MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A. I. Memo No. 451 November 1977

A THEORY OF HUMAN STEREO VISION

by

D. Marr and T. Poggio**

SUMMARY. An algorithm is proposed for solving the stereoscopic matching problem. The algorithm consists of five steps: (1) Each image is filtered with bar masks of four sizes that vary with eccentricity; the equivalent filters are about one octave wide. (2) Zero-crossings of the mask values are localized, and positions that correspond to terminations are found; (3) For each mask size, matching takes place between pairs of zero-crossings or terminations of the same sign in the two images, for a range of disparities up to about the width of the mask's central region; (4) Wide masks can control vergence movements, thus causing small masks to come into correspondence; (5) When a correspondence is achieved, it is written into a dynamic buffer, called the $2\frac{1}{2}$-D sketch.

It is shown that this proposal provides a theoretical framework for most existing psychophysical and neurophysiological data about stereopsis. Several critical experimental predictions are also made, for instance about the size of Panum's area under various conditions. The results of such experiments would tell us whether, for example, cooperativity is necessary for the fusion process.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-75-C-0643.

**Max-Planck-Institut fur Biologische Kybernetik,
74 Tubingen 1, Spemannstrasse 38, Germany.

O Introduction

In a recent article, Marr & Poggio (1976) analyzed the computational structure of the stereo correspondence problem for stereo vision, and derived a cooperative algorithm for extracting disparity information from stereo image pairs. Although the problem addressed there was not directly related to the question of how our brains extract disparity information, the algorithm they described, summarized here in figure 2, has a natural interpretation in terms of neural structures.

One characteristic of this algorithm is its lack of dependence on eye-movements, so a critical preliminary question for its relevance to biology concerns the relative importance of neural fusion and of eye-movements for stereopsis (Marr & Poggio 1976). Various kinds of evidence suggest that eye-movements play an important role in stereo-vision, suggesting that a rather different kind of algorithm may be involved in human stereopsis.

In this article, we review the computational structure of the stereo disparity problem, and briefly outline existing approaches to solving it. We then review the available neurophysiological and psychophysiological evidence, and point to some of the empirical questions left unresolved in the literature. Finally, we formulate an algorithm designed specifically as a theory of the matching process in human stereopsis, and present a theoretical framework for the overall computational problem of stereopsis. We show that our theory accounts for most of the available evidence, formulate the predictions to which it leads, and describe some critical experiments. A computer implementation of the algorithm, and the results of some of these experiments, are described by Grimson & Marr (1978), and by Richards & Marr (1978).

1 Computational structure of the stereo-disparity problem

Because of the way our eyes are positioned and controlled, our brains usually receive similar images of a scene taken from two nearby points at the same horizontal level. If two objects are separated in depth from the viewer, the relative positions of their images will differ in the two eyes. Our brains are capable of measuring this disparity and of using it to estimate depth.

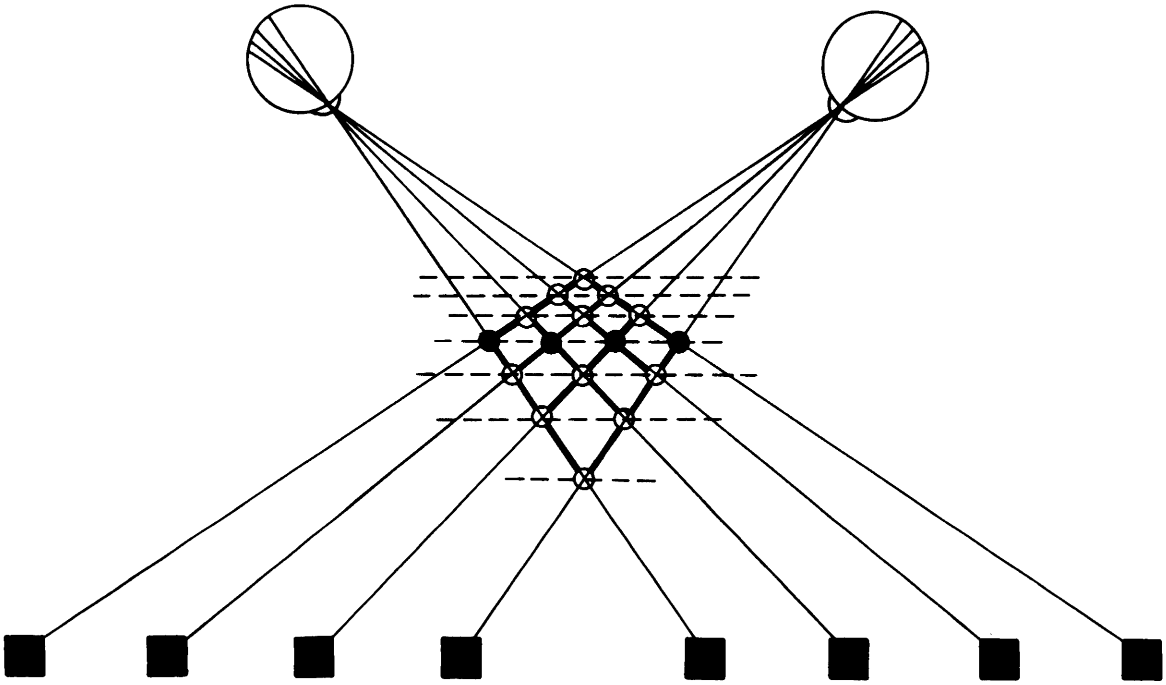
Three steps (S) are involved in measuring stereo disparity: (S1) a particular location on a surface in the scene must be selected from one image; (S2) that same location must be identified in the other image; and (S3) the disparity in the two corresponding image points must be measured.

If one could identify a location beyond doubt in the two images, for example by illuminating it with a spot of light, steps S1 and S2 could be avoided and the problem would be easy. In practice one cannot do this (figure 1), and the difficult part of the computation is solving the correspondence problem. Julesz found that we are able to interpret random-dot stereograms, which are stereo pairs that consist of random dots when viewed monocularly but fuse when viewed stereoscopically to yield patterns separated in depth. This might be thought surprising, because when one tries to set up a correspondence between two arrays of random dots, false targets arise in profusion (figure 1). Even so and in the absence of any monocular or high level cues, we are able to determine the correct correspondence.

