ZERO-CROSSINGS:
SYMBOLIC VISION PRIMITIVES
EMULATING PHYSIOLOGIC ENCODING SCHEMES

Daniel P. Lulich
B.S., Portland State University, 1981

A thesis submitted to the faculty
of the Oregon Graduate Center
in partial fulfillment of the
requirements for the degree
Master of Science
in
Computer Science & Engineering

Dec, 1985

The thesis "Zero-crossings: Symbolic Vision Primitives Emulating Physiologic Encoding Schemes" by Daniel P. Lulich has been examined and approved by the following Examination Committee:

Richard B. Kieburtz, Thesis Research Advisor
Professor and Chairman,
Department of Computer Science and Engineering
Oregon Graduate Center

Daniel W. Hammerstrom, Thesis Committee Chairman
Associate Professor,
Department of Computer Science and Engineering
Oregon Graduate Center

David Maier
Associate Professor,
Department of Computer Science and Engineering
Oregon Graduate Center

Kent A. Stevens
Associate Professor,
Department of Computer and Information Science
University of Oregon

ACKNOWLEDGEMENTS

This thesis would not have been possible without the help and encouragement of several individuals. I would like to thank them at this time.

Dr. Richard Kieburtz provided the research atmosphere. His encouragement to work on a problem that was personally exciting motivated this thesis. Dr. Kent Stevens turned "hacking" into directed research. Without his guidance and intellectual input this work would still be a pile of unrelated pieces. Lise Storc and Harry Porter contributed immensely to the completion of this thesis.

I would also like to thank Dr. Dan Hammerstrom and Dr. David Maier for their participation on my thesis committee. In addition, several members of the faculty and students of the Department of Computer Science and Engineering at the Oregon Graduated Center were willing subjects in several perceptual experiments. All of these individuals deserve a special thanks.

To all of you thanks a lot!

ABSTRACT

Zero-crossings:
Symbolic Vision Primitives
Emulating Physiologic Encoding Schemes

Daniel P. Lulich, M.S.
Oregon Graduate Center, 1985

Supervising Professor: Richard B. Kieburtz

David Marr has proposed a computational model of early vision. This model uses the current understanding of the physiology of the human visual system as an intuitive basis for the discovery of the algorithms necessary for machine early vision [Marr82]. This thesis will describe how to implement Marr's model by analysis of each computational task comprising early vision. The first task is sampling the visual world. This computation builds a discrete two-dimensional intensity array. Next, digital filtering techniques are used to construct symbolic primitives which form an intermediate representation, from which the ***raw primal sketch*** is built [Marr76]. These primitives are the zero-crossings of the second directional derivative taken in all orientations and at a number of different spatial scales throughout the intensity array. Marr *et al.* have argued that images encoded with zero-crossing primitives contain sufficient symbolic information to reconstruct the original visual image, and that these primitive symbols are formed into tokens for manipulation by

higher-order vision algorithms [CrMP80].

At the finest available spatial scale, how are the shapes of objects represented by the visual system? The author has investigated the above question by performing computer and psychophysical experiments. Preliminary results show that zero-crossing primitives are consistent with the shapes perceived by human subjects. Therefore, zero-crossings may be available for scrutiny at the finest scale of visual resolution. This result is consistent with Marr's model.

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

What does it mean for a man or machine to see? Seeing is the creation of internal descriptions of the physical world [RiMa81]. As a sensation, seeing is effortless. Therefore, we often take for granted the enormous complexity involved in processing descriptions of what we see. Imagine sitting in a field and observing nature. Notice the subtlety of color and myriad textures that make up all that is seeable. The sheer quantity and variety of physical objects along with the richness of our perceptions give the impression that seeing is magic. How can a machine be constructed to collect all of this detailed information and then act quickly and intelligently upon the internalized representation? It is much harder to build a seeing machine than it first appears.

David Marr during the mid-seventies observed that vision is primarily an ***information processing task*** [Marr76]. He noted that a vision machine would have to process large and complex blocks of real world data. The implication of Marr's operational definition is that the tools of Artificial Intelligence, Digital Signal Processing, and Systems Design can be brought to bear upon vision problems [Brad82].

Marr's approach is to first partition vision into a group of smaller processes. However, where should the partitions fall? With vision there is a

model machine — the human visual system — that solves the problem well. Thus, Marr's approach was to partition the design of a seeing machine following closely the natural partitions of the human visual system. This thesis will consider the first set of natural visual partitions. Sampling and symbolic encoding are the first operations of a vision system and are aptly called ***early vision.***

Exactly how the human visual system encodes and decodes the physical world is still a matter of much speculation [Greg73]. The neurophysiologists and perceptual psychologists trying to understand vision are asking many of the same questions as computer-vision researchers. Thus, not only can computer-vision researchers use historical knowledge from the physiology of vision, but their research may in turn contribute to the fundamental understanding of human vision [Marr82]. Contributions to the basic understanding of human vision is one of the primary goals of Marr's information processing approach. Portions of this approach will be implemented and evaluated here.

Since vision can be viewed as an information processing task, what is the nature of the information acted upon? The physical world is composed of surfaces that reflect light, emit light, or act with optical properties on light. It is the stream of photons from or through these surfaces which are the input to light sensors in the human eye. This intensity information describes the physical properties of light-manipulating surfaces.

The surfaces acted on by light are complex, made up of spatially elaborate structures (i.e., they are not necessarily smooth everywhere). For example, from a distance a grass lawn looks like a smooth continuous green surface. As you approach it, the individual blades of grass come into view and the surface is textured by a multitude of stick-like structures. At close range, an individual blade is a smooth surface with a prominent feature running up the middle. Each of these viewing distances presents a light transducer with different intensity data describing the spatial elaboration of the surface. It is interesting to note that at each particular viewing distance, objects generating the intensity data tend to be more like each other than at smaller or larger distances [Hild80]. For example, blades of grass look similar when viewed up close.

The physical descriptions of the objects which compose the world is the information a vision processor must extract. In early vision, this extraction process begins by encoding (i.e., sampling) the intensity of light at a point in the input scene. The human eye gracefully performs intensity encoding, and a simple television camera or charge coupled device could be used by a machine. Having sensed and quantized the intensity data, have we really seen anything? According to Aristotle, "Vision is to know what is in the world after looking at it." All we have gathered is a group of intensity values, and these values do not mean surfaces or blades of grass. However, using Marr's approach we now have a knowledge representation (i.e., intensity information) that can be used to recover the properties of surfaces of real-world objects.

In order to know what we see, it is necessary to distinguish between the intensity data of different objects and their boundaries. We have to analyze this data and use the results of this analysis to construct a usable symbolic representation of the original object. This set of symbols can then be manipulated to discover the object's relationship to the rest of the world. Marr's approach encodes intensity information into a set of symbols called ***zero-crossings,*** which mark the positions of significant intensity changes.

A set of zero-crossing symbols may be manipulated by the brain to complete our perception of the world. These perceptual tasks are higher-order visual processes, which include the discovery of surface shape from intensity, recognition of familiar objects, stereopsis, the analysis of texture and motion. These problems have intrigued artificial vision researchers for two decades, and many interesting algorithms and application-dependent vision systems have solved portions of these problems. However, vision researchers have removed from consideration most natural objects and have substituted a few well-chosen representative objects. Unfortunately, this approach limits the generality of the algorithms proposed [Brad80]. Marr's information processing approach views the human visual system as an example of a general vision processor and seeks algorithms to emulate it.

The purpose of this thesis is to implement Marr's information processing approach to vision, and to test this implementation for consistency with the human visual system. We will limit our discussion to early vision. In the next two chapters the sampling of intensity data and the construction of zero-

crossings will be discussed. In the third chapter we will add a smaller zero-crossing detector which will improve the sensitivity to fine details of Marr's algorithm. We will also examine the limits of visual acuity of humans, and ask how can such fine acuity be explained by our computer model?

In the last chapter, the implementation will be evaluated by examining the response of the model to very fine detail. At the limits of fine spatial resolution we find that Marr's model predicts that zero-crossing artifacts will occur. A computer-based experiment will be performed that will locate and specify the size and shape of these artifacts. The finding of these artifacts prompted the question, "Will these artifacts also be perceived by human visual system?" Resolving this question is the primary goal of this thesis. The motivation for the question extends from Marr's use of the human visual system as the primary model. If the computer experiments find artifacts, then humans may see similar artifacts. This will be true if we have modeled the human visual system correctly. A psychophysical experiment is performed using human subjects, which demonstrates that artifacts do indeed appear. The shapes and sizes of these artifacts correspond well with the shapes and sizes of zero-crossing artifacts found in the computer experiment. These results will be discussed in detail and suggestions for further research proposed.

SAMPLING VISUAL INFORMATION

In the act of vision, an image of the physical world is presented to each eye. The image is composed of intensity changes caused by light reflecting from objects making up the original scene. These intensity changes can be expressed mathematically as a continuous two-dimensional function limited by the pupil size and resolving power of the lens of the eye. The quantization of the amplitudes of individual intensity changes at discrete positions along this continuous function is a ***sampling*** operation. After sampling, a two-dimensional intensity array has been created to represent the visual information about the physical world [CrMP80]. If the sampled array of intensity values is displayed as a matrix of dots, where the brightness of each dot represents the value at the corresponding x-y position in the intensity array, the image would appear similar to a photograph of the visual scene.

When taking a photograph, certain constraints must be met to insure that an adequate picture results. A good photograph results when the resolution and speed of the film are chosen properly. If the speed of the film is too slow, it is impossible to capture a moving object. If the resolution of the film is too low, crucial details in the image are lost. Losing visual information is unacceptable to any vision machine unless the portion of information lost is

extremely small or is useless to the higher-order vision algorithms [Baxe84]. Therefore, the resolution and speed of the sampling process must be chosen just as carefully as photographic film. The approach used here is to choose values for speed and resolution greater than or equal to the worst case performance of the human visual system.

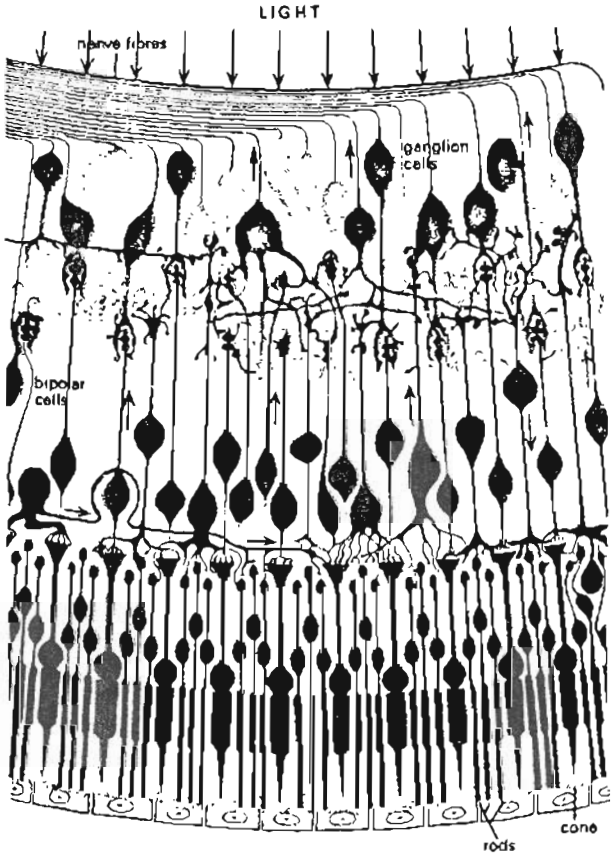


Fig. 1. Cross-section of the retina.
Light travels through layers of blood vessels, nerve fibers and supporting cells to light-sensitive cells (rods and cones). These lie at the back of the retina, which is functionally inside-out [Greg73].

The sensory portion of the human retina contains an array of light-sensitive cells, the rods and cones (see Fig. 1). The purpose of these cells is to

sample the light intensity at each point in the visual field in varying levels of background illumination. The cones are active in high levels of illumination, whereas the rods are tuned for low levels [Greg73]. This array of cells is not

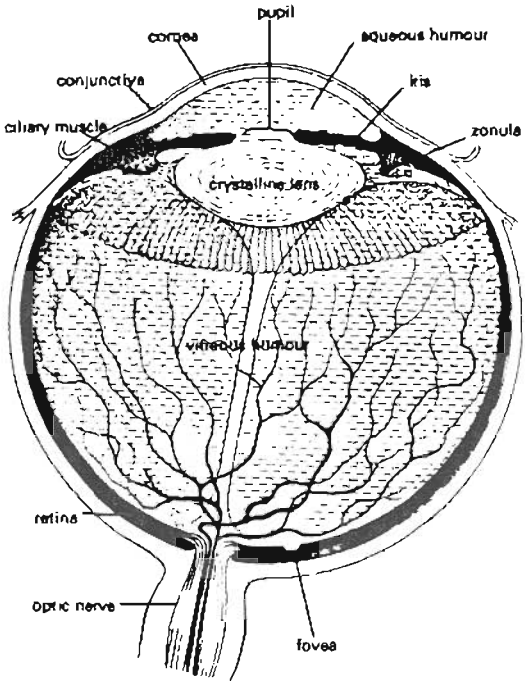


Fig. 2. Cross-section of human eye.
The retina is a three-dimensional structure
located at the back of the globe. The fovea
lies in the central portion of the retina [Greg73].

rectangular but has a three-dimensional shape that is mapped onto the inside of a hemisphere. In the central region of the retina lies the fovea, which is a small area populated by densely packed cones (see Fig. 2). High-visual-acuity tasks such as reading are performed after projecting the image onto the fovea. Since vision researchers are interested primarily in high visual acuity tasks, the

cone cell is the transducer modeled.

In order to capture the topographic arrangement of cone cells, an imaginary two-dimensional array is laid as a grid across the central portion of the retina. The grid is centered upon the fovea. Each element in the grid corresponds to one or more cone cells, thus setting the lower bound of resolution. There are arguments against a rectangular-transducer mapping based upon the anatomy of the cone cells. When packed closely together, these cells do not map well to a grid-like structure. The shape of a single cone and its neighbors would best map to a hexagonal array [CrMP80]. However, the two-dimensional array is a well understood data structure and, when properly dimensioned, serves as a satisfactory topology for light-sensitive transducers. Therefore, a two-dimensional array is used as the output of the sampling operation.

In the retina the size of a single cone is small, on the order of 25″ of arc†, which guarantees a highly resolved input image [MaHP79]. Even at this resolution, information begins to be lost as soon as detailed structures in the input image project smaller than a single cone. Computer representations of images suffer similar problems when attempting to represent too much of the input image as a single element of the intensity array. To prevent loss of information, the parameters described in the Sampling Theorem are used to sample the intensity values from a visual image.

