



University of Pennsylvania
ScholarlyCommons

Technical Reports (CIS)

Department of Computer & Information Science

May 1984

Computational Models of Visual Hyperacuity

Eric Paul Krotkov
University of Pennsylvania

Follow this and additional works at: https://repository.upenn.edu/cis_reports

Recommended Citation

Eric Paul Krotkov, "Computational Models of Visual Hyperacuity", . May 1984.

University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-84-43.

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/cis_reports/726
For more information, please contact repository@pobox.upenn.edu.

Computational Models of Visual Hyperacuity

Abstract

The process of visual hyperacuity is described and analyzed in the terms of informative theory. It is shown that in principle, the detection and representation of both luminance and edge features can be performed with a precision commensurate with human abilities.

Algorithms are formulated in accord with the different representational methods, and are implemented as distinct computer models, which are tested with vernier acuity tasks. The results indicate that edge information, encoded either in the manner proposed by Marr and his colleagues (as zero-crossings in the Laplacian of a Gaussian convolved with the image) or when encoded as a simple filtered difference allows finer spatial localization than does the centroid of the intensity distribution.

In particular it is shown that to judge changes of relative positions with a precision of 0.1 sec arc in two and three dimensions, it is sufficient to represent the displacement of an edge by the difference of two Laplacian-Gaussian filters rather than by the difference between interpolated zero-crossings in them. This method entails no loss of relative position information (sign), allows recovery of the magnitude of the change, and provides significant economies of computation.

Comments

University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-84-43.

**COMPUTATIONAL MODELS OF
VISUAL HYPERACUITY**

**Eric Paul Krotkov
MS-CIS-84-43
GRASP LAB 15**

**Department Of Computer and Information Science
Moore School
University of Pennsylvania
Philadelphia, PA 19104**

May 1984

Acknowledgements: This research was supported in part by DARPA grants NOOO14-85-K-0018 and NOOO14-85-K-0807, NSF grants MCS8219196-CER, MCS-82-07294, 1 RO1-HL-29985-01, U.S. Army grants DAA6-29-84-K-0061, DAAB07-84-K-F077, U.S. Air Force grant 82-NM-299, AI Center grants NSF-MCS-83-05221, U.S. Army Research office grant ARO-DAA29-84-9-0027, Lord Corporation, RCA and Digital Equipment Corporation.

UNIVERSITY OF PENNSYLVANIA
THE MOORE SCHOOL OF ELECTRICAL ENGINEERING
SCHOOL OF ENGINEERING AND APPLIED SCIENCE

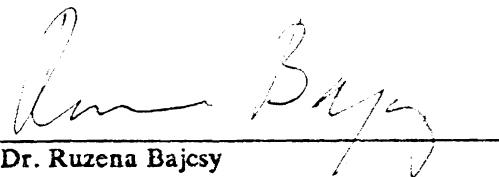
COMPUTATIONAL MODELS OF VISUAL HYPERACUITY

Eric Paul Krotkov

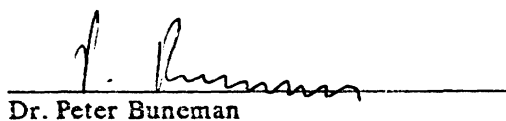
Philadelphia, Pennsylvania

May 1984

A thesis presented to the Faculty of Engineering and Applied Science of the University of Pennsylvania in partial fulfillment of the requirements for the degree of Master of Science in Engineering for graduate work in Computer and Information Science.



Dr. Ruzena Bajcsy



Dr. Peter Buneman

ABSTRACT

The process of visual hyperacuity is described and analyzed in the terms of information theory. It is shown that in principle, the detection and representation of both luminance and edge features can be performed with a precision commensurate with human abilities.

Algorithms are formulated in accord with the different representational methods, and are implemented as distinct computer models, which are tested with vernier acuity tasks. The results indicate that edge information, encoded either in the manner proposed by Marr and his colleagues (as zero-crossings in the Laplacian of a Gaussian convolved with the image) or when encoded as a simple filtered difference allows finer spatial localization than does the centroid of the intensity distribution.

In particular it is shown that to judge changes of relative positions with a precision of 0.1 sec arc in two and three dimensions, it is sufficient to represent the displacement of an edge by the *difference* of two Laplacian-Gaussian filters rather than by the difference between interpolated zero-crossings in them. This method entails no loss of relative position information (sign), allows recovery of the magnitude of the change, and provides significant economies of computation.

ACKNOWLEDGEMENTS

This paper would never have been possible without the considerable assistance and energies of Dr. Kenneth Knoblauch. He has contributed many original ideas, mathematical rigor, and his solid knowledge of the many aspects of vision.

I thank Dr. Ruzena Bajcsy for her sound judgment, creative ideas, and her support and guidance.

Peter Allen has always helped.

I am very grateful to Dr. Ellen Hildreth and Dr. Roger Watt, who provided valuable criticisms of earlier drafts.

TABLE OF CONTENTS

I. Introduction	1
II. Visual Acuity	
1. Image Formation and Quality	2
2. Types of Acuity	4
3. The Factors Underlying Resolution	6
4. What Factors Underlie Localization?	9
III. Hyperacuity as a Computational Problem	
1. Different Levels of Explanation	11
2. What Needs to be Computed?	12
3. Computation Strategies	16
4. Representation of Position Information	20
5. Algorithms	32
6. Hardware Implementation	34
7. Summary	35
IV. Implementation and Results	
1. Implementation	37
2. Methods	38
3. Results	38
V. Discussion	45
Notes	47

I. INTRODUCTION

One of the central issues in vision research concerns spatial relations and location. The human perceptual process of visual hyperacuity, the ability to perceive with extreme precision spatial position information, both laterally and in depth (for example reading a vernier), poses some profound and as yet unanswered questions about how a visual system acquires and represents very fine-grained spatial information.

These questions have traditionally been posed in the languages of psychology, psychophysics, and neurophysiology. In this paper they are asked in a different language; here they are considered from within the paradigm of information theory.

Such an approach is not new, nor is it suggested to supplant traditional approaches, but it does offer certain advantages. Foremost among these is the explicitness with which hypotheses about how a system represents and processes precise spatial information can be formulated and tested. Here, three such hypotheses are developed and implemented as computer models, which are developed and presented as follows.

First, the various visual acuities are defined both in terms of different types of visual tasks and in terms of their limiting physical and psychophysical principles. A difference of an order of magnitude between the threshold limits of resolution and localization is observed.

Second, the process of visual hyperacuity is explicitly cast in a computational framework. Viewing the process as a computational problem, questions about the nature, purpose, form and implementation of computations performed upon visual input are discussed. Specific computational mechanisms are developed and formulated as algorithms.

Third, the actual implementation of these algorithms using simulated data, and natural data in two and three dimensions is described.

The last section interprets the results, compares them with the performance of the human visual system, and discusses implications both for theories of human hyperacuity and for a number of computer vision issues.

II. VISUAL ACUITY

Before the question of what mechanism is responsible for visual hyperacuity can be raised, the process itself must be well defined. To do this requires an understanding of visual acuity in general, which is in turn impossible without first considering the physics and psychophysics of image formation, which is the subject of this section.

1. Image Formation and Quality

Here we will only consider images formed by devices such as the eye and the camera, which use a converging lens to focus an inverted real image on a surface behind the lens.

The image formation process is a conversion from a continuous function to a discrete function, effectively describing the image as samples at discrete points. We shall formally describe this conversion with the *delta function*, which may be defined by:

$$\delta(x) = \begin{cases} 0 & \text{when } x \neq 0 \\ \infty & \text{when } x = 0 \end{cases}$$
$$\int_{-\infty}^{+\infty} \delta(x) dx = 1$$

This does not represent a function in the sense in which the word is used in analysis (to stress this fact Dirac called it an "improper function"), and the above integral is not a meaningful quantity until some convention for interpreting it is declared. Here (after Bracewell (1978)) it is interpreted as the limit of a set of functions:

$$\delta(x) = \lim_{n \rightarrow \infty} \delta_n(x)$$

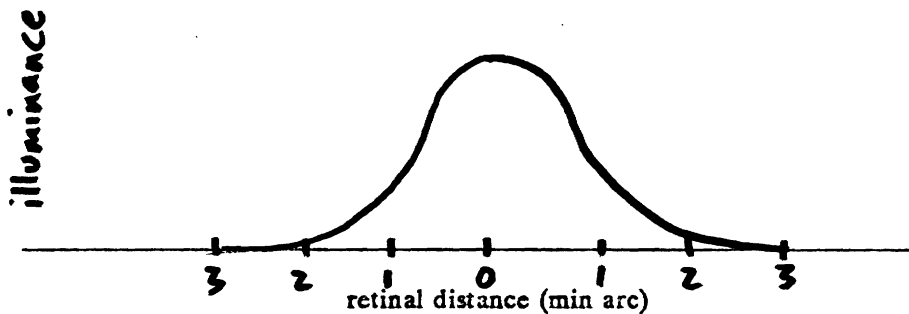
where

$$\delta_n(x) = \begin{cases} n & \text{if } |x| < 1/2n \\ 0 & \text{otherwise} \end{cases}$$

A continuous image may be multiplied by a two-dimensional "comb", or array of delta functions, to extract a discrete *sample* for each delta function. After sampling an image may be described as a discrete function $f(x,y)$ giving the light intensity (gray-level) at each point (x,y) on the surface behind the lens (*image plane*).

A fundamental description of the *quality* of an image produced by any optical system is the point spread function, i.e., the distribution of light in the image plane of a point object. This function can be regarded as the spatial probability distribution (in the image plane) of a single photon emitted from a point source. FIGURE 1 illustrates a reasonable estimate of the point spread function of the normal human eye in good focus.

FIGURE 1.



Point spread function of the normal human eye.

An alternative description of the quality of an image can be given, in the language of electrical engineering, by the *modulation transfer function* of the optical system. In a linear system the point spread function can be derived from the (spatial) frequency response by Fourier analysis, so the two are equivalent: Just as we characterize the quality of an amplifier by the way it handles a train of pure sine wave inputs (where the output sine waves emerge as sine waves with a change in amplitude and phase which depends upon their frequency), so do we characterize the quality of an optical system by the way it takes a sinusoidal grating and images it as a sinusoidal grating with a reduction in amplitude of modulation and a change in phase. The plots of modulation and phase versus spatial frequency of sinusoidal gratings are respectively called the *modulation transfer* and *phase transfer* functions, and together they contain the same information as the point spread function.

The significance of these functions lies in the fact that once either is known, it is possible to predict what is present on the image plane by determining what the optics of a visual system may have done to a target. Since we will not be using sinusoidal gratings, we here

adopt the point spread function as our measure of the quality of an image.

2. Types of Acuity

There are many types and measures of visual acuity, but four subdivisions of this field are traditionally drawn. Each presents visual acuity as a threshold which is measured in the spatial domain; for example, the size of a feature in the visual field is changed until the subject can make a correct response. Unless otherwise specified, we shall be concerned with *foveal* rather than *perifoveal* or *peripheral* acuities in discussions of the human visual system.

The minimum visible

The minimum visible refers to the minimum size necessary for a feature to be detected. The kinds of tests used in experiments on the detection of small objects include: (a) bright objects against a dark background, (b) dark objects against a bright background, and (c) low contrast objects. Riggs (1965) indicates that this is primarily an incremental luminance detection task.

Recognition

Recognition tasks require the subject to name the test object. This task is used in clinical studies, in which wall charts and test plates commonly present progressively smaller printed symbols to be recognized. The subject is then scored on the minimum width of line, gap, or other characteristic of the object correctly identified.

Despite the popularity of this method, the results it gives are difficult to interpret theoretically and few experimental investigations of acuity have made use of symbols of this sort.

The minimum resolvable

The minimum resolvable refers to the minimum size necessary for an internal differentiation of an object to be made. (e.g., Is this a single or a double star? Is this an O or a C?) The test objects have in common the fact that each single element of the pattern would

be clearly identified if it were presented alone.

Visual acuity, in this sense, is the reciprocal of the angular separation between two elements of the test pattern when the two images are resolved. This measure is comparable to the "resolving power" of a camera or a telescope. The theoretical limit of this resolving power is a function of the wavelength of the light and the diameter of the aperture (see Section II.4).