In order to formulate the correspondence computation precisely, we have to examine its basis in the physical world. Two constraints (C) of importance may be identified (Marr 1974): (C1) a given point on a physical surface has a unique position in space at any one time: and (C2) matter is cohesive, it is separated into objects, and the surfaces of objects are generally smooth compared with their distance from the viewer.

These constraints apply to locations on a physical surface. Therefore, when we translate them into conditions on a computation we must ensure that the items to which they apply in the image are in one-to-one correspondence with well-defined locations on a physical surface. To do this, one must use image predicates that correspond to surface markings, discontinuities in the visible surfaces, shadows, and so forth, which in turn means using predicates that correspond to changes in intensity. One solution is to obtain a primitive description of the intensity changes present in each image, like the primal sketch (Marr 1976), and then to match these descriptions. Line and edge segments, blobs, termination points, and tokens, obtained from these by grouping, usually correspond to items that have a physical existence on a surface.

The stereo problem may thus be reduced to that of matching two primitive symbolic descriptions, one from each eye. One can think of the elements of these descriptions as carrying only position information, like the black dots in a random-dot stereogram, although



1. Ambiguity in the correspondence between the two retinal projections. In this figure, each of the four points in one eye's view could match any of the four projections in the other eye's view. Of the 16 possible matchings only four are correct (filled circles), while the remaining 12 are "false targets" (open circles). It is assumed here that the targets (filled squares) correspond to "matchable" descriptive elements obtained from the left and right images. Without further constraints based on global considerations, such ambiguities cannot be resolved. Redrawn from Julesz (1971, figure 4.5-1).

for a full image there will exist rules that specify which matches between descriptive elements are possible and which are not. The two physical constraints C1 and C2 can now be translated into two rules (R) for how the left and right descriptions are combined:

(R1) Uniqueness. Each item from each image may be assigned at most one disparity value. This condition relies on the assumption that an item corresponds to something that has a unique physical position.

(R2) Continuity. Disparity varies smoothly almost everywhere. This condition is a consequence of the cohesiveness of matter, and it states that only a small fraction of the area of an image is composed of boundaries that are discontinuous in depth.

In practise, R1 cannot be applied simply to grey-level points in an image, because, a grey-level point is in only implicit correspondence with a physical location. It is in fact impossible to ensure that a grey-level point in one image corresponds to exactly the same physical position as a grey-level point in the other. A sharp change in intensity, however, usually corresponds to a surface marking, and therefore defines a single physical position precisely. The positions of such changes may be detected by finding peaks in the first derivative of intensity, or zero-crossings in the second derivative.

2 Current approaches to the matching problem

One of the most significant advances in modern psychophysics was Julesz's (1960) invention of the random-dot stereogram. In addition to its various applications, this invention posed one of the few clear problems for neurophysiology and psychology, because it showed that stereoscopic fusion is a relatively early and independent computation.

Julesz's (1971) subsequent studies have helped to shape the character of experimental and theoretical approaches to this problem. One of his most influential suggestions has been the notion that the computation of stereo disparity depends on competing excitatory and inhibitory influences between nearby items with the same and different disparities (Julesz 1971, page 220 last paragraph). This suggestion arose out of his belief that binocular fusion is a *cooperative* process, a belief whose foundation we shall examine critically below. Apart from AUTOMAP (Julesz 1962), all the models we now examine were attempts at realizing this idea.

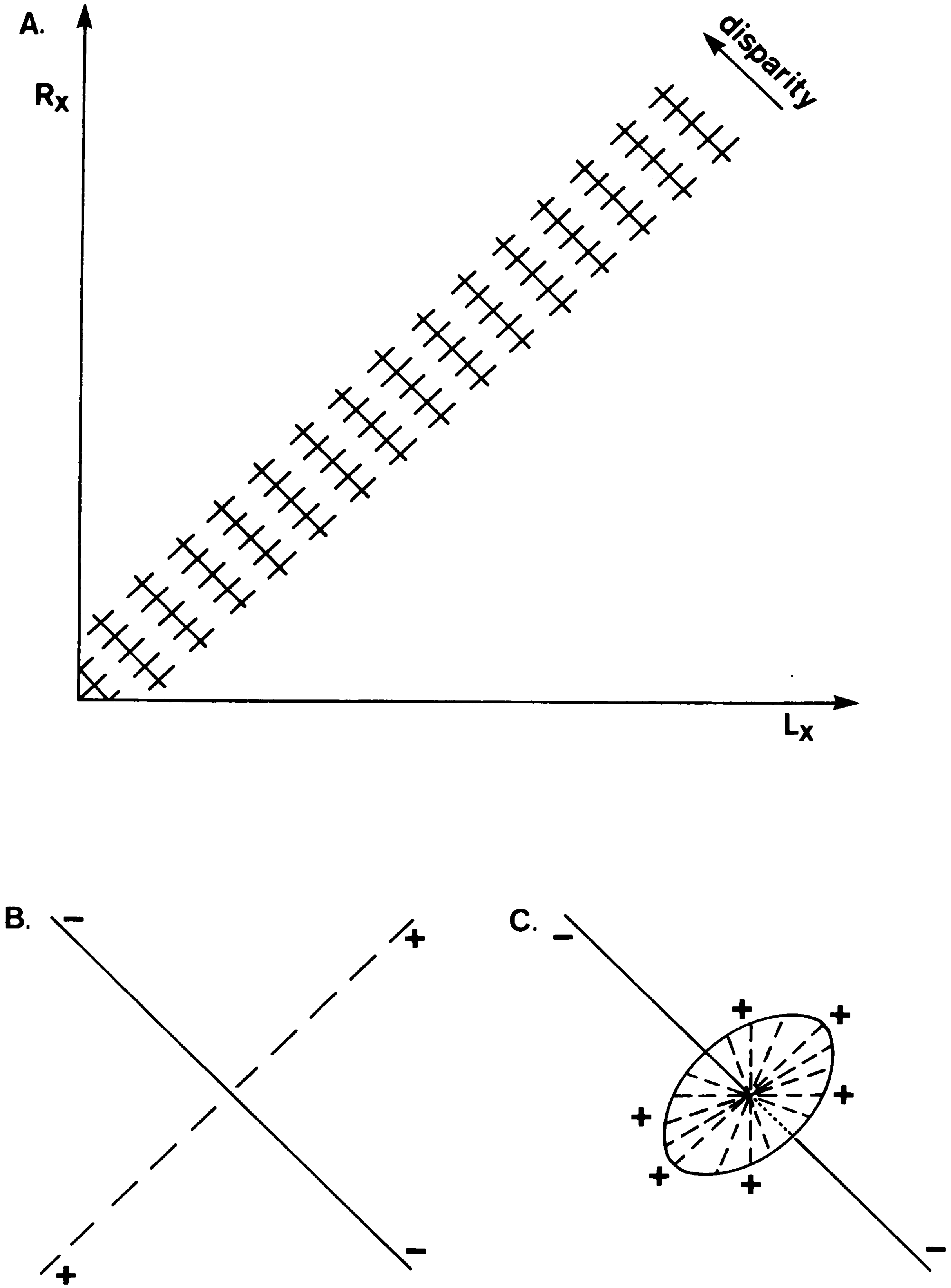
AUTOMAP (Julesz 1962) is a cluster-seeking program that operates in various layers shown in figure 2. Of the two rules formulated in the last section, it implements R2 (continuity) implicitly (because it detects only clusters), but it fails to implement R1. Hence in an ambiguous stereogram, both organizations will be detected simultaneously.