† Seconds of arc is an angular measure which can be calculated by $angle = arctan\ d/D$. The image of a 1 inch line segment viewed from 10 feet subtends 28.7′ minutes of arc on the retina. This is about 70 cone diameters.

The Sampling Theorem in Two Dimensions :
(Gabor, Nyquist, Shannon, Whittaker)

A band-limited continuous function of two variables $r(x,y)$ can be recovered ***exactly*** from a rectangular array of its sampled values $r_s(x,y)$ by interpolating the original values from the sampled values. Exact recovery is guaranteed if the sampling frequency ω_s (Nyquist frequency) is greater than twice the maximum spatial frequency $2\omega_{max}$ (Nyquist rate) of the original function:

$$\omega_s > 2\omega_{max} \tag{1}$$

The proof of the Sampling Theorem is widely available [Brac65][Hamm77][OpWi83]. The Sampling Theorem gives the specific requirements for sampling a continuous form of an image. Using the Sampling Theorem, all of the original image detail can be recovered from a few well chosen samples. Therefore, a properly designed sampling algorithm is a reversible encoding scheme that preserves information content.

When considering the properties of the Sampling Theorem, the concept of a ***signal*** is useful. The intensity change information in a image is a continuous function of two independent positional variables. This function is a signal. A signal has real numbers as values, and these values are characteristics of physical phenomena [OpWi83]. In our common conception of signals (e.g., radio signals), a signal exhibits properties such as moving between sender and receiver, the ability to be transformed from one form of energy to another, and a power

or strength that is measurable. Intuitively, the light which conveys intensity information exhibits all of these properties.

Signal processing is the application of mathematical techniques to signals in order to transform them into more usable or informative forms [Hamm77]. Thus, both the eye and vision machines are signal processors. The Sampling Theorem shows how to sample a continuous function by the application of a sampling function. This is a signal processing operation. The sampling function has the property of making the continuous input function everywhere discrete. The sampling function used to build a rectangular sampled array of intensity values is:

$$r_s(x,y) = r(x,y) \cdot \mathrm{comb}(x/X) \cdot \mathrm{comb}(y/Y) \tag{2}$$

where:

$$\mathrm{comb}(x) = \sum_{n=-\infty}^{\infty} \delta(x-n) \tag{3}$$

This sampling function is a two-dimensional array of δ functions spaced at intervals of width X in the x direction and width Y in the y direction (see Fig. 3).

The function δ used throughout this thesis is the unit impulse function† and describes a point in sampling space where a sample is to be taken [OpWi83]. The δ function consists of a spike (i.e, impulse) whose amplitude is

† The unit impulse is a function common to digital signal processing. The mathematical properties of this function are not very well founded, but have been thoroughly studied as part of the class of generalized functions. The description of the function given here is more than adequate as a practical method for implementing the sampling operation. For a complete development of the unit impulse function see Bracewell [Brac65].

one. On either side of the this spike the value of the function is zero. Sequences of δ functions are often constructed and describe how samples are chosen across the range of an input function. The interval between impulses is always assumed to be equal. The operation of sampling is the multiplication of the input signal by the sampling function. Thus, the result of sampling scales the amplitude of the δ impulses so that they are equal to the amplitude of the input signal at that position (see Fig. 4).

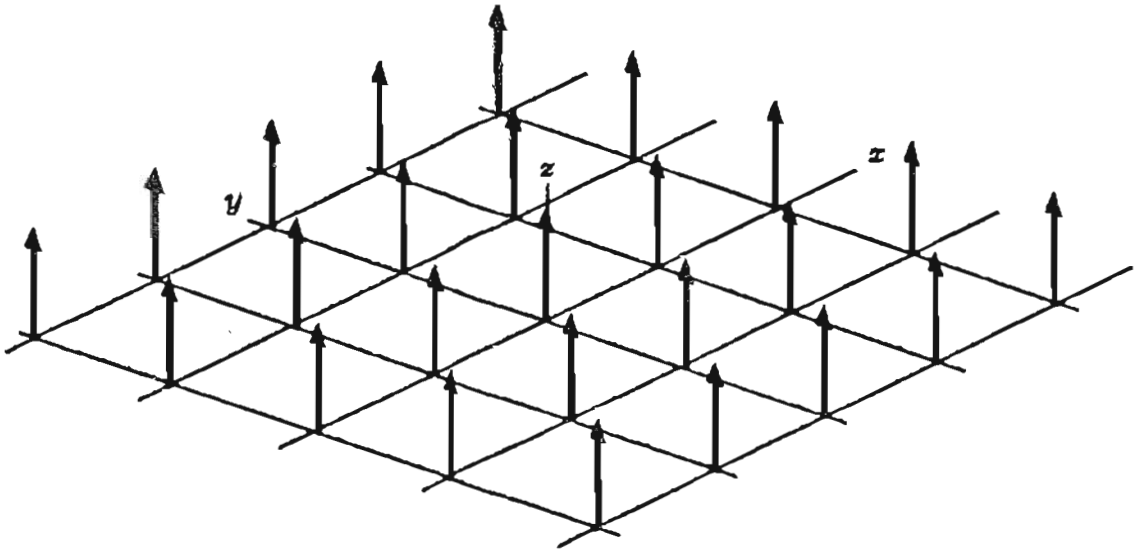


Fig. 3. Two-dimensional array of δ functions.

Next, the amplitudes of the δ functions are quantized to complete the machine representation of the intensity array. We would like to model this quantization process after the transduction process of a single cone. The cones

are sensitive to higher levels of illumination. Therefore, a threshold of illumination exists below which the cones are insensitive. Also, the level of background illumination influences the range of intensity change to which cones are sensitive.

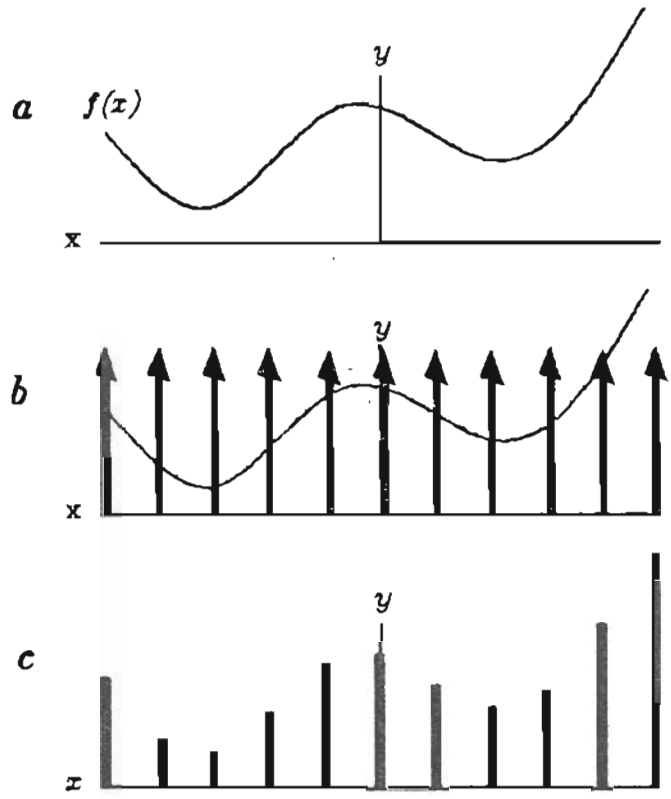


Fig. 4. Sampling operation in one dimension.
(a) $f(x)$ is a one-dimensional continuous function
(b) Sampling operation: a row of δ functions is multiplied with $f(x)$.
(c) The result of sampling.

This suggests that our quantization algorithm must dynamically adjust intensity thresholds as a function of physical variables (e.g., background illumina-

tion). This added level of complexity is a research issue. The intensity array values are quantized, modeling the general range of human ***perceived*** intensity. The human visual system perceives intensity changes in normal illumination in a logarithmic manner (i.e., the dark intensities values map to a larger range of the total intensity scale than bright intensity values) [Greg73][Baxe84]. The mathematical precision of quantization is usually determined by the number of bits conveniently available in the implementation hardware. All images used in this work are sampled with eight bits of precision.

The Sampling Theorem states explicitly how close the δ functions must be to one another to prevent information loss. To understand this result we define the spatial frequency of a signal. The frequency of a visual signal can be described through an example. An image constructed with alternating black and white stripes spaced closely together has a higher spatial frequency on average. Conversely, an image constructed in the same manner with stripes which are much wider has a lower average spatial frequency (see Fig. 5). Spatial frequency can be roughly defined as the rate at which an image changes intensity. The spatial frequency can be computed with a transformation that maps the signal's physical components (time domain) to its frequency components (frequency domain). This is done mathematically with the well-known Fourier transform [Brac65][Hamm77][OpWi83].

$$F(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-i2\pi(ux+vy)}dxdy \qquad (4)$$

and its inverse:

$$f(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}F(u,v)e^{i2\pi(ux+vy)}du\,dv \tag{5}$$

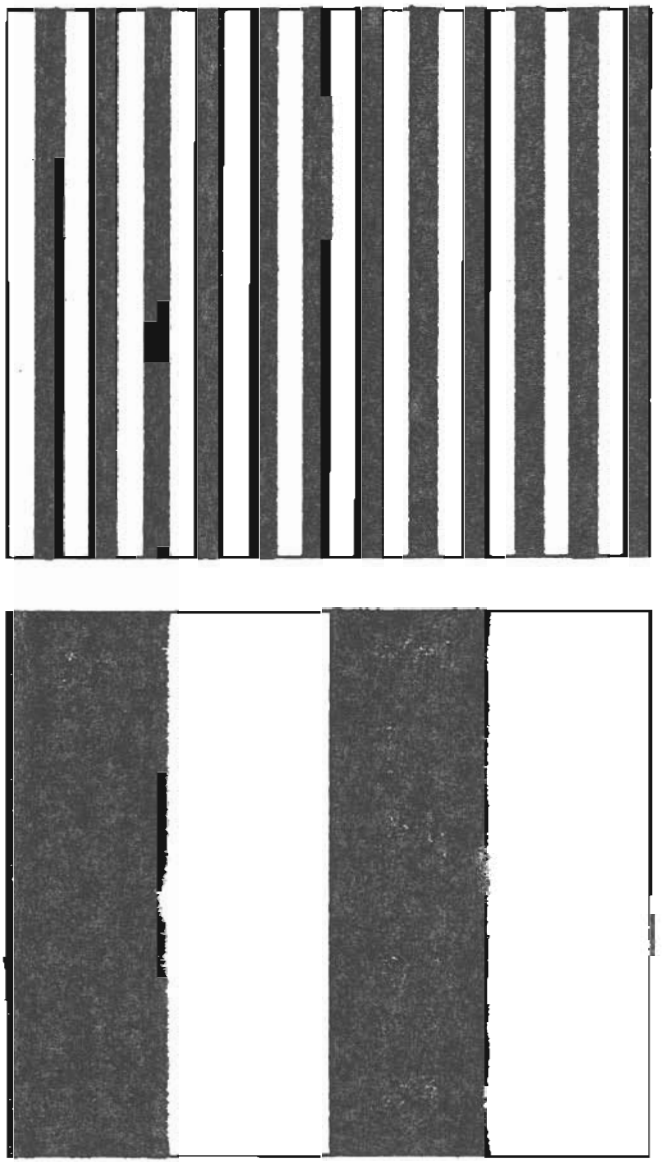


Fig. 5. Two-dimensional spatial frequency example.

The Sampling Theorem states that the input signal must be band-limited. For a signal to be band-limited there must exist an upper bound on the signal's highest frequency component. If the maximum frequency of the input signal is known, the Sampling Theorem tells how the sampling rate (the distance between δ functions) is chosen. The sampling rate must satisfy the Nyquist criterion, which states that the sampling rate must be twice the highest spatial frequency of the two-dimensional input function.

After the reflectance function has been sampled according to the Nyquist criterion, simple interpolation can be used to reconstruct the intensity information. If the sampling rate is chosen to be several times greater than the maximum input frequency, simple linear interpolation is adequate [OpWi83]. However, increasing the sampling frequency will increase the number of samples used to represent intensity data. To reduce storage, the Sampling Theorem is used to find the minimum number of samples necessary to successfully reconstruct the data. Thus, proper sampling can be viewed as a data compaction technique.

If only the minimum number of samples is acquired, a more sophisticated interpolation function will be needed. Such a function can be a well-chosen filter function, for example:

$$\mathbf{sinc}(x) = \sin \pi x / \pi x \tag{3}$$

In order to recover the original continuous-intensity data, the sampled intensity data must be filtered with a filter function similar to the above. Filtering

is a signal processing technique that here takes a weighted average that is specified by the filter function, and applies (i.e. convolves) this average to each point of the sampled data (see Appendix A). This filtering operation is constrained, and the correctness of the recovered data is dependent upon the amplitude and frequency characteristics of the filter function. However, If the operation is applied carefully the original continuous-intensity data can be totally recovered. Even though filtering with the above function is a more sophisticated operation, the end result is comparable to a simple linear interpolation of sampled data.

Does the human visual system take advantage of the results of the Sampling Theorem? Some vision researchers believe that it does [Barl78][MaHP78]. For example, the retinal image is band-limited by optics to about 60 cycles per degree. The size and spacing between cones cells is sufficiently close to guarantee that the Nyquist Criterion is met. It has also been suggested that layer $4C\beta$ of the striate cortex could be the site of image reconstruction [CrMP80]. Layer $4C\beta$ contains 50 times more processing cells than input cells. It has been hypothesized that a point-for-point reconstruction of the visual image could be performed here. However, there are not enough cells acting as inputs to $4C\beta$ to perform a one-to-one mapping of sampled intensity data. One possible solution is that a much more compact representation of intensity data is input $4C\beta$. To implement such a solution, further encoding of sampled visual information is necessary. In the next chapter, we will discuss a more compact representation of the visual image. The compaction will be done by localizing

discontinuities in the image intensity. This can be implemented by finding the zero-crossings in the second non-directional derivative. Later, we shall see that a reconstruction of the original image at the cortical level of the human brain may be necessary to explain a perceptual task called hyperacuity.

CONSTRUCTING ZERO-CROSSING PRIMITIVES

The intensity array of sampled values is the input to a higher-level algorithm for symbolic encoding. This algorithm is performed during the process called ***early vision***. One purpose of early vision is to construct a set of symbolic primitives that will further compact the representation of the physical world. These primitives mark or denote the areas of meaningful change in the intensity array.