Localization

Localization refers to the minimum detectable difference in the relative location of objects. (e.g., Is the upper line to the right or left of the lower line? Is the upper line in front of or in back of the lower line?) It is interesting to note that the human visual system is actually very poor at judging absolute distances in the absence of very strong cues.

Both vernier acuity and stereoacuity are commonly tested by the use of a straight line broken in the middle. The task is to detect small displacements either laterally or in depth of one line segment as shown in FIGURE 2.

FIGURE 2.



The task of localization as illustrated by vernier acuity.

(To be precise it must be noted that localization concerns shifts in the position of arbitrarily large images, rather than images the minute size of hyperacuity thresholds).

This experiment yields very small thresholds. For example, Berry (1948) reports thresholds of about 2 seconds, and other observers report similar values. It should be noted that a

2-second displacement of the test line amounts to about 0.01 mm seen at a distance of 1 meter! The exact thresholds are influenced by the characteristics of the test object (including target length, gap size, target orientation and target curvature) and the background.

Restriction of terms

Now that the types of acuity have been defined we shall restrict our attention to resolution and localization, since it has just been shown that they are the acuities which concern the types of tasks in which we are interested. It remains to see what physical principles and perceptual mechanisms account for these two acuities, what different processes might be engaged in each of these tasks, and how the results for each compare. Resolution will be discussed first.

3. The Factors Underlying Resolution

The threshold figure of one minute of arc for the human visual system has been widely accepted for many years. For instance, Westheimer (1977) has found that, with practice on the classical two-point test, an angular separation of about 1 minute of arc can be distinguished with 75% success. As before, threshold measurements are made in the domain of space (distance), but because retinal distances are most conveniently expressed in terms of visual angle, the units of seconds or minutes of arc are employed.

There are a large number of factors which must be considered in determining the limits of the resolution of an optical system, including (1) eye movements (2) contrast effects, (3) intensity effects, (4) stimulus duration, (5) state of adaptation, (6) the dimensions of the receptor mosaic, and (7) aperture size.

Although the impact of each of the first five factors is significant, we shall assume that conditions have been optimized with respect to them. So that under normal testing circumstances, namely constant daylight illumination, test objects of nearly 100 per cent contrast, and no significant eye movements, these factors are not limiting.

The dimensions of the receptor mosaic are significant to the extent that, given the unidimensional nature of receptor output, there is no simple explanation for seeing something placed, say, 1/7 of the way between the positions of two receptors. The distance separating two receptors must then be considered. In the human central fovea, where the inner segment of a single cone covers 20 seconds of arc and the centers of cones are separated by about 30 seconds of arc, a cone is (approximately) placed at each node and antinode of the highest spatial frequency passed by the optics of the eye (Snyder and Miller (1977)), which is one minute of arc. Thus the optical and receptor mosaic size factors converge on the same limit of one minute. If the human receptor mosaic were coarser, as it is in the parafovea, then it would be the primary limit to the resolution achieved by the human eye.

The size of the aperture is an important and complex factor in resolution. A large aperture allows more light energy to stimulate the receptors and diminishes the blur due to the diffraction of light. A small aperture, on the other hand, diminishes the effects of spherical and chromatic aberrations in the lens. An ideal system would map object points into image points rather than into a distribution such as given by the point spread function.

To determine the parameters of this distribution in a physically ideal optical system we must find the absolute limit of resolution as given by the diffraction theory of light. This establishes a lower bound on the spatial resolution of a visual system without aberrations.

We begin by considering Fraunhofer diffraction of a point source by circular apertures. The projection of a circle is an ellipse, but results obtained from elliptical projected apertures of small eccentricity will be quite similar to those from circular projected apertures. Without loss of generality we assume that the aperture is circular.

In this case, the intensity of light energy in the image plane is (after Ford (1973)) given by

$$I = I_0 \left[\frac{\sin(a \sin\theta/2\lambda)}{(a \sin\theta/2\lambda)} \right]^2 \quad (1)$$

where a is the diameter of the aperture, λ is the wavelength of light in air, and θ is measured

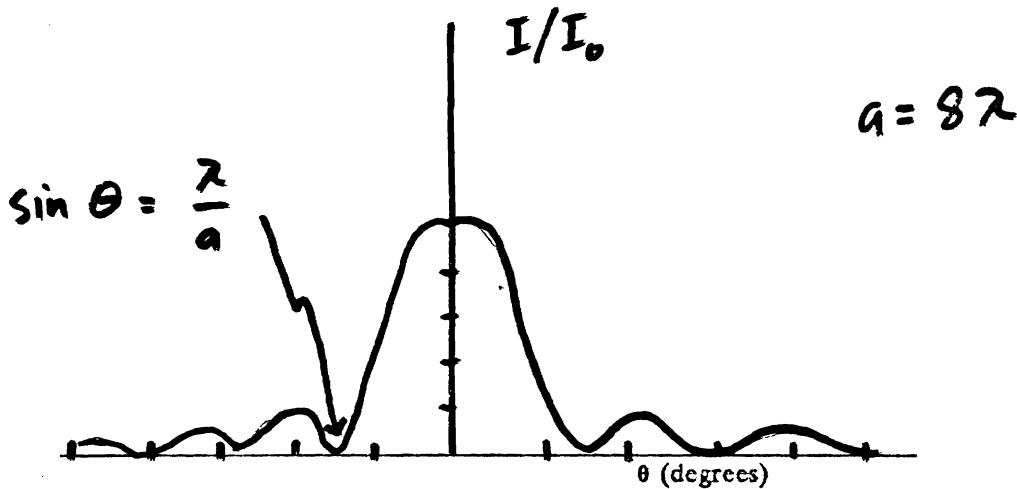
in object space.

Since $(\sin x)/x \rightarrow 1$ as $x \rightarrow 0$, Equation (1) gives $I = I_0$ at the central peak of the diffraction pattern. The angle of the first minimum next to the central maximum is given by

$$\theta = 1.22 \frac{\lambda}{a} \quad (2)$$

(with $\frac{\lambda}{a} \ll 1$). FIGURE 3 shows a graph of I versus θ . Eighty-four percent of the total area in this intensity pattern is in the first maximum, or *Airy disk*.

FIGURE 3.

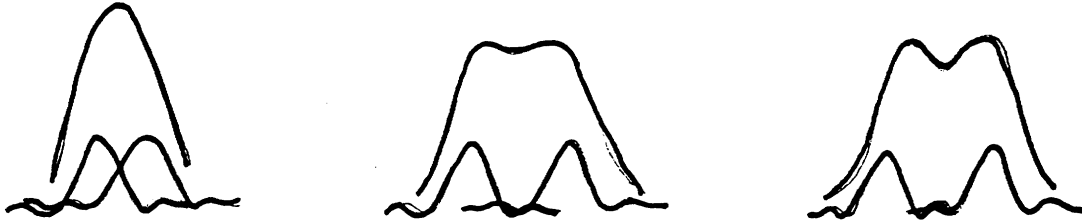


Intensity in Fraunhofer diffraction pattern vs. angle of observation.

FIGURE 4 pictures several light distributions of two equally bright incoherent point sources separated by θ . Rayleigh has suggested that two points are resolved when the center of the Airy disk of one falls exactly at the first zero of the second, i.e., when their angular separation is $1.22 \frac{\lambda}{a}$ radians. Hence Equation (2) can be used as an expression of the resolving limit of an optical system.

The normal, emmetropic human eye has a point spread function which approximates the Airy disk for a 2.3 mm aperture. Equation (2) then yields a value of almost exactly 1 minute of arc at a wavelength of 555 nm.

FIGURE 4.



Schematic illustration of light distributions when two point sources are presented with three different angular separations. The depth of the dip determines the limits of resolution.

Westheimer (1976) points out that this is not an absolute limit, since resolution is possible as long as the dip ("dimple") between the peaks of the spread functions is detectable and since it is only an approximation to identify the point spread function as an Airy disk rather than some other function such as a Gaussian or exponential. If the sensitivity of the observer is high then much smaller dips in the bivariate intensity distribution may be identified. Thus the physical resolution limit of the eye may be less than the one minute of arc given by Rayleigh's criterion, but not by a significant amount, especially when the aberrations present in a non-ideal system are considered.

This is an arbitrary description of resolution, in the sense that there are in principle no limits on the detection of a minimum between two maxima in the light distribution except those imposed by noise. Despite its arbitrariness, it is reasonable and is consistent with psychophysical evidence of the type Westheimer has presented.

4. What Factors Underlie Localization?

Let us reconsider the case of a single point source. If it is so dim that only a single photon is absorbed by a given receptor, its location in object space can only be determined within the bounds of the probability spread given by the point-spread function. With increasing

intensity of the source, however, more and more photons will be absorbed, so that the shape and position of the point-spread function will become more and more distinct. To borrow an analogy from Westheimer (1976): "The situation may be likened to the scattering onto a plane of grains of sand that are fed through a funnel held some distance above the plane. A heap is formed whose center can be determined with greater and greater precision as the quantity of sand increases."

When there exist two point sources so close together that their point-spread functions overlap, there can be no discrimination of which photon originated in which point source. This constitutes the bottom line of the diffraction limit of resolution. But diffraction, while limiting the resolution of two point sources, does not limit the localization of a single point source. Once we are no longer concerned whether there is one feature or two, the diffraction limit does not apply.

Precision of localization then essentially becomes a problem of output comparison among photoreceptors, in the sense that the question asked is "What is the relative position of the feature?" rather than "Are there one or two features?"

As noted previously, localization thresholds in the detection of alignment errors (Is this feature to the left or to the right, in front or behind?) have been reported to be as low as 2 or 3 seconds of arc. These visual tasks have thresholds as much as a full order of magnitude smaller than the threshold given by the diffraction theory of light. Further, these thresholds are much finer than the sampling mosaic of the retina, where cones in the fovea are separated by at least 20 sec arc.

Westheimer (1975) has coined the term "hyperacuity" to emphasize this difference in scale between resolution and localization. Although the exact hyperacuity threshold values are dependent upon the criterion and measurement techniques, the outstanding fact is that these thresholds can not *prima facie* be reconciled with the diffraction limit of the eye. Given this fact, a framework in which such a reconciliation can be made must be sought.

III. HYPERACUITY AS A COMPUTATIONAL PROBLEM

The previous section has motivated the need for an explanation of how a visual system can make extremely fine judgments of relative position, judgments an order of magnitude more precise than those of absolute position. We shall pursue such an explanation from the perspectives of computational and information theory. In his book Vision David Marr (1982) treats at length the form and nature of an adequate computational theory, and there he sets out some clear standards for a rigorous methodological approach.

1. Different Levels of Explanation

One of Marr's central aims in Vision is to formulate rigorous computational theories of various perceptual processes, theories which must specify why a perceptual process is undertaken (what it is for) and how it proceeds (what it does, what it computes). Such a theory is said to be computational because it provides an explanation of a perceptual process in terms of the activity of an information processing or computing device. This device must be understood on at least three different levels.

At the top level, the performance of the device is characterized as a mapping from one kind of information to another. The abstract properties of this mapping are defined precisely, and its appropriateness and efficacy for the task at hand are demonstrated. The questions to be answered by this top level include: What is the goal of the computation? Why is it appropriate? What is the logic of the strategy by which it can be carried out?

At the intermediate level, choices of a representation for the input and output of the information processing device as well as the algorithms to be used to transform one into the other are made. The questions raised at this level include: How can this computational theory be implemented? What representation of the information will be employed? What useful operations can be performed upon this representation?

At the bottom level, the details of how the representation and algorithms are physically realized are at issue. Questions raised at this level concern neurophysiology for the human

visual system and machine architecture and organization for computer vision systems.

An explanation for any particular perceptual phenomenon can be provided at one of these levels, but the important point is that no explanation is considered complete until it addresses the issues raised at *each* level. Accepting Marr's criteria for an adequate explanation of a perceptual process each of these explanatory levels will be considered from within the context of hyperacuity.

2. What needs to be computed?

From an information-processing point of view the level of computational theory is of critical importance. For in trying to understand the nature of the computations that enable visual hyperacuity, it is far easier to think in terms of the kind of computational problems that must be solved than in terms of the complex ocular and neural hardware in which their solutions are implemented, just as it is easier to think in terms of the integers than in terms of signals propagating through AND-NOT circuits when trying to understand the process of addition. The question, now, is "What needs to be computed?" rather than "How?"

