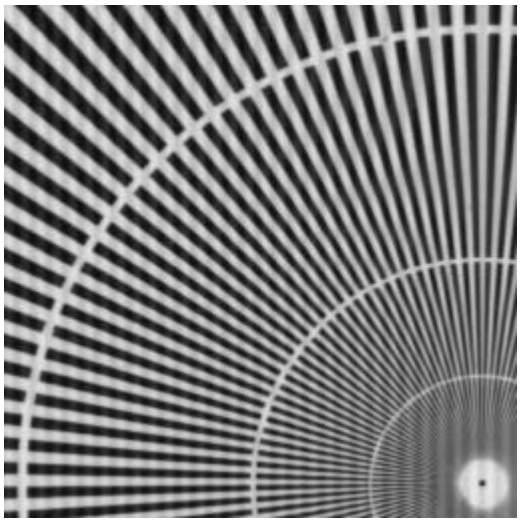

FIGURE 27. Star with Vertical Sine Waves (8 Cyc/Deg)

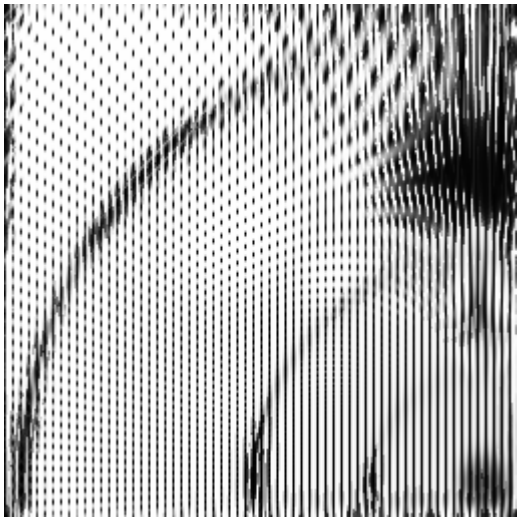

FIGURE 28: Daly: Detection Map of Star with Vertical Sine Waves

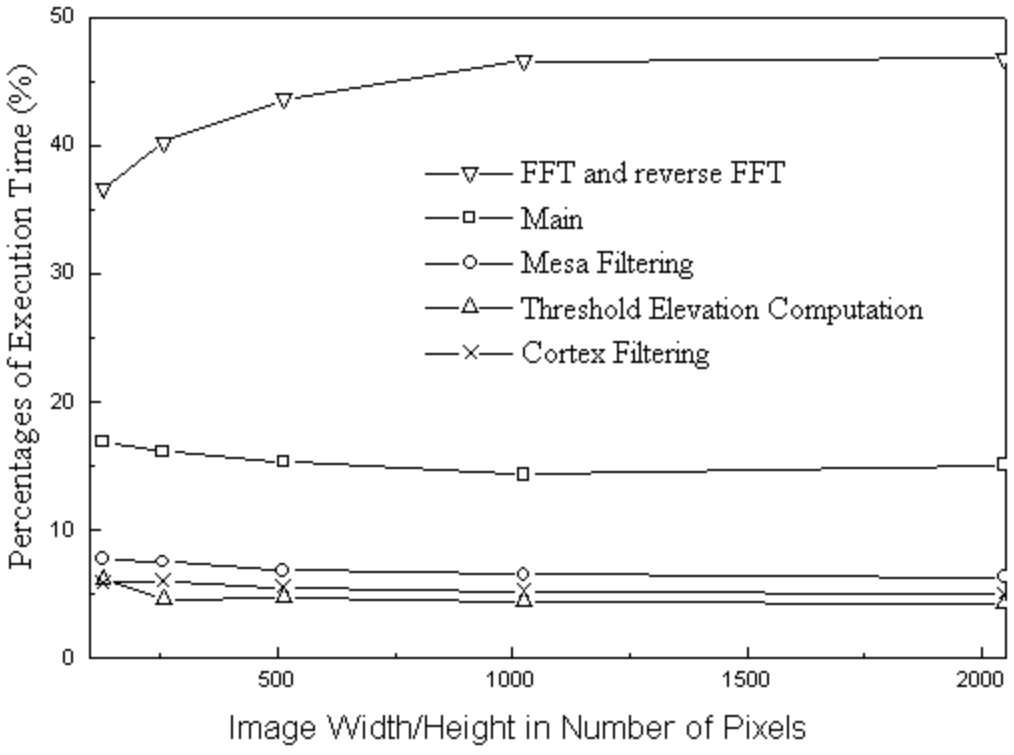

FIGURE 29: Execution Profile of the Daly Model Implementation

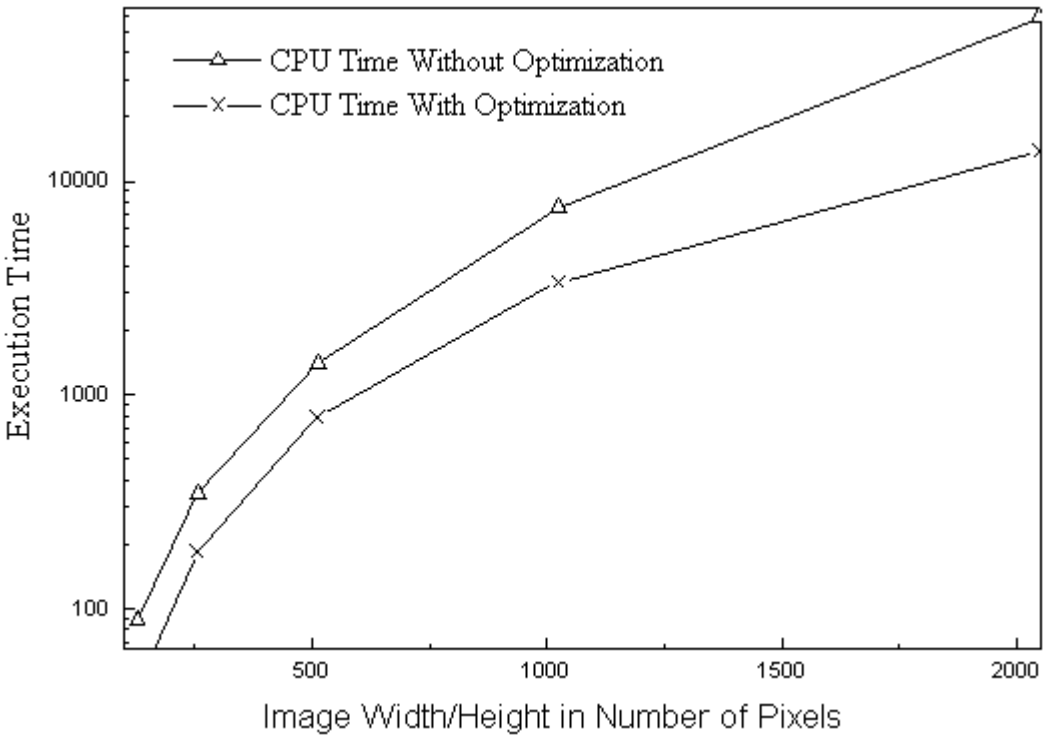

FIGURE 30. Complexity of the Daly Model

CHAPTER IV

SARNOFF MODEL

The Sarnoff model has been designed for physiological plausibility as well as speed and simplicity. While the Daly model is an example of a frequency domain visual model, the Sarnoff model represents another category: spatial domain visual models. Its key elements include resampling, steerable pyramid channeling, a transducer for just noticeable difference (JND) calculations and final refinement (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. The just noticeable difference is the threshold stimulus needed to discriminate between the combination of the stimulus and the background, and the background by itself. The JND is also referred to as a difference limen (DL) by some vision scientists. *DL* can be used as the unit of the JND (Schwartz, 1994).

The Sarnoff model was designed to predict visibility within a wide viewing angle, where peripheral vision, as well as foveal vision, plays an important role. Predicting display visibility in a cockpit environment is one of its applications. The eccentricity of the human optical system should be taken as one of the important parameters for describing non-foveal vision. In an image quality measurement environment, however, the images we are interested in are always in full attention.

In other words, they are always in focus. Thus peripheral vision can reasonably be ignored under this circumstance. Similarly, we can also assume that for image quality assessment, the fixation depth of the optical system is the same as the image depth.

The design of the Sarnoff model employs the same set of psychophysical facts as those that form the basis of the Daly model. A general psychophysical background was given in Chapter III. In the following paragraphs, functionality and specific psychophysical justifications for the Sarnoff model are discussed in detail.

Model Description

In this section, the influence and function of each stage of the Sarnoff model are addressed. These stages are stimuli, optics, resampling, steerable pyramid, phase independent energy response, CSF normalization, transducer, disc-shaped kernel blurring, and JND distances summation. The general structure of the Sarnoff model is shown in Figure 31.

Stimuli

Two digital images, the original image (or the reference image) and the reconstructed image, are the input stimuli for the Sarnoff model. In addition, viewing distance and resolution (or internal sampling size) of the input images must be specified as well. As mentioned above, eccentricity and fixation depth can be ignored in the case of image quality measurement. The resolution of most computer ter-

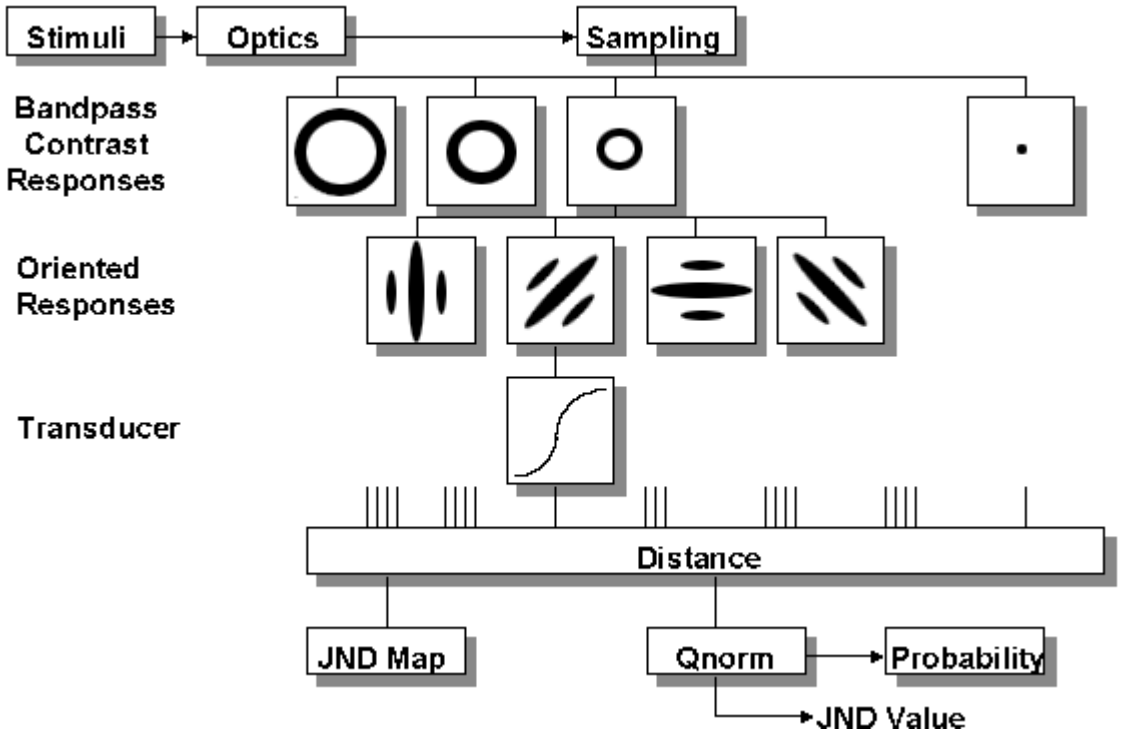

FIGURE 31. Sarnoff Model

minals is about $100dpi$. The viewing distance is within a meter. The maximum luminance of a terminal is usually no more than $50cd/m^2$ according to photometer measurements. These are the default parameter values and they can be overridden.