Historically, both computer vision and image analysis systems have located significant intensity changes and marked them as primitives [Brad82]. The primitives have traditionally been called edges, since they roughly correspond to the physical boundaries between objects in the image. Many algorithms are available for detecting edges [Hild80][Baxe84][Winst84]. Are edges the only primitives that need to be labeled in the early stage of vision? Clearly, there are other physical phenomena that give rise to intensity changes, such as reflections, shadows and fine texture. A representation of the world consisting only of ideal edges can not account for human perception. A general vision processor needs more than edge information to reconstruct an image of the physical world. Marr has proposed that lines, bars, and blobs, which can be composed from raw edge information, may be the intermediate symbolic

tokens used in vision [Marr82].

The method chosen to detect intensity changes is dependent upon the physical phenomena that give rise to these changes. Many intensity changes are sharp, such as those on the borders of physical objects. Other changes are gradual, such as shadows cast on a dark surface. There are still other cases where both sharp and gradual intensity changes lie on top of one another. To detect such a variety of intensity change types requires a detection algorithm to isolate changes at different spatial scales.

Fig. 6. Intensity array: Woman wearing a hat.

Notice that in Figure 6 there are a number of object surfaces producing intensity changes at different scales. For example, the feathers on the hat are higher in spatial frequency than shadows on the cheek, and this property

identifies the spatial scale at which these objects can be found. Intensity changes at a particular scale are usually produced by objects of the same type [Marr82]. Since other types of objects will be found by detecting intensity changes at other scales, the relationship between objects of different types can be discovered by comparing intensity changes across scales [Marr82].

The rate of change of the function at a given point may be calculated by taking the derivative of the function. The first derivative will produce a maximum value at the greatest rate of change. The second derivative has a zero-crossing at this maximum (see Fig. 7). Intensity changes are calculated by taking the second derivative across the intensity array. This differentiation is performed by the application of a digital filter with second derivative properties. This filter can also be called a second derivative operator.

Digital filters designed for edge detection have traditionally been directionally sensitive [MaUl179]. The peaks or zero-crossings are determined only when the operator is properly oriented to an edge. For example, the Sobel operator is maximally sensitive to intensity changes for which the intensity difference is orthogonal to the x axis [Brad82]. In natural images, intensity changes are rarely organized neatly. Therefore, a second derivative operator that is non-directional is needed to produce zero-crossings independent of direction of intensity change. The only linear second derivative operator that is non-directional is the Laplacian operator [Hild80].

Any operator which is used to process natural images must be tunable to different spatial scales (i.e., spatial frequencies). If a scale is chosen so that

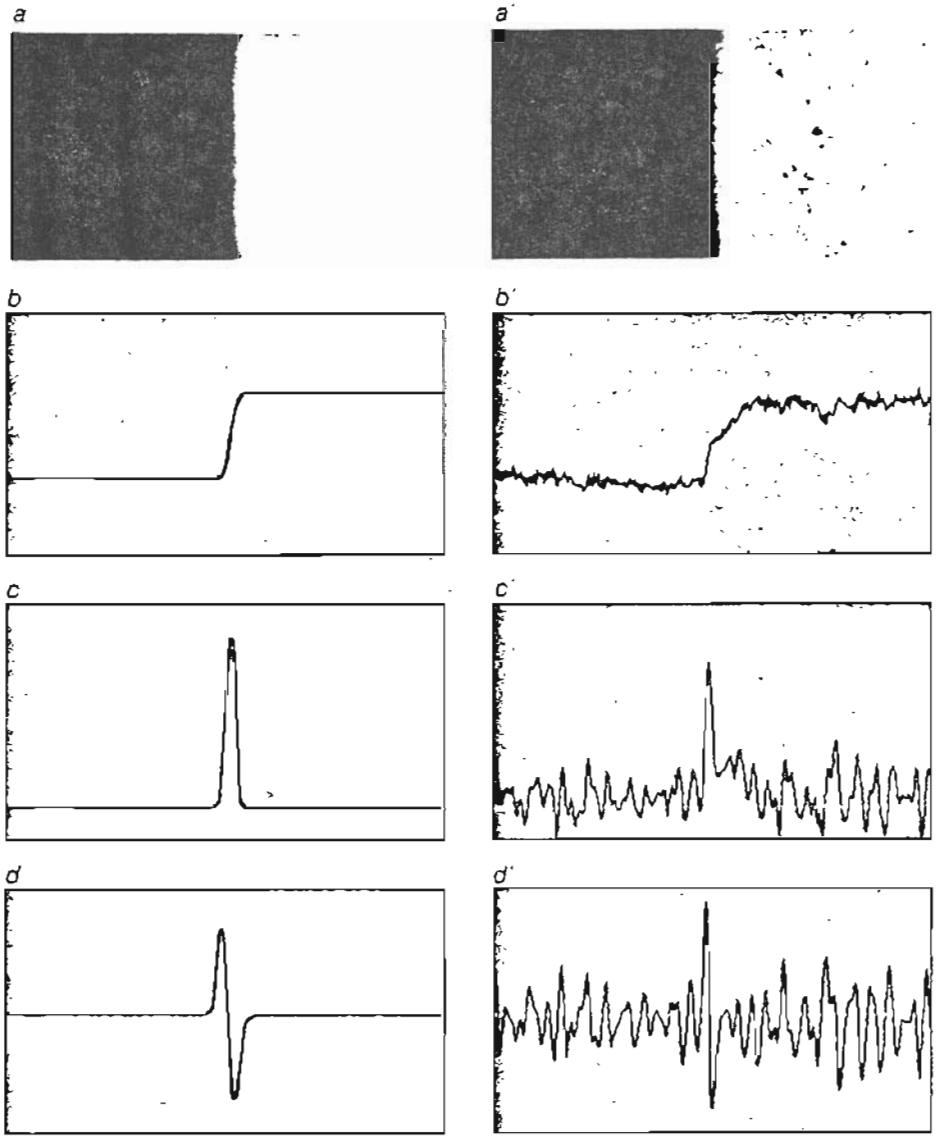


Fig. 7. Spatial derivatives

(a) Ideal spatial step. (b) One-dimensional cross-section of (a).
(c) First derivative of (a). (d) Second derivative with zero-crossing.
(a') Noisy spatial step. (b') Cross-section of step (b').
(c') First derivative of (a'). (d') Second derivative of (a').

sharply detailed objects are to be isolated, the filter should not be sensitive to soft, fuzzy objects. To isolate objects in the intensity array at a particular

scale, the intensity array is first smoothed with a Gaussian operator. The application of the Gaussian operator can be viewed as the averaging of an intensity value across neighboring intensity values (i.e., local neighborhood). The averaging is performed in such a way that closest neighbors are heavily weighted and progressively lower weights are assigned to neighbors further away. The effect of this sort of averaging is smoothing or filtering of unwanted high spatial frequencies in the locality of any intensity value. The size of the neighborhood determines the amount of local smoothing.

Algebraically, the two operators, the Laplacian and the Gaussian, can be combined into a single operator that retains the properties of both [Hild80]. In two dimensions this combination operator is called the Laplacian of a Gaussian and is given as:

$$\nabla^2 G(x,y) = [2 - (x^2+y^2)/\sigma^2]\, e^{-(x^2+y^2)/\sigma^2} \tag{4}$$

The $\nabla^2 G$ operator is a Mexican-hat-shaped operator, where ω is the width of the positive-going center of the operator (see Fig. 8). The spatial constant σ of the $\nabla^2 G$ equation is related to ω by:

$$\sigma = \omega/2\sqrt{2} \tag{5}$$

The width ω is called the ***excitatory center*** of the operator because it increases the value of intensity data positioned over this region. This center is surrounded by a negative annulus, called the ***inhibitory surround***. Data values positioned over this region are negatively weighted.

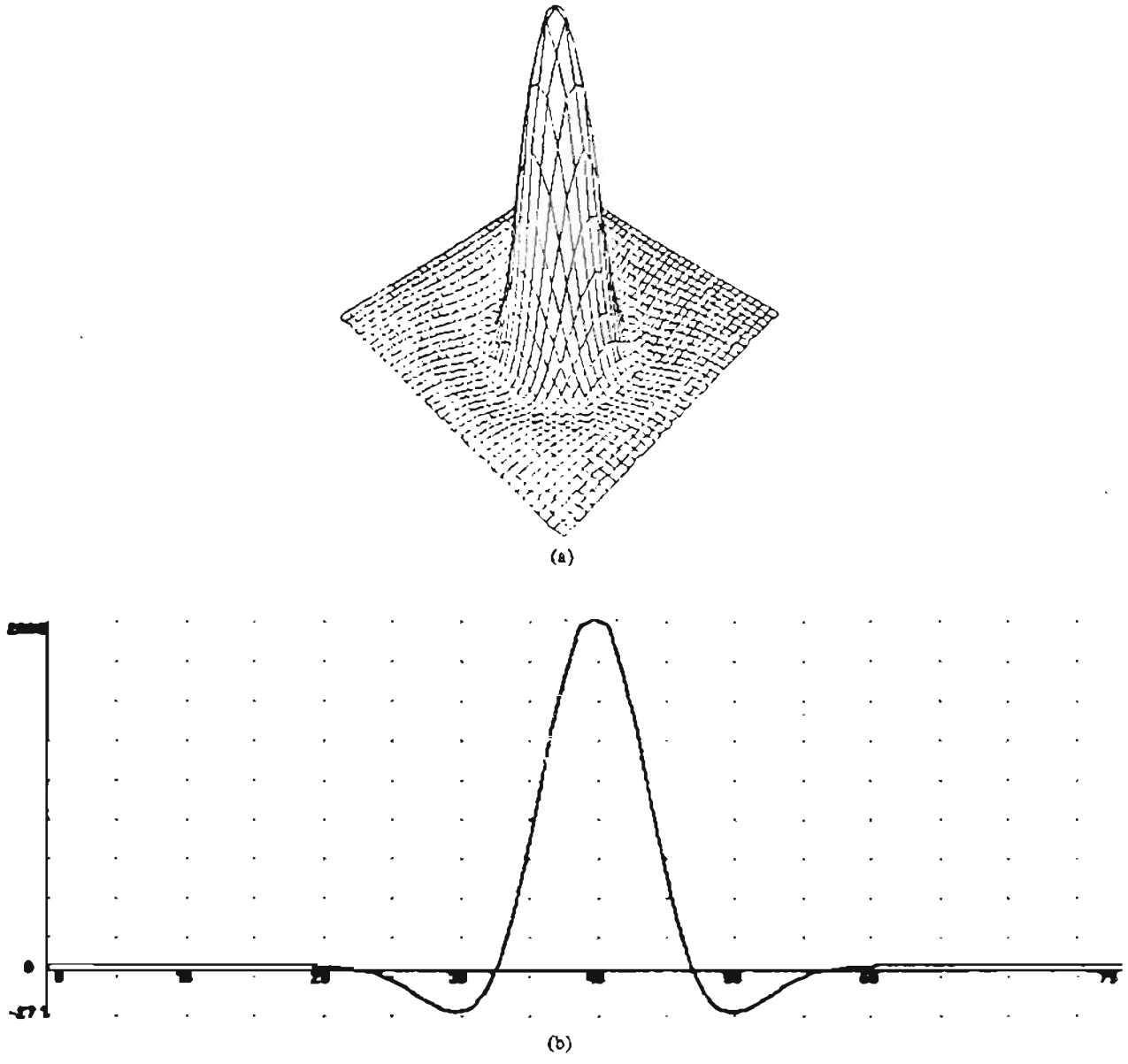


Fig. 8. The Laplacian of a Gaussian operator.
(a) Two-dimensional plot of the Mexican-hat-shaped operator.
(b) A one-dimensional cross-section of the operator.

A recent controversy in the literature has pointed out the need for careful engineering of this operator [GrHi85]. To properly construct an impulse response version of the continuous Laplacian of a Gaussian function, the $\nabla^2 G$ function must be sampled over a range that is at least 2ω on both sides of the

origin. This sampling will insure that the brim of the Mexican hat, or skirt of the filter, is wide enough to preserve the properties of the operator. This circularly symmetric operator must be balanced such that the total area under the sampled function integrates to zero. A balanced operator is less sensitive to high-frequency noise without losing any smoothing properties of the underlying Gaussian [GrHi85]. To maintain balance, the individual values of the impulse response filter are tweaked. In this work, the precision of the quantized version of the $\nabla^2 G$ is 1 part in 2048 using integer values (all values 1/2048 of the maximum amplitude or less are set to zero). This impulse response version of the filter is convolved with the intensity array to give a raw zero-crossing output (see Appendix A). The convolution of $\nabla^2 G$ with an intensity array can be expressed as:

$$\nabla^2 G(x',y') * I(x,y) \tag{6}$$

An example of the resulting image can be seen in Figure 9.

In this form, the zero-crossing information is available but not apparent. A simple algorithm is applied to the raw convolution data to enhance the locations of the zero-crossings. The algorithm can be stated as:

```
if convolution_value[x,y] > 0
   convolution_value[x,y] := 1;
else
   convolution_value[x,y] := 0;
```

This intermediate representation of the image is called the binarized image [Marr82]. In the binarized image, the locations of the zero-crossing are places where changes between zeros and ones occur. To mark the zero-crossings, the

binarized image is walked from right to left and top to bottom comparing adjacent elements. When a change is found, the location is marked (see Fig. 11).

Fig. 9. Fig. 6 convolved with the Laplacian of a Gaussian.
Zero is represented by a neutral gray shade. Lighter grays are positive. Darker grays are negative. Notice the overall smoothing which isolates features at one spatial scale.

How do we choose ω to set the sensitivity of the $\nabla^2 G$ operator to the desired spatial frequency scale? Physiological properties of the human retina are used to guide our intuition. Perceptual experiments and electrophysiologic recordings from the retinal ganglion cells (see Fig. 1) verify the shape and bound the sizes of the $\nabla^2 G$ operators [MaHP78]. In addition, Wilson and oth-

ers have proposed a set of psychophysical ***channels*** sensitive to spatial frequency as components of the human visual pathway [WiBe79][WiGi77]. The receptive fields of the retinal ganglion cells form the physiological front end of these channels. These fields are circularly symmetric and are composed of an excitatory region with an inhibitory surround. At each point on the retina, overlapping receptive fields are tuned to different spatial frequencies. The sizes of these fields grow with eccentricity from the foveal region. Therefore, fields in the retinal periphery are predominantly sensitive to lower spatial frequencies. Receptive fields are architecturally composed of ganglion cells connected to groups of cones via a variety of interneurons [Pogg84] (see Fig. 12). One of

Fig. 10. Binarized representation of Fig. 9.
White is 1 and black is 0.