The dipole model (Julesz 1971, page 203ff, and Julesz & Chang 1976). Each position on each retina is associated with a magnetic dipole, whose polarity is determined (in the case of random-dot stereograms) by the retinal intensity value. Spring coupling between the tips of adjacent dipoles implements the continuity rule R2. The orientation of a dipole represents a disparity value, and the fact that each dipole can have only one orientation at a time provides an implementation of the uniqueness rule R1. Notice that unlike the other models we shall discuss, this one does not represent explicitly all possible states of figure 2, since each horizontal or vertical line in that figure corresponds to the angular range in position of a single dipole. Hence taken literally, this model would correspond to a scheme in which disparity at each position is signalled by the rate of firing of a single neuron and can therefore be thought of as a one-pool model. It would be interesting to see a computer implementation of such a model.

Sperling (1970). This model is based on correlation between two grey-level images (his eq. [7] p.471, and see also p.483 lines 9-12). Its approach is unsatisfactory for two reasons; firstly, as we have already seen, grey-levels are an inappropriate domain for the matching function, and secondly the area and disposition of the neighbourhoods over which the correlation is taken is crucial and left unspecified. Sperling's work does however make an interesting point of the connexion between stereopsis and vergence movements.

2. The explicit structure of the two rules R1 and R2 for the case of a one-dimensional image is represented in (a), which also shows the structure of a network for implementing the algorithm described by eq. 1. L_x and R_x represent the positions of descriptive elements in the left and right images. The continuous vertical and horizontal lines represent lines of sight from the left and the right eyes. Their intersection points correspond to possible disparity values. R1 states that only one match is allowed along any given horizontal or vertical line; R2 states that solution planes tend to spread along the dotted diagonal lines, which are lines of constant disparity.

In a network implementation of these rules, one can place a "cell" at each node; then solid lines represent "inhibitory" interactions, and dotted lines represent "excitatory" ones. The local structure at each node of the network in (a) is given in (b). This algorithm may be extended to two-dimensional images, in which case each node in the corresponding network has the local structure shown in (c). (From Marr & Poggio 1976 fig. 2).



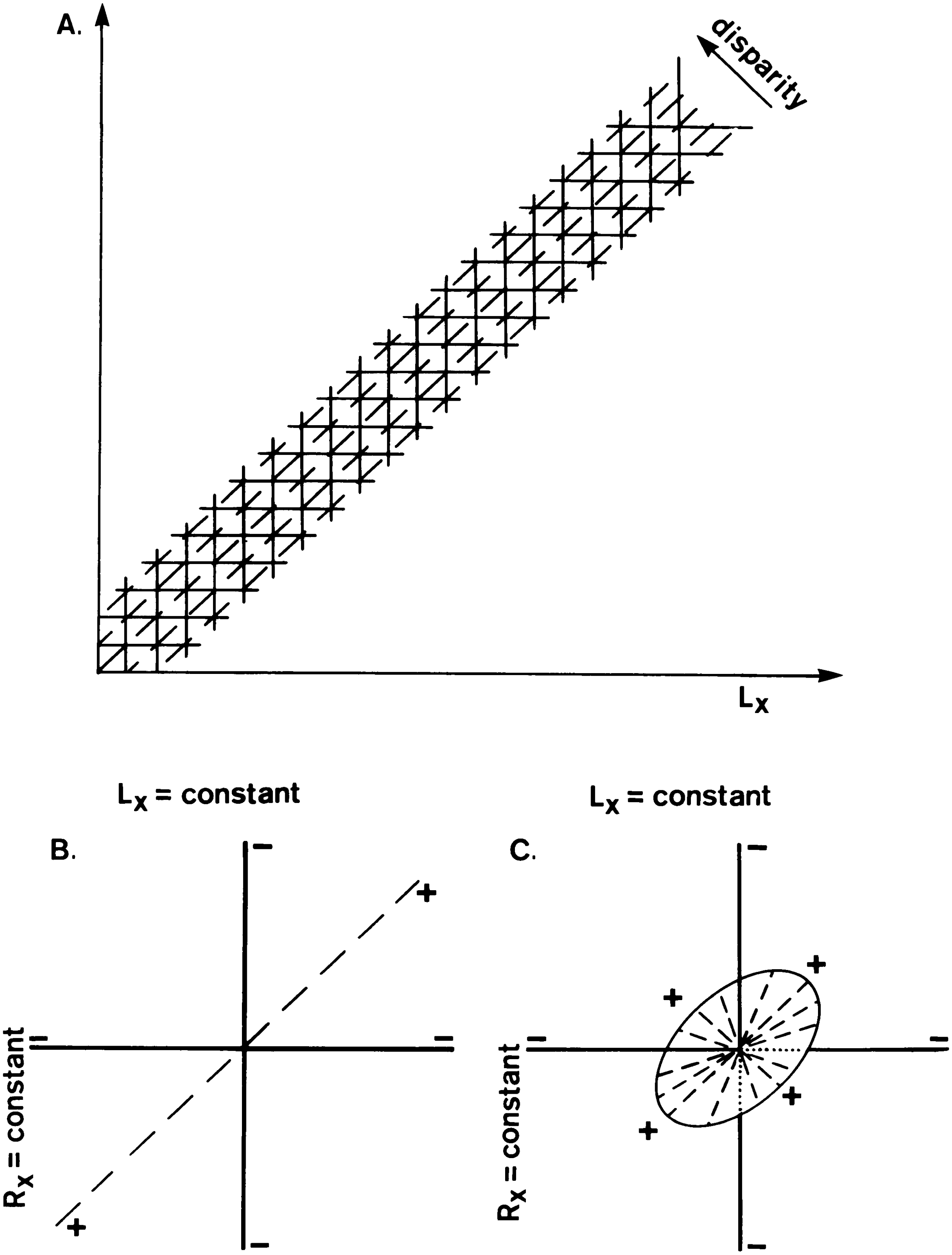
3. The uniqueness rule R1 gives rise to two sets of inhibitory interactions, along the lines of sight from each eye, as illustrated in figure 2. Several of the algorithms that are described in the text use inhibitory connexions like those illustrated here. Roughly speaking, these algorithms search for solutions that can be regarded as a single-valued, continuous vector field in a plane.