2.1. The goal of the computation

The evolutionary or ethological significance of human visual hyperacuity is, of necessity, a matter for speculation. But a very delicate sensitivity to changes in the foveal visual field may have played the role of an early-warning system, signalling danger in the form of camouflaged predators ahead, or may alternatively have played the role of a sophisticated prey-detection system in predators.

It is interesting to observe that quicker and more accurate judgment of position is an accompaniment to more rapid locomotion. This relationship is consistent with the biological facts that acuities are more highly developed in the most mobile animals, namely many birds and some of the most active mammals, and are more developed in predators like hawks and owls than in prey. It is also possible that hyperacuity developed along with stereopsis (for which it is clearly useful), and was not particularly useful for vernier type tasks.

At any rate, hyperacuity is a response to a *change* in location, and as such, implies a judgment of *relative* position. For the hunter, the item of interest is that of the difference between the past and present positions of the prey. For a psychophysical experiment, the item of interest is that of the difference between the past and present positions of the target stimuli.

The task, or goal of the computation, can be formulated more precisely by observing that in both cases the situation is that of having a test target, where the task is to determine whether or not the test target has been displaced between two views, or frames. Since the frames are presented at different times, there must be a "memory mechanism" which can accurately recall the location of the test target in the first frame. In a psychophysical experiment this introduces a new random variable, for which a control must be provided. One solution is to provide a stationary reference in both frames, where the task is to identify the position of the test target in a test frame relative to the position of the test target in the reference frame. In this task, the position of the reference target need not be stored or recalled, and is the same in both reference and test frames. The stimulus arrangement defining the task is illustrated in FIGURE 2.

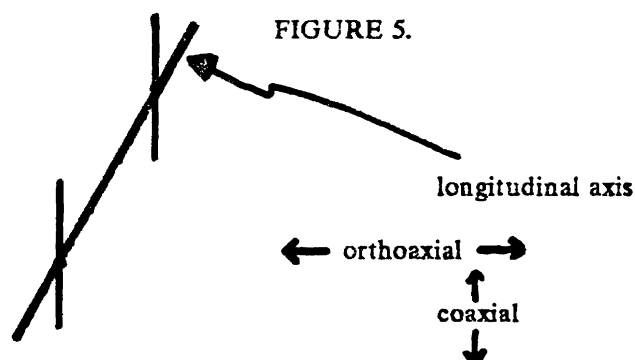
2.2. Different types of position

Clearly the test target in frames 1 and 2 may be related by any combination of translation, rotation, or deformation. This calls into question what exactly our notion of position is. It is clear that the parameters relevant to position judgments include offset, orientation and shape, and it is equally clear that the capacity to identify 'where something is' is complex and ambiguous until some metric is adopted.

One group of researchers approaches this issue by describing a single conceptual framework (contour analysis) subsuming different mechanisms (for slope and position). From experiments with blurred targets, Watt, Morgan and Ward (1983) conclude that there are two (and possibly three or more) distinct mechanisms involved in vernier acuity. One is responsible for the discrimination of absolute slope cues, and is employed in tasks requiring judgments of the

shapes of curved lines (there may be a distinct mechanism operating on highly curved lines). The second is sensitive to relative positional differences.

One statement of the difference between these two mechanisms (see Watt and Andrews (1982) for a more precise account) is that the first makes use of position information along a common longitudinal axis defined by the target (*coaxial*) (see FIGURE 5) while the second uses only information orthogonal to the same axis (*orthoaxial*). Another interpretation is that the first mechanism concerns both deformations (changes in curvature) and rotations (changes in slope) while the second concerns translations (changes in position *per se*).



Since both slope and curvature are derivatives (mathematically and figuratively) of position, the second mechanism appears to address the more primitive notion of position. Parenthetically it is remarked that by deforming, the identity of the target itself is changing. This is not strictly a change in position, i.e., it is not the same target in two different positions. Similarly, a rotation is an angular displacement, not strictly a spatial displacement.

Since we are interested in the more primitive notion and its relation to the first mechanism, we can now define 'relative position' as the spatial displacement of a test target between two frames relative to the longitudinal axis determined by the reference (stationary between frames) and test targets.

2.3. Simplifications

In the above analysis a number of simplifications have been introduced which must be made explicit.

First, it is not specified *when* the computation is to be made. Perhaps the high-precision information is always available and always computed; perhaps it is only available "on demand". Regardless, the focus here is on the process performing the computation, rather than on the process(es) deciding if and when it is necessary.

Second, it has been assumed that the relevant stimuli in the two frames can be discriminated from the irrelevant. In reality, there must be some kind of high-level mechanism which selects the important parts of the frame to analyze.

Third, the fact that the image moves over the receptor mosaic, either because of unintended eye movements or because of failure to accurately track a moving object, presents problems which are important and unresolved. But the topic of spatiotemporal determination of position is beyond the scope of this paper.

Fourth, a topic which is not addressed here concerns the fact that in the human visual system the photoreceptors perform a log transformation of intensity, which may effect the subsequent position measurement. It would be interesting to consider psychophysical experiments which address this issue, and to experiment with images with and without such log transformations (perhaps by using different digitizing devices).

These four simplifications make the research more tractable without making it trivial. A computational model which also accounts for the unsimplified issues would certainly be richer, but not more fundamental or profound with respect to furnishing an explanation of how high precision relative positions can be recovered by a visual system.

2.4. Summary

The goal of the computation is to determine whether or not a change in relative position (in the narrow sense) exists. If so, in what direction (and possibly by how much)? The computation can now be viewed as a transformation taking an input of two frames into an output of a vector. Since this is a computation of relative position, the output need only be a unit vector.

3. Computation Strategies

Now that the answer to the question "What needs to be computed?" has been determined to be relative position, the logic of a strategy to make the computation must be examined. Two strategies will be discussed: the first based on the theory of mean local sign; the second on interpolation. Other strategies are certainly possible. For any strategy, the important question is whether it can correctly and effectively compute relative position.

3.1. Mean Local Sign

In 1899 Hering pointed out that a point on the retina might actually be localized within a region smaller than that of any single photoreceptor, since an "averaging process" could act to fill the gaps between discrete photoreceptors. Hering's account of this "averaging" relied upon the assumption of an extremely regular spatial arrangement of cones in the retina. Even though cones are arranged fairly regularly in the fovea due to their dense packing, Hering could not explain how averaging occurred when stimuli did not fall precisely in a rigid pattern on the cones, which modern histology shows are not perfectly regularly spaced.

After Andersen and Weymouth (1923) this hypothesis was elaborated as that of "mean local sign", in which localization is derived from a combination of samples taken along the area or strip of receptors stimulated by the target. The local sign of each receptor is presumably either on for stimulated or off for unstimulated. When on, a spatial value inherent in the receptor which represents the whereabouts of that receptor is available to whatever processes are interested.

Since the receptors are distributed randomly, in the long run equal numbers of them will lie in all parts of the strip, and the center of the strip will represent the "center of gravity" of all receptors stimulated. The average or mean of these receptors is therefore not restricted to units such as inter-receptor distance or receptor diameter, but may be accurate to a small fraction of these units.

Thus the average of the positions of the stimulated receptors is accurate to a higher pre-

cision than any of the measures entering into its formation; the overall estimate of position, based on the combination of samples, will improve on any individual estimate. Thus the overall precision of localization will be limited only by the number of samples and their variances, and in principle can be on the scale of hyperacuity, e.g., an order of magnitude better than that of any sample. Relative position may then be accurately determined by computing the vector corresponding to the difference of the mean local signs of the test target between frames.

3.2. Interpolation

Since localization is an order of magnitude more accurate than resolution (Section II.5), hyperacuity obviously demands that the visual system somehow estimate the optical image lying between neighboring receptors. This estimation is analogous to the interpolation process for drawing a continuous curve through a discrete set of data points in order to estimate the value of a point lying between samples.

A continuous optical image is sampled at a set of discrete points by the photoreceptors on a surface behind the lens. If these samples are taken sufficiently close to each other, the samples provide an accurate representation of the original continuous image, to the extent that that pattern can be reconstructed by interpolation. The limits on the 'closeness' of the samples are precisely expressed by the (Whittaker-Shannon) sampling theorem, which states that a band-limited function $g(x,y)$ can be recovered exactly from a rectangular array of its sampled values as long as $g(x,y)$ contains no spatial frequencies greater than one-half the sampling frequency.

An optical system will not transmit spatial frequencies higher than ω_0 , at which the modulation transfer function is zero:

$$\omega_0 = \frac{d}{\lambda f}$$

where λ is the wavelength of light, f the focal length of the lens, and d the diameter of the lens. A diffraction-limited optical system thus produces an image which is bandlimited, since

the effect of the optics of the system is that of a low-pass spatial filter with some cut-off frequency, ω_0 . In the human eye this limit is about 60 cycles per degree of visual angle when the pupil is at its smallest (about 2mm) in bright light, and lower values when the pupil is larger. The signal must be lowpass filtered *before* sampling in order to avoid overlap of the sidelobes in the Fourier spectrum (aliasing). Lowpass filtering *after* sampling cannot always avoid aliasing. Since a spatial cut-off frequency ω_0 is guaranteed, to apply the sampling theorem and to guarantee that no information is lost we must insure that the distance between the samples does not exceed the Nyquist limit $\frac{1}{2\omega_0}$ at any sampling level (e.g., photoreceptors, ganglion cells, cortex, etc.). In particular this requires having a receptor at each node and antinode of the highest spatial frequency passed by the optics-- a condition which is found approximately in the central foveas of a number of animals (Snyder, 1979) where the sampling frequency is 120 cycles/degree.

When these conditions are satisfied the theorem guarantees that it is possible to reconstruct the function from the set of samples using some process of filtration. What in the transform domain is filtering amounts to interpolation in the function domain. The two are equivalent; the original function may be reconstructed either by spatial interpolation or by spatial filtering. The effect of sampling is to replicate the original spectrum in an infinite number of side lobes. Spatial interpolation is accomplished by filtering out all side lobes but the central one, which is the original spectrum.

The classical spatial interpolation scheme employs the *sinc* function, but others may be used, for example the *circ* function, and simple linear interpolation by the *triangle* function

$$\text{tri}(x) = \begin{cases} 1 - |x| & |x| < 1 \\ 0 & |x| > 1 \end{cases}$$

An ideal spatial filtering scheme (corresponding to an infinite sum of *sinc* functions) employs a filter with transfer function

$$R(\omega_x) = \begin{cases} k & \text{for } |\omega_x| < \omega_{xl} \\ 0 & \text{otherwise} \end{cases}$$

Whatever scheme is used it should be clear that, once the continuous function is reconstructed, spatial locations can be computed with an arbitrary accuracy. In principle, then, this scheme allows the determination of localization finer than the sampling mosaic. By reconstructing continuous functions from sampled functions, picking a convenient point (on the target) and comparing its values in both continuous functions, a judgment of relative position can be made.

The difficulty with this scheme as presented is that there may be no way to implement an ideal spatial filter such as the *sinc* function. In particular, the human visual system can only approximate the *sinc* function (which extends infinitely in space), and the receptive field corresponding to even a truncated approximation of this filter is likely to be very complex. One alternative is to search for a filter which provides a good approximation to the exact reconstruction which can be simply implemented. Another alternative is to reconstruct some important feature of the image (or target) rather than the original continuous image function. These alternatives will be explored in the next section.

3.3. Comparison

It has been shown that both computational strategies can in principle compute relative position. The actual conditions under which the principles apply are summarized here.

The computation of mean local sign will have accuracy limited by the number of samples and the blur introduced by the optics. This implies that if the light distribution is very narrow, or if too small a region of the light distribution is sampled, accurate localization may not be possible. The major computational costs are: collecting the local signs in both frames; computing the mean in both frames; and computing the difference of the means. This is evidently a very simple computation to perform.

Interpolation will fail if the sampling rate exceeds the Nyquist limit, i.e., if the distance between the samples is twice that of the highest spatial frequency passed by the optical system. The major computational costs are in filtering in the transform domain (or alternatively, in

applying an interpolation function in the function domain) and in selecting features of the filtered image to which position may be assigned.

4. Representation of Position Information

The input to the process is two arrays of intensity values, which can be described by two image functions (as in Section II.1). The output of the entire process is a judgment of relative location: left, right, above, below, front, back (with respect to the axis of the target). Effectively, this is the sign (+ , - , 0) of displacement in a given direction. In addition, a numerical quantity denoting the magnitude of position offsets may be derived. Thus the output of the process can be completely specified as a vector.