Optics Stage: Blurring

When considering the optical system of the human eye, the retina as well as the optical lens should be taken into account. The retina is where the virtual image forms through the optical lens of the eye. It's biological characteristics have a direct influence on retina image formation.

As pointed out earlier, a single point spread function (PSF) can be used to

predict the foveal performance of the two-dimensional optics of the eye, under the assumption that this PSF is circularly symmetric. In such a case, the PSF can be easily derived from the line spread function (LSF) (Westheimer, 1986). When the image of interest is in good focus (foveal vision) and when the pupil diameter is near $3mm$, the even-symmetric LSF of Westheimer (Westheimer, 1986) is:

$$L_i = 0.47e^{-3.3i^2} + 0.53e^{-.93|i|}$$

where i refers to position on the retina specified in terms of minutes of visual angle.

However, the PSF of the human eye is actually not circularly symmetric. Not all orientations can be in good focus due to the astigmatism of the eye. Theoretically, when the PSF is circularly asymmetric, two in-focus 1-D systems of different orientations are needed. Any intermediate angles can be predicted from the contributions of the 2 in-focus 1d systems (Wandell, 1995). A typical LSF and PSF are shown in Figure 2.

In the Sarnoff model, a single circularly symmetric function is used as the PSF. Although it might not be theoretically robust, this simplification is acceptable when the impact of the astigmatism of the eye is not very noticeable. The PSF utilized in the Sarnoff model is as follows:

$$Q(\rho) = 0.952e^{-2.59|\rho|^{1.36}} + 0.048e^{-2.43|\rho|^{1.74}}$$

where

ρ: Distance in minutes of arc from a point of light.

$Q(\rho)$: Intensity of light at a distance ρ, relative to the maximum.

For every pixel we can obtain the ρ values of its neighbors when we know the viewing distance and image resolution. Discrete convolution can then be applied using the kernel $Q(\rho)$ shown in Figure 32. The effect of the PSF convolution is blurring of the input images. The output of the optics stage is assumed to be what activates the retinal cones and rods.

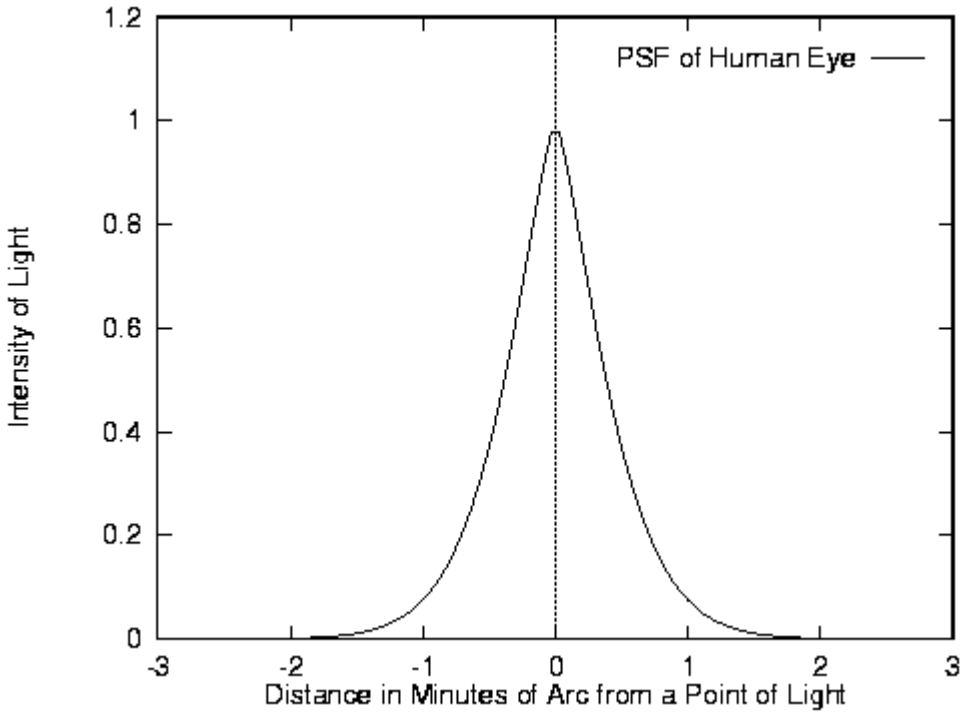

FIGURE 32. The PSF Used in Sarnoff Model

Resampling

This stage simulates the spatial resampling that is done by the cone and rod mosaic in the retina. When the processed luminance signals leave this stage they have a fixed resolution due to the the fixed density of the cone and rod mosaic. Foveal viewing has a resolution of 120 pixel/degree. The neuron pathway takes this value as the ultimate resolution thereafter. This precomputation is necessary because the signal channeling scheme has been developed under the assumption that all input signals always have the same resolution. 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.

In our simulation, resampling is done by Gaussian convolution followed by a point sampling. If the resolution of the original image is smaller than 120 pixel/degree, the convolution and point sampling will "expand" the image by inserting more interpreted values within the input matrix of luminance values. If the resolution of the original image is higher than 120 pixel/degree, the convolution and point sampling will "shrink" the image. However, for both cases the output resolution will remain constant at 120 pixel/degree. If the expanding (or shrinking) ratio is an integer then convolution is easy. If it is not, discrete convolution will have difficulty achieving a continuous result.

If the original images are too big, then the local image quality cannot be assessed in a single glance. This naturally leads to a block-dividing process in which

a big image will be divided into N smaller blocks. For each sub-image (or block), the convolution and point sampling sequence can be applied without exceeding the fixed upper resolution. The rule is to make sure that the output of this stage is constant at 120 pixel/degree. The final detection image can be obtained by simply mosaicing the JND maps of all sub-images from the N runs.

The input images are now ready for the pyramid break down. This step assumes that input signals are all within the same frequency range (namely, the foveal resolution).

Steerable Pyramid

In the Sarnoff model, the Laplacian pyramid (Burt and Adelson, 1983A) is used to store the wavelet representation of the resampled input image, while a quadrature mirrored pair of convolution kernels is used to record information along each of the four orientations. After this stage, the raw luminance signal is converted to units of local contrast.

A Laplacian pyramid is used to record decomposed information for all seven band-pass levels. Limited by the spatial domain convolution approach, the peak frequency of each level has to be a power of 2. The seven bandpass levels have peak frequencies from 32 to 0.5 cycle/degree, where each level is separated from its neighbors by one octave. The contrast pyramid operation is expressed as follows:

$$\hat{c}_k(\vec{x}) = \frac{I(\vec{x}) * (G_k(\vec{x}) - G_{k+1}(\vec{x}))}{I(\vec{x}) * G_{k+2}(\vec{x})}$$

where

$\vec{x}$: Two-dimensional position vector.

$I(\vec{x})$: Input image processed by the PSF and resampling.

$\hat{c}_k(\vec{x})$: Contrast at pyramid level k.

$G_k(\vec{x})$: Gaussian convolution kernel.

With each band-pass level, there is a Gaussian convolution kernel of different shape (i.e. distribution range). The higher the k, the lower the frequency of the band, and the flatter the Gaussian convolution kernel.

$$G_k(\vec{x}) = \frac{1}{(\sqrt{2\pi}\sigma_k)^2} e^{\frac{-(x^2+y^2)}{2\sigma_k^2}}$$

where

$$\sigma_k = 2^{k-1}\sigma_0$$

For reasons of simplicity, a steerable pyramid was actually used in both the Sarnoff model and our implementation. In addition, the steerable pyramid has a better performance. The steerable pyramid is a multi-scale, multi-orientation, image decomposition. It consists of both frequency decomposition and orientation decomposition (Karasaridis and Simoncelli, 1996). This image decomposition scheme is also called image channeling. It is a very common technique in early vision analysis and image-processing applications. Image channeling takes into account frequency

selectivity and orientation selectivity effects that occur in the cortex.

The wavelet transformation in the steerable pyramid is a tight frame (i. e. self-inverting). It has the advantage of being both translation-invariant and rotation-invariant (Karasaridis and Simoncelli, 1996). In addition, the filter bank used to construct the pyramid is polar-separable in the frequency domain.

The decimation of the filters is handled in a recursive manner. With this algorithm, a filtered image is scaled down by a factor of 2 before it enters the next decomposition level. The requirement for steerability of the orientation filters constrains the orientation tuning. A set of filters forms a steerable basis if 1) the filters are rotated copies of each other, and 2) a copy of the filter at any orientation may be computed as a linear combination of the basis filters (Appendix A). The simplest example for steerable basis filters is a set of N+1 N*th*-order directional Gaussian derivatives (Karasaridis and Simoncelli, 1996).

Figure 33 demonstrates the four decomposed orientational contrast representations of the original image. Each pyramid relates to one of the four basis orientations. For each orientation, a number of levels of the odd-phase bandpass subimages are shown with decreasing resolutions.

Phase-Independent Energy Response

There is a frequency level and an orientation associated with each channel of the image. Two sets of coefficients can be obtained by convolving the quadrature mirrored pair filters (odd-phase filter and even-phase filter) with a certain frequency

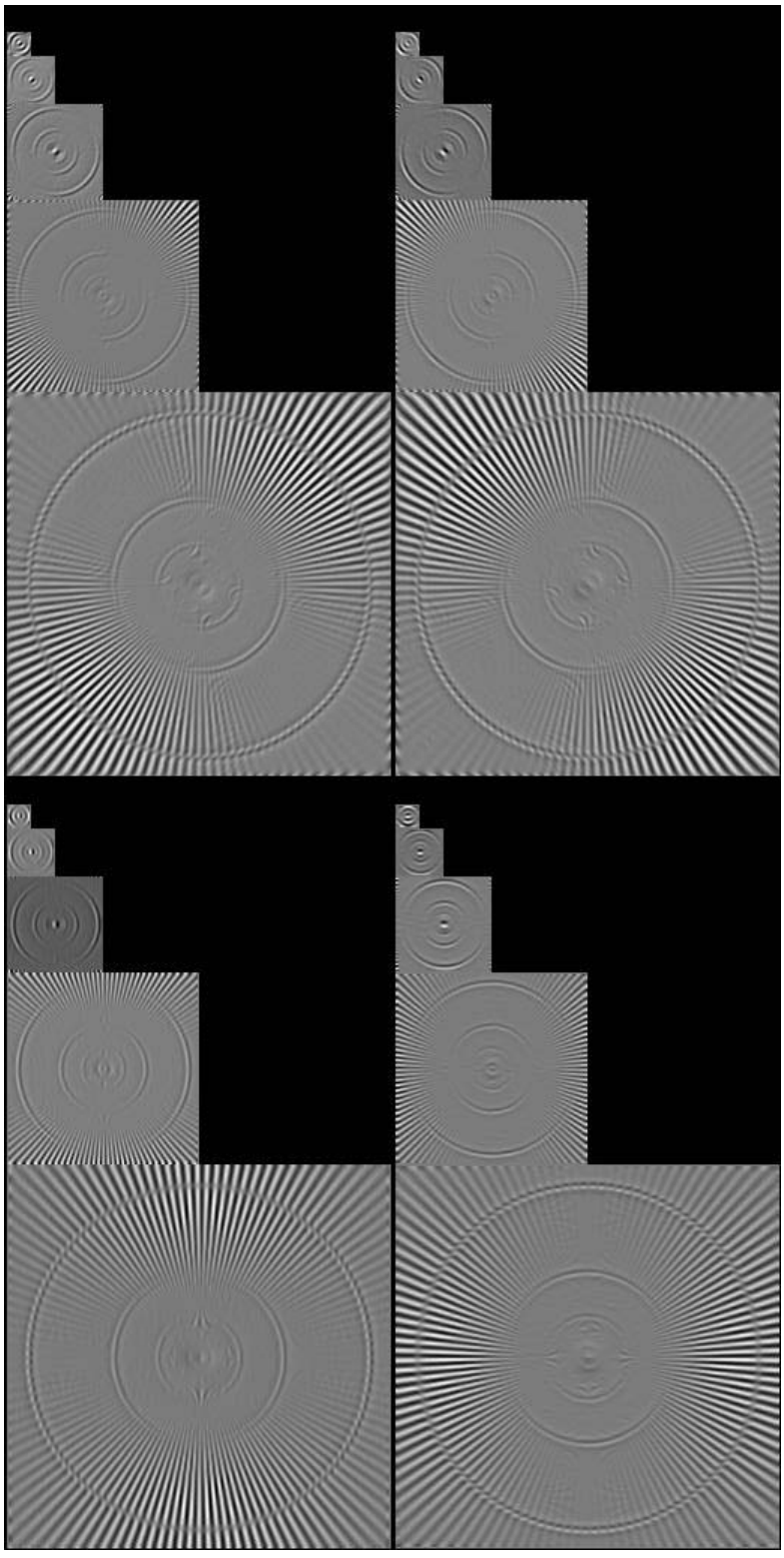

FIGURE 33. Contrast Pyramids: Orientation 0 - Orientation 3

band (Lubin, 1995). These two sets of coefficients contain both frequency and orientation information for a channel. The even-phase filter is the Hilbert transform of the odd-phase filter, which means they have the same frequency response but differ in phase response by 90 degrees.