Nelson (1975) dilated on the ideas of Julesz, making two proposals which could implement our two rules. In terms of figure 2, the geometry of the lines of inhibition is left unclear, but it probably corresponds more to the inhibition shown in figure 3 than that of figure 2, and so does not precisely implement rule R1. Nelson gave no precise algorithm, nor did he implement any form of his ideas.

Dev (1975) was one of the first to formulate a precise algorithm that attempted to embody Julesz's ideas [Dev 1975, eq. (1) & (2), p. 515]. In terms of our rules, the algorithm realizes R2, but an incorrect version of R1 (see figure 3). Dev's algorithm is not cooperative, however, because it is linear (her equations 1 and 2). Dev writes (p. 526 lines 18-19) of applying a threshold to the results of the linear operation, but she did not say how to set such a threshold appropriately, and this is, of course, not a trivial problem[1].

Hirai & Fukushima (1976) constructed a neural model that correctly implemented the uniqueness rule R1 [their function (1) p 48], but did not implement rule R2, preferring instead a network that favoured solutions with lower parallax. This is an interesting idea, and a form of it plays a role in our theory (cf. figure 9).

Sugie & Suwa (1977) proposed a new and complex (non-linear and iterative) model that implements part of rule R2, but apparently uses


A.
disparity
L_X
L_X = constant
B.
R_X = constant
+
+
−
−
−
−
L_X = constant
C.
R_X = constant
+
+
+
+
−
−
−
−

the incorrect (figure 3) version of rule R1. The presence of an AND gate on their BN4 neurons (their fig. 4b) prevents their model from exhibiting "filling-in" phenomena, and calls into question the exact nature of the rule R2 that their network realizes.

Marr & Poggio (1976) formulated the iterative algorithm

$$C^{(t+1)}_{x,y;d} = \sigma\left\{ \sum_{x',y',d'\varepsilon S(x,y,d)} C^{(t)}_{x',y';d'} - \varepsilon \sum_{x',y',d'\varepsilon O(x,y,d)} C^{(t)}_{x',y';d'} + C^{(0)}_{x,y;d}\right\}$$

where $C_{x,y;d}$ denotes the state of the cell corresponding to position *(x,y)*, disparity *d* and time *t* in the network of figure 2; *S(x,y,d)* is a local excitatory neighborhood confined to the same disparity layer, and *O(x,y,d)*, the inhibitory neighborhood, consists of cells lying on the two lines of sight (figure 2c). ϵ is an inhibition constant, and σ is a threshold function. The initial state C^0 contains all possible matches, including false targets, within the prescribed disparity range. The rules R1 and R2 are implemented through the geometry of the inhibitory and excitatory neighborhoods *O* and *S* (figure 2c). This algorithm was shown to solve random-dot stereograms successfully (Marr & Poggio 1976 figures 3-6). In a mathematical analysis of the algorithm (Marr, Palm & Poggio 1978), it was demonstrated that states obeying the two rules R1 and R2 were stable states of the algorithm, and it was shown that, for a wide range of parameter values, the algorithm converges.

The nature of the input to which the algorithm is applied was not specified in detail. The success of the algorithm was demonstrated only for random-dot stereograms, where this problem does not arise. In addition, the algorithm has no dynamics in this form, and therefore exhibits no hysteresis.

3 Common characteristics of these algorithms

Apart from AUTOMAP and Sperling (1970), all of these algorithms are based on Julesz's proposal that fusion in human stereopsis is a cooperative process. An essential feature of these algorithms is that they are designed to select correct matches in a situation where false targets occur in profusion. That is, apart possibly from early versions of Julesz's dipole model, they do not critically rely on eye movements, since in principle, they have the ability to interpret a random-dot stereogram without them.

At the level of neurophysiology, these algorithms (with the exception of Julesz's dipole model, which we discussed above) all require many disparity "layers". This would imply (i) the existence of many "disparity-detecting" neurons, whose peak sensitivities cover a range of disparity values that is much wider than the tuning curves of the individual neurons, and which are rather insensitive to the nature of the descriptive element (*e.g.* edge, termination) to which they may refer; (ii) organization of these units into disparity layers (or stripes or columns); (iii) the presence of reciprocal excitation within each layer; and (iv) the presence of reciprocal inhibition between layers. For Marr & Poggio's algorithm, the inhibition should exhibit the characteristic "orthogonal" geometry of the thick lines in figure 2 (the lines of sight).

We turn now to an examination of the available empirical evidence.