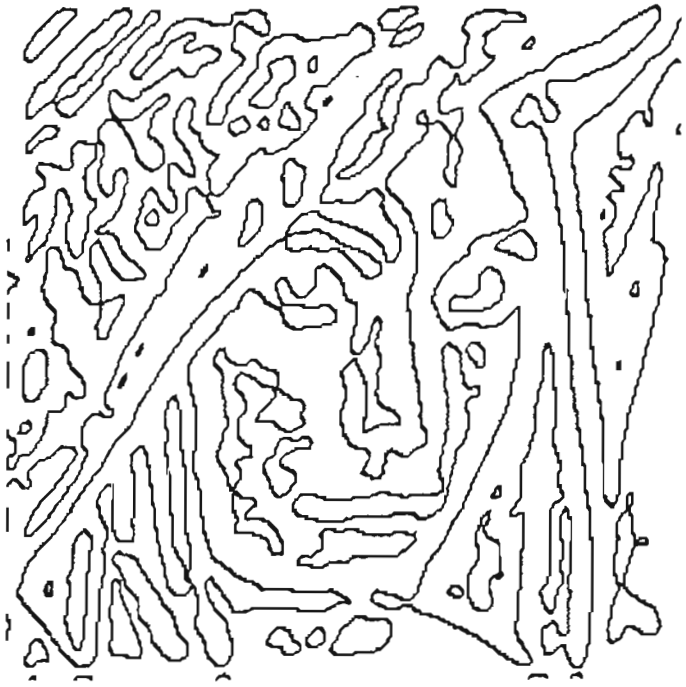


Fig. 11. Zero-crossing representation of Fig. 6.

the functions of a ganglion cell seems to be the differentiation of local intensity changes across the group of cones to which it is connected. The larger the number of cones, the larger the receptive field of the ganglion cell and the lower the spatial frequency sensitivity of the channel. It should be noted that there are probably a variety of receptive fields with different functional characteristics.

Wilson has modeled the local detection of intensity change of spatial channels as the difference of two Gaussian distributions (DOG) [WiBe79]. Marr and Hildreth have argued that the DOG function is a good engineering approximation to the $\nabla^2 G$ operator [MaHi80]. Therefore, the author used the sizes of the visual channels found by Wilson to choose ω in the $\nabla^2 G$ operator. These

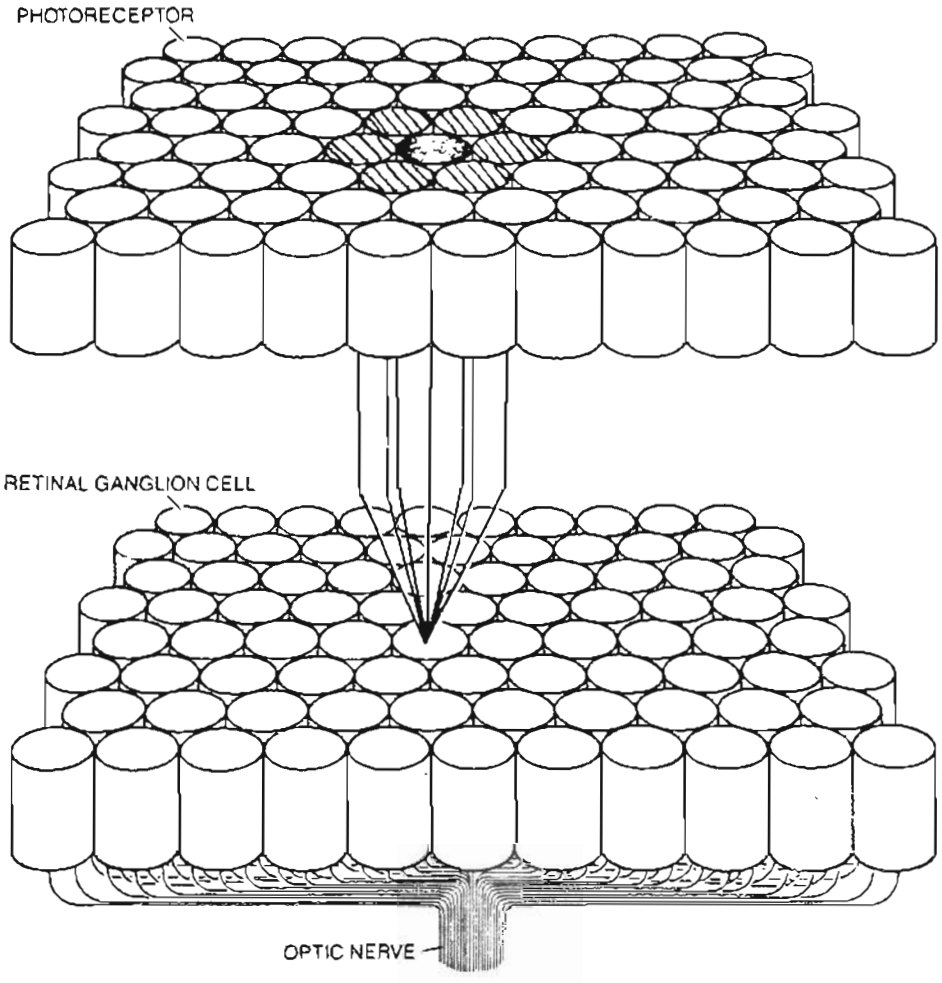


Fig. 12. A Receptive field of the human retina.

A single ganglion cell connects to a group of cones via intermediate nerve cells (not shown). The receptive field has an excitatory center and inhibitory surround. Excitation and inhibition are defined by weighted connections of the ganglion cell dendrites upon a neighborhood of photoreceptors.

channels, which are labeled N, S, T, and U have positive excitatory centers of 4.4′, 8.7′, 16.5′, and 29 6′, which correspond to the ω values of 4, 9, 17, 30 pixels†. The T and U channels found by Wilson are not functionally the same as N and S, and it is probably incorrect to assign their sizes as values of ω for a

† This is a measurement of distance that corresponds to the number of lighted dots on a graphics display device. The units of measure are pixels (picture elements). It is assumed there exists a one-to-one mapping from intensity array elements to pixels. It is also assumed that each pixel corresponds to approximately two cone cells of diameter 25″. Therefore, an image is focused on the fovea of an imaginary retina. All images in this chapter should be viewed from about 4.5 feet to guarantee the above relationships.

$\nabla^2 G$ operator [WiBe79]. However, lacking a better description for the low frequency channels these values were used.

To complete the zero-crossings representation from intensity data gathered from the physical world, a $\nabla^2 G$ operator of each size is convolved with the intensity array. Since four sizes of operators have been chosen to cover high, medium, and low spatial frequency features, four zero-crossing representations of the input image are created. From an aesthetic viewpoint, it is interesting to note that only a few different-size operators with elegant mathematical properties are necessary to process zero-crossing symbols. This observation is consistent with perceptual and physiological data from the retina.

Marr suggests that zero-crossings are organized into an rich and more informative representation called the ***raw primal sketch*** [Marr82]. In the raw primal sketch, information from each channel contributes to the construction of higher-level groups of tokens, such as line segments, bars, and blobs. In addition, the position and orientation of these symbols are tagged. These tokens are likely the input higher-order vision processes.

The construction of the raw primal sketch is computationally intensive. In this work, four complete convolutions of the intensity array are necessary. This is not as bad as it first seems. There are methods to reduce the amount of computation. The first is to look closely at the $\nabla^2 G$ operator. We have seen that this operator is an engineering approximation to the difference of two

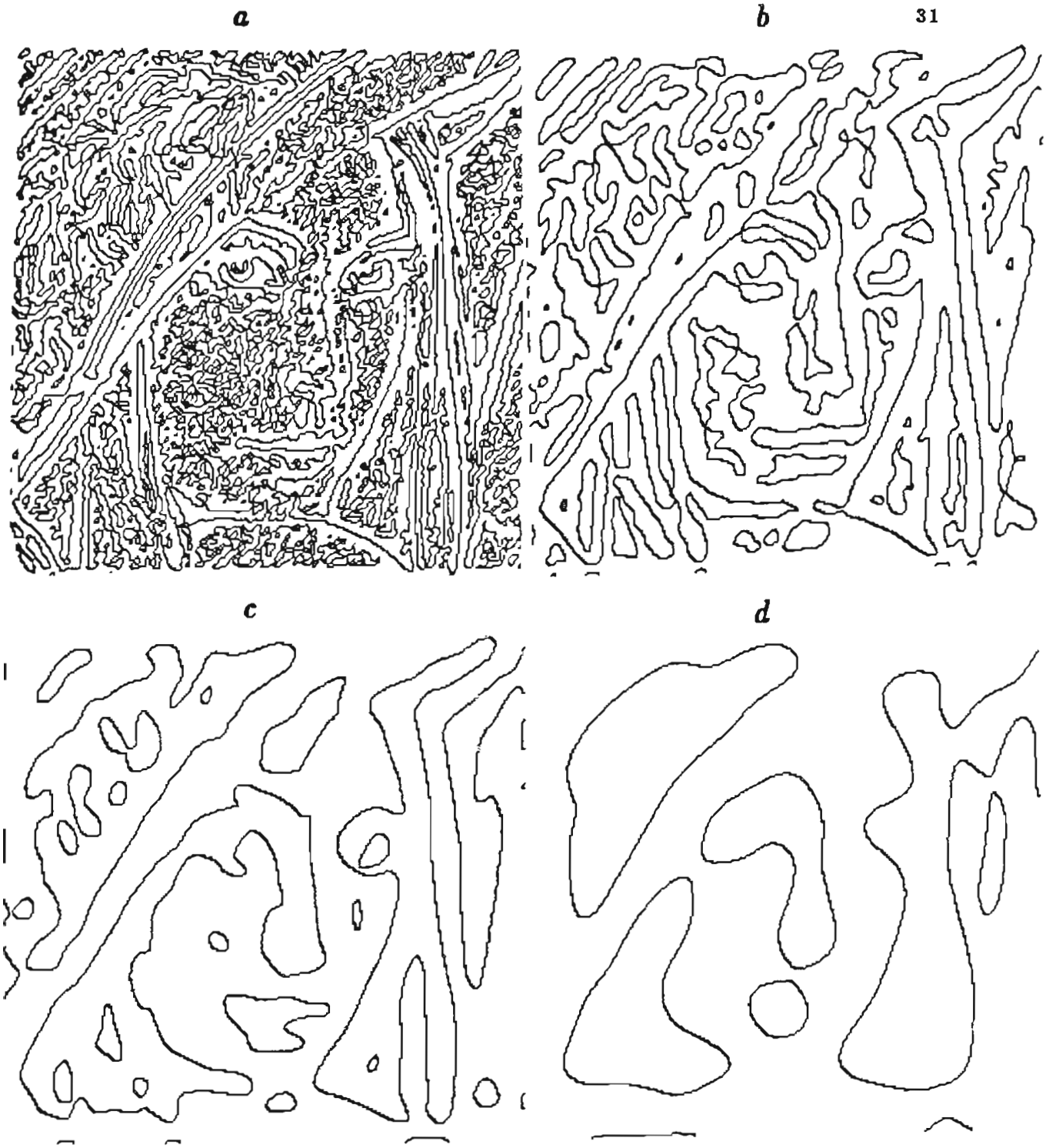


Fig. 13. Raw primal sketch

(a) Fig. 1 convolved with a $\nabla^2 G$ operator with ω equal to 4 pixels,
(b) ω equal to 9 pixels (c) ω equal to 17 pixels (d) ω equal to 30 pixels.
(1 pixel = 50″ at 4.5 feet).

Gaussians suggested by physiological experiments. Therefore, the DOG operator can be substituted for the $\nabla^2 G$ operator (see Fig. 13). The DOG operator has some distinct advantages over $\nabla^2 G$. A computational improvement is achieved by decomposing the two-dimensional form of the operator into four one-dimensional forms of the Gaussian function (see Appendix B for a description and proof of separability of Gaussians). Successive convolution using the one-dimensional forms requires a significantly lower number of operations than the two-dimensional DOG [CrPa84]. The time complexity of an individual convolution of the intensity array with a filter of a specific size is $\mathbf{O}(n^2)$ with the $\nabla^2 G$ operator. Using the DOG operator, this can be reduced to 4N [CrPa84] (see Appendix B). Even though the Difference of Gaussians operator can provide a significant computational improvement, the solution to the problem of real-time early vision remains elusive. An architectural clue to solving this incredible computational dilemma is found by examining the human visual system. This system is a massively parallel machine. Thus, to achieve real-time performance parallel computation must be used. The use of systolic array architectures and the high-density implementation of such architectures in VLSI are currently achieving real-time vision pre-processors [Kung84].

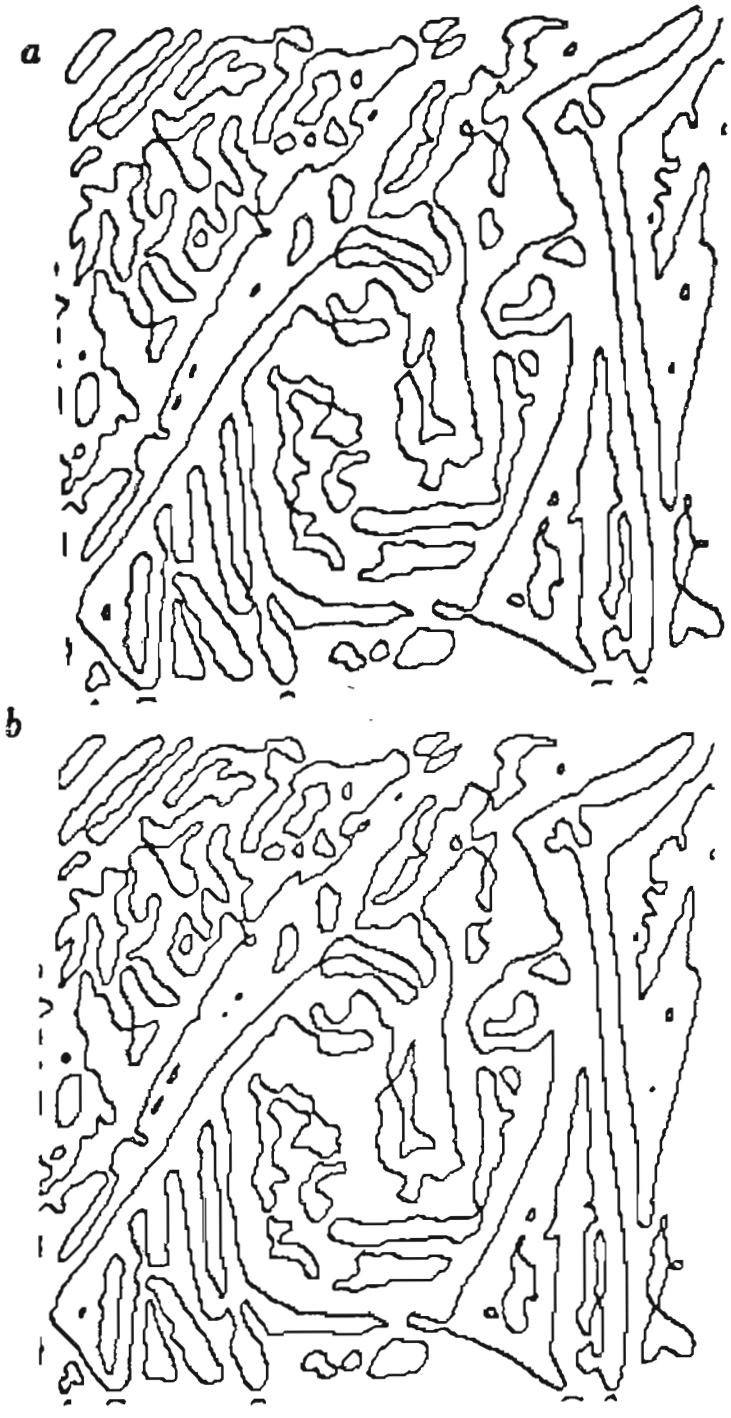


Fig. 14. Comparison of operators
(a) $\nabla^2 G$ operator with ω equal to 9 pixels.
(b) DOG operator with ω equal to 9 pixels.
(1 pixel = 50″ at 4.5 feet).