Energy response is the sum of the squared odd-phase coefficient and the squared even-phase coefficient. It is phase independent, because it sums up the contribution from both odd-phase and even-phase energy response. Odd-phase energy stems primarily from the edges in the image while even-phase energy mainly comes from lines or bars.

Theoretically, two sets of coefficients (even phase coefficients and odd phase coefficients) are needed to calculate the overall energy response. After convolving a bandpass image with the Hilbert pair kernels, outputs from these two filterings are squared and summed.

$$e_{k,\theta}(\overline{X}) = (o_{k,\theta}(\overline{X}))^2 + (h_{k,\theta}(\overline{X}))^2$$

where θ indexes over the 4 orientations, and

$\overline{X}$: Two-dimensional position vector.

θ: Orientation.

k: Pyramid level.

o: Oriented operator.

h: Hilbert transform of the oriented operator.

However, in practice only one set of coefficients is used. The overall energy response is approximated by slightly blurring the square of the coefficients from the odd-phase (or odd-symmetric) filtering (personal correspondence with Simoncelli, E. P., 1997). A narrow Gaussian kernel is used in the convolution for the blurring effect.

CSF Normalization: Precomputation for the Transducer

The normalization stage, as a preprocess to the transducer stage, is the counterpart to the contrast sensitivity function normalization in the Daly model. The energy measure $e_{k,\theta}(\overline{X})$ is normalized by the square of the modulation transfer function (M_t), which is the reciprocal of the CSF. According to the formula originally proposed by Barten (Barten, 1989), M_t is not only a function of the local mean of the current frequency band, but also a function of the environment luminance, image width, and other things. The formula is as follows:

$$\frac{1}{M_t(v)} = ave^{-bv}\sqrt{1 + ce^{bv}},$$

where

$$a = \frac{540(1 + \frac{0.7}{L})^{-0.2}}{1 + \frac{12}{w(1+\frac{v}{3})^2}}$$

$$b = 0.3(1 + \frac{100}{L})^{0.15}$$

$$c = 0.06$$

M_t is close to the grating contrast detection threshold for the current pyramid level with peak frequency v_k and local luminance L_k. The envelope of the M_t's of all pyramid levels (from level 1 to level 7) should be adjusted to fit the contrast sensitivity functions (CSF). The CSF normalization function can be expressed as:

$$E_{k,\theta}(\overline{X}) = \frac{e_{k,\theta}(\overline{X})}{(M_t(v_k, L_k(\overline{X})))^2}$$

where

v_k: Peak frequency for the pyramid level k.

L_k: Local luminance value used in the contrast calculation.

$\overline{X}$: Image matrix.

$e_{k,\theta}(\overline{X})$: Energy response of frequency level k and orientation θ.

$E_{k,\theta}(\overline{X})$: Normalized energy response of frequency level k and orientation θ.

The scale of $E_{k,\theta}(\overline{X})$ is DL^2 (DL: difference of limen). The normalization contributes to contrast detection, while the next stage contributes to contrast discrimination.

Transducer

A transducer is used to refine the JND map by taking the spatial masking dipper effect into consideration (Section II.8). The dipper shape reflects the behavior

of the contrast discrimination function. This stage features the transformation of a sigmoid non-linearity with two variables n and w.

$$T(A) = \frac{2A^n}{A^{n-w} + 1}$$

where

A: $\sqrt{E_{k,\theta}(\overline{X})}$ The square root of the normalized energy response from channel k, θ.

n: A real number around 2.

w: A real number around 0.2.

Given the equation above we can do the following analysis:

Let $A = \sqrt{e}$.

Case 1

When $A \ll 1$,

$$T(A) \propto \frac{2A^n}{1} \propto A^n$$

Case 2

When $A = 1$,

$$T(A) = \frac{2}{1+1} = 1$$

Case 3

When $A \gg 1$,

$$T(A) \propto \frac{2A^n}{A^{n-w}} \propto A^w$$

From the cases above, we see that when the square root of the normalized contrast energy A is small, maximum transducer output from pyramid level k accelerates as A^n; when A is large, the function is compressive as A^w. For intermediate values of A at the contrast detection threshold for frequency v_k, the transducer output is about 1.

Disc-Shaped Kernel Blurring

According to Lubin (Lubin, 1995), the following characteristic of the oriented filters should also be modeled. For a single filter at a single spatial position (i.e. for each of the 28 bands), given a sine wave to which the filter is optimally tuned, the output as a function of the number of cycles in the patch will asymptote at a little more than one cycle. Foveal human sensitivity increases when the number of cycles reaches five. A disc-shaped filter kernel of a diameter of five can be applied to each band to account for this effect. The filter used in our implementation is not a perfectly disc-shaped window, but instead a 5 by 5 matrix square with each entry taking the value 0.04. Through our implementation, we have found out that this pooling stage makes a surprisingly big difference in the output of the JND map.

Distance of M-Dimensional Vector Summation

After getting the JND difference map for each channel, the last stage is devoted to putting together the contributions from all channels. This leads to the concept of a space of multiple dimensions. We have 28 channels to perform the summation: number of pyramid levels (7) times the number of different orientations (4) gives 28. For each spatial position, the final JND distance can be regarded as the distance between two 28-dimension vectors. Each of the 28 JND maps obtained from previous stages describes the difference along one dimension. The JND distance between two points in such a 28-dimension space can be calculated with the following formula:

$$D(\overline{X}_1, \overline{X}_2) = (\sum_{i=1}^{28} (P_i(\overline{X}_1) - P_i(\overline{X}_2))^Q)^{\frac{1}{Q}}$$

where

$\overline{X}_1, \overline{X}_2$: Two input images.

Q: Parameter currently set at 2.4. If $Q = 2$, this expression corresponds to Euclidean distance.

Calibration

Calibration is used to avoid selecting the model parameters on a case by case basis. This procedure is based on the assumption that the shape the transducer function is independent of the pyramid level or orientation channel, except for a scale factor that is determined by the CSF.

The procedure is divided into two steps. The first step makes sure the CSF fits the psychophysical data. The second step adjusts the variables in the transducer function so that its outputs match those from human vision.

As part of the first step, informal tests were performed with human subjects to get the CSF data for our test environment. From running the Sarnoff model, we feel that there is not a large difference between the results obtained with the CSF data mentioned above and the CSF data first measured by Schade (Schade, 1948). However, we have found that the calibration of the second stage of the transducer has a large impact on the accuracy of the final detection results. The output of the transducer stage $T(A)$ is the the final JND for each pixel in each channel. The JND is expressed as:

$$JND = \frac{A}{T_e} = T(A)$$

where

A: Normalized contrast. A is approximately the square root of the energy response (see Section IV.1 and IV.1).

T_e: The threshold elevation normalized by CSF.

$T(A)$: The transducer output.

The formula for the transducer output $T(A)$ is given in Section IV.1. A new expression of T_e that only contains variables n and w is obtained.

$$T(A) = \frac{2A^n}{A^{n-w}+1} = \frac{A}{T_e}$$

$$\Longleftrightarrow T_e = \frac{A}{T(A)} = \frac{A^{n-w}+1}{2A^{n-1}}$$

Spatial masking is well described in terms of threshold elevation, which serves as a guide in the calibration process. Two variables n and w are open to calibration. The expressions for $T(A)$ and the newly derived T_e are plotted out with different values of n and w shown in Figures 34 to 37. The scales of those graphs are exactly the same.

Generally, variable w has more influence on the curve segment when the normalized mask contrast A is bigger than 1.0, while variable w has more influence on the curve segment when A is smaller than 1. By looking at Figures 34 to 37, one can see that different w's and n's lead to different slopes in the two segments respectively. The proper combination of the values for n and w can not only produce a partial dipper effect but also can reflect the masking effect when A is higher than 1.0.

From psychophysical tests mentioned in Chapter III we know that for the high mask contrast the slope of T_e is within a range from 0.7 to 1 (see Section III.1). Thus a range of $[0.1, 0.3]$ seems reasonable for w. From Figures 34 to 37, it can be seen that the bigger n, the bigger the dipper effect becomes. At the same time, however,