The choice of a representation for position is important to the extent that it determines what information is made explicit, and consequently, the ease and speed with which that information can be accessed, and the types of operations which can be performed upon it. In essence, this choice will determine the image features to which position is assigned. Two representations are discussed, one for each strategy of computation.

4.1. Luminance Features

Based on the theory of mean local sign, one obvious proposal is that location be assigned to the "center of gravity", or arithmetic mean, of the light distribution. By the centroid of $f(x,y)$ we mean the point (\bar{x}, \bar{y}) which gives the ratio of the first moment to the area of $f(x,y)$:

$$\frac{\int x f(x) dx}{\int f(x) dx}$$

or equivalently, the slope at the zero frequency component of the Fourier transform of the distribution divided by the amplitude of the zero frequency component:

$$\frac{F'(0)}{2\pi i F(0)}$$

Roughly speaking (\bar{x}, \bar{y}) tells where a function is mainly concentrated; in statics (\bar{x}, \bar{y}) is the center of gravity of a beam whose mass density is $f(x,y)$. The representation should not be in the frequency domain, however, because of the expense of performing the Fourier transform.

Westheimer (1979) has provided a great deal of evidence consistent with this proposal. In one experiment, Westheimer and McKee presented the observer with a stimulus composed of two strips about 2.4' wide and 6.4' high, abutting vertically. Each strip was composed of 9 bands, 14" wide spaced 3" apart. Observers cannot resolve the bands. To each strip was added a 10th band with the same characteristics as the others. The added bands could be either vertically aligned (centers of gravity match) or not (center of gravity offset). According to Westheimer and McKee the observers were not able to detect an inhomogeneity within either strip, but were able to indicate at 75% correct when the center of gravity had shifted by on average 4.7". While the centroid is not the only cue in the above distributions, the evidence is very suggestive.

The representation of position by the centroid implies that a number of potentially salient features of the intensity distribution are ignored, only the mean is extracted. In effect this representation replaces the lines of the target with a single point. While some information loss is inherent in such a transformation, the issue is whether *position* information is lost. In principle the centroid of a function can be determined exactly, and because the centroid is invariant under translation no relative position information is lost, in the sense that any change in the distribution will be reflected by a change in the mean.

Hence, the information loss can only be introduced by the representation of the mean local signs. More precisely, the computed mean is a *sample mean* \bar{x} , which is only an approximation of the continuous mean μ of the intensity population presented as a stimulus. The reliability of \bar{x} as an estimate of μ is often measured by the standard error of the mean, σ/\sqrt{n} . However, this metric is based upon the Central Limit Theorem which requires that the samples be independent. Given the structure in the image, the intensity samples are not independent. A metric based on paired samples, which does not require independence, uses the Student-t distribution. However, this assumes that the intensity samples are drawn from a normal population, and that the sample variance is known. Neither assumption is tenable in general. Because the normality and variance of the intensity distribution are not known *a priori*, little

statistical leverage on the accuracy of the approximation is readily available.

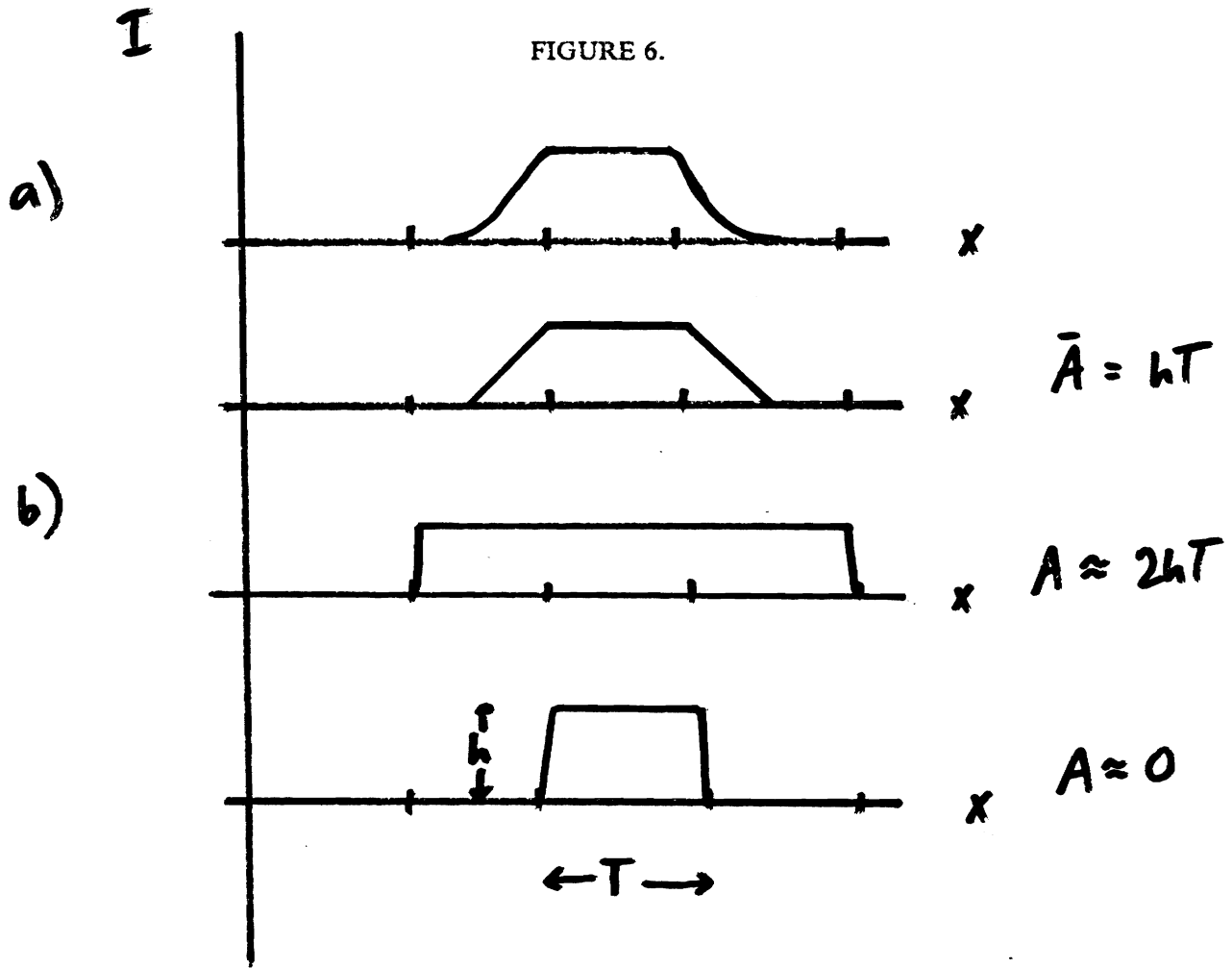
An upper bound on the error of the approximation of the area underneath the continuous function using the rectangle rule is

$$E = \frac{f''(x)(b-a)^3}{12}$$

but tighter upper bounds can be derived. The intensity distribution is determined by the physical characteristics of the target. FIGURE 6 illustrates the distribution assuming a uniform background, perfect contrast, a regular sampling frequency, and a homogeneous target. In calculating the centroid, a computation of area underneath the distribution is performed. In this ideal environment, the error in the discrete approximation \bar{A} of the area A underneath the continuous function is less than T , the sampling period.

If the edges are blurred, as they are by the optics, then the error is much smaller. As shown in FIGURE 7, the area \bar{A} closely approximates A , since the area discrepancies cancel each other, as \bar{A} either overestimates or underestimates A . Evaluating the accuracy of the approximation depends critically upon the edge blur, which is given by the point spread function. Taking several spread functions, TABLE 1 shows the magnitude of the errors when the area \bar{A} is calculated with a rectangular rule during the computation of the centroid. Clearly, the error is smaller than 10% of T , as required by the need for precise localization.

From both FIGURE 7 and TABLE 1 it is clear that sign information is preserved in all cases except when edges are not blurred at all. Thus it can be concluded that transforming the lines of the target into a single point representing the centroid does not entail loss of enough position information to curtail localization. Further, it is evident that the higher the decay of the point spread function, the coarser the approximation of the sampled centroid to the continuous centroid will be. Moreover, the conditions on the area over which information is necessarily gathered for accurate localization are that it include a target at least one sampling period wide, its blurred edges, and that the sampling region must be consistent between frames.



(a) Intensity distribution of idealized target.

(b) Three continuous distributions all fitting sample points.

Area \bar{A} is computed using the rectangle rule. A may vary from \bar{A} .

$\bar{A} = hT$ $0 < A < 2hT$ $|A - \bar{A}| < hT$ $|A - \bar{A}| < T$,
since in high contrast $h=1$.

Psychophysical experiments have shown that more stringent conditions are imposed by the human visual system. Westheimer and McKee have demonstrated that there is a region, extending either side of the target and parallel to its major axis, about 5° wide with a longitudinal span of 30° , within which information for vernier judgments may be collected and presumably summed to advantage. This area certainly meets the conditions outlined above.

The mode and median of the light distribution are alternative metrics, but they are not germane to the theory of mean local sign developed in Section III.3.1, and are not as robust as the mean in the face of small fluctuations.

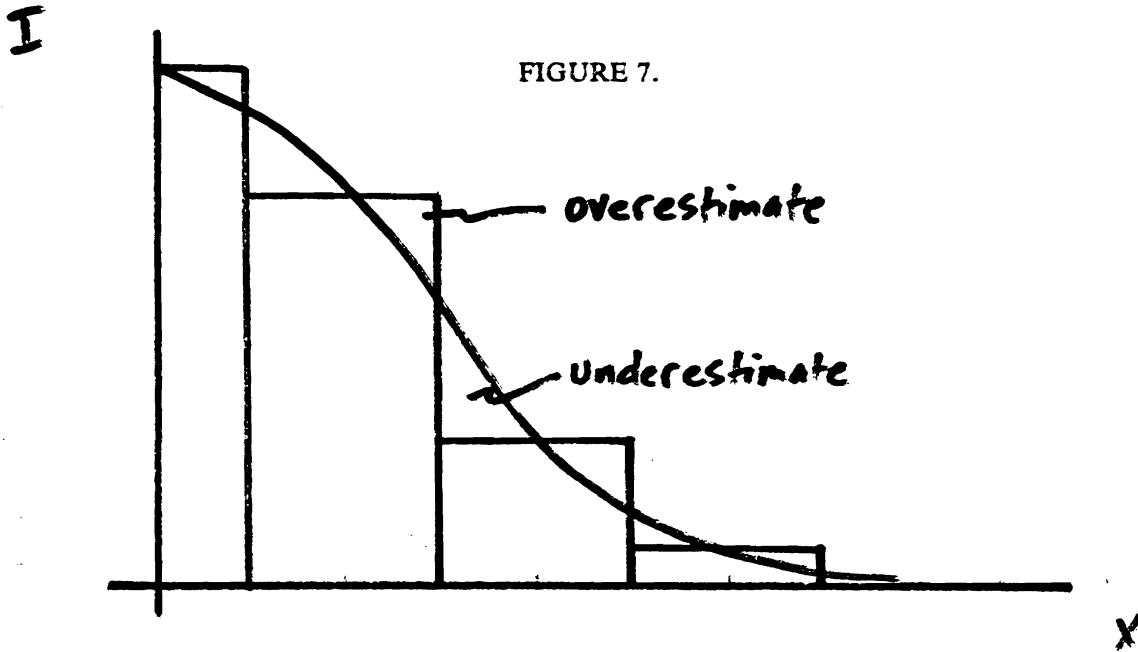


FIGURE 7.

Area discrepancies cancel each other, allowing higher accuracy in determining centroid.

TABLE 1.

x	exp(-x)	human
0.0	0.000000	0.000000
0.1	0.002010	-.001275
0.2	0.001068	-.000716
0.3	-.000885	0.000468
0.4	-.001827	0.001027
0.5	0.000182	-.000248
0.6	0.002200	-.001532
0.7	0.001266	-.000981
0.8	-.000679	0.000194
0.9	-.001615	0.000745
1.0	0.000401	-.000540

TABLE 1 shows errors involved in discretely approximating the continuous centroid given two kinds of edge blur (see Appendix 2 for precise definition of blurring functions). Error is measured by $C-\bar{C}$. C =continuous centroid \bar{C} =discrete centroid. Offset is measured in units of inter-receptor distance. An offset of 0.7 units means that the light distribution has been shifted 7/10 of the way to the next receptor. The third column illustrates that more edge blur improves accuracy.