ACUITY AND THE FIFTH CHANNEL

An implementation of the raw primal sketch has been described. This intermediate representation is a rich set of primitive symbols that mark local discontinuities in the intensity array. It has been suggested that a set of zero-crossings at different spatial scales is used to extract the size, orientation, and positions of visual objects. We have also described the existence of spatially-tuned channels in the human visual system and used this evidence to set the size of the Laplacian of a Gaussian operator. The size of these operators bound the spatial frequency response of the zero-crossing encoding algorithm. Sensitivity to high spatial frequency in humans sets the limit of visual acuity along with other physical factors. These factors include the optics of the eye and the spacing between the retinal cones. If our model is an emulation of human visual processes then it should be consistent with human acuity measurements. Shortly, we will see that this is not the case and we will use results of Marr and others to overcome this weakness in the model.

Fine visual acuity is dependent upon the size of the smallest spatial channel. The smallest channel described by Wilson is the N channel. This channel has a size of $4.4'$ of arc in the fovea. Marr *et al.* have shown that the N channel is too large to explain fine visual acuity [MaHP79]. In order, to

understand these results we first must examine human acuity measurements.

An example of measuring the upper bound of normal acuity in humans is the two-point acuity test [WiEE84]. A subject views two dots from a fixed distance. The dots are moved closer and closer to each other until the subject cannot resolve them as distinct. Then the distance between the dots is measured. Humans can distinguish between dots with as little as $1'$ of arc of separation. Relative positions of dots can be resolved to less than $1'$ of arc. This performance is called hyperacuity. A typical test for hyperacuity is the three-dot test [WaAn82]. A subject views three dots from a fixed distance. The dots are aligned either horizontally or vertically. The object of the test is for the subject to determine when the dots are out of alignment. The outer two dots remain fixed, while the middle dot is moved. With practice, subjects can perceive misalignment with deflections of the middle dot of only $2''$ to $5''$.

Human acuity testing identifies two independent tasks that describe the localization power of the visual system. From an information processing point of view, two different algorithms might be used to perform these tasks, even though both algorithms might solve portions of the acuity problem with shared resources. Also, if different algorithms exist, then the algorithms may operate upon different representations of intensity information. We shall see that with the addition of a smaller spatial frequency channel the zero-crossing representation can explain perceived fine visual acuity, but is inadequate to explain hyperacuity.

Marr, Hildreth and Poggio have proposed a smaller channel which has a central diameter of 1'20" [MaHP79]. The diameter of the proposed channel is sufficiently small to the detect separations of 1' found in fine acuity. Research on human spatial channels does not preclude the existence of a smaller channel. Channels of such a small size have not been investigated. It has been suggested that such a channel could be constructed from a midget ganglion cell projecting to a single cone [MaHP79]. The implications of adding this new channel to our previous repertoire of four is that the model now conforms to measured human acuity. It also implies that objects or features of objects which subtend smaller than 1.3' zero-crossings will not be detected as we would expect, since they are beyond the resolution of our model.

Addition of the smaller channel cannot explain hyperacuity The proposed fifth channel, with a diameter of 1.3' of arc, is still too large to detect changes of 2" to 5" of arc. An even smaller channel is not a plausible solution to hyperacuity, since we have reached the size limitation of a single cone. A single cone is approximately 25" of arc in width. Therefore, small changes in position are not detected absolutely, but might be interpolated from raw convolution values. Such a mechanism has been suggested but has not been verified [CrMP80]. If interpolation is used to detect very small changes in an object's position from zero-crossing input, then hyperacuity most likely is performed higher up the visual pathway. This suggests a reconstructed version of raw convolution data is made available for scrutiny by a group of yet unknown operators. The function of such operators could be to extract small features or

changes in features that which have been lost due the limited size of the smallest channel.

Obviously, this interpolation and reconstruction mechanism is an added computational expense. However, if hyperacuity is viewed as a demand-driven process this expense may be warranted. Hyperacuity is a perceptual skill and may be learned. Early on, the visual system may learn that certain methods of interpolating data from the zero-crossing representation gives a reliable mechanism for positioning objects close to one another. We can consider this hyperacuity representation as a bit map of reconstructed convolution values from the zero-crossing representation. As previously mentioned, layer $4C\beta$ of the striate cortex could be the site of such a reconstruction process.

The construction of a hyperacuity bit map representation is a sampling problem. The discussion of sampling in chapter two provides a theory to deal with this problem but this theory is incomplete. It is important to verify that the zero-crossings alone are sufficient to completely reconstruct the convolved intensity information. If the convolved image can be reconstructed without loss of information, all of the original details are present and available for scrutiny by vision analysis algorithms. A theorem by B.F. Logan has shown that a one-dimensional signal can be reconstructed from its zero-crossings [CrMaP80]. This theorem provides a theoretical basis for the preservation of information by zero-crossings.

Logan's Theorem

> If a one-dimensional analytic function (a) is a bandpass of bandwidth one octave or less, and (b) has no complex zeroes in common with its Hilbert transform, then the function is completely determined, up to an overall multiplicative constant, by its real zero-crossings.

Logan's Theorem is limited to one-dimensional signals. A two-dimensional result is needed for our purposes. Mathematicians have not been able to use Logan's method of proof to extend his theorem to two dimensions, but this is an active area of research [CrMaP80]. The conditions of Logan's theorem constrain the method of computation of zero-crossings. The second condition can be ignored since it can be shown to hold only for pathological signals [MaU179]. In general, condition (b) will be satisfied by all visual signals. The first condition is the most interesting. It determines the size of the filters used to construct zero-crossings. Therefore, the Laplacian of a Gaussian filter must have a spatial frequency bandwidth of no larger than one octave. Using Logan's result it can be speculated that raw convolution values can be reconstructed by zero-crossings alone, if filtered with enough differently scaled operators, where each operator has a bandwidth of one octave or less.

Unfortunately, physiologic measurement of spatial channels in the human retina do not meet the ideal one octave bandwidth condition. They are considered to be about an octave and a half. The values of our operators have

been chosen to approximate this scale. Marr argues that the human visual system does not seem to lose visual information because Logan's bandwidth condition is not exactly met. So, it is reasonable to assume that Logan's Theorem can be relaxed, or that another mechanism is involved in completing the representation. Both are probably true [Marr82]. We can relax the size constraints of the operators. In fact, at one and a half octaves the failure rate of image reconstruction is about 8% [CrMP80]. Another mechanism which may provide additional intensity-change information is the gradient calculated at the zero-crossing. This vector supplies information about the contrast, width, and spatial orientation of the intensity change, and this may be sufficient to properly reconstruct the convolved image [Hild80]. Therefore, gradient information maybe be packaged along with zero-crossing symbols and used in layer $4C\beta$ of the striate cortex to construct a hyperacuity bit map representation.

We have suggested mechanisms and modifications to the original model that may be able to account for both fine visual acuity and hyperacuity measurements in humans. We have introduced a fifth visual channel with a central diameter small enough to take care of fine visual acuity. We have also speculated that a bit map representation of convolution values may be available for scrutiny by other operators and this may account for hyperacuity. In the next chapter we will attempt to validate the existence of zero-crossing representations and evaluate these representations near and beyond the limits of fine acuity. When image features are smaller than the diameter of the smallest channel artifacts of the convolution process should appear. If zero-crossing artifacts

are found and zero-crossings are the representation used by the human visual system then subjects should see these artifacts.

ZERO-CROSSING ARTIFACTS

The finest spatial scale of Marr's computational model is determined by the size of the fifth visual channel. This is the smallest channel carrying the convolution of the image with a $\nabla^2 G$ operator whose ω is $\mathbf{1.3}'$. If intensity changes occur in an image that subtend an angle smaller than ω of the smallest channel, it would be expected that this operator could not reliably detect these changes and any zero-crossing reported by an algorithm using this operator is an artifact. Rather than considering zero-crossing artifacts as unwanted information produced by $\nabla^2 G$ operators, we can use this property to explore the consistency of many aspects of Marr's model with the human visual system. In this chapter we will describe a computer experiment which locates zero-crossing artifacts. Then we will examine the perception of human subjects to see if similar artifacts are present.

Throughout this thesis, we have used the human visual system as a guideline for the implementation of the model. Therefore, as we begin the testing and verification stage of the implementation, we will evaluate the model against the performance of the human visual system. If we look at the perception of humans near the limits of visual acuity, zero-crossing artifacts should appear. Stevens [1985, personal communication] has shown that this prediction

holds with informal psychophysical experiments. In collaboration with Stevens, we have systematically explored these phenomena. The results of this investigation are that artifacts do appear to humans and the general shape of the artifacts are similar to the artifacts of the $\nabla^2 G$ operator found in computer experiments. The implication of these results are the following:

- A zero-crossing representation may be used by the human visual system.
- Spatial channels and $\nabla^2 G$-like operators are used to derive a zero-crossing representation.
- The size of the smallest channel in the visual system is roughly the size of the proposed fifth channel.
- Zero-crossing artifacts are present in the human visual system.
- A zero-crossing representation may be the input to higher-order vision processes (e.g., hyperacuity).
- Marr's model is consistent with some of the perceptual experiences of human subjects.

Computer Experiment Method:

The experiment was conducted on a DEC VAX-11/780. The software was written by the author in the "C" programming language. This program conformed to the details of the implementation of Marr's model previously described. Images were displayed with a a Metheus Omega 440 graphics controller and a Tektronix 690SR monitor. An Imagen 8/300 laser printer was used for hard copy output. A $\nabla^2 G$ operator with a central excitatory diameter of 4 pixels

(i.e., ω of $1'40''$). was constructed to approximate Marr's smallest channel. It was assumed that one cone of diameter ($25''$) mapped to one pixel in this experiment. This $\nabla^2 G$ operator was convolved with three checkerboard patterns one pattern at a time. Each checkerboard was composed of 16x16 alternating black and white squares. The size of the individual squares were 4 pixels ($1.6'$), 2 pixels ($50''$), and 1 pixel ($25''$). The checkerboards were set in a intermediate gray background (see Fig. 15(1-3a)). Zero-crossings and the gradient across the zero-crossings were computed.

Computer Experiment Results:

When the size of a square was approximately equal to the width of the operator, the gradient at the zero-crossings at the intersection of squares fell off steeply, and zero-crossing artifacts near the border squares were noticed. These artifacts consisted of elongation of border squares and rounding of corner squares. The border squares were approximately $50''$ larger than an expected square and rounded at the ends. The corner squares bloomed by roughly $25''$. (see Fig 15(1c)). As the size of the squares became smaller than the width of the operator, the artifacts of convolution were elongation of the squares at the borders and rounding of corners. The border squares were $100''$ larger and the corner square bloomed $50''$. (see Fig. 15(2b)). As the square-size approached the smallest value, the zero-crossings in the interior of the checkerboard were no longer marked. However, the corners were still marked as small circles. The diameters of the circles was roughly $75''$. (see Fig. 15(3b)).