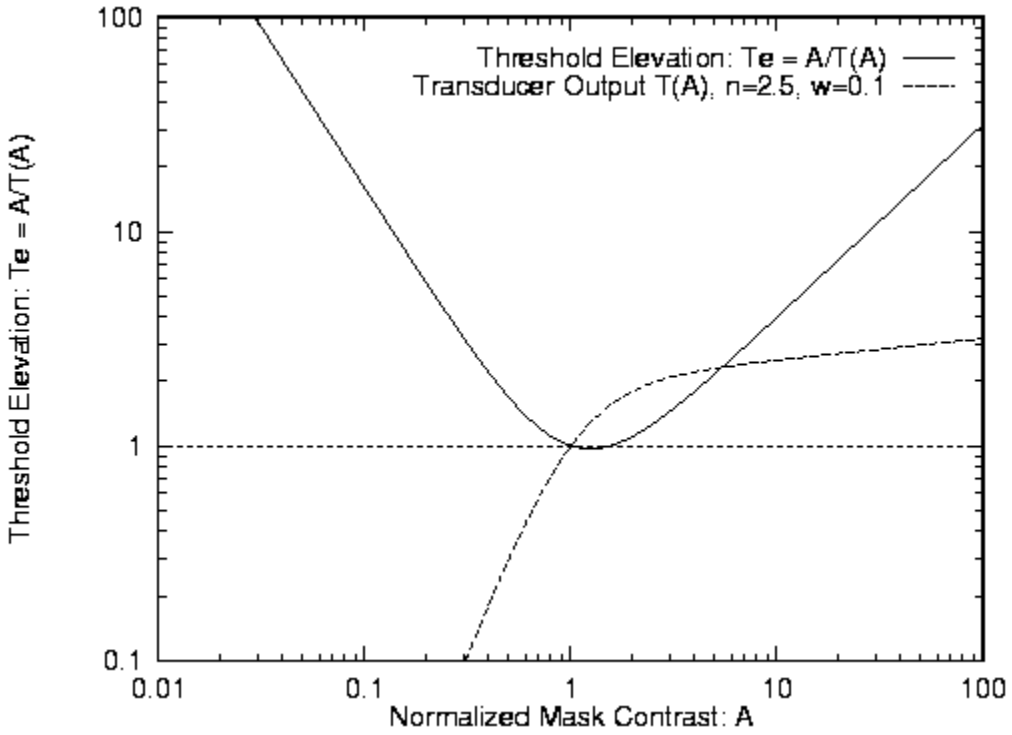

FIGURE 34. Variable n, w Calibration: $n = 2.5$, $w = 0.1$

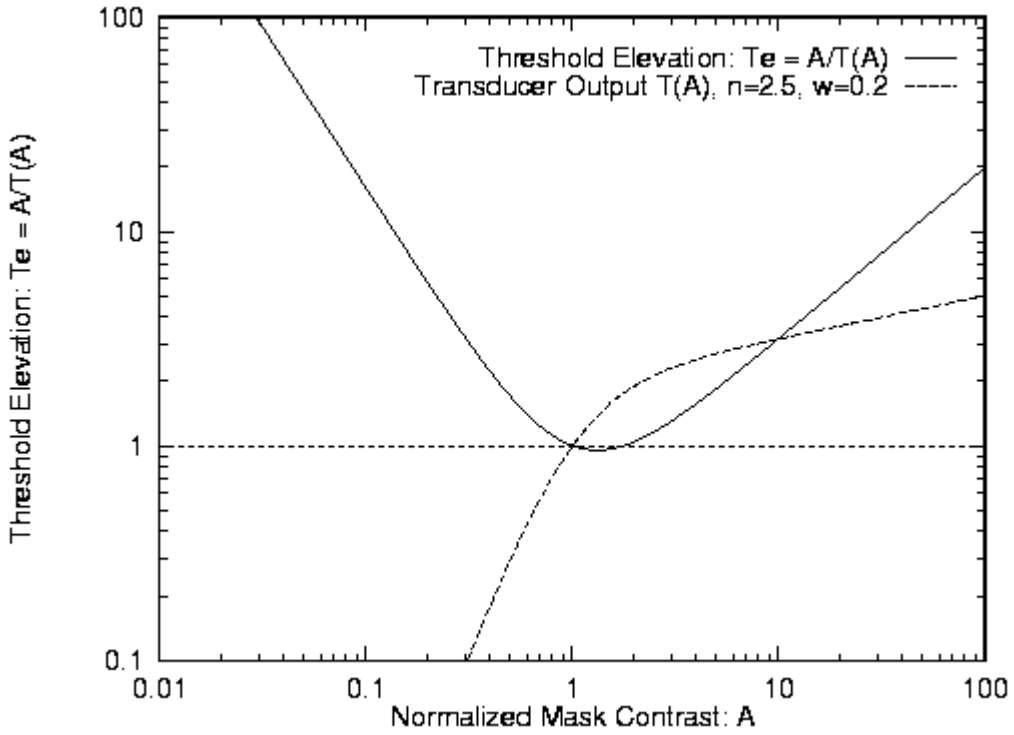

FIGURE 35. Variable n, w Calibration: $n = 2.5$, $w = 0.2$

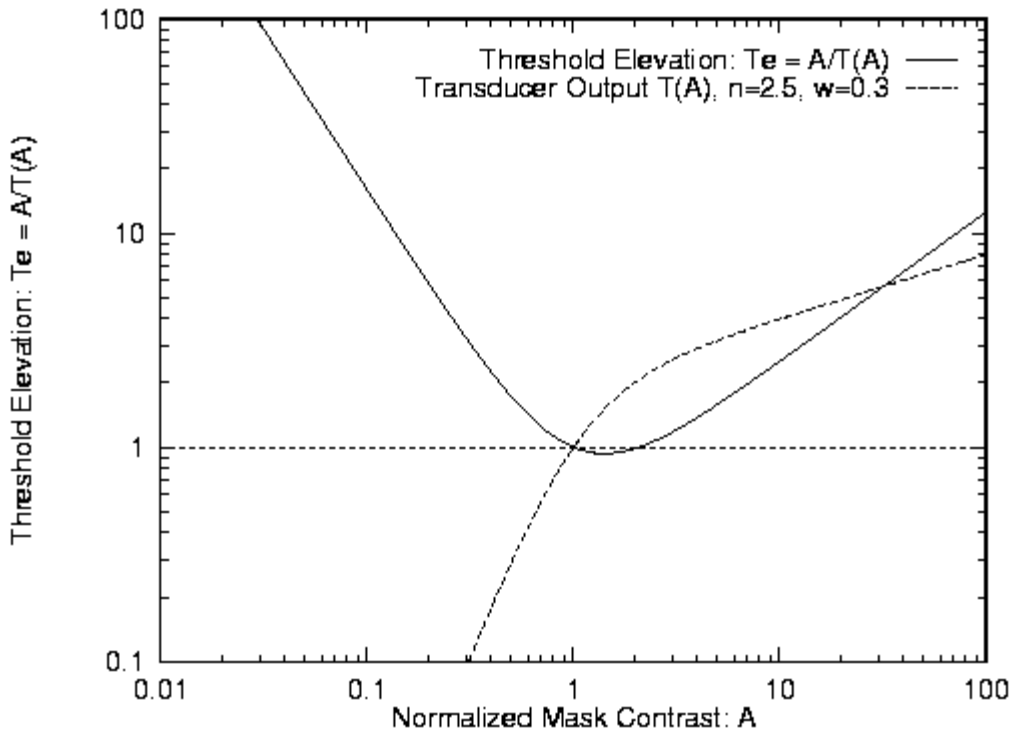

FIGURE 36. Variable n, w Calibration: $n = 2.5$, $w = 0.3$

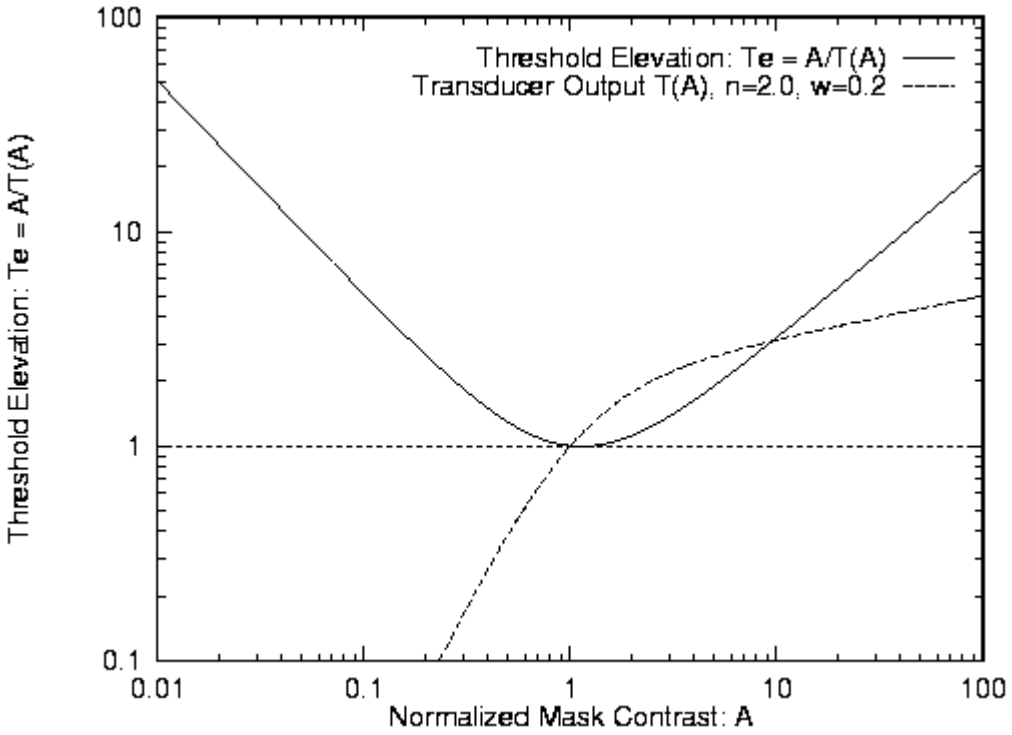

FIGURE 37. Variable n, w Calibration: $n = 2.0$, $w = 0.2$

bigger distortion is introduced when n grows larger. Values between 2.0 and 2.5 are reasonable for variable n.

In the second stage of calibration, informal experiments with human subjects have been conducted. The data collected is used to verify the choices of the n and w values. Sine wave patterns with and without noise are generated as test images. All peak frequencies of the seven band-pass levels of the steerable pyramid are used as frequencies for these sine waves patterns. Different combinations of n and w are used. The results of these psychometric experiments have approximately matched the detection results obtained from running the Sarnoff model. We have realized that when n is between 2 and 2.5 and w is about 0.2, the detection results are generally the best. This matches the theoretical prediction of the value range for n and w.

Performance

Detection Results

For comparison, the same input images that were used to test the Daly model were also used to evaluate the Sarnoff model (Section III.2). In addition, more test images were used for further demonstration and discussion (Figures 38, 44 , 45, and 46). 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 about $50cd/m^2$. The viewing distance was about 0.8 meter. The reason for choosing

0.8 meter and not 0.5 meter as in the Daly model testing 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. As mentioned in Section IV.1, convolution interpolation in resampling is easier with an integer expansion rate.

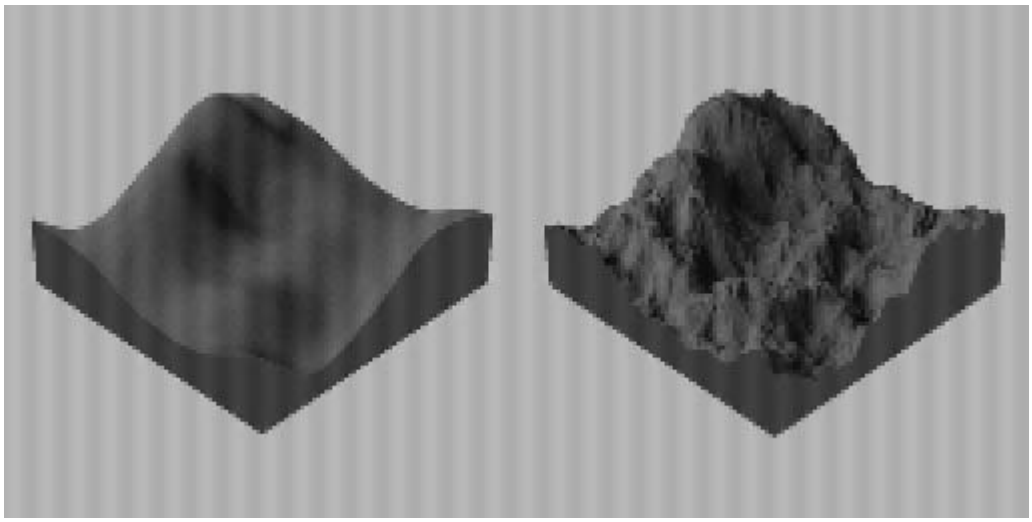

FIGURE 38. Mountains with Sine Waves (8 Cyc/Deg)

The reconstructed image in Figure 38 has a sine wave mask of 8 cycles/degree, which is 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 38. This distorted image is fed into the Sarnoff model along with the original mountain image in Figure 16. As shown in Figure 39, the masking effect is more accurately predicted than in the Daly model. The noise pattern in the background is also properly detected. The maximum JND of this detection map is 4.33. The mean JND is 2.36.