4 Evidence from neurophysiology and psychophysics

4.1 Neurophysiology

The questions of interest to us can be formulated clearly, as follows: (i) Are there disparity detectors? (ii) If so, how finely tuned are they, and what range of disparities is covered by their peak sensitivities? For example, are there many, or are there just two or three (crossed, uncrossed and possibly zero disparity)? (iii) Are they organized into layers or columns of equal disparities? What are their excitatory or inhibitory relationships to one another? (iv) Are the disparity detectors sensitive to specific spatial features (*e.g.* oriented edges, oriented bars, or terminations)?

Let us now examine the evidence that is available about these four points.

(i) Although most physiologists believe that disparity detectors do exist, there is apparently some disagreement about the cortical area involved. Barlow, Blakemore & Pettigrew (1967) originally reported the existence of disparity sensitive units in the primary visual cortex of the cat, a finding substantiated by several subsequent articles, for example, Pettigrew, Nikara & Bishop (1977) and Nelson, Kato & Bishop (1977). Hubel & Wiesel (1970) failed to find depth-sensitive neurons in area 17 of the macaque monkey, though they did find them in area 18. They later stated that this situation is also true in the cat (Hubel & Wiesel 1973). Recently, Poggio & Fischer (1977) reported depth sensitive neurones in areas 17 and 18 of the alert macaque monkey. They were unable to offer an explanation for the difference between their's and Hubel & Wiesel's results, apart from the difference in the state of the animals. In studies on the cat, Hubel & Wiesel usually used barbiturates, whereas all the other investigations were carried out under nitrous oxide.

(ii) Barlow *et al.* (1967 figure 3) reported disparity sensitive cells in the cat that had a range of about 6.3 degrees at 5 - 15 degrees eccentricity. Pettigrew *et al.* (1968 figure 11) described cells at an eccentricity of 8 degrees tuned to a disparity of about 3 degrees (see also figure 9 of Nikara, Bishop & Pettigrew 1968). In the monkey, Poggio & Fischer (1977) found a range of *optimal* disparity sensitivity of for example ±0.3 degrees at 1 degree eccentricity (see their figure 8).

Little is certain about sharpness of disparity tuning. In the cat, figure 10 of Nelson *et al.* (1977) exhibits the response of a disparity-sensitive cell that is tuned to an unknown disparity value, with an accuracy of ±0.5 degrees. Bishop, Henry & Smith (1971 figure 6c) described a cell that was about twice as finely tuned. In the monkey, Poggio & Fischer (1977) found four types of depth-sensitive cell in areas 17 and 18; (a) cells excited by and narrowly tuned to stimuli at the depth of the plane of fixation; (b) cells whose response was essentially the complement of (a) (also described in the cat by Pettigrew *et al*. 1968 p.406); (c) *near* neurons, that were stimulated by stimuli in front of the fixation plane and were suppressed by those behind it; and (d) *far* neurons, the opposite of *near* ones. Some of the class (a) cells had a disparity tuning as sharp as 3', whereas some of the class (b) cells exhibited a total range of binocular interaction that could extend to more than ±1 degree of disparity.

There are hints of a monotonic relationship between eccentricity and optimal disparity (Poggio & Fischer 1977 figure 8), and between receptive field size and the sharpness of disparity tuning (Pettigrew *et al*. figure 11).

(iii) Hubel & Weisel (1970) in the macaque remarked that cells representing a given stereoscopic depth relative to the surface of fixation are grouped together, possibly into columns (p 41).

There is no evidence about the physiological connections among these cells. Almost all disparity-sensitive cells have, however, been reported to have inhibitory flanks for disparities lying outside the tuning range.

(iv) Little is known about the spatial features to which disparity neurons are sensitive, for instance, whether they have "bar-shaped" or "edge-shaped" receptive fields. Hubel & Wiesel (1970) imply that in the monkey, most binocular depth cells have vertically oriented, elongated receptive fields. Nelson *et al.* (1977) found that binocular neurons in the striate cortex of the cat are rather insensitive to differences in the orientations of slits or bars in the two eyes. They concluded that binocular units do not detect tilt directly.

Comments

Up to now, the questions we set out have not received direct attention, and the answers are at best uncertain, partly because of contradictory conclusions from different laboratories. Nevertheless, it is not clear why, for example, disparity tuning curves were not measured quite soon after Barlow *et al.'s* (1967) original work. Apparently, only Poggio & Fischer (1977) have provided careful evidence on this point, and it is unfortunate that their findings contradict the opinions implied by previous workers.

An important factor contributing to this state of affairs has been the considerable technical difficulty involved in experiments on stereoscopic disparity, for example, the precise control of eye position and stimulus eccentricity. The apparent need to use moving stimuli and slits that are more than a few minutes of arc wide make somewhat uncertain the interpretation of even those measurements that can be obtained. We believe that flashed stimuli, of the type used in psychophysics to avoid eye movements, may help in studies of these cells, despite the extra difficulties they introduce.

Taken altogether, the physiological evidence is not compelling about even the basic question ((ii) above) of whether there are two pools of disparity detectors, or several. The recent work of Poggio & Fischer seems to support Richards' (1970, 1971) two pools idea.

4.2 Psychophysics

It is impossible to give a brief review of the entire psychophysics of stereo vision, and the interested reader may turn to the book by Julesz (1971) or the review by Richards (1975), which also contain extensive bibliographies. In this article, we restrict our attention to points that we regard as important for our analysis of the computational structure of the stereo disparity problem. We divide our survey into four sections, each dealing with a different aspect of the problem.

The relevance of eye-movements.

As we stated earlier, one of the critical preliminary questions about the information processing structure of human stereopsis concerns the relative importance of neural fusion and of eye movements. Unfortunately, almost without exception (Fender & Julesz 1967, Evans & Clegg 1967, Richards 1977) all studies using random-dot stereograms proceeded by viewing the pairs with free eye movements (Julesz 1971) even though the smallness of Panum's fusional area (Fender & Julesz 1967) suggested that eye movements must be important.