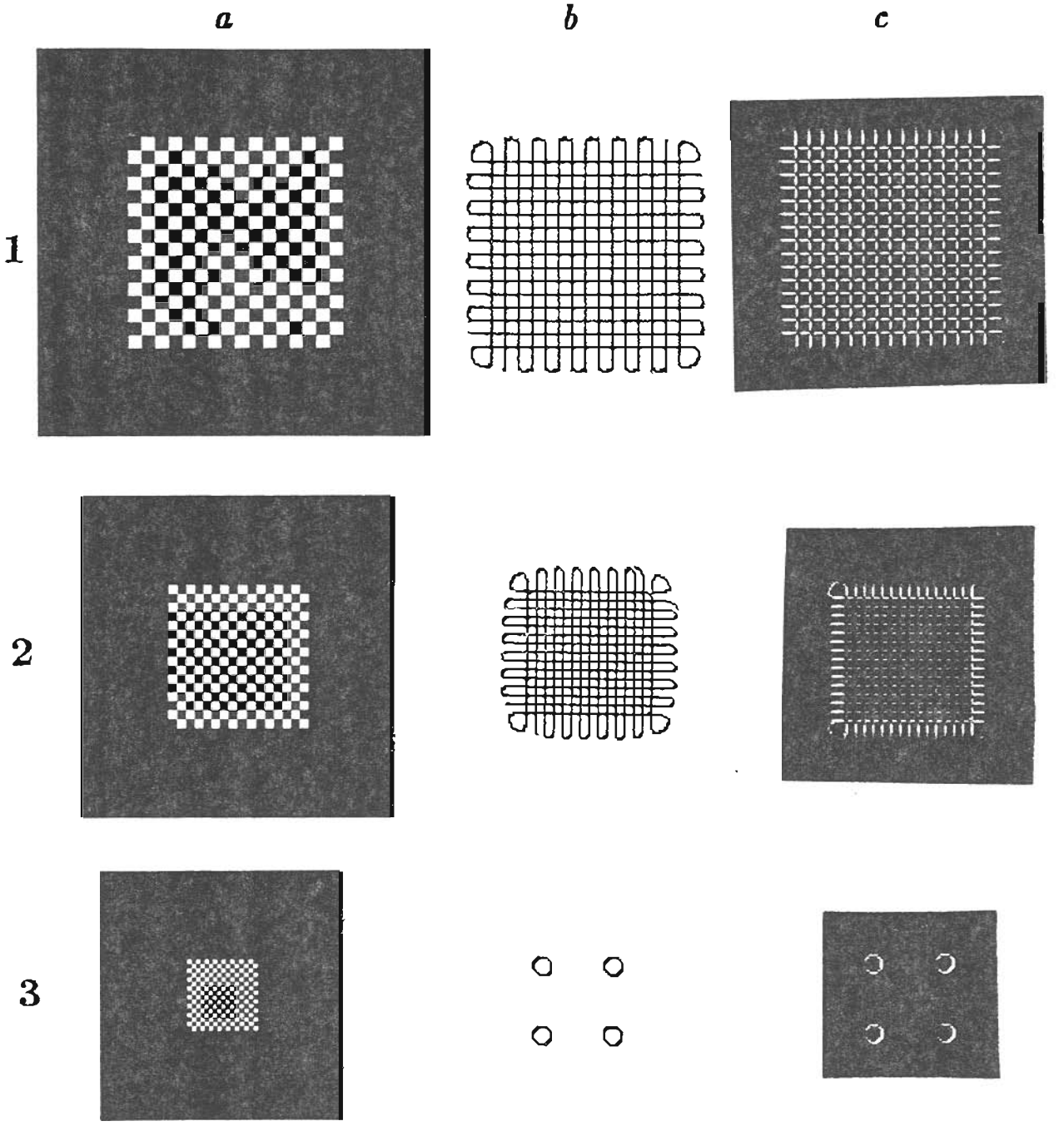


Fig. 15. Results of computer experiment.
(1a) Checkerboard 4 pixels. (1b) Zero-crossings. (1c) Gradient at zero-crossings.
(2a) Checkerboard 2 pixels. (2b) Zero-crossings. (2c) Gradient at zero-crossings.
(3a) Checkerboard 1 pixels. (3b) Zero-crossings. (3c) Gradient at zero-crossings.

Computer Experiment Discussion:

The $\nabla^2 G$ operator will not detect intensity changes reliably if the intensity changes occur over a region smaller than the width of the operator. If this situation occurs then zero-crossing artifacts can be expected. These artifacts consist of gradient changes in interior squares when the size of the operator is near the size of the square. Border squares become progressively longer and rounded as the size of squares fall smaller than ω. The corner squares become progressively rounder and bloom until this artifact is all that remains. Why do the borders seem most affected while the interior zero-crossing contours result in the expected grid?

In order to understand this phenomena, it is necessary to understand what Marr calls the Condition of Linear Variation [Hild80]. If the condition of Linear Variation holds then zero-crossings align well with the edge where an intensity change occurs. This Condition is an assumption about the local intensity change that states the intensity change near and parallel to the line of zero-crossings should be locally linear. At the borders, the linear variation condition does not hold (relative to the size of the operator). It is important to remember that these operators are sensitive over a region of the intensity array much larger than ω, since the actual diameter of the whole operator is about 3.5ω. It is probably the symmetry in the interior of the checkerboard that guarantees the grid-like output down a certain scale. Therefore, it is the size of the operator relative to the squares and the failure of linear variation

within the scope of operator that generate zero-crossing artifacts.

Hildreth has reported that channels larger than the square size are sensitive to artifacts, and she used this phenomena in an attempt to explain the perception of squares being organized along the diagonal. She found that output of the zero-crossing algorithm for squares, which subtended $\mathbf{3}'$, by a channel with a ω of about two times the square size, resulted in a zero-crossing map that looked like a herringbone pattern [Hild80]. Stevens psychophysical data predicts elongation of border squares, corner rounding and disappearance of interior zero-crossings when the smallest channel operator is used and square sizes varied from $\mathbf{2}'$ to $\mathbf{.8}'$ [1985, personal communication].

Perceptual Experiment Method:

A checkerboard pattern set in a intermediate gray background was placed at viewing height on a wall, under approximately normal lighting conditions. The checkerboard was computer-generated and exact in all details (except size) to the pattern used in the computer experiment. The size of individual squares were .0247x.0247 inches. A calibrated rule was used to mark distance on the floor. 10 human subjects with a mean age of 30 years participated. The subjects were asked whether they considered their vision to be normal. The experimental results of those who responded negatively were thrown out. All subjects were instructed, and allowed to practice if they desired to do so. The subjects were told that artifacts would be seen, and that it was their task to find the distance from the stimulus where the artifacts appeared and disappeared.

The effect of stepping away from the stimulus is analogous to processing a smaller sized version of the pattern, as was done in the computer experiment. This measurement was logged.

Perceptual Experiment Results:

subject	diagonals	elongated	gray	gone
1	3.2'	1.9'	1.3'	1.04'
2	4.2'	3.1'	1.7'	1.04'
3	2.9'	1.9'	1.3'	0.85'
4	3.8'	1.8'	1.4'	0.93'
5	2.2'	1.5'	1.1'	0.93'
6	3.9'	1.6'	1.0'	0.85'
7	3.1'	1.8'	1.6'	0.94'
8	2.9'	1.8'	1.2'	0.94'
Mean	3.3'	1.9'	1.3'	0.94'
S.D.	0.6'	0.5'	0.2'	0.07'

The four convolution artifacts predicted by the computer experiment were recognized by all subjects. These artifacts are: the grouping of squares on the diagonals, elongation of border squares, the disappearance of interior squares, and the disappearance of corner features. The numerical entries in the table are the degrees of arc an individual square subtends on the retina. Mean and standard deviation values are shown.

Perceptual Experiment Discussion:

In this experiment, many variables such as illumination, precision of distance measure, and contrast of stimulus were not adequately controlled. In spite of this, the results are significant. It may be argued that the subjects were

coached into recognizing artifacts. This does not seem to be a real problem because we were searching for evidence of fine curvature in the artifacts. The subjects were informed of the presence of artifacts, but independent of any coaching, reported the attributes and features of the artifacts. All subjects reported artifacts consisting of curved features and squares that were organized along diagonals. These results are consistent with observations made by Hilderth and Stevens.

Many of the observed artifacts were rounded. The radius of curvature of rounded zero-crossing artifacts were measured and varied from $6''$ to less than $3''$. These figures place the curvature of rounded artifacts in the hyperacuity range. All subjects reported rounded features at this scale suggesting that zero-crossings may be available for scrutiny by a hyperacuity mechanism.

With the above data and the observations of other vision scientists we feel that this set of experiments has shown: 1) that a zero-crossing representation may be used by the human visual system, 2) that spatial channels and $\nabla^2 G$ operators are the constructors of this representation, 3) that the size of the smallest channel is approximately $1.3'$, and 3) that tasks requiring discrimination of shape at the limit of resolution likely have have a zero-crossing map available for scrutiny. All of the above findings give evidence that Marr's zero-crossing model of early vision is consistent with many perceptual experiences of humans.

SUMMARY AND FURTHER RESEARCH

An information processing approach to computer vision based on work previously done by Marr has been implemented and evaluated [Marr82]. Our implementation focused on early vision, which consisted of sampling intensity data and the construction of zero-crossings.

Sampling was shown to be a demanding computation. If sampling is not performed correctly, errors can occur. Through the use of the Sampling Theorem, we could guarantee that the original intensity data could be reconstructed from the sampled intensity array. Therefore, the sampling operation is a method of data compaction without the loss of information. By examining the physiology of the human visual system, it was suggested that reconstruction properties laid out in the Sampling Theorem apply to the human visual system. The algorithm used by our implementation was roughly equivalent to sampling properties of cones in the fovea.

After sampling, Marr constructs zero-crossing primitives. Zero-crossings mark significant intensity changes in the intensity array at different spatial scales. The Laplacian of a Gaussian operator is used to detect the zero-crossings in all orientations. Zero-crossings are detected by convolving four $\nabla^2 G$ operators of different sizes with the intensity array. This completes

Marr's raw primal sketch. The sizes of the operators were chosen based on the size of measured spatially tuned channels in the human visual system. It was shown that a computational speedup could be achieved by replacing $\nabla^2 G$ with an approximately equivalent operator the Difference of Gaussians (DOG). The time complexity could be reduce from $O(N^2)$ to $O(4N)$.

The high visual acuity of the human visual system cannot be emulated by the smallest of the four $\nabla^2 G$ operators used in this thesis. Therefore, Marr *et al.* proposed a smaller operator whose is sized to be sensitive to zero-crossings at the limit of resolution of a single cone. This operator was implemented and tested. The following question was raised, "What happens to zero-crossing contours when the size of the intensity change is smaller than the width of the smallest operator?" A computer experiment was performed that showed zero-crossing artifacts were the result. In this experiment a checkerboard pattern was convolved with $\nabla^2 G$ function scaled to correspond to the smallest operator. Artifacts were located and their sizes and shapes determined. These artifacts consisted of elongation and rounding of border and corner squares, and the disappearance of interior zero-crossings at the limits of resolution.

We then asked, "If artifacts at the limits of fine spatial resolution appear in the computer model and if the computer model is an emulation of the human visual system, would humans also perceive such artifacts?". Human subjects were tested at the limits of acuity and reported artifacts of similar size and shape. The measurements of viewing distance and artifact sizes were

consistent with results of the computer experiment and the observations of Hildreth and Stevens. The implications of these results are the following:

- Zero-crossing artifacts are perceived by humans.
- Spatial channels and $\nabla^2 G$-like operators are used by the human visual system to derive zero-crossings.
- The size of the smallest channel in the visual system is roughly the size of a single cone.
- The radius of curvature of perceived artifacts is small enough to suggest that zero-crossings may be available for scrutiny by a hyperacuity mechanism.
- Marr's model is consistent with the perceptual experiences of human subjects at the limits of resolution when checkerboard patterns are used.

The author feels that this work not only provides insight into the functionality of second derivative operators near their limit of resolution, but also suggests that this model may be used to predict the behavior of the human visual system at that limit.

Further research into the problems of acuity and hyperacuity should be carried out. A simple approach to examining the limits of resolution of the visual system relative to the optimum size of operators is to give the computational model the two- and three-point acuity test. This experiment would provide important computer data to compare with the wealth of psychophysical data available. It would be interesting to see if these results would be consistent with the findings here.

Recent perceptual research on acuity by Watt and Morgan suggests an alternate set of primitives that could be used in addition to zero-crossings. These primitives are constructed from a second derivative operator and are the local maximum (i.e., peak) and minimum (i.e., trough) near the intensity change. These primitives can be computed with little change to the computational model and may be the precursors of the zero-crossing computation in the human visual system.

The nervous system encodes inhibition and excitation as different discharge rates of individual neurons. The output of the inhibitory surround of a spatial channel is encoded as the trough, and the output of the excitatory center of the channel is encoded as the peak. Watt and Morgan argue that the zero-crossing can be computed by looking halfway between these two peaks. Also, the slope across the zero-crossing is easily determined using the peak and trough.

The researchers claim that peak and trough information is necessary to explain the ability of human subjects to discriminate the extent of blurry edges with a high degree of precision [WaMo83]. A blurry edge will lie between the peak and trough. Unfortunately, this model of primitives was developed with one-dimensional tests and ideas. Extending the peak and trough primitives to two dimensions may produce interesting problems, for instance, how is the case handled where the peaks and troughs from several intensity changes overlap in two dimensions? Nevertheless these different primitives are worth investigation.

Simple experiments can be designed to do comparison testing between zero-crossings and the peak and trough counterparts. It can be easily determined by marking an image with peaks and troughs whether significant intensity changes are detected. Also, the computer model of the peak and trough primitives will quickly expose any problems that may occur in two dimensions. It is a natural extension of this work to seek other possible symbolic primitives that may be involved in early vision. We have given arguments here for peak and trough primitives, but more primatives may be discovered.

REFERENCES

[Baxe84]

Baxes, G.A., Digital Image Processing: A Practical Primer, Prentice-Hall, 1984.

[Brac65]

Bracewell, R., The Fourier Transform and Its Applications, McGraw-Hill, 1965.

[Brad82]

Brady M., "Computational Approaches to Image Understanding," *Computing Surveys*, vol. 14, pp. 3-71, 1982.

[Barl78]

Barlow H.B., "The efficiency of detecting changes of density in random dot patterns," *Vision Research*, vol. 18, pp. 637-650, 1978.

[CrMP80]

Crick, H.C., Marr, D., and Poggio, T., "An information processing approach to understanding the visual cortex," AI Tech. Report 557, Artificial Intelligence Lab., MIT, Cambridge, Mass., 1980.

[CrPa84]

Crowley, J.L., and Parker, A.C., "A representation for shape based on peaks and ridges in the difference of low pass transform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, pp. 156-171, Mar. 1984.

[Greg73]

Gregory, R.L., Eye and Brain: the Psychology of Seeing, McGraw-Hill, 1973.

[Grim80]

Grimmson, W.E.L., "A computer implementation of a theory of human stereo vision," AI Tech. Report 565, Artificial Intelligence Lab., MIT, Cambridge, Mass., 1980.

[Grim85]

Grimmson, W.E.L., "Computational Experiments with a feature based stereo algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-7, no. 1, pp. 17-34, Jan. 1985.

[GrHi85]

Grimmson, W.E.L., and Hildreth, E., "Comments on "Digital step edges from zero-crossings of second directional derivatives"," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-7, no. 1, pp. 121-126, Jan. 1985.

[Hamm77]

Hamming, R.W., Digital Filters, Prentice-Hall, 1977.

[Hild80]

Hildreth, E.C, "Implementation of a theory of edge detection," AI Tech. Report 579, Artificial Intelligence Lab., MIT, Cambridge, Mass., 1980.

[JuSc81]

Julesz, B., and Schumer, R.A., "Early visual perception," *Ann. Rev. Psychol.*, vol. 35, pp. 575-627, 1981.

[Kung84]

Kung, H.T., "Systolic algorithms for the CMU Warp Processor," in Proceedings of the Seventh International Conference on Pattern Recognition, International Association for Pattern Recognition, pp. 570-577, 1984

[MaHi80]

Marr, D., and Hildreth, E., "Theory of edge detection," *Proc. Roy. Soc. London*, ser. B, vol. 207, pp. 187-217, 1980.

[MaHP78]

Marr, D., Hildreth, E., and Poggio, T., "Bandpass channels, zero-crossings, and early visual information processing," AI Memo 491, Artificial Intelligence Lab., MIT, Cambridge, Mass., 1978.