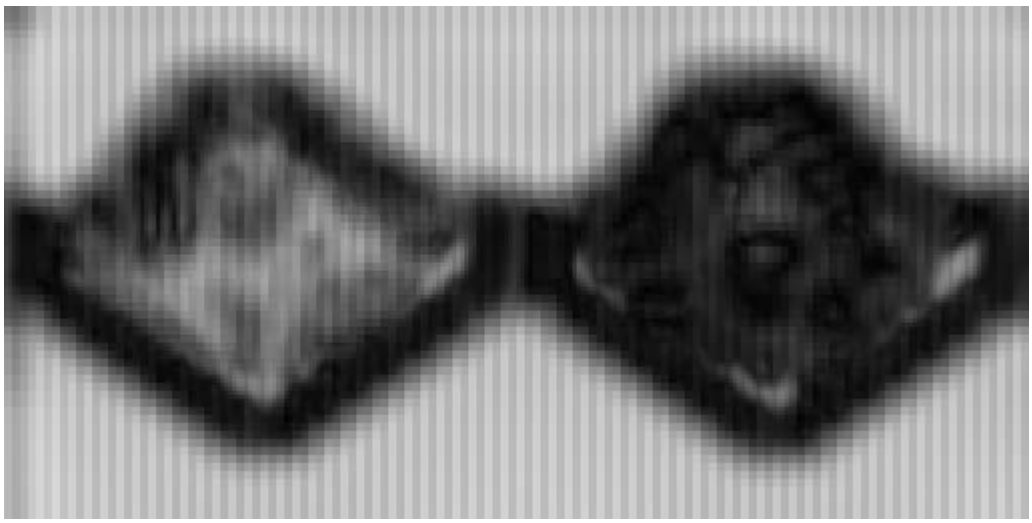

FIGURE 39: Sarnoff: Detection Map of Mountains with Sine Waves (8 Cyc/Deg)

The quantized mountain image at 4 bits/pixel (Figure 19) and the original mountain image at 8 bits/pixel (Figure 16) are 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 model. The detection map is shown in Figure 40 with a maximum JND of 2.86 and an average JND of 0.31.

The detection map of the blurred chapel (Figure 22) is shown in Figure 41. As in the Daly model, the most distorted part of the image, i.e. 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 human eyes. The maximum and average JND of this detection map are 3.92 and 1.04 respectively.

The detection map in Figure 42 for the input image pair of the original chapel

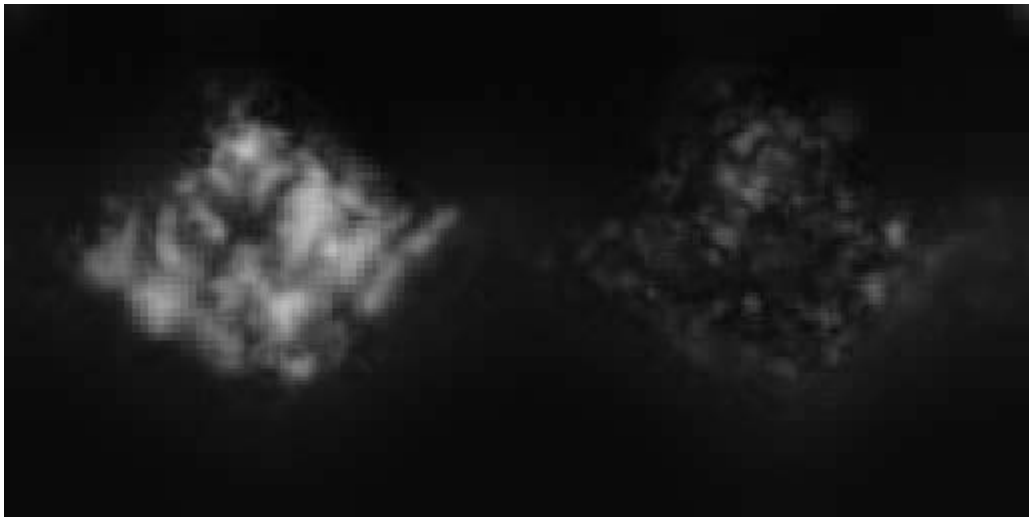

FIGURE 40. Sarnoff: Detection Map of Quantized Mountains

(Figure 21) and the chapel with sine waves (Figure 24) shows a better prediction of the masking effect than in the Daly model. 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.

The star pattern is again used to test the orientational and frequency selectivities of spatial masking. The detection map for the image pair in Figure 26 and 27 is shown in Figure 43. According to the model's prediction, less distortion is seen in the rings with relatively high but uniform luminance, and also in the area where the orientations and frequencies of both signal and noise match. However, the model is not perfect. The black region on the left, two fifth's of the distance from the bottom of the prediction map, shows that the distortion in this region is under-predicted. The maximum JND of this detection map is 1.99 and the average JND is 0.63.

Finally, the model was tested on images of quantized peppers. Different quan-

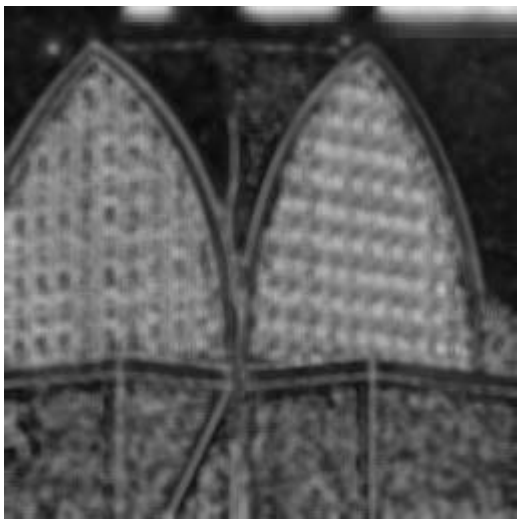

FIGURE 41. Sarnoff: Detection Map of Blurred Chapel

tization levels were used to test the sensitivity of the model and to determine how accurately it could predict a difference. The original pepper image is shown in Figure 44. The quantized pepper images have a grey scale depth of 3 bits/pixel (Figure 45) and 4 bits/pixel (Figure 46). Stronger contour aliases are observed when quantization is strong (Figure 45). The detection maps for these two quantized images are shown in Figure 47 and 48. The maximum and mean JND of the detection map in Figure 47 are 4.39 and 1.12 respectively, and for Figure 48 they are 3.96 and 0.70 respectively. Generally, for these images the predictions from the model correctly reflect what is seen by the human visual system.

Speed and Memory

The execution time of the Sarnoff model has been measured on an HP 9000/700 computer without compiler optimization. For example, when the resampling stage

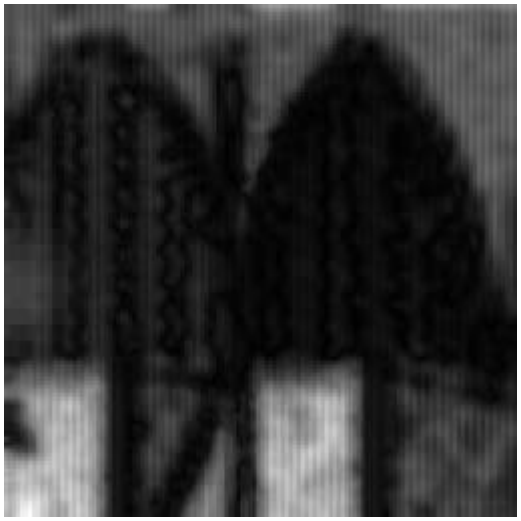

FIGURE 42. Sarnoff: Detection Map of Chapel with Sine Waves

of the model does not resize the input images, the running time of the Sarnoff model is about 100 seconds for input images of size 256x512. For input images of size 512x512, the running time is about 200 seconds.

The Sarnoff model functions purely in the spatial domain with simple operations. The modeling of each perception 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 point-by-point 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 our performance measurements mentioned above.

Recall that in the Daly model the upper bound of the complexity is $O(n^2)$

FIGURE 43: Sarnoff: Detection Map of the Star with Vertical Sine Waves

due to the time-consuming Fourier transformations. If FFT and FFT^{-1} are used, the overall complexity can be reduced to $O(n \log n)$. Therefore when the size of the input images increased, the Sarnoff model's superiority in terms of speed becomes more and more obvious.

However, the Sarnoff model gains its speed at the cost of memory. It has to record data of all frequency levels and of all orientations. The generating and maintaining of the wavelet pyramids, local mean pyramids and contrast pyramids takes a lot of memory. Whereas in the Daly model, only one representation of the image in the frequency domain is needed.

FIGURE 44. Original Pepper Image

FIGURE 45. Quantized Pepper (3 Bits/Pixel)

FIGURE 46. Quantized Pepper (4 Bits/Pixel)

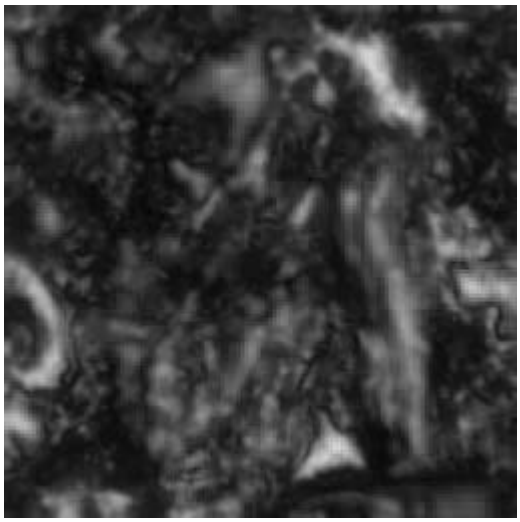

FIGURE 47: Sarnoff: Detection Map of quantized Pepper (3 Bits/Pixel)

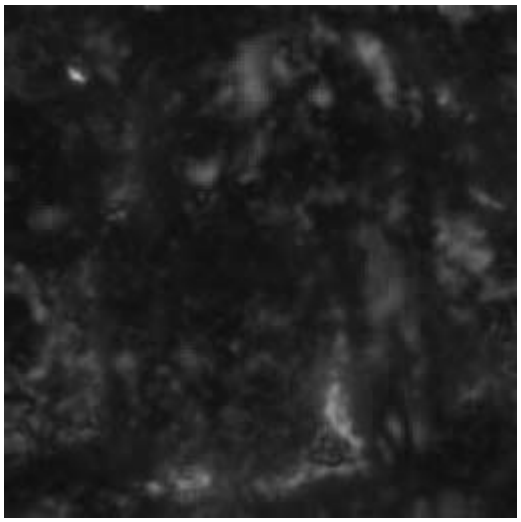

FIGURE 48: Sarnoff: Detection Map of the quantized Pepper (4 Bits/Pixel)

CHAPTER V

COMPARISON OF THE DALY AND THE SARNOFF MODEL

The Daly model and the Sarnoff model have their respective advantages and drawbacks. The differences between the two models come from 1) the different approaches they represent (i. e. the spatial domain approach and the frequency domain approach respectively); 2) emphasis on different aspects of human visual perception; and 3) different implementation techniques.

Advantages of the Daly Model over the Sarnoff Model

The Daly model, like several other psychophysical analyses, 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 more easily be 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 frequencies. It is more natural to make use of this function in the frequency domain. The advantage of frequency domain models, such as the

Daly model, is to have a precise and continuous CSF normalization. In the Sarnoff model, CSF normalization is approximated by performing it in only seven discrete frequency bands (levels). For each band, its single peak frequency is used to get the CSF values.

The Daly model has a finer simulation of the orientation selectivities. Six orientation filters are used for each frequency band. Although this might be slightly over-complete, six *fan* filters can produce more accurate results. A choice of six *fan* filters is better especially when the input images are small and when resources are not a major concern. In the Sarnoff model, only four orientation channels are used. This is acceptable but it introduces some degradation in the final results.

When two images are compared and assessed, the mask cannot be derived solely from any one of them. Mutual masking is adopted in the Daly model to produce more plausible threshold elevation maps for all bands.

For the Daly model, there is no power-of-2 limitation to the size of the image. However the FFT (Fast Fourier Transform) performs best when the base of the image size is a prime number (e. g. 2, 3, 5, ...). On the other hand, in the Sarnoff model the size of the input image (actually the image size after resampling) needs to be a power of 2.

Advantages of the Sarnoff Model over the Daly Model

While the Daly model has advantages in frequency domain operations, a more straight-forward simulation of each stage of visual perception is incorporated in the

Sarnoff model. This includes optics, sampling, channeling, and cortex spatial masking. Since there is no physiological evidence that the HVS performs any Fourier domain processing, the spatial domain model is a more suitable mathematical description of the underlying neural process. The Sarnoff model tries to reproduce the same functions that happen along the brain path.