Although some observers can see depth in simple random-dot stereograms that are presented in a flash or under stabilized image conditions, eye-movements (or the associated retinal motion) are essential for many observers for simple stereograms, and even then the perceived depth may be ambiguous or inappropriate (Richards 1977), except possibly in the disparity range 0-13' (Mayhew & Frisby 1978). For complex stereograms such as Julesz's spiral (1971 fig. 4.5-4), eye-movements are probably essential (Frisby & Clatworthy 1975, Saye & Frisby 1975).

Two pools or many disparity layers?

Richards (1970, 1971) and Richards & Regan (1973) proposed that the mechanisms underlying stereoscopic depth perception are organized into at least two pools, roughly corresponding to crossed and uncrossed disparities. Richards based this proposal on a study of "stereoanomalous" observers, who are able to process one of these kinds of disparities more strongly than other.

Although these data do not rule out the existence of many functional layers, they suggest the genetic importance of the two pools idea, and this itself hints at its functional implications.

Disparity detectors and spatial features

In the monocular situation, the visibility of a one-dimensional sinusoidal grating remains unchanged in the presence of masking noise filtered so as to contain no spectral components nearer than two octaves to the spatial frequency of the grating (Stromeyer & Julesz 1972). The equivalent finding holds also for two-dimensional patterns (Harmon & Julesz 1973). Kaufman (1964) and Julesz (1971, 3.9 & 3.10) found that one can simultaneously experience both binocular rivalry and fusion of different spectral components in a stereogram. Julesz & Miller (1975) recently put this finding on a quantitative basis. They selected masking noise bands, containing equally effective noise energy, such that their bands either overlapped the stereoscopic image spectrum or were two octaves distant. The first case resulted in rivalry, but in the second, stereoscopic fusion (and the consequent perception of depth) could be maintained despite the presence of strong binocular rivalry caused by the masking noise.

This raises the possibility that disparity information might at some stage be conveyed by independent stereopsis channels, tuned to different spatial frequencies, and roughly one octave wide. Mayhew & Frisby's (1976) interesting results using rivalrous texture stereograms are also consistent with this idea.

The available psychophysical evidence about the orientation sensitivity of these channels suggests that it is poor (Julesz 1971 p.

89), which is consistent with the neurophysiological findings we reviewed earlier (Nelson *et al*. 1977).

The channels found by Julesz & Miller are probably the same as those analyzed by several investigators (e.g. Campbell & Robson 1968). In spatial terms, such channels probably correspond to receptive fields that are bar-shaped rather than edge-shaped or more like gratings (see figure 5b in the next section).

Interestingly, it appears that line terminations can also be matched (Julesz 1971 p. 80, see also p. 92, Frisby & Julesz 1975). This raises the question of whether the matching of terminations relies on their explicit extraction from an image, or whether it is an epiphenomenon attributable to the existence of narrowly tuned frequency channels [cf. Cowan's (1977) discussion of Shapley & Tolhurst's (1973) results].

There is independent evidence that disparity mechanisms make bar-by-bar correlations as opposed to edge-by-edge correlations (Felton, Richards & Smith 1972). Using a modification of the Blakemore & Campbell (1969) adaptation technique, Felton *et al*. presented high-contrast sine-wave gratings binocularly at and off the plane of fixation. Under these conditions, the greatest rise in threshold following adaptation occurs for test gratings presented in the same plane as the adapting grating. They found that the adapted mechanisms have narrow rather than broad spatial-frequency tuning curves.

Another very important question which these results naturally raise is whether these independent spatial channels are also separated in the disparity domain, that is, whether they are further distinguished by the range of disparity values they can convey. Felton, Richards & Smith (1972) again provided evidence on this point, concluding that over the 1.0 degree disparity range they examined, narrow bar detectors feed small disparity mechanisms whereas wide bar detectors feed large disparity mechanisms. We shall propose later that wide bar detectors can in fact detect small disparities, but with poorer resolution than small bar detectors.

We feel that indirect evidence about this point may come from the differing reports of the size of Panum's fusional area. Fender & Julesz (1967) gave a figure of 6' for Panum's fusional area for random-dot stereograms with a dot size of about 2'. In subsequent experiments however, Julesz & Chang (1976) routinely flashed stereograms with a range of disparities, implying that those up to ±18' were fused. An attractive explanation for this discrepancy is that Panum's area depends on the dot size, since Julesz & Chang's were 6' square. This hypothesis is clearly in the same spirit as the conclusion of Felton, Richards & Smith.

Hysteresis and cooperativity

In a seminal paper[2], Fender & Julesz (1967) demonstrated the existence of hysteresis in stereopsis. They studied the fusion of binocularly stabilized random-dot stereoscopic images, and found that once fused (in the 6' Panum area), images could be pulled apart symmetrically by about 2 degrees in the horizontal direction without loss of stereopsis or fusion. This finding provided the basic reason for a widespread belief that binocular fusion is a cooperative process. Later, Julesz adduced in its further support the phenomena of (i) disorder-order transitions and multiple stable states in stereopsis (Julesz and Chang 1976 p. 117), (ii) the pulling effect with ambiguous random-dot stereograms (Julesz & Chang 1976), and (iii) Julesz's conclusion (1971 p. 200) that "stereopsis is a parallel process in which each depth plane is simultaneously processed." The idea that stereopsis is cooperative formed the starting point for all attempts at constructing neural models for this computation (*cf.* section 1).

In their original paper, Fender & Julesz concluded that the labelling of corresponding points can occur only within Panum's fusional region, but that under appropriate conditions, these labels could then be preserved for large retinal image shifts. Their original data, however, only indicated the presence of a "simple memory process" which they then chose (p. 829) to call hysteresis. There is no evidence that this hysteresis is intrinsic to the labelling process itself, an hypothesis which is essential if stereopsis is to be regarded as a simple cooperative phenomenon. On the contrary, the phenomenon of hysteresis appears over a disparity range of 2 degrees, which is much greater than even the largest estimates of Panum's fusional area. Futhermore, Fender & Julesz (p. 829) actually "suggest the existence of three different processes in stereopsis. A labelling process, which is operative in Panum's fusional region, establishes correlation between corresponding areas in the left and right images having various disparities. A cortical-registration process preserves these labels even if the left and right images are pulled apart on the retinas... [and] convergence motion of the eyes, which compensate for large or rapid errors of disparity."