Another proposal is to assign location to the position of the most active receptor, that is, to the peak of the light distribution. Andrews, Butcher and Buckley (1973) have shown that by quantizing position in this manner, precision should be as good as, if not better than that actually achieved by subjects. However, Watt and Morgan (1983) have rejected this

model since their data demonstrate that the human visual system assigns location on the basis of the entire light distribution, rather than just isolated local features. Also, from a computational viewpoint, peaks can be difficult to localize with high precision, particularly if the distribution is flat on top.

4.2. Edge Features

As discussed in Section III.3.2 interpolation of the image function by filtering can achieve exact reconstruction only when an ideal, completely bandpass filter is used. So rather than interpolate the image function we consider interpolation of some important image feature. Since it is difficult to imagine how a hyperacuity threshold can be observed in the absence of detectable contours, this section will treat representing the spatial position of the target by its contours, but is not as sensitive to blur.

In this case the only condition on the area over which position information is extracted is that it include one part of the test target in both frames. This area is consistent with that used in the centroid computation.

Zero Crossings

The first spatial derivative of an edge has a maximum, and the second derivative has a zero-crossing at the point where the edge is located. Thus, the zero-crossings of the second derivative correspond to locations of significant intensity discontinuities in the image, which in turn correspond to physically significant features such as edges.

Marr and his colleagues (Marr and Hildreth (1980), Crick, Marr and Poggio (1980)) have suggested that an effective and efficient mechanism for encoding spatial contour information is to smooth the sampled image and then detect points of inflection or zero-crossings in the second derivative of the result.

Smoothing is important because a major difficulty with natural images is that changes in intensity occur over a wide range of scales. It follows that one should consider separately the changes occurring at different scales, since no single filter can be simultaneously optimal at all

scales. The fact that there appear to be bandpass (scaled) channels in the human visual system lends credence to this scheme. The scale of the filter is given by its Gaussian space constant (standard deviation) σ . In this application, a small σ be used, since intensity changes over a very small spatial area are to be detected.

Detecting edges thus requires convolving the discretely sampled image function with a second order smoothing filter, for example the (isotropic) Laplacian of a Gaussian or the difference of two Gaussians (or DOG, as suggested by Wilson and Bergen (1979)). The Laplacian of a Gaussian is given by:

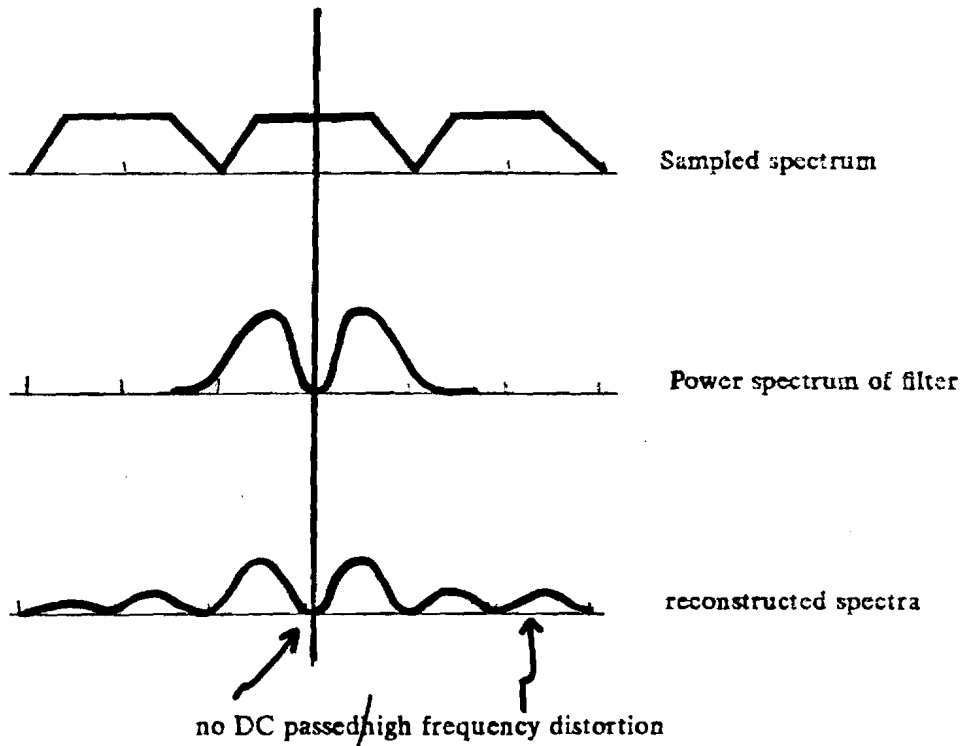
$$\nabla^2 G(r, \theta) = \frac{r^2 - 2\sigma^2}{\sigma^4} e^{-\frac{r^2}{2\sigma^2}}$$

The shape of this filter has a center-surround structure which corresponds to the receptive fields of some neurons (see Section III.6).

This filter is not necessarily being used as an interpolation function. In fact, because it is not a bandpass filter of width one octave, it cannot provide an exact reconstruction (see FIGURE 8). But this filter does faithfully preserve the spatial frequencies at which intensity discontinuities occur, assuming that σ is appropriately chosen. An ideal filter contains those frequencies present in the stimulus, but not their higher harmonics introduced by sampling. If the bandwidth of the filter is too broad, these higher harmonics will be included, thus interfering with the signal and reducing the accuracy with which it can be represented. Morgan and Watt (1982) suggest that this is exactly the case for the human visual system. They suggest that the DOG filter is adequate to explain the precision of interpolation found in their psychophysical experiments and in particular that zero-crossing features are preserved.

Interpolation can be employed not with the explicit aim of reconstructing the image, but to find with high precision some feature of the convolution profile. Of the stationary points in this profile (peaks, troughs, zero-crossings) the latter are the easiest to localize, since the location of a flat peak is hard to determine. Hildreth (1980) performed statistical experiments on a wide variety of intensity profiles and compared the performance of different interpolation

FIGURE 8.



Interpolation with $\nabla^2 G$.

functions in positioning the zero-crossings: an ideal extended *sinc* function, a truncated *sinc* function, a Gaussian, and a triangular function (linear interpolation). She found that the stimuli typically used in hyperacuity experiments would not distinguish between the different functions. She also points out that the size of the support required for their computation is lower for Gaussian and linear functions, which are very simple, local functions.

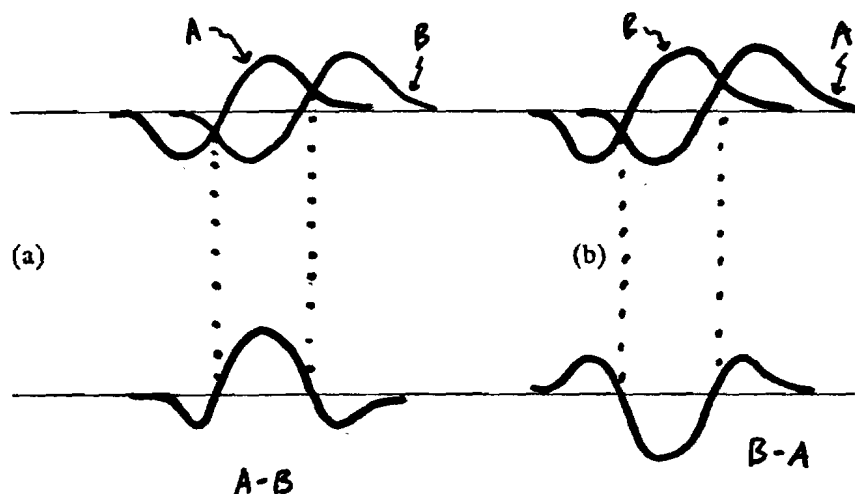
Once zero-crossings are found in both filtered images, to compare the difference in their locations (their relative position) between the two filtered images, corresponding points in the two frames must be matched, which introduces the correspondence problem. Alternatively, a mean zero-crossing location may be calculated for each filtered image and the difference of the means calculated. However this alternative introduces its own errors, as discussed with respect to the centroid in the previous section.

Filtered Differences

To locate N zero-crossings, N interpolations must be performed, which can be computationally expensive. Along the same lines of the above analysis we can use, not zero-crossings in and of themselves, but the *difference* between contours of the test and reference targets, in order to perform fewer interpolations.

Marr makes this point in Vision in discussing the neural implementation of stereo fusion, but does not expound upon its significance for hyperacuity. In this approach the signals to be combined by the difference operation originate in the test and reference targets, rather than in the left and right eyes. The two signals and their differences are illustrated schematically in FIGURE 9.

FIGURE 9.



The sign of the slope at the zero-crossings in the difference of two signals A and B uniquely determines the direction of position offset: right in (a) and left in (b).

From this diagram it is clear that the sign of the slope at either zero-crossing in the difference uniquely determines the direction of position change in the original signal (signal A). In Marr's terminology, this is equivalent to using spatial and temporal gradients to determine the direction of movement of a zero-crossing (see his Figure 3-33). This amounts to detecting a phase difference in the power spectra of the two images.

There are four possible assignments of zero-crossing location: the pixels to the left and the right of either (actual) zero-crossing, four altogether. In principle any one of these may be chosen; in practice the difference should be evaluated at that zero-crossing which will least degrade absolute position recovery and which requires the least searching. The important point is that only one zero-crossing need be found.

It should be noted that the phase of the original signals is important. If the signals occur too far apart their difference will be zero. If this is the case then Gaussian filters with a larger σ must be employed, which corresponds to searching for intensity differences over a larger spatial region. Once a filter with σ appropriate for the size of displacement is selected, then the phase difference cannot be zero.

Marr goes on to note that for too closely occurring zero-crossings or for very different contrasts in the two eyes, this mechanism can be unreliable. As our experiments show, however, zero-crossings probably do not occur too closely to invalidate the sign information, although such crowding may distort absolute position information. Contrast effects have not been explored.

Absolute position information may be encoded in one of two ways, either in the slope of the convolution signal at a particular zero-crossing, or in the height of the peaks and troughs. These features may be extracted directly, or some indirect measure, for instance the mean of the zero-crossings, may be used. Both approaches involve the introduction of further errors in position information; because we are primarily interested in relative position, neither approach is adopted here.

4.3. Comparison

For the centroid computation strategy there is little alternative to assigning position to a single high-precision numerical quantity. This represents the centroid as a statistical feature of the light distribution. This is clearly an efficient and convenient representation, whose form is dictated by the computation strategy.

For the interpolation strategy, position can be represented either by zero-crossings in the filtered images or simply by the difference between them. The former representation requires the detection of all zero-crossings, which requires interpolation, and that the difference in zero-crossing locations between the two filtered images be determined by matching. The latter representation of position requires a simple difference operation and the detection of a single zero-crossing in this difference and so is considerably more efficient and economical than the former, and is thus preferred.

In sum, the representations of position used by the two strategies are both high-precision numerical quantities. However, they stand for very different features. For the centroid, it is a statistical feature of the light distribution; for interpolation, it is the difference between two filtered images evaluated at a single zero-crossing.