Since there is no single frequency representation in the Sarnoff model, it is possible to represent the CSF normalization as a function of location. The MTF term used in the Sarnoff model is a function of the local mean of each pyramid level (Section IV.1). Theoretically, an MTF with phase information (i.e. as a function of pixel positions) should simulate local luminance adaptation better. Practically speaking this refined MTF does not show a remarkable improvement over the MTF obtained with a single adaptation luminance according to our tests.

The Sarnoff model uses a better approach to determine the local luminance mean which is needed to compute the contrast. In this model, the local luminance mean of each pixel is the average of the luminance of neighboring pixels. In the Daly model, the luminance of the pixel itself is used as the local luminance mean under the assumption of an arbitrarily close viewing distance. This assumption is physiologically untrue, whereas the local mean concept in the Sarnoff model is a better approximation.

As discussed in Section IV.2, the performance of the Sarnoff model is faster especially when the size of the input images is big. The Sarnoff model only operates in the spatial domain. It avoids the expensive FFT and FFT^{-1} transformations

which take up to 40% of the run time in the Daly model. The execution time of the Sarnoff model is $O(n)$, where n is the number of pixels. For comparison, the complexity of the Daly model is $O(n \log n)$. The choice of 4 orientation filters instead of 6 speeds up the process as well.

In the Sarnoff model the CSF normalization is done after the contrast pyramid is obtained. Therefore, distortion introduced by the non-linear MTF cannot interfere with the channeling. On the other hand in the Daly model 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 (Legge, 1980), an important feature of the masking model is the ordering of its elements. The linear filter is better to be placed before the nonlinearity transducer.

Common Features

While the root mean squared error (RMSE) measure tends to treat the entire human visual system as a "black box," both the Daly model and the Sarnoff model 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.

Both the Daly model and the Sarnoff model use JND's as the metric to quantify the quality of 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 summation mechanism. 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.

Each stage of both the Daly model and the Sarnoff model, which are typical examples of mechanistic models, can be extensively 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, corresponding refinements of the mechanistic models can be easily done without major changes to their basic architecture.

Other Differences

In the Daly model the optics point spread function (PSF) is not modeled as an element of the human visual system to avoid a shift-variant nonlinearity and the accompanying problem of noninvertibility. As a tradeoff, the blurring effect from convolving the PSF with the input images could have lead to a better approximation of the adapted luminance in the retina. It is a coarse approximation, although the process is invertible which is what signal processing usually prefers. In the Sarnoff

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

Although eccentricity is used as an input parameter in the Daly model, 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 model 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 averaging effect in the pooling stage of the HVS is simulated in the Sarnoff model when the output of the transducer is convolved with a disc-shaped kernel (Section IV.1). 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 model does not consider this special property of the HVS.

The two visual models have different ways of visualizing the detection results. In the Sarnoff model, the final JND map is shown directly as the final output. In the Daly model, a psychometric function is used to convert the JND values into detection probabilities. As a result, the final output visualization is a map of the

detection probabilities as a function of location.

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

Since the Sarnoff model operates solely in the spatial domain, its ability to select signals of an arbitrary frequency is limited. As shown in Section V.4, the Sarnoff model 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 16 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 is not as good.

To illustrate this, sine wave noise of different frequencies has been introduced into the original mountain image (Figure 16). Three distorted input images are shown in Figures 38, 17, and 49. The sine wave frequencies in these three input images are respectively 8, 9, and 16 cycles/degree.

The detection results are shown in Figures 39, 51, and 50. In the same order, the maximum JND's are 4.33, 4.18, and 3.51. The mean JND's are 2.36, 2.27, and 1.27.

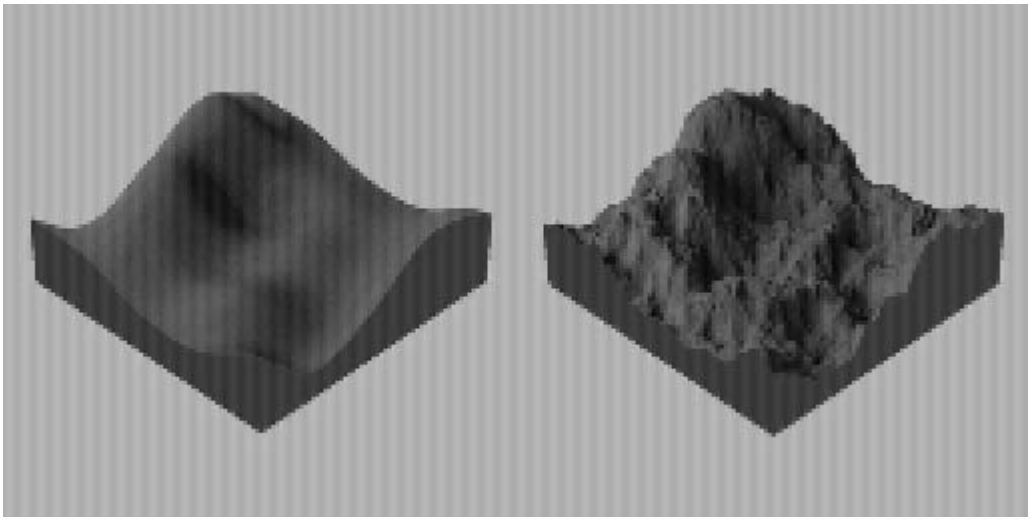

FIGURE 49. Mountains with Sine Waves (16 Cyc/Deg)

Common Problems

Although the mechanism used to handle the local luminance mean in the Sarnoff model is more appropriate than the one in the Daly model (Section V.2), it is still not robust. For example, 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 incorrect. In our current implementation, a non-zero local luminance mean is found by increasing the number of neighboring pixels for averaging. An adjustment can automatically be made by substituting the zero luminance mean with the new non-zero value.

Both models face difficulties in finding a correct general CSF representation. In the Daly model, the peak sensitivity P is hand-picked for different environments.

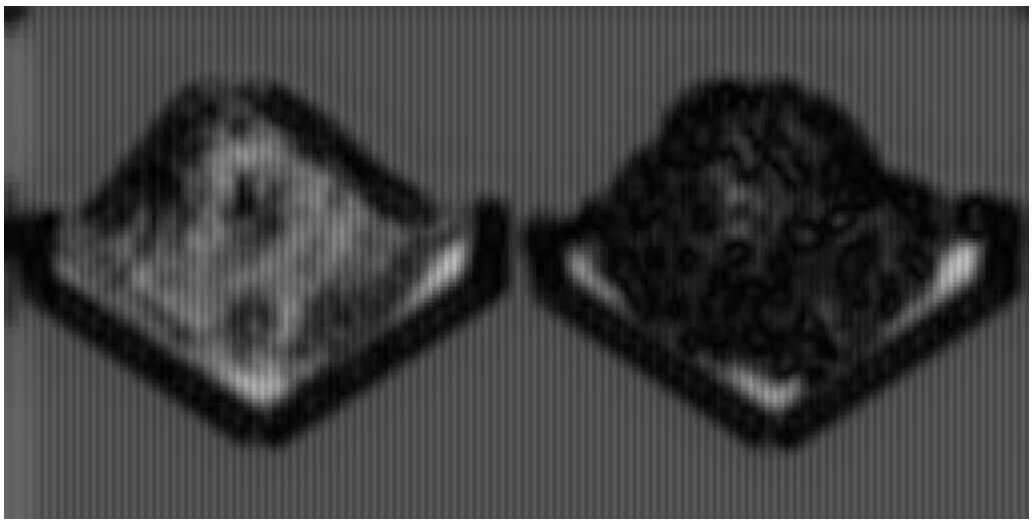

FIGURE 50: Sarnoff: Detection Map of Mountains with Sine Waves (16 Cyc/Deg)

This parameter adjustment has to be done before each run of the Daly model. In the Sarnoff model, calibration is done for CSF normalization. However, in different luminance environments, CSF's are different and so are the CSF normalizations. Therefore, the question boils down to the following: At which environment/adaptation luminance level should the CSF test and calibration should be done to get optimal results for all situations?

The number of orientation filters used in these two models is either more than sufficient or just barely sufficient (Section V.1 and V.2). A hybrid of the two could be adopted: 4 different orientation filters could be used for lower frequency bands where orientation selectivities are relatively weak, and 6 different orientation filters or more could be used for higher frequency bands where orientation selectivities are stronger.

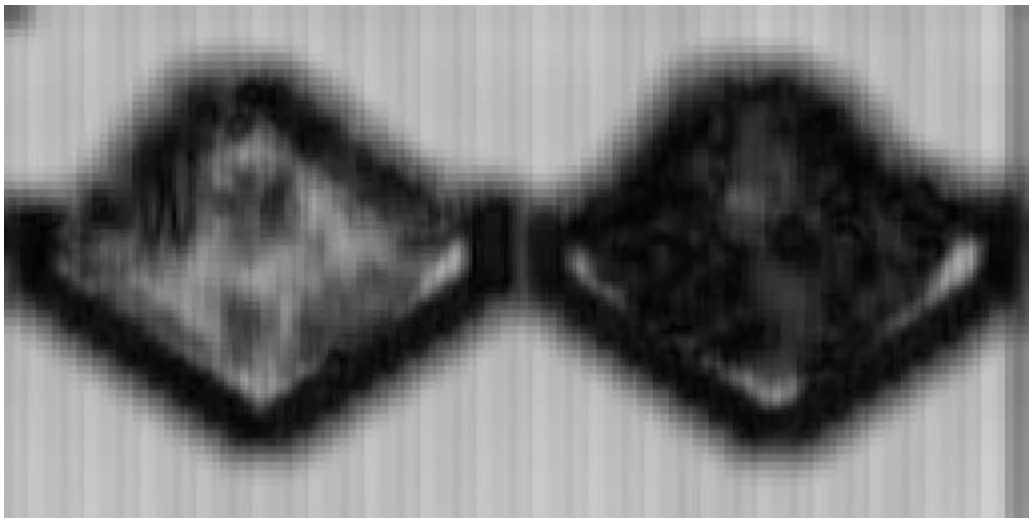

FIGURE 51: Sarnoff: Detection Map of Mountains with Sine Waves (9 Cyc/Deg)

CHAPTER VI

CONCLUSIONS AND FUTURE WORK

Conclusions

The Daly and Sarnoff models are working mechanistic visual models. These models accomplish their results by attempting to simulate the functionalities of each element along the visual perception pathway from the optics of the eye to the brain. Generally the predictions from these models match those of the human visual system. The most important contribution of these models is that spatial masking effects are detected correctly. The Sarnoff model is particularly successful in this regard.

The key element in the mechanistic visual models is channeling. This is where the original input images are decomposed into different bands. Each band has a different frequency range and a different orientation. In this way, spatial tuning and orientational tuning are performed as a normalization stage for spatial masking, which simplifies the spatial masking mechanism to a single lookup function. In the Daly model, this function is described as the threshold elevation equation. In the Sarnoff model it is given as the transducer formula. Another key element in these models is the CSF, which describes contrast detection as a function of spatial

frequencies. The CSF normalization is always coupled with the spatial masking mechanisms. Only after they have been normalized by the CSF can the contributions from all frequency bands be regarded as equal. In the Daly and Sarnoff models, the order of these two key elements (i.e. channeling and CSF normalization) are arranged differently.

The most distinguishable difference between these two models lies in the way that they choose to represent the information contained in the visual field. The Daly model uses the frequency domain approach, while the Sarnoff model is an example of the spatial domain approach. Understanding the mechanisms of both models and their strengths and weaknesses helps us understand vision and visual models in general. Better visual models can be designed by combining the strengths of the two.

Potential Problems and Future Work

Some potential problems were found in these two models through our implementations and tests. The common problem is the choice of local means and the choice of the number of orientations in orientational tuning. Furthermore, both models face difficulties in finding a correct general CSF representation. These issues has been discussed in Section V.5.