[MaHP79]

Marr, D., Hildreth, E., and Poggio, T., "Evidence for a fifth, smaller channel in early human vision," AI Memo 451, Artificial Intelligence Lab., MIT, Cambridge, Mass., 1979.

[MaPH80]

Marr, D., Poggio, T., and Hildeth E., "Smallest channel in early human vision," *J. Opt. Soc. Am.*, vol. 70, pp. 860-870, 1980.

[MaPg76]

Marr, D., and Poggio, T., "From understanding computation to understanding neural circuitry," AI Memo 357, Artificial Intelligence Lab., MIT, Cambridge, Mass., 1976.

[MaPo76]

Marr, D., and Poggio, T., "Cooperative computation of stereo disparity," *Science*, vol. 194, pp. 283-287, 1976.

[MaPo77]

Marr, D., and Poggio, T., "A theory of human stereo vision," AI Memo 451, Artificial Intelligence Lab., MIT, Cambridge, Mass., 1977.

[Marr76]

Marr, D., "Early processing of visual information," *Phil. Trans. Roy. Soc.*, B 275, 483-524, 1976.

[Marr82]

Marr, D., Vision, W.H. Freeman, 1982.

[MaUl79]

Marr, D., and Ullman S., "Bandpass channels, zero-crossings, and early visual information processing," *J. Opt. Soc. Am.*, vol. 69, pp. 914-916, 1979.

[MaUll79]

Marr, D., and Ullman, S., "Directional selectivity and its uses in early visual processing," AI Memo 524, Artificial Intelligence Lab., MIT, Cambridge, Mass., 1979.

[Nuss78]

Naussbaumer, H.J., "New algorithms for convolution and DFT based on polynomial transforms," in IEEE 1978 Intern. Conf. Acoust., Speech, Signal Proc., pp 638-641, 1978.

[OpWi83]

Oppenheim, A.V., Willsky, A.S., and Young I.T., Signals and Systems, Prentice-Hall, 1983.

[Pogg84]

Poggio, T., "Vision by man and machine," *Scientific American*, vol. 250, no. 4, pp. 106-117, 1984.

[RiND82]

Richards, W., Nishihara, H.K., and Dawson, D., "Cartoon: a biologically motivated edge detection algorithm," AI Memo 668, Artificial Intelligence Lab., MIT, Cambridge, Mass., 1982.

[RiMa81]

Richards, W., and Marr, D., "Computational algorithms for visual processing," Final Report NSF-77-07569, Artificial Intelligence Lab., MIT, Cambridge, Mass., 1981.

[WaAn82]

Watt, R.J., and Andrews, D.P., "Contour curvature analysis: hyperacuities in the discrimination of detailed shape," *Vision Res.*, vol. 22, pp. 449-460, 1982.

[WaMo83]

Watt, R.J., and Morgan, M.J., "The recognition and representation of edge blur: evidence for spatial primitives in human vision," *Vision Res.*, vol. 23, pp. 1465-1477, 1983.

[WaMo84]

Watt, R.J., "Further evidence concerning the analysis of curvature in human foveal vision" *Vision Res.*, vol. 24, pp. 251-253, 1984.

[WiEE84]

Williams, R.A., Enoch, J.M., and Essock, E.A., "The resistance of selected hyperacuity configurations to retinal image degradation," *Invest. Ophthalmol. Vi. Sci.*, vol. 25, pp. 389-399, 1984.

[WiBe79]

Wilson, H.R., and Bergen, J.R., " A Four Mechanism Model For Threshold Spatial Vision," *Vision Res.*, vol. 19, pp. 19-32, 1979.

[WiGi77]

Wilson, H.R., and Giese, S.A., " Threshold visibility of frequency gradient patterns," *Vision Res.*, vol. 17, pp. 1177-1190, 1977.

[Winst84]

Winston, P.A., Artificial Intelligence, Addison-Wesley, 1984.

APPENDIX A: CONVOLUTION

The convolution operation (also called the composition product, the running mean, smoothing, blurring or smearing) describes how to take a weighted mean over a narrow range of a function. In image processing, convolution is used to apply a digital filter to an input image. If a Gaussian filter is used, the results of convolution are smoothing or smearing of the input image [Braxe83]. The filter is also called the kernel, operator, or mask. A computational vision processor calculates the composition product of a two-dimensional kernel and the sampled reflectance data [Grim85]. A detailed explanation of the methods presented here appears in Bracewell's book [Brac65].

In two dimensions, the continuous form of the convolution integral for functions $f(x,y)$ and $g(x,y)$ can be written:

$$f * g = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x',y')\, g(x-x',y-y')\, dx'dy' \qquad (1a)$$

where the $*$ denotes convolution. Since the reflectance data $r_s(x,y)$ is discrete after sampling and this function is convolved with a discrete impulse response filter $f_s(x,y)$, the discrete form of the convolution integral is used:

$$r_s * f_s = \sum_{x'=-\infty}^{\infty} \sum_{y'=-\infty}^{\infty} r_s(x',y')\, f_s(x-x',y-y') \qquad (2a)$$

Calculating the convolution of two functions is simple when performed as serial products. An example of how to hand-calculate a one-dimensional convolution with this method is as follows: Each function will be represented as a sequence of integers, function $A=\{1,4,5,3\}$ and function $B=\{1,3,3\}$. A is written in a column and B is written on a strip of paper in ***reverse*** order (see Fig. A1). Function B will slide downward next to column A during the calculation. At each step an element in B is multiplied with the corresponding element in A and the results of the multiplications are summed and written down. If B does not have a corresponding element in A, multiply by zero. Next, slide the strip of paper one element down and repeat the multiplication and summation process. Continue sliding the strip of paper until column B has slipped past column A. See (Fig. A1(b-h)) to trace this example.

The method of serial products extends naturally to two dimensions. When calculating a two-dimensional convolution in a vision processor, the filter is centered over each point in the intensity array. All of the values which lie under the filter are multiplied and summed with the corresponding filter values. Next, the mask is moved over a neighboring point in the intensity array and the computation repeated. The mask continues to slide until every value of the intensity array is visited once. Since all of the convolution kernels used in this paper are circularly symmetric, reversing the kernel is not necessary.

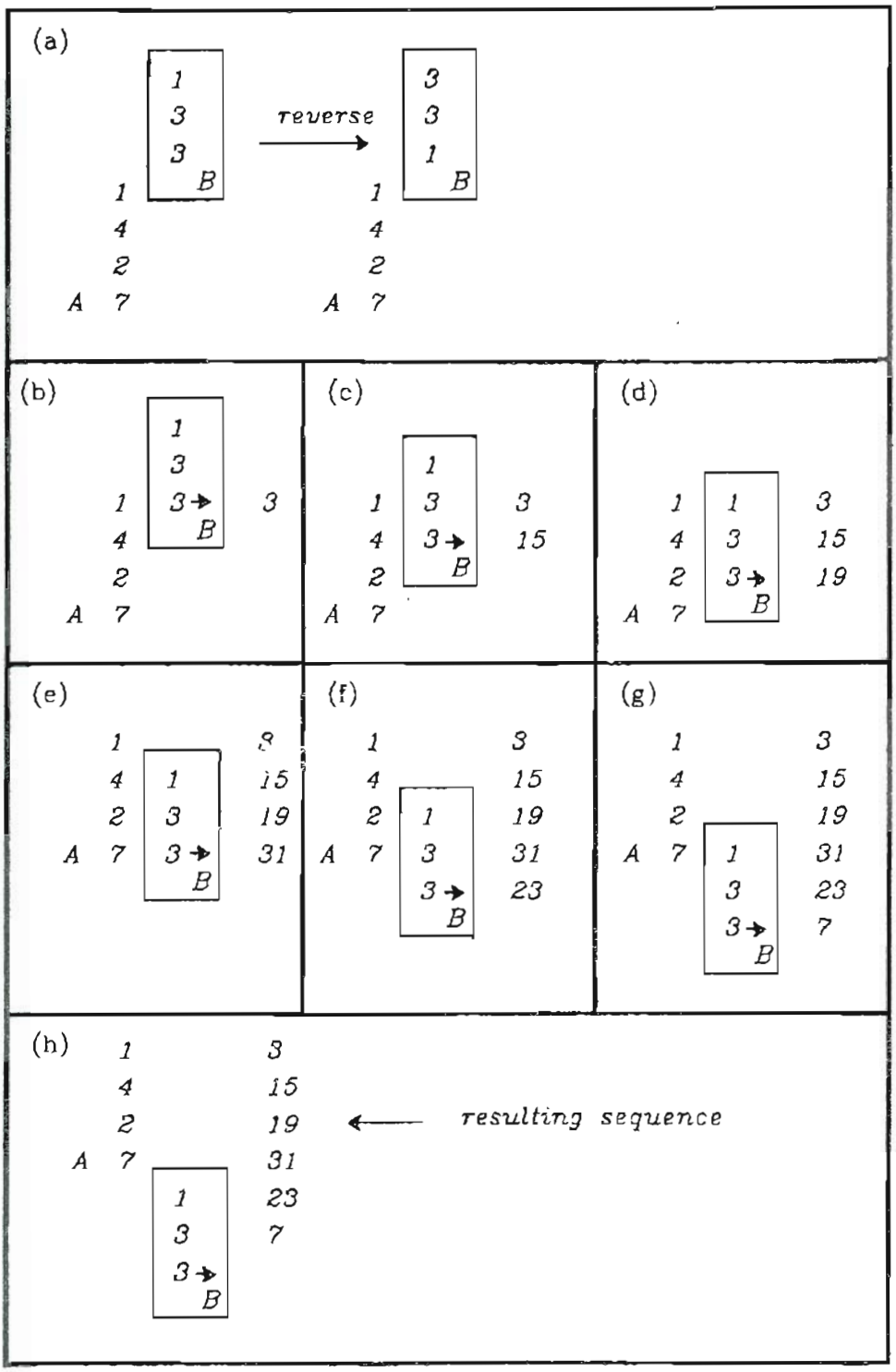


Fig. A1. Hand trace of serial products method

The time complexity of the two-dimensional serial products method is dependent upon the size of both the intensity and filter arrays. Since the size of the kernel *(M)* is usually much smaller than the size of the image being convolved *(N)*, the algorithm is effectively $\mathbf{O}(n^2)$ [CrPa84]. More elegant convolution methods exist. These include taking advantage of the convolution theorem, which changes the convolution operation to multiplication in the frequency domain [Hamm77]. Before convolution degenerates to multiplication, the Fourier transform of both the filter and intensity arrays is taken. After multiplying the frequency domain values of the filter and intensity arrays together, the inverse Fourier transform is used to get back to the time domain. A time penalty is paid for taking the transform, but there are fast transformation algorithms available [OpWi83]. Also, methods of transforming convolution data into polynomials can take advantage of polynomial algebra to significantly reduce the number of multiplications and additions [Nuss78].

APPENDIX B: SEPARABILITY PROOF

The difference of Gaussians function (DOG) is computed by subtracting two Gaussian functions. This is a two-dimensional function, and when convolved with an image requires $\mathbf{O}(n^2)$ time. The Gaussian function is the only two-dimensional function which is both circularly symmetric and separable into one-dimensional components [CrPa84]. Since the difference of Gaussians function is built from two Gaussian functions, this function can be separated into four one-dimensional forms of the Gaussian. Each of these one-dimensional forms can be convolved with the image and the result will be equal to convolution with the two-dimensional form. This will require $4N$multiplications and additions per point in the image compared to N^2, where N is the number of values along one side of the mask.

The proof of the separability of Gaussians lies in the frequency domain. Therefore, to understand the proof it is important to visualize the functions in both the frequency and time domains (how to take a Fourier transform of a function will not be discussed). Almost all of the functions used in this proof look the same in both the time and frequency domains, so it is not painful to move back and forth. The reason for going to the frequency domain is that multiplication can be used rather than convolution (see the convolution

theorem [Brac65]).

The impulse function δ is used to sample the continuous two-dimensional form of the Gaussian. In this case we are assuming that we are taking infinitely many samples spaced at infinitely small distances. The intention is to maintain as closely as possible the continuous form of the function being sampled. Here sampling can be viewed as taking a slice of the function.

The difference of Gaussians function is given by:

$$DOG(x,y) = 1/\sigma_e\, e^{-(x^2+y^2)/2\sigma_e^2} - 1/\sigma_i\, e^{-(x^2+y^2)/2\sigma_i^2} \qquad (1b)$$

where the ratio of space constants $\sigma_e:\sigma_i$ determines the amplitude and width of the function. If the ratio of the space constants is 1.5, this function closely approximates the Mexican-hat-shaped Laplacian of Gaussian function [Hild80]. The proof of separability of the Gaussian function extends directly to the DOG.

The two-dimensional form of the Gaussian function is given by:

$$G(x,y) = e^{-(x^2+y^2)/2\sigma^2} \qquad (2b)$$

By the laws of exponents equation (2b) is equal to:

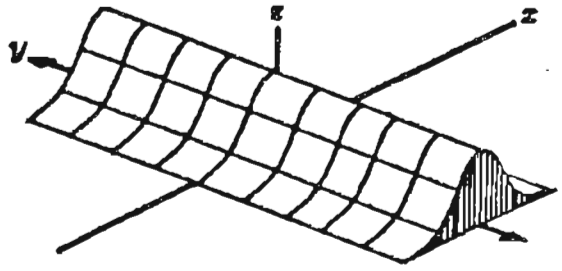
$$G(x,y) = e^{-x^2/2\sigma^2} \cdot\ e^{-y^2/2\sigma^2} \qquad (3b)$$

The right hand side of equation (3b) can be renamed:

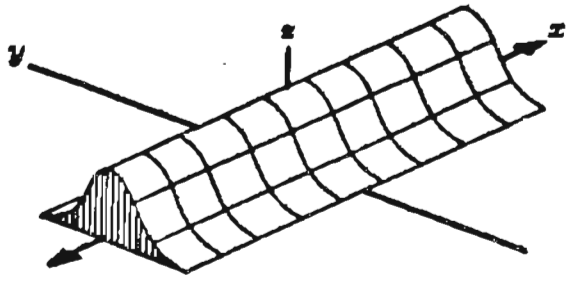
$$G(x,y) = G_x(x,y) \cdot\ G_y(x,y) \qquad (4b)$$

Figure (B1) shows the time domain plots of the terms of equation (4b). Two impulse functions are introduced. These functions are used to take a slice of the two-dimensional Gaussian tubes. This sampling process will create the

$G_y(x,y)$



$G_x(x,y)$



$G(x,y)$

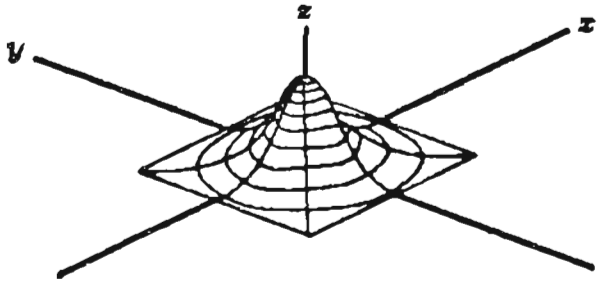


Fig. (B1). Time domain plots of the terms of equation (4b).

one-dimensional Gaussian cross sections. We will prove that convolution with these cross sections is equivalent to convolution with the two-dimensional function.