4.4. Depth Position

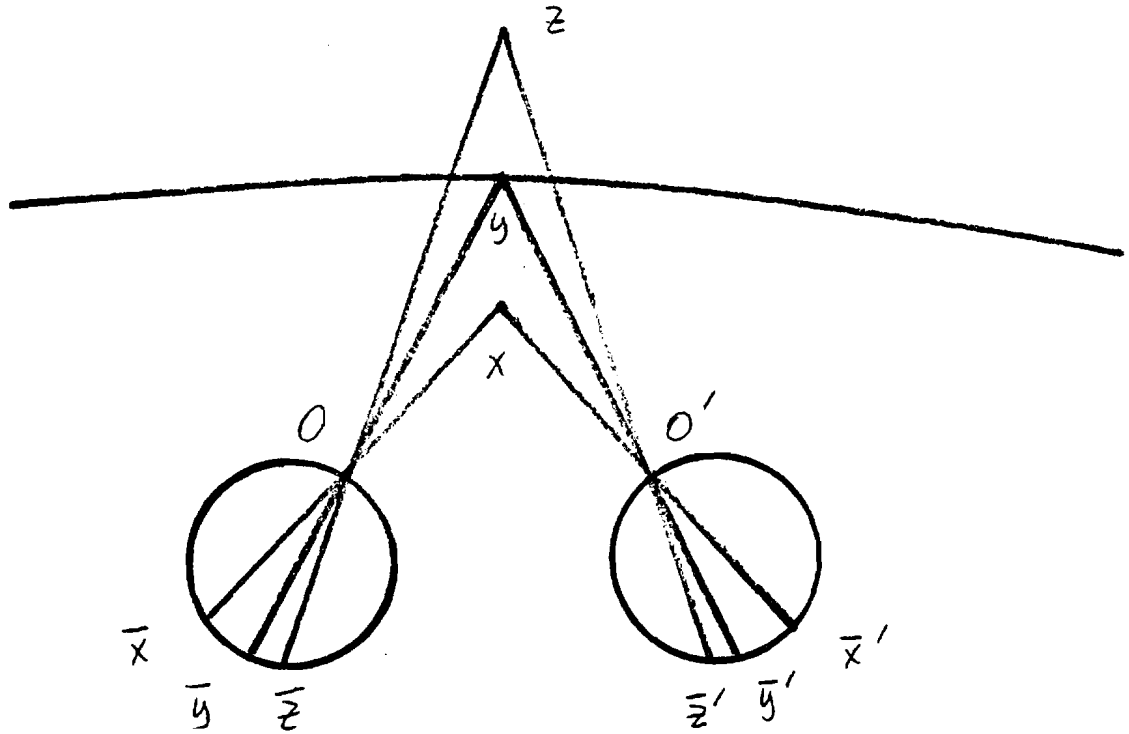
Now that several methods for representing position information in two dimensions have been identified, we shall consider how any one of these methods can be extended to represent locations in three dimensions. Again, because of the nature of hyperacuity tasks, we are concerned with representations which will allow judgments of *relative* depth.

Here attention is restricted to *binocular* depth cues, since other depth cues (interposition, accommodation, shading, etc.) do not rely on position information *per se*. Stereopsis, the perceptual process exploiting binocular information to determine the distance of points in the visual field to the observer, involves the detection of differences in the images recorded by the left and right eyes and using these differences to infer relative distance and surface orientation. These differences depend only upon position information; more precisely, such a difference will be called a *disparity*, which refers to an angular difference in position of a point imaged on the two eyes.

We shall consider a process akin, but not equivalent, to stereopsis. In this process the relative judgments of two-dimensional positions, rather than disparities as defined above, are

used to determine the sign of a change of position in depth--either towards or away from the observer, either nearer or farther.

FIGURE 10.



The Vieth-Muller horopter.

FIGURE 10 illustrates the geometrical construct called the (Vieth-Muller) horopter, which is useful for explaining singleness and doubleness of vision with two eyes. The points O and O' represent the optical nodes of the two eyes as well as their centers of rotation, and H represents the horopter circle, the locus of object points which lie at the intersection of two lines, one drawn from each retina through the nodal point. Muller maintained that singleness of vision existed only when an object lay on the horopter circle. It now appears that singleness of vision exists for points lying sufficiently (within Panum's fusional area) close to the horopter.

This construction of the horopter is oversimplified (because the optical nodes and centers of rotation may not coincide, and because the notion of corresponding retinal points is not precise) but for our purposes it is sufficient, for we are not explicitly concerned with

singleness or doubleness of vision. In Figure 6 point y is fixated, and the projection \bar{z} of z in the left eye is to the left of \bar{y} , and \hat{z} is to the right of \hat{y} . Similarly, \bar{x} is to the right of \bar{y} , and \hat{x} is to the left of \hat{y} . These differences in sign (leftness, rightness) are therefore in principle sufficient to determine relative depth.

5. Algorithms

5.1. Centroid

This computation is of the centroid of the light distribution as described in Section III.4.1. If $f(x,y)$ denotes the intensity at the receptor at (x,y) then the (i,j) th central moment is given by

$$m_{ij} = \sum_x \sum_y x^i y^j f(x,y)$$

and the centroid by

$$(\bar{x}, \bar{y}) = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}} \right)$$

where m_{00} is the total 'mass' in the distribution.

Due to the nature of the vernier target, the area of the image over which the centroid must be computed is consistent, i.e., it does not change with the nature of the local intensity distribution. This area meets the conditions specified in III.4.1.

Consider two distinct light distributions. These may be separate either in space, for example two abutting vertical lines, or in time, for example the same vertical line viewed on two different occasions. If two centroids $\alpha = (\bar{x}, \bar{y})$ and $\beta = (\hat{x}, \hat{y})$ are computed for the two distributions, then a judgment of their relative positions can be formed on the basis of the difference $\alpha - \beta$. Furthermore, the magnitude of this difference may accurately indicate the absolute spatial position offset.

5.2. Difference of Laplacians of Gaussians

This computation is of the sign of the difference of the convolution of the light distribu-

tion with the second derivative of a Gaussian evaluated at a particular zero-crossing. Again consider two distinct light distributions F_1 and F_2 . If the integral coordinate (x,y) is the first zero-crossing in a row, and $f_i(x,y)$ represents the intensity at the receptor at (x,y) , the sign of the difference

$$\nabla^2 G * f_1(x,y) - \nabla^2 G * f_2(x,y)$$

will determine the direction of position offset, as illustrated in FIGURE 9. Since $\nabla^2 G$ is a linear operator, a costly convolution can be saved by evaluating

$$\nabla^2 G * (f_1(x,y) - f_2(x,y))$$

A positive sign is interpreted as a shift to the right, and negative to the left. As previously noted, (x,y) may be chosen from four alternatives; the first occurrence is here selected to minimize searching. This choice may not allow the best recovery of absolute offset information.

5.3. Depth Determination

Let I_{ij} denote an image where $i \in \{\text{Left,Right}\}$ and $j \in \{1,2\}$ (time 1 and time 2, or target region 1 and target region 2). Let $S_i = \text{sgn}(I_{i1} - I_{i2})$, where the difference is computed by any one of the three methods detailed above.

Then the movement in depth is determined by TABLE 2, where the sign of a position to the left of the retinal position of the projection of a fixated point is by convention negative.

TABLE 2.

Sleft	Sright	movement
-	-	none
-	+	towards viewer
+	-	away from viewer
+	+	none

6. Hardware Implementation

We must now consider how the representations and algorithms discussed above might be physically realized in the information-processing device. The physical realizations are fully specified when the device employed is a digital computer. Consequently, the issue of interest is how the computations are implemented in the hardware of the human visual system. Here we shall provide a coarse and rather unoriginal treatment of this fascinating topic.

The neural image from the foveal region of the retina is represented by the outputs of X and Y ganglion cells, which are neurons. Impulses from these cells are transmitted by the optic nerves to the optic chiasma, where the optic nerves from the right and left eyes partially decussate. The impulses then travel along fibers in the optic tracts to the lateral geniculate nucleus (LGN); LGN cells project by the visual radiation to various layers in area 4C of the striate cortex (Brodmann's Area 17). Here, as in the LGN, there appears to be point-to-point correspondence between specific regions and specific areas in the retina.

6.1. Centroid

A specific mechanism for accomplishing the centroid computation can not yet be clearly identified, but it is clear that it would differ greatly from those proposed for interpolation operations. Knoblauch (1983) has suggested a possible receptive field organization based on a computation of the first moment divided by the area (see III.3.1), which relies on compressive transformations (logarithms) and lateral inhibition (for subtraction) to calculate the quotient.

The shape of the receptive field is given by the second derivative of a Gaussian (center-surround). Interestingly, this shape is quite broad spatially with respect to hyperacuity thresholds. A simulation shows that the receptive field produces a monotonic function of the position of the centroid.

6.2. Interpolation

A biological zero-crossing detector might not really detect the zeros of the convolution output, but could infer their presence and location from the activity occurring adjacently in

the image. So a neural implementation of zero-crossing detection may not yield a position measurement which corresponds precisely with the position of the ideal, theoretical zero-crossing.

Marr and Hildreth (1979) and Marr, Poggio and Ullman (1979) have proposed physiological schemes for how simple cells in the striate cortex may detect and represent oriented zero-crossing segments. Similar mechanisms may perform a simple difference operation upon the representation of the output of the DOG convolution.

Barlow (1979) and Crick (1980) suggested that, since there are 30 to 100 times as many granule cells per unit area in layer 4C β as there are terminating optic radiation fibers, a filtered version of the visual image passed from ganglion cells to LGN is reconstructed there in a fine-grained version of the original. According to this view, the representation of the visual data which is accessed by the process of hyperacuity is performed by granule cells in layer 4C β of the striate cortex.

Whether this finer position information is always explicitly represented or is computed only "on demand" (for example, using compiled "visual routines" as Ullman describes) is unclear. In either case, the question of how this information is computed is still central.

7. Summary

The perceptual process of hyperacuity has been treated as a computational problem. Two different approaches have emerged both of which provide explanations of how relative position judgments an order of magnitude more precise than absolute position judgments can be made. The two computations both transform two input images into an output vector representing their difference in relative position, which in a narrow sense is the spatial displacement of a test target relative to an axis defined in terms of the target.

The centroid computation is based on the theory of mean local sign, which essentially states that the center of gravity of a reasonable light distribution can be localized in units finer than the receptor mosaic. The representation and algorithmic computation of the cen-

troid are particularly simple, employing only simple arithmetic on high-precision numerical quantities. A hardware implementation of this approach has not been clearly envisioned. Overall, the centroid computation is simple and efficient, but depends critically upon the nature and sampling of the light distribution in the original images, especially edge blur.

The difference of gaussians computation is based upon interpolation, which essentially requires that the original images have their discretely sampled values 'close' together. This requirement can be met by most diffraction-limited visual systems, since the optics impose a bandlimit on the spatial frequency of the images. The assignment of position to edge features requires expensive filtering, but there are compelling arguments that the filtering is performed for other reasons as well. A hardware implementation of this computation is suggested, but can not be considered complete. Overall, the difference of gaussians computation is more expensive than the centroid, but is more robust with respect to the nature of the light distribution.

Both approaches can be easily extended from two to three dimensions by using a stereo pair of images. Thus relative 3D positions can be computed with the same accuracy as relative 2D positions.

This will suffice as a treatment of hyperacuity as a computational problem. Two theories have been advanced; their virtues may now be discriminated empirically.

IV. IMPLEMENTATION AND RESULTS

1. Implementation

Synthetic images of a vernier target are constructed, in which the reference target is not explicitly represented. In the first, the test target is represented as a bar whose edges are blurred by a function given by Gubisch's (1967) expression of the line-spread function of the human eye in terms of the sum of a Gaussian distribution and an exponential decay function:

$$f(x) = .47e^{-33x^2} + .53e^{-93|x|} \quad (1)$$

Associating one cone with one pixel and assuming a regular receptor distribution, this gives a spatial sampling rate of 30 sec arc/pixel, which is comparable to the spacing of cones in the fovea, and thus comparable to the sampling rate of the human eye.

In a second target image, only the tip of the test target is explicitly represented, as a "point" which is given by the point-spread function which is the two-dimensional extension of the line-spread:

$$f(x,y) = .47e^{-33(x^2+y^2)} + .53e^{-93(x^2+y^2)^{1/2}}$$

The spatial sampling rate is the same as above. Results are the same for the point target as for the bar target; for simplicity, we will discuss only the bar target.

There are obvious differences between this simulated data and the real human retina. The receptors in the simulation are square while cones are round, and the receptors are arranged in a highly regular square array, while the cones are arranged in a loosely structured not necessarily square array. Nevertheless, with respect to the computation of relative position, these differences are negligible, and present no obstacles to testing the two computational approaches.

An important issue is what method of fixation is employed, which poses the question of how the 'interesting' subimage is selected. The answer, in this implementation, is to employ *a priori* knowledge of the nature and orientation of the target: it is assumed that the target is two vertical line segments separated by at least one row of pixels. With this knowledge it is a

simple matter to find the two line segments and to establish a window on the picture in the area where the two lines abut. The same window is used for both algorithms, consistent with the conditions specified in III.4.