Another problem with the Sarnoff model arises in transducer calibration. The calibration process is time-consuming and environment/device dependent. Two variables (n and w) are tuned in the calibration process. The best values for these two

variables are on average the best fit but may not be the best fit for all conditions. It is difficult to precisely match the prediction results with the psychophysical results . To find an efficient method for calibration remains a topic for future work.

Both the Daly and the Sarnoff models restrict themselves to analyzing the luminance difference between two static images. In reality, people are also interested in color. Incorporating color factors into current luminance models would expand the application of visual modeling significantly. It would also be very useful to add the time dimension into the visual models. Video signals, not only static frames, can be assessed. This could find important application in video coding testing and related areas.

The work reported in this thesis has lead to a better understanding of the underlying mechanisms of visual perception and of the design of visual models. When it comes to perception or the processes in the human brain, there is no end to possible explorations.

APPENDIX

QUADRATURE MIRROR FILTERS

Gaussian derivatives are important functions for image analysis. A steerable quadrature pair of them would be useful for many vision tasks (Freeman and Adelson, 1991). In this chapter we first give the steerable filter formula for the second Gaussian derivative of any directions. As an example, four steerable quadrature pairs are derived and illustrated. These four steerable quadrature pairs are used in constructing the steerable pyramid in the Sarnoff model.

Steerable Filter Formula

According to Freeman and Adelson (Freeman and Adelson, 1991), three basis functions G_{2a}, G_{2b}, G_{2c} are sufficient for interpolating a second Gaussian derivative G_2 in any direction θ:

$$G_2(\theta) = K_a(\theta) \cdot G_{2a} + K_b(\theta) \cdot G_{2b} + K_c(\theta) \cdot G_{2c}$$

where

$$G_{2a} = 0.9213(2x^2 - 1) \cdot \exp[-(x^2 + y^2)]$$

$$G_{2b} = 1.843xy \cdot \exp[-(x^2 + y^2)]$$

$$G_{2c} = 0.9213(2y^2 - 1) \cdot \exp[-(x^2 + y^2)]$$

$$K_a(\theta) = cos^2(\theta)$$

$$K_b(\theta) = -2cos(\theta)sin(\theta)$$

$$K_c(\theta) = sin^2(\theta)$$

The Hilbert transform of G_2 can be approximated by using a third-order odd parity polynomial, which is steered by four basis functions: H_{2a}, H_{2b}, H_{2c}, and H_{2d}. The Hilbert transform approximation H_2 is given as follow:

$$H_2(\theta) = K_a(\theta) \cdot H_{2a} + K_b(\theta) \cdot H_{2b} + K_c(\theta) \cdot H_{2c} + K_d(\theta) \cdot H_{2d}$$

where

$$H_{2a} = 0.9780(-2.254x + x^3) \cdot \exp[-(x^2 + y^2)]$$

$$H_{2b} = 0.9780(-0.7515 + x^2) \cdot \exp[-(x^2 + y^2)]$$

$$H_{2c} = 0.9780(-0.7515 + y^2) \cdot \exp[-(x^2 + y^2)]$$

$$H_{2d} = 0.9780(-2.254y + y^3) \cdot \exp[-(x^2 + y^2)]$$

$$K_a(\theta) = cos^3(\theta)$$

$$K_b(\theta) = -3cos^2(\theta) \cdot sin(\theta)$$

$$K_c(\theta) = 3cos(\theta) \cdot sin^2(\theta)$$

$$K_d(\theta) = -sin^3(\theta)$$

Four Steerable Quadrature Filter Pairs

With the formula given above, G_2 and H_2 can be shifted arbitrarily in both phase and orientation. As an example, four steerable quadrature pairs are calculated and plotted in Figure 52, 53, 54, and 55.

First Pair - Orientation of 0°:

$$G_2(0^\circ) = G_{2a}$$

$$H_2(0^\circ) = H_{2a}$$

Second Pair - Orientation of 45°:

$$G_2(45^\circ) = 0.5G_{2a} - G_{2b} + 0.5G_{2c}$$

$$H_2(45^\circ) = \frac{\sqrt{2}}{4}H_{2a} - \frac{3\sqrt{2}}{4}H_{2b} + \frac{3\sqrt{2}}{4}H_{2c} - \frac{\sqrt{2}}{4}H_{2d}$$

Third Pair - Orientation of 90°:

$$G_2(90^\circ) = G_{2c}$$

$$H_2(90^\circ) = -H_{2d}$$

Forth Pair - Orientation of 135°:

$$G_2(135^\circ) = 0.5G_{2a} + G_{2b} + 0.5G_{2c}$$

$$H_2(135^\circ) = \frac{\sqrt{2}}{4}H_{2a} - \frac{3\sqrt{2}}{4}H_{2b} - \frac{3\sqrt{2}}{4}H_{2c} - \frac{\sqrt{2}}{4}H_{2d}$$