The impulse function that samples along the x axis is:

$$\delta_x(x,y) = \begin{cases} 1 & x = 0 \\ 0 & otherwise \end{cases} \tag{5b}$$

The impulse function that samples along the y axis is:

$$\delta_y(x,y) = \begin{cases} 1 & y = 0 \\ 0 & otherwise \end{cases} \tag{6b}$$

It is convenient to rename some of the functions given:

$$G_x(x,y) = G_x \tag{7b}$$

$$G_y(x,y) = G_y \tag{8b}$$

$$\delta_x(x,y) = \delta_x \tag{9b}$$

$$\delta_y(x,y) = \delta_y \tag{10b}$$

To prove that the convolution of the two-dimensional Gaussian function with an image is equal to the convolution of separated one-dimensional Gaussian forms of the original Gaussian (the Gaussian tube functions sampled with impulse functions) we must show the following:

$$G(x,y) * I(x,y) = (G_x \cdot \delta_y) * ((G_y \cdot \delta_x) * I(x,y)) \tag{11b}$$

where $I(x,y)$ is the image (see Fig. (B2)).

By the associativity of convolution, it is sufficient to show:

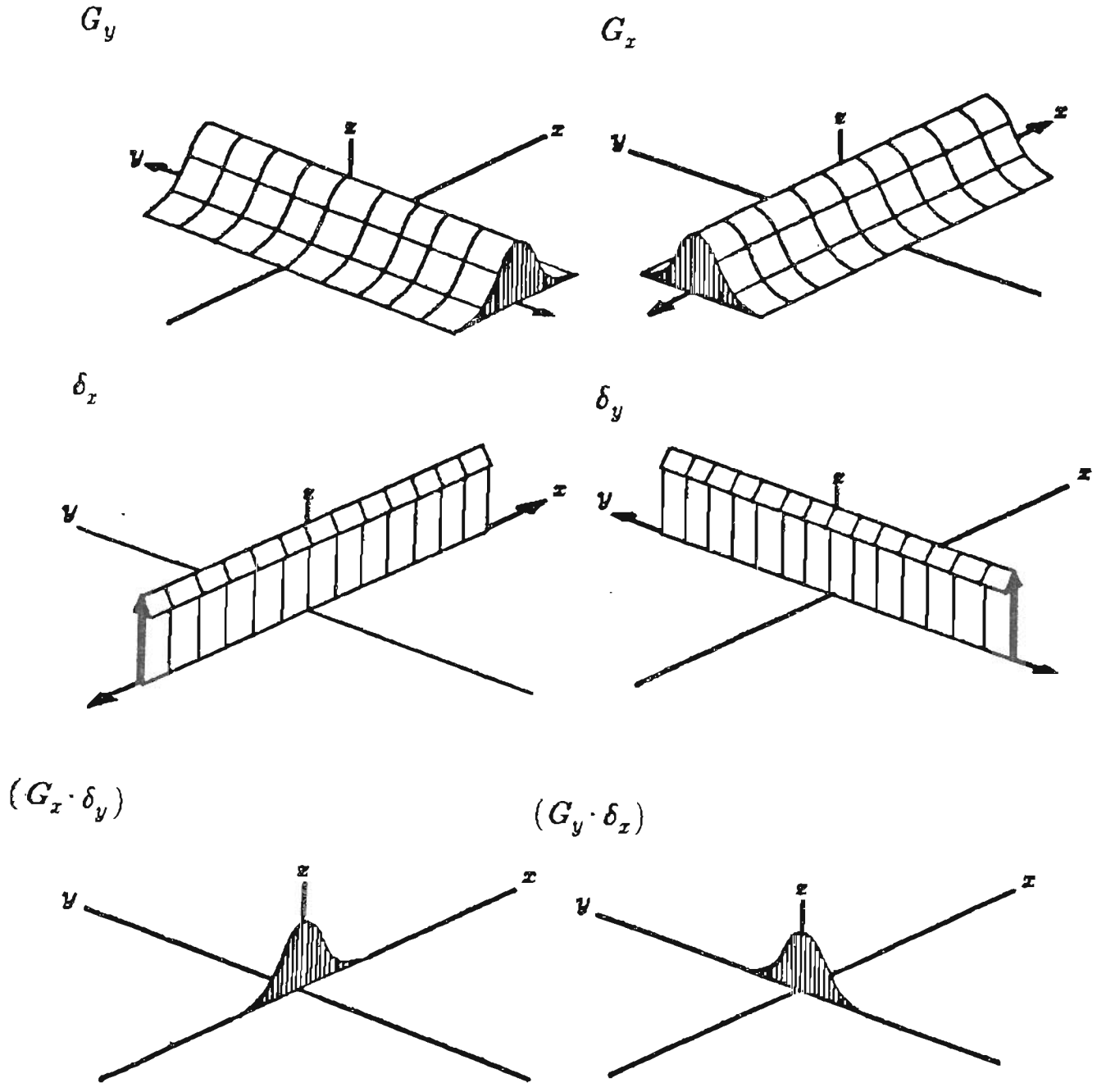


Fig. (B2). Time domain plots of sampling process.

$$G(x,y) = (G_x \cdot \delta_y) * (G_y \cdot \delta_x) \tag{12b}$$

Consider the Fourier transform of the right side of equation (12b):

$$\overline{(G_x \cdot \delta_y) * (G_y \cdot \delta_x)} \tag{13b}$$

By the convolution theorem, equation (13b) can be written as (see Fig. (B3)):

$$\overline{(G_x \cdot \delta_y)} \cdot \overline{(G_y \cdot \delta_x)} \tag{14b}$$

Since G_x depends only on x and δ_y depends only on y , the Fourier transform of their product is separable (similarly for G_y and δ_x) [p. 245 Brac65]. Therefore, equation (14b) can be rewritten as:

$$\overline{(G_x)} \cdot \overline{(\delta_y)} \cdot \overline{(G_y)} \cdot \overline{(\delta_x)} \tag{15b}$$

Regrouping terms gives (see Fig. (B4)):

$$\overline{(G_x)} \cdot \overline{(G_y)} \cdot \overline{(\delta_x)} \cdot \overline{(\delta_y)} \tag{16b}$$

Recombining separable products yields:

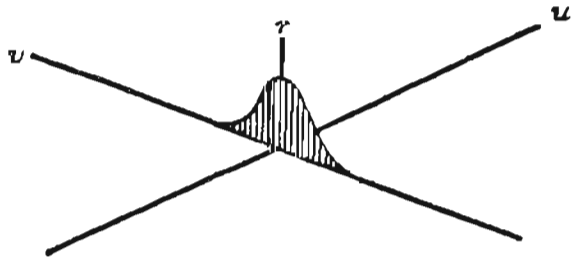
$$\overline{(G_x \cdot G_y)} \cdot \overline{(\delta_y \cdot \delta_x)} \tag{17b}$$

By the properties of exponentials and performing multiplication of the impulse functions, equation (17b) is rewritten (see Fig. (B5)):

$$\overline{G(x,y)} \cdot \overline{\delta(x,y)} \tag{18b}$$

Since $\overline{\delta(x,y)}$ is the unit plane, the equation (18b) is equivalent to:

$\overline{(G_x \cdot \delta_y)}$



$\overline{(G_y \cdot \delta_x)}$

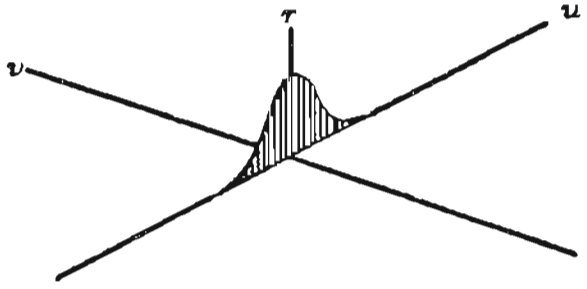
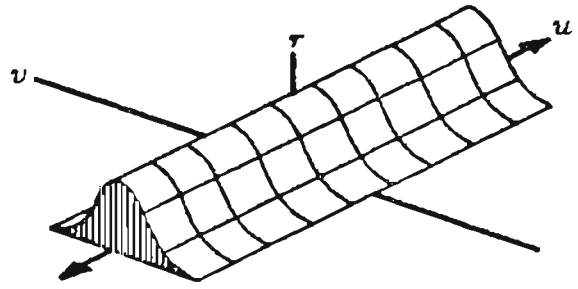
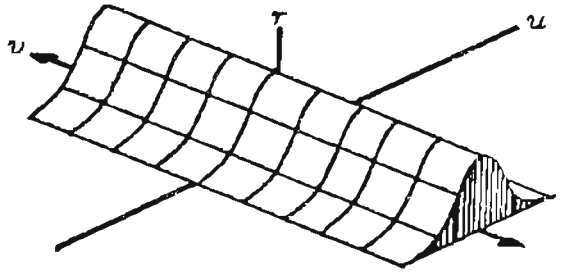


Fig. (B3). Frequency Domain Plots of terms of equation (14b).

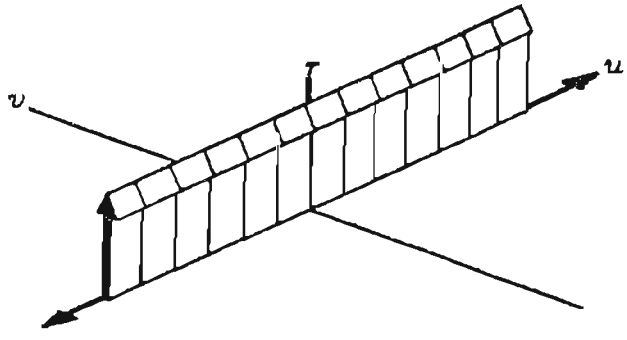
$\overline{(G_y)}$



$\overline{(G_x)}$



$\overline{(\delta_y)}$



$\overline{(\delta_x)}$

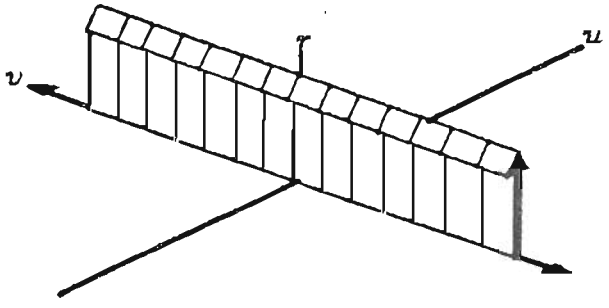


Fig. (B4). Frequency Domain Plots of terms of equation (16b).

$$\overline{G(x,y)} \tag{19b}$$

Taking the inverse Fourier transform yields:

$$G(x,y) \tag{20b}$$

It is now clear that:

$$G(x,y)=(G_x\cdot\delta_y)*(G_y\cdot\delta_x) \tag{21b}$$

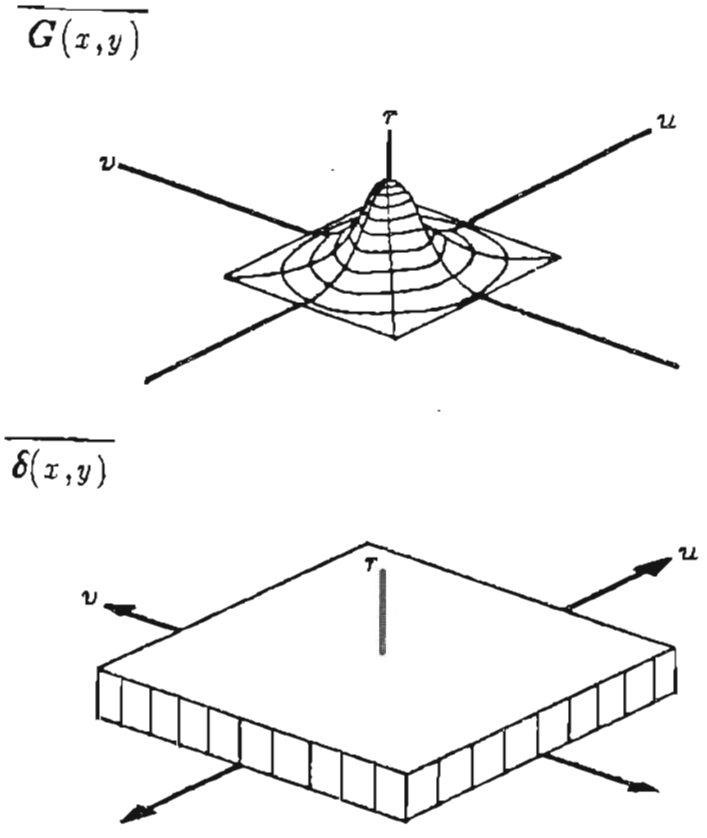


Fig. (B5). Frequency Domain Plots of terms of equation (18b).

Biographical Note

The author was born on August 9, 1953 in the northwest section of Portland, Oregon. He attended both grade school and high school in Portland. In 1971, he graduated from Jesuit High school and the following fall attended Portland State University. During his undergraduate career, the author studied Philosophy, Physiological Psychology, and Pre-Medicine.

In 1973 with a degree in Philosophy completed, the author accepted a fellowship to study neuroscience at the Neurological Sciences Institute of Good Samaritan Hospital and Medical Center. After the fellowship was completed the author was asked to remain at the Neurological Science Institute as a research assistant. After three years, the Good Samaritan Hospital offered the author a position as Director of the Clinical Laboratory of Visual Electrophysiology. This position was accepted and the author spent five years as director.

In 1981, the degree in Physiological Psychology was completed at Portland State University. At this time the author became interested in computer engineering. In the fall of 1983, full time graduate study was begun at the Oregon Graduate Center. During his study, several quarters were spent as tutor of the Computer Graphics Laboratory. In December of 1985, the requirements for the degree of Master of Science were completed. The author plans to pursue a doctoral degree in Computer Science.