It seems to us that, while the notion of hysteresis can certainly be applied to the registration process, because it is essentially a memory, there is no direct evidence for cooperativity in the labelling process itself.

4.3 Conclusions

Many of these findings cast doubt on the relevance of cooperative algorithms to the question of the fusion process in human stereo vision. The principal points are (a) the apparently crucial role played by eye-movements in human stereo vision, (b) the ability of some subjects to tolerate a 15% expansion of one image (Julesz 1971 figure 2.8-8), (c) the findings about independent spatial-frequency-tuned channels in binocular fusion, of which our tolerance to severe defocussing of one image is a striking demonstration (Julesz 1971 figure 3.10-3), (d) the physiological, clinical and psychophysical evidence about Richards' three-pools hypothesis, and (e) the size of Panum's fusional area (6' - 18') which seems surprisingly small to have to resort to cooperative mechanisms of neural fusion for the elimination of false targets.

Finally, we may mention that none of the theories based on cooperativity gives a clear indication of the nature of the spatial features that should be matched.

5 A theory of stereopsis

Taken together, these findings indicate that an approach of a quite different kind to the problem is probably necessary. In this section, we present an alternative theory, describing firstly a rough outline and the ideas that led us to it, and after that we formulate the theory in detail.

5.1 An outline of the theory

The basic computational problem in binocular fusion is the elimination of false targets, and the difficulty of this problem is in direct proportion to the range and resolution of the disparities that are considered. The problem can therefore be simplified by reducing either the range, or the resolution, or both, of the disparity measurements that are taken from two images. An extreme example of the first strategy would lead to a diagram like figure 2 in which only three adjacent disparity planes were present (e.g. +1, 0, -1) each specifying their degree of disparity rather precisely. The second strategy, on the other hand, would amount to maintaining the range of disparities shown in figure 2, but reducing the resolution with which they are represented. In the extreme case, only three disparity values would be represented, crossed, roughly zero, and uncrossed.

These schemes, based on just three pools of disparity values, substantially eliminate the false targets problem at the cost on the one hand of a very small disparity range, and on the other, of poor disparity resolution. Thus the price of computational simplicity is a trade-off between range and resolution.

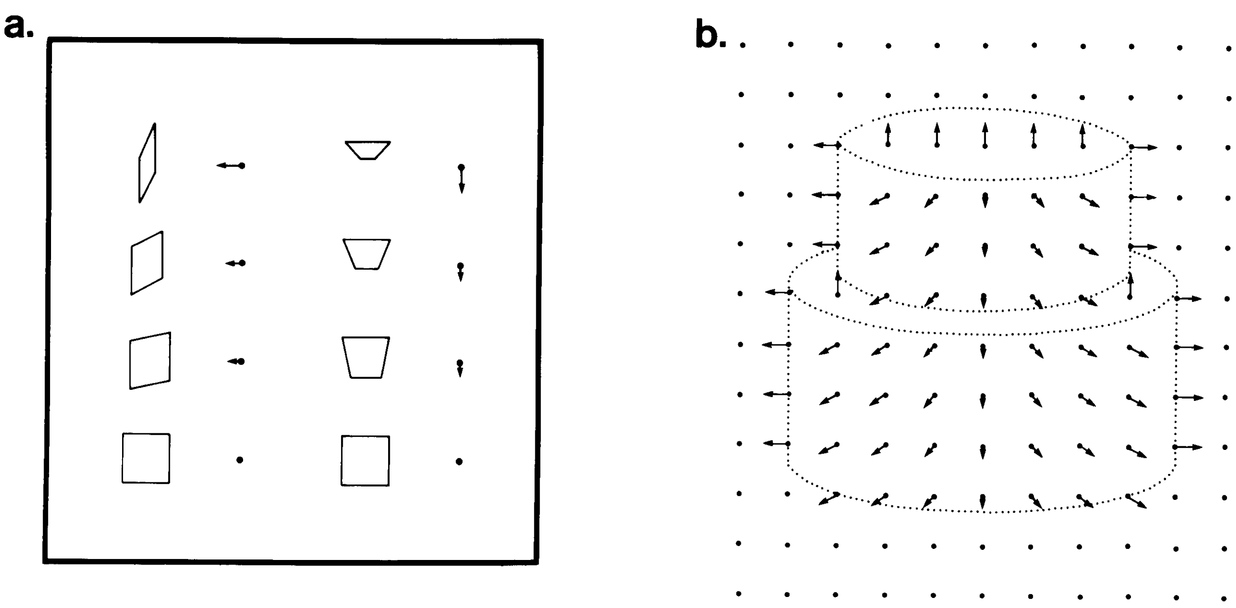
One would, however, expect the human visual system to possess both range and resolution in its disparity processing. In this connection, the existence of independent spatial-frequency-tuned channels in binocular fusion is of especial interest, because it suggests that several copies of the image, obtained by successively finer filtering, are used during fusion, providing increasing and, in the limit, very fine disparity resolution at the cost of decreasing disparity range.

A notable feature of a system organized along these lines is its reliance on eye-movements for building up a comprehensive and accurate disparity map from two viewpoints. The reason for this is that the most precise disparity values are obtainable from the high-resolution channels, and eye-movements are therefore essential so that each part of a scene can ultimately be brought into the small disparity range within which high resolution channels operate. The importance of vergence eye-movements is especially attractive in view of the recent evidence about their role in human stereopsis (see section 4.2 above), and the extremely high degree of precision with which they may be controlled (Riggs & Neihl 1960, Rashbass & Westheimer 1961a).

These observations suggest a scheme for solving the fusion problem in the following way: (1) Each image is analyzed through channels of various coarsenesses, and matching takes place between corresponding channels from the two eyes for disparity values of the order of the channel resolution. (2) Coarse channels control vergence movements, thus causing finer channels to come into corresondence.