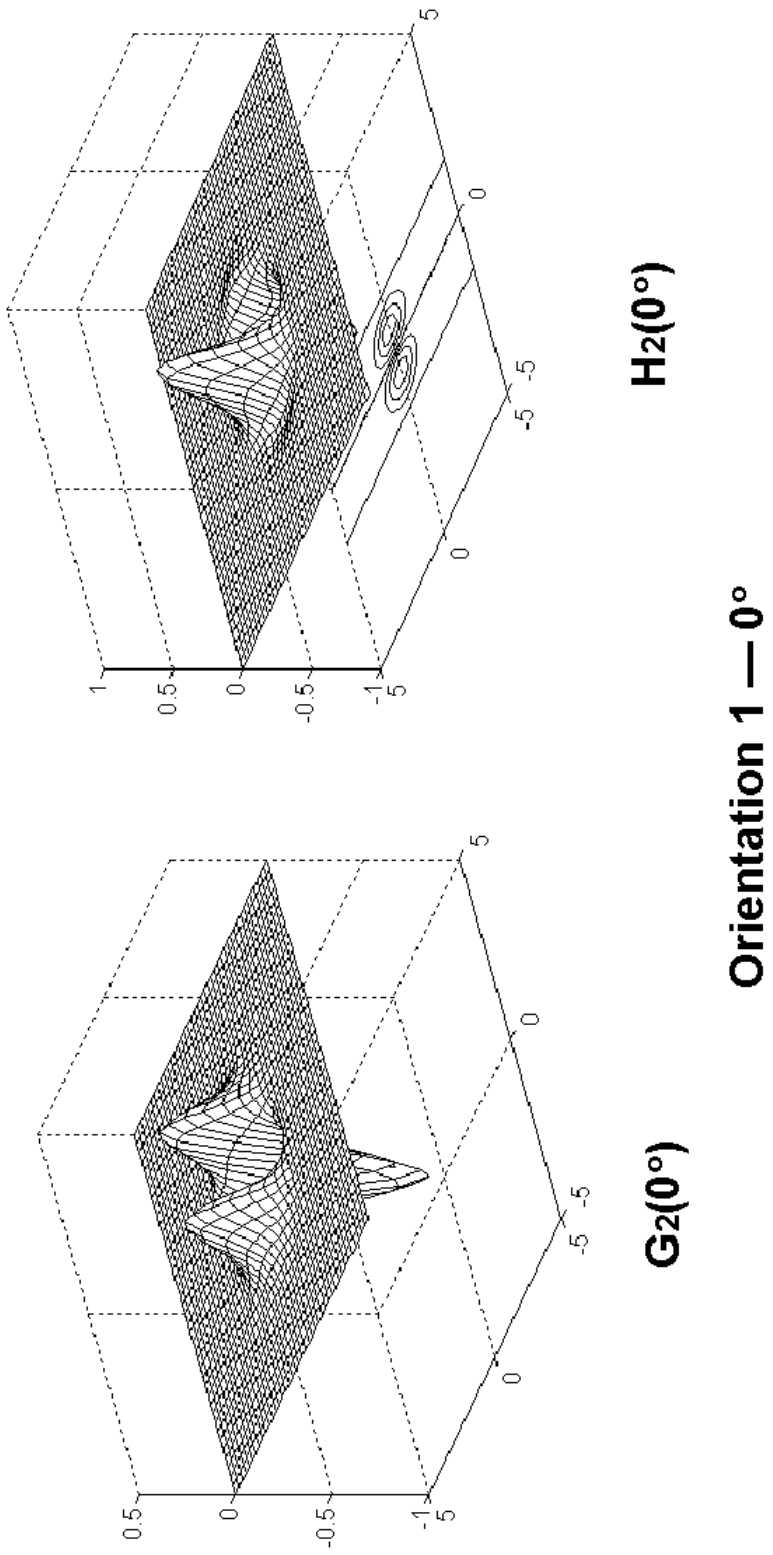

FIGURE 52. Quadrature Pair at 0°

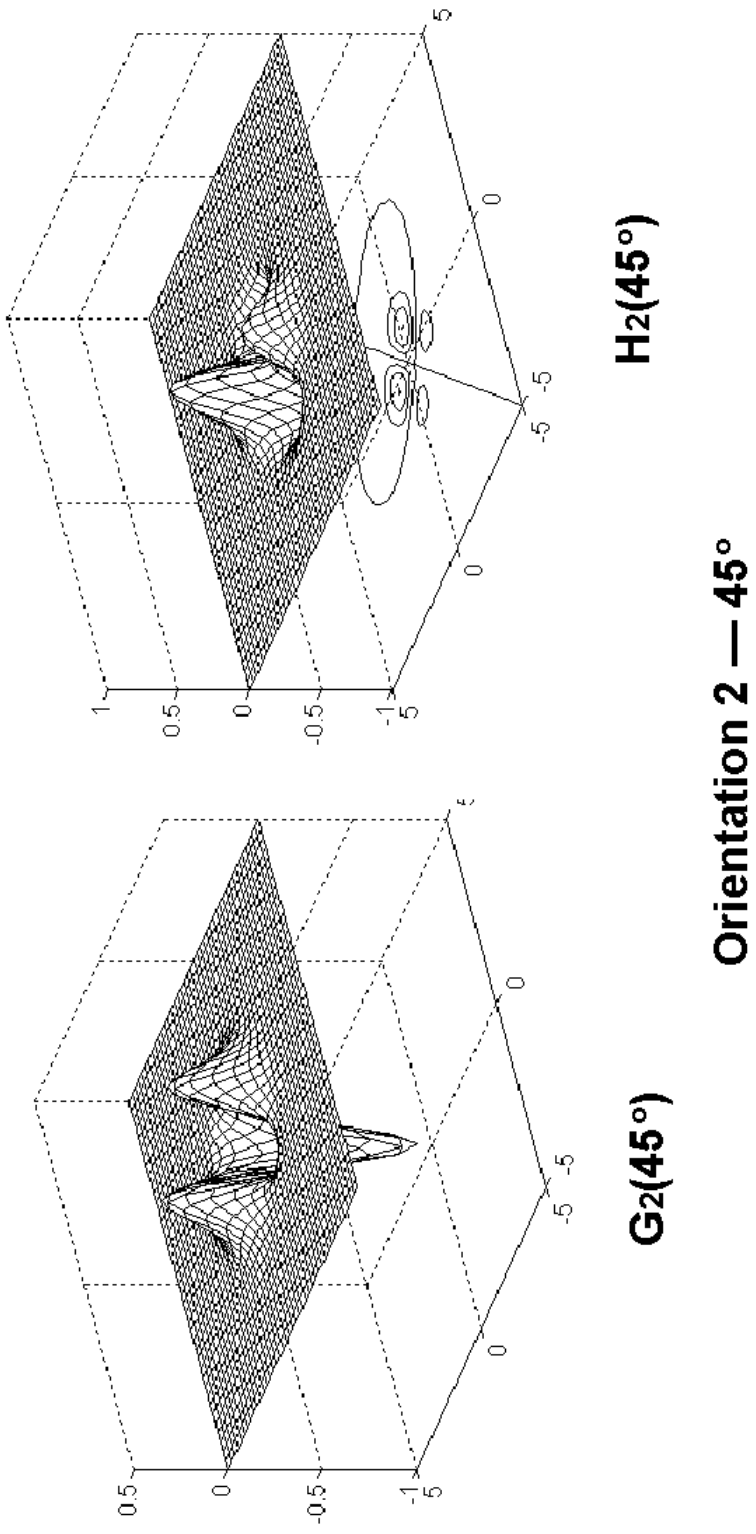

FIGURE 53. Quadrature Pair at 45°

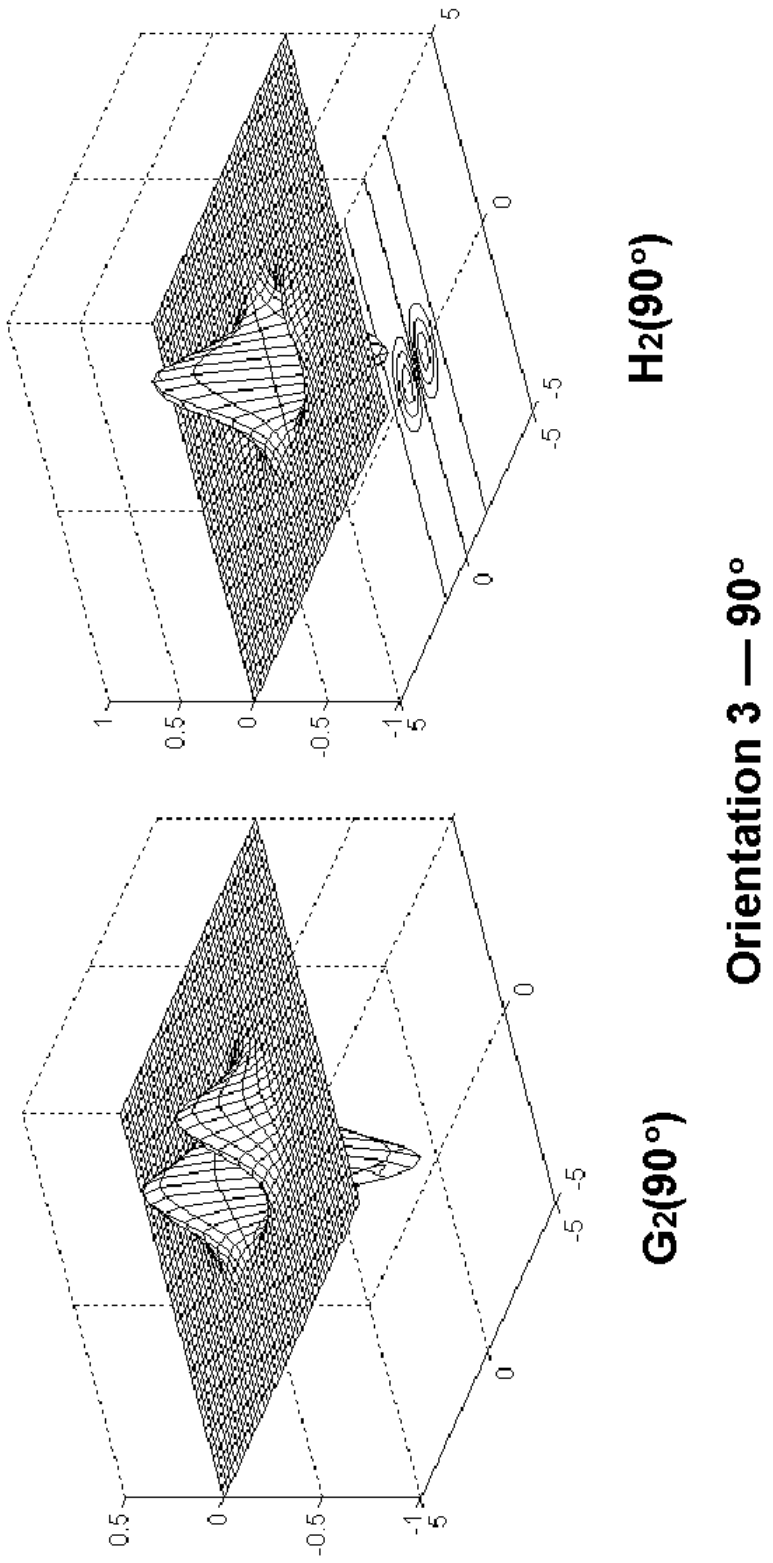

FIGURE 54. Quadrature Pair at 90°

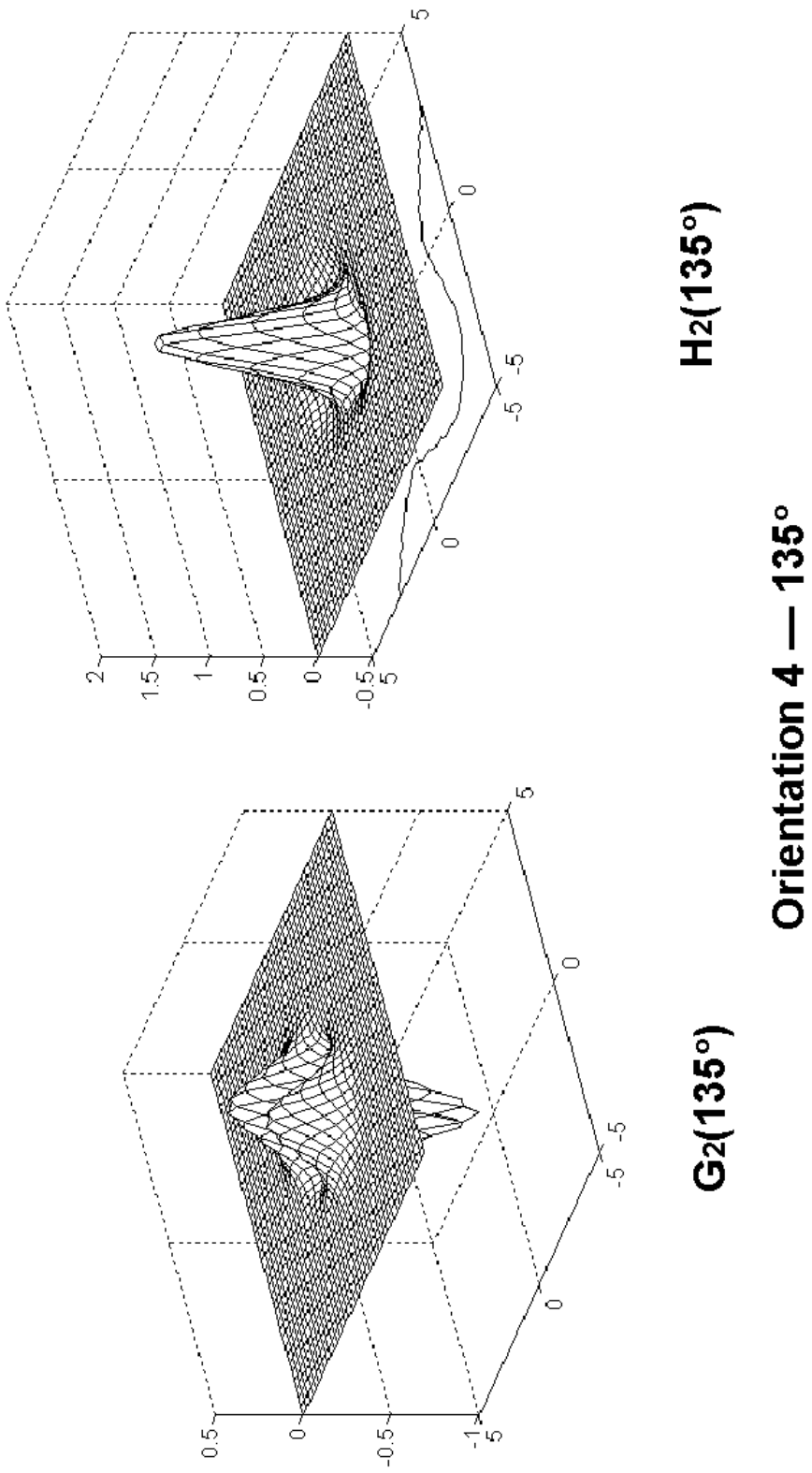

FIGURE 55. Quadrature Pair at 135°

BIBLIOGRAPHY

Barten, P. G. J. (1989). 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*, SPIE Vol. 1077, pp. 73-82.

Bolin, M. R. and Meyer, G. W. (1995). A Frequency Based Ray Tracer. In *Proceedings ACM SIGGRAPH'95*, pp. 409-418.

Burt, P. J. and Adelson, E. H. (1983A). The Laplacian Pyramid as a Compact Image Code. In *IEEE Transactions on Communications COM-31*, pp. 532-540.

Burt, P. J. and Adelson, E. H. (1983B). A Multiresolution Spline With Application to Image Mosaics. In *ACM Transactions on Graphics*, Vol.2, No.4, pp. 217-236.

Carlson, C. R. and Cohen, R. W. (1980). A Simple Psychophysical Model for Predicting the Visibility of Displayed Information. In *Proceedings of the Society for Information Display*, Vol.21/3, pp. 229-245.

Cornsweet, T. N. (1970). *Visual perception.* Academic Press.

Daly, S. (1993). The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity. In *Digital Images and Human Vision*, Watson, A. B., editor, MIT Press, pp. 179-206.

Freeman, W. T. and Adelson, E. H. (1991). The Design and Use of Steerable Filters. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13, Num. 9, pp. 891-906.

Hall, E. L. (1979). *Computer Image Processing and Recognition.* Academic Press.

Karasaridis, A. and Simoncelli, E. (1996). A Filter Design Technique for Steerable Pyramid Image Transforms. In *Proceedings of ICASSP-96.*

Landy, M. S. and Movshon, J. A. (1991). *Computational Models of Visual Processing.* Bradford.

Legge, G. E. and Foley, J. M. (1980). Contrast Masking in Human Vision. In *Journal of Optical Society of America*, Vol. 70, pp. 1458-1470.

Lubin, J. (1993). The Use of Psychophysical Data and Models in the Analysis of Display System Performance. In *Digital Images and Human Vision*, Watson, A. B., editor, MIT Press, pp.163-178.

Lubin, J. (1995). A Visual Discrimination Model for Imaging System Design and Evaluation. In *Vision Models for Target Detection and Recognition*, Peli, E., editor, World Scientific, pp. 245-283.

Mitsa, T., Varkur, K. L., and Alford, J. R. (1993). Frequency-Channel-Based Visual Models as Quantitative Quality Measures in Halftoning. In *SPIE Vol. 1913*, pp. 390-401.

Peli, E. (1990). Contrast in Complex Images. In *Journal of Optical Society of America*, A/Vol.7, pp. 2032-2040.

Peli, E. (1995). *Vision Models for Target Detection and Recognition.* World Scientific.

Schade, O.H., (1948) Electro-Optical Characteristics of Television Systems. I. Characteristics of Vision and Visual Systems. In *RCA Review 9*, pp. 5-37.

Schreiber, W. F. (1993). *Fundamentals of Electronic Imaging Systems - Some Aspects of Image Processing.* Springer-Verlag.

Schwartz, S. H. (1994). *Visual Perception - Clinical Orientation.* Appleton and Lange.

Simoncelli, E. P., Freeman, W. T., Adelson, E. H., and Heeger, D. J. (1992) Shiftable Multi-Scale Transforms. In *IEEE Transactions on Information Theory*, Vol. 38(2), pp. 587-607.

Wandell, B. A. (1995). *Foundations of Vision.* Sinauer Associates, Inc.

Watson, A. B. (1987). The Cortex Transform: Rapid Computation of Simulated Neural Images. In *Computer Vision, Graphics, and Image Processing 39*, pp. 311-327.

Westheimer, G. (1986). The Eye as an Optical Instrument. In *Handbook of Perception and Human Performance*, Boff, K. and Thomas, J., editors, Wiley and Sons, New York, NY.