2. Methods

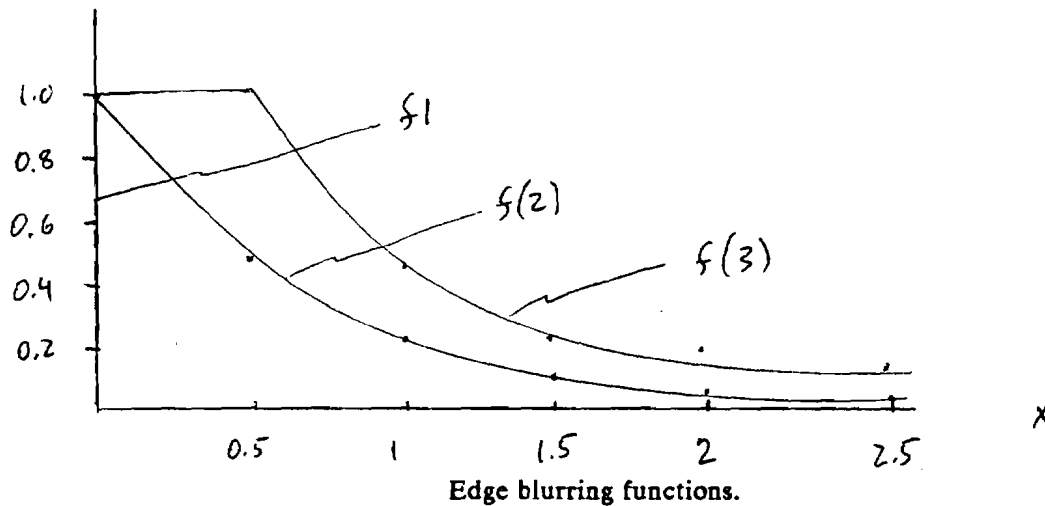
In addition to determining the overall accuracy of the two algorithms, the effect of target width and edge blur are investigated. It has repeatedly been shown that there is some type of interaction between 2 close contours (Flom (1963); Sullivan (1972); Westheimer and Hauske (1975); Marr (1982, p. 154)); target width is clearly an important variable. As shown in the discussion of the accuracy of the centroid computation, the degree of edge blur is also significant.

A digitized synthetic image of a vernier target is acquired, in which the diameter of the bar and the blur of the edges are variable. FIGURE 11 illustrates the different blurring functions used. The two algorithms for determining location are then executed, and the actual and computed position offsets are recorded. Actual position is known by construction of the target.

Several experiments have been conducted using real data. The importance of using real data is to demonstrate the practical feasibility of the suggested representation methods. Because of errors introduced by measurement, and noise introduced by illumination, a detailed analysis of these experiments is not presented. It will suffice to say that thresholds on the order of those recorded for synthetic data have been measured, and that it is unlikely that the results for natural data would differ significantly from those for simulated data except of course for repeatability.

3. Results

FIGURE 11.



Blurring functions:

$$f1(x) = \text{delta}(x)$$

$$f2(x) = .47e^{-3.3x^2} + .53e^{-.93|x|}$$

$$f3(x) = .47e^{-1.65x^2} + .53e^{-.465|x|}$$

Blur function f2 is the line spread function of the human eye (Equation 1). Blur function f3 is similar in structure to f2 but has much slower decay.

Centroid

As illustrated in TABLE 3, target width has no significant effect upon the accuracy of the computation of the centroid. Errors in the centroid computation do not vary significantly as target width is varied from 300 to 1.5 sec arc (10 pixels to 0.05 pixels).

Edge blur proves to be significant in the centroid computation; here it is evident that larger blur improves the accuracy of the computation. The data is tabulated in TABLE 4 and plotted in FIGURE 12.

The overall accuracy of the centroid computation is determined by using images with a large diameter and edge blur as in the human eye. This corresponds to column 2 of TABLE 3 and column three of TABLE 4, and is replotted as FIGURE 13. The overall accuracy of the centroid computation is on the order of 1 sec arc, since errors are considerably higher when smaller offsets are used (not plotted).

The centroid algorithm exhibits linear behavior, which is significant because it means

TABLE 3.

offset	300	200	100	50	10
0	0	0	0	0	0
6	0.001881	-.004469	-.004688	-.005512	.019512
12	0.008845	0.000336	0.002036	-.004326	.001958
18	0.001033	0.000786	0.002673	-.003523	-.00465
24	0.001903	0.003144	0.004406	0.004878	-.00468
30	0	0	0	0	0

Bar Width Table.

Offset and target width are measured in seconds of arc. Columns represent error in centroid computation given diameter of target, with edges blurred with point spread function of human eye. Error is the unnormalized quantity (actual-real), and is antisymmetric around offset of 30 sec arc. Error is 0 at this offset because the overestimates and underestimates in area underneath function exactly cancel (cf. FIGURE 7).

that absolute as well as relative position information can be extracted over the range of shifts in position up to 30 sec arc. Overall, it has been shown that in the ideal case of a simulation, the centroid provides enough information to make judgments of localization with an accuracy commensurate with that exhibited by the human visual system.

Filtered Differences

Varying target width and edge blur in the same ways as discussed above, the filtered difference method showed no variability in accuracy. The direction of motion was determined correctly for all offsets of magnitude greater than 1/1000 of a pixel, after which no motion at all is detected. Thus this difference operation is clearly capable of determining position changes far finer than those in the range of hyperacuity.

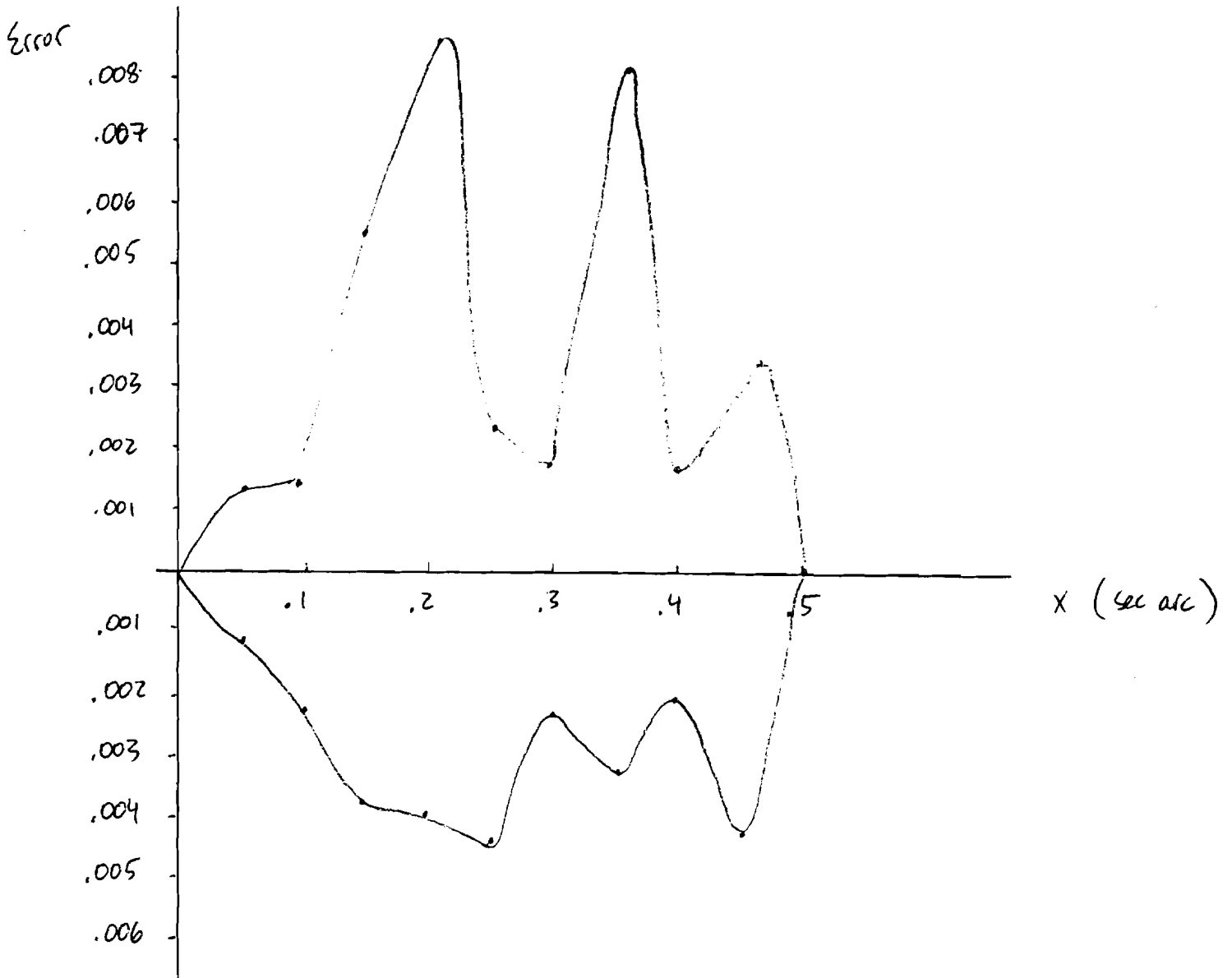
The behavior of the difference algorithm is more difficult to interpret when it is used to recover the absolute magnitude of the offset. The four different curves in FIGURE 14 represent the difference evaluated at the four possible positions which can be chosen as a zero-crossing on any row of the image. There are four possible choices because there are two changes in sign, and unless the change in sign falls precisely on the center of a pixel, there

TABLE 4.

offset	f1	f2	f3
0	0.0	0.0	0.0
6	0.40	0.001881	-.003660
12	0.30	0.008845	-.003963
18	0.20	0.001033	-.002588
24	0.10	0.001903	-.0023
30	0.0	0.0	0.0

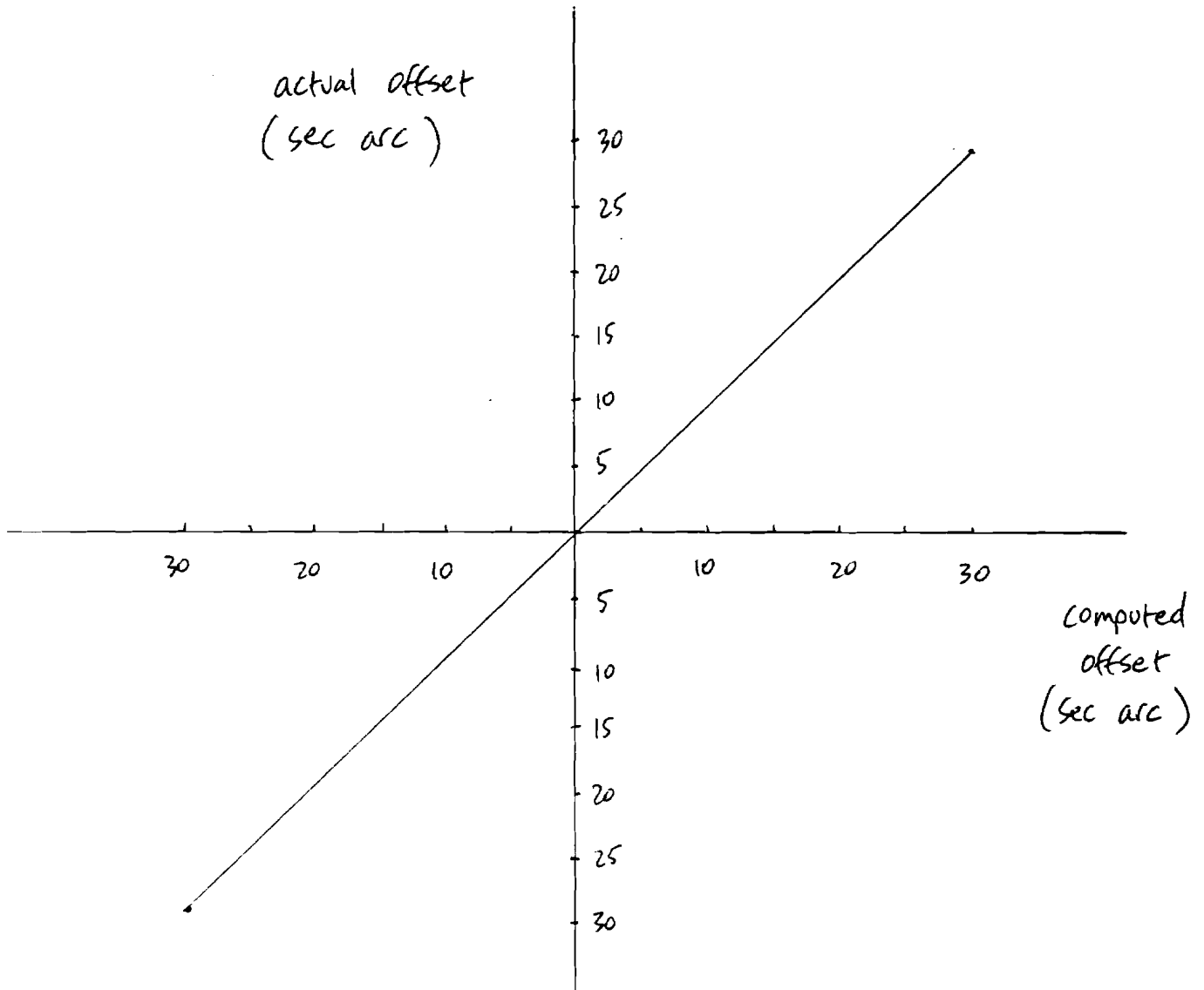
Offset is measured in seconds of arc. Errors calculated as in TABLE 3. Target is 300 sec arc wide. Functions defined in FIGURE 11.