This scheme raises a puzzle. Since it contains no hysteresis, it provides no explanation for the basic findings that led Julesz to conclude that binocular fusion is a cooperative process. Recent work in the theory of intermediate visual information processing argues on computational grounds that a key goal of early visual processing is the construction of something like a "depth map" of the visible surfaces round a viewer, (Marr & Nishihara 1977 figure 2, Marr 1977 section 3). The motivation for this proposal is that a description of objects' shapes has to be derived *via* a description of their visible surfaces, and information about these is obtainable by a number of different and probably independent processes, which extract disparity, motion, shading, texture gradient and contour information. These different types of information need to be combined, in a buffer somewhere. One proposal for carrying this out is the construction of a representation that makes explicit the depth and orientation of visible surface elements, and contours of surface discontinuity, in a coordinate frame that is centered on the viewer (Marr 1977 Table 2). Marr & Nishihara called this representation the $2\frac{1}{2}$-D sketch (see figure 4).

The important point here is that the $2\frac{1}{2}$-D sketch is in some sense a *memory*, and it is this idea, together with the remarks of Fender & Julesz that we quoted above, that offers a possible solution to our



4. The 2-½-D sketch represents depth, surface orientation and contours of discontinuities in these quantities. A convenient representation of surface orientation is illustrated in (a). The orientation of the needles is determined by the projection of the surface normal on the image plane, and the length of the needles represents the dip out of that plane. A typical 2-½-D sketch appears in (b), although depth information is not represented in the figure.

puzzle. Suppose that the hysteresis Fender & Julesz observed is not due to a cooperative process during fusion, but is in fact the result of using a memory buffer in which to store the depth map of the image as it is discovered. Then, the fusion process itself need not be cooperative (even if it still could be), and in fact it would not even be necessary for the whole image ever to be fused simultaneously, provided that a depth map of the viewed surface were built and maintained in this intermediate memory.

Our scheme can now be completed by adding to it the following two steps: (3) when a correspondence is achieved, it is held and written down somewhere (e.g., in the $2\frac{1}{2}$-D sketch); (4) there is a backwards relation between the memory and the masks, perhaps simply through the control of eye-movements, that allows one to fuse any piece of a surface easily once its depth map has been established in the memory.

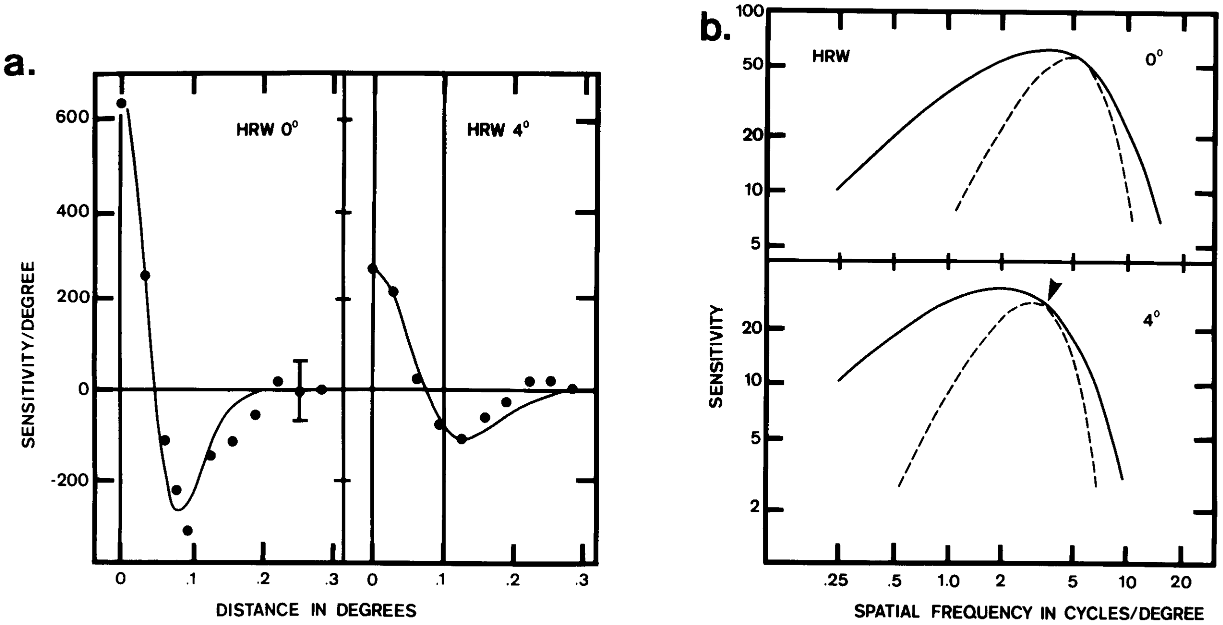
We turn now to a more detailed analysis of these ideas.

5.2 The nature of the channels

The articles by Julesz & Miller (1975) and Mayhew & Frisby (1976) establish that spatial-frequency-tuned channels are used in stereopsis and are independent. Julesz & Miller's findings imply that two octaves is an upper bound for the bandwidth of these channels, and suggest that they are the same channels as those previously found in monocular studies (Campbell & Robson 1968, Blakemore & Campbell 1969). Although strictly speaking it has not been demonstrated that these two kinds of channel are the same, we shall make the assumption that they are. This will allow us to use the numerical information available from monocular studies to derive quantitative estimates of some of the parameters involved in our theory.

The idea that there may be a range of different size or spatial frequency tuned mechanisms was originally introduced on the basis of psychophysical evidence by Campbell & Robson (1968). This led to a virtual explosion of papers dealing with spatial frequency analysis in the visual system. Recently, Wilson & Gieze (1977) and Cowan (1977) integrated these and other anatomical and physiological data into a coherent logical framework. The key to their framework is (a) the partitioning of the range of sizes associated with the channels into two components, one due to spatial inhomogeneity of the retina, and one due to local scatter of receptive field sizes; (b) the correlation of these two components with anatomical and physiological data about the scatter of receptive field sizes and their dependence on eccentricity.

On the basis of detection studies, they formulated an initial model embodying the following conclusions: (1) at each position in the visual field, there exist "bar-like" masks (see figure 5a), where tuning curves have the form of figure 5b, and which have a half-power bandwidth of about an octave. (2) The bandwidth of the local sensitivity function at each eccentricity is about three octaves. Hence the range of receptive field sizes present at each eccentricity is about 4:1. In other words, at least three and probably four