FIGURE 12



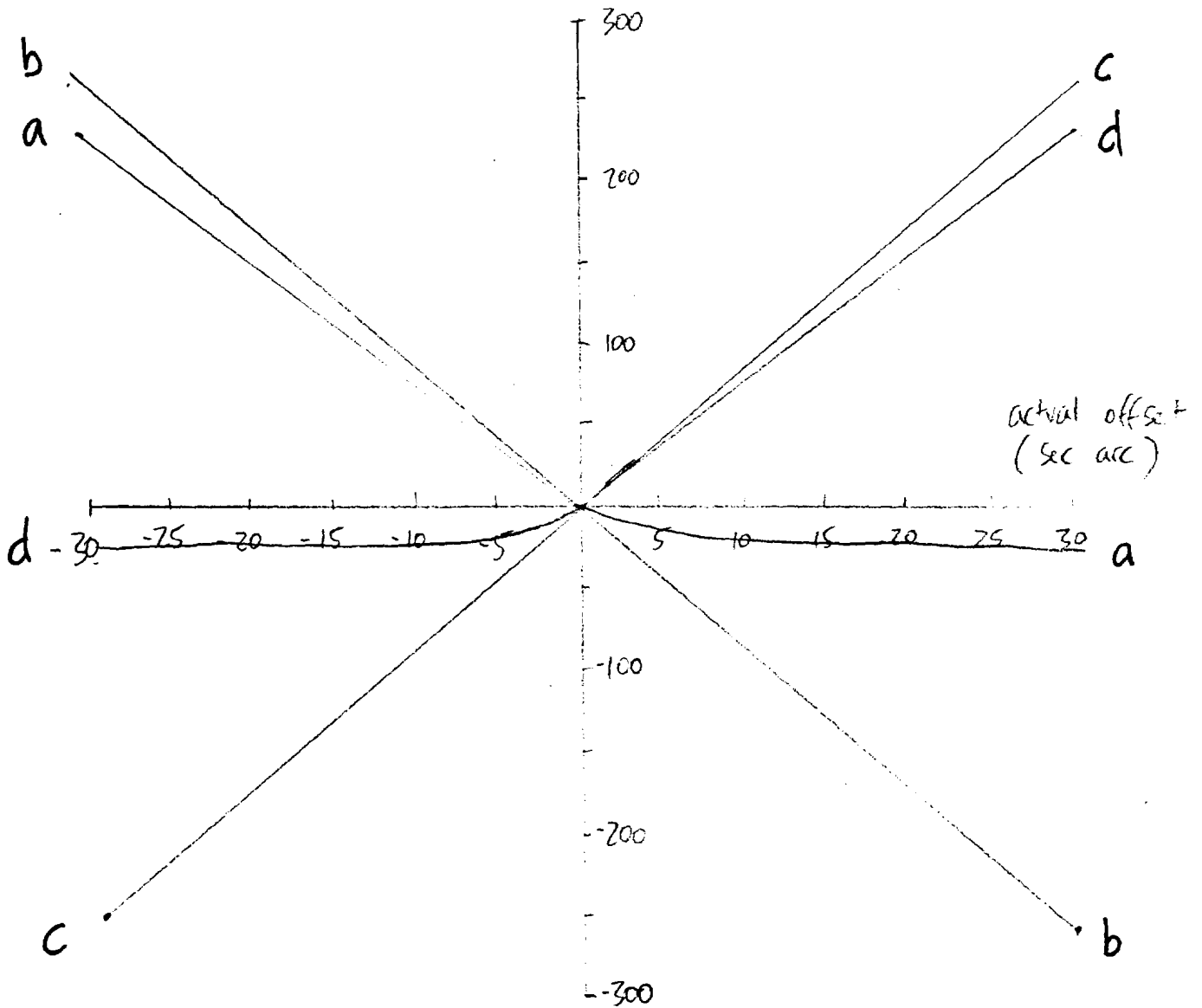
exist two pixels which fall on either side of the actual position of the change in sign.

FIGURE 13



As is evident from FIGURE 14 the best candidates for zero-crossings at which to evaluate the difference are the two which exhibit linear behavior over the range of shifts of -1 to 1 pixels (30 sec arc in either direction). The reason for the non-linear behavior of the other two is not immediately obvious. At any rate, accurate sign information is preserved and extracted by this operation, and magnitude information is available to an as yet unknown pre-

FIGURE 14.



Difference operation applied to line-spread simulation data. The four lines represent the difference evaluated at 4 different possible zero-crossing positions: a, b, c, and d are respectively the column positions in a given row in the difference matrix which correspond to the possible zero-crossing locations.

cision.

V. DISCUSSION

This section briefly summarizes the results, which are derived from a simulation of the human visual system subjected to stimuli of the type presented in vernier acuity tasks.

It has been established that in principle at least two different methods are capable of detecting and representing changes in position in the hyperacuity range of 2-5 sec arc. In the simulation they have indeed allowed accurate judgments of *changes* in position on this scale, but have also exhibited the capability to provide accurate judgments of the *magnitude* of the changes in position.

The centroid computation has a precision on the order of 1 sec arc (1/10 pixel), and is adversely affected by diminishing degree of edge blur, but unaffected by target width. To this extent, the nature of the intensity distribution is important, and the centroid computation is expected to be less robust in the face of changes in the quality of the image. It is possible to directly extract absolute as well as relative position information from the representation of the centroid, which provides economy of storage and time.

The filtered difference computation is more expensive than the centroid, but has a higher precision, on the order of .1 sec arc (1/100 pixel). This precision is an order of magnitude better than that provided by the human visual system, at least for vernier acuity tasks. Judgments of relative motion are unaffected by degree of edge blur and target width; judgments of absolute position do depend upon target width. Perhaps most significant is the discovery of the savings in complexity afforded by the filtered difference representation of edge features, without having to find zero-crossings. Our primary goal in further research in this area is to determine the accuracy of the recovery of absolute position information using filtered differences alone.

Many applications of the kinds of processing discussed are possible. Extraction of high-precision relative position information from relatively coarse data can be useful in graphics (to defeat aliasing), in tool control, manipulator positioning, stereo matching, analysis of aerial images, optical motion detectors, and many other tasks. The choice of which approach to take

will depend upon the speed and accuracy requirements of the task, but both can provide extremely fine spatial localization.

In conclusion, a computational treatment of the problem of extracting spatial position information with an accuracy far finer than that afforded by relatively blunt optical instruments, has shown that two different kinds of processing can in principle account for hyperacuity thresholds. It is still unknown how the human visual system performs with such efficiency and proficiency in these tasks, but it is the mystery of these small miracles which demands further research on the nature of the spatial sense of the eye.

NOTES

Anderson and Weymouth (1923), "Visual Perception and the Retinal Mosaic," *Am. J. Physiol.*, Vol. 64, ppl. 561-594.

Andrews, Butcher and Buckley (1973), "Acuities for spatial arrangement in line figures: human and ideal observers compared", *Vision Res.*, Vol. 13, pp. 599-620.

Barlow (1979), "Reconstructing the visual image in space and time," *Nature*, Vol. 279, pp. 189-190.

Berry (1948), "Quantitative relations among vernier, real depth and stereoscopic depth acuities", *J. Exp. Psychol.*, 38:708.

Bracewell (1978), *The Fourier Transform and Its Applications*, p.70.

Crick, Marr and Poggio (1980), "An Information Processing Approach to Understanding the Visual Cortex," MIT AI Memo No. 557.

Fahle and Poggio (1981), "Visual hyperacuity: spatio-temporal interpolation in human vision," *Proc. R. Soc. London, B* 213, pp. 451-477.

Flom, Weymouth and Kahneman (1963), "Visual resolution and contour interaction," *J. Opt. Soc. Am.*, Vol. 53, pp. 1026-1032.

Ford (1973), *Classical and Modern Physics*, Vol. 2, p. 941.

Gubisch (1967), "Optical performance of the human eye," *J. Opt. Soc. Am.*, Vol. 57, pp. 407-415.

Hildreth (1980), "Implementation of a Theory of Edge Detection," MIT AI-TR-579.

Knoblauch (1983), "Suggestions for simple neural models that extract centroids," personal communication.

Marr (1982), *Vision*, Freeman.

Marr and Hildreth (1980), Theory of Edge Detection, *Proc. R. Soc. London, B* 207, 1980, pp. 187-217.

Marr, Poggio and Hildreth, "Smallest channel in early human vision," *J. Opt. Soc. Am.*, Vol 70, No. 7, pp. 868-870.

Marr, Poggio and Ullman (1979), "Bandpass channels, zero-crossings, and early visual information processing," *J. Opt. Soc. Am.*, Vol 69, pp. 914-916.

Morgan (1980), "Analogue models of motion perception," *Phil. Trans. R. Soc. London*, B290, pp. 117-135.

Morgan and Watt (1982), "Mechanisms of interpolation in human spatial vision," *Nature*, 299, pp. 553-555.

Morgan, Watt and McKee (1983), "Exposure Duration Affects the Sensitivity of Vernier Acuity to Target Motion," *Vision Research*, Vol. 23, No. 5, pp. 541-546.

Poggio, Nishihara and Nielsen (1982), "Zero-crossings and Spatiotemporal Interpolation in Vision: aliasing and electrical coupling between sensors," MIT AI-Memo No. 557.

Pratt (1978), *Digital Image Processing*, Ch. 4, Wiley.

Riggs (1965), "Visual Acuity," in *Vision and Visual Perception*, ed. Graham, pp. 323-324.

Snyder and Miller (1977), "Photoreceptor diameter and spacing for highest resolving power," *J. Opt. Soc. Am.*, Vol. 67, pp. 696-697.

Sullivan, Oatley and Sutherland (1972), "Vernier acuity as affected by target length and separation," *Percept. Psychophysics*, Vol. 12, pp. 438-444.

Ullman (1983), "Visual Routines", MIT AI-Memo No. 723.

Watt and Andrews (1982), "Contour curvature analysis: Hyperacuties in the discrimination of detailed shape," *Vision Res.*, Vol. 22, pp. 449-460.

Watt and Morgan (1983), "Mechanisms Responsible for the Assessment of Visual Location: Theory and Evidence", *Vision Research*, Vol. 23, p. 97.

Westheimer (1975), "Visual Acuity and Hyperacuity", *Invest. Ophthal. Visual Sci.*, 14, pp. 570-572.

Westheimer (1976), "Diffraction theory and visual hyperacuity", *Am. J. Optom.*, Vol 53, p.362.

Westheimer (1979), "The spatial sense of the eye," *Invest. Ophthalmol. Visual Sci.*, pp. 893-912.

Westheimer and Hauske (1975), "Temporal and spatial interference with vernier acuity," *Vision Research*, Vol. 15, pp. 1137-1141.

Westheimer and McKee (1975), "Visual Acuity in the Presence of Retinal-Image Motion," *J. Opt. Soc. Am.*, Vol 65, No. 7, pp. 847-850.

Wilson and Bergen (1979), "A four mechanism model for spatial vision," *Vision Res.*, Vol. 19, pp. 19-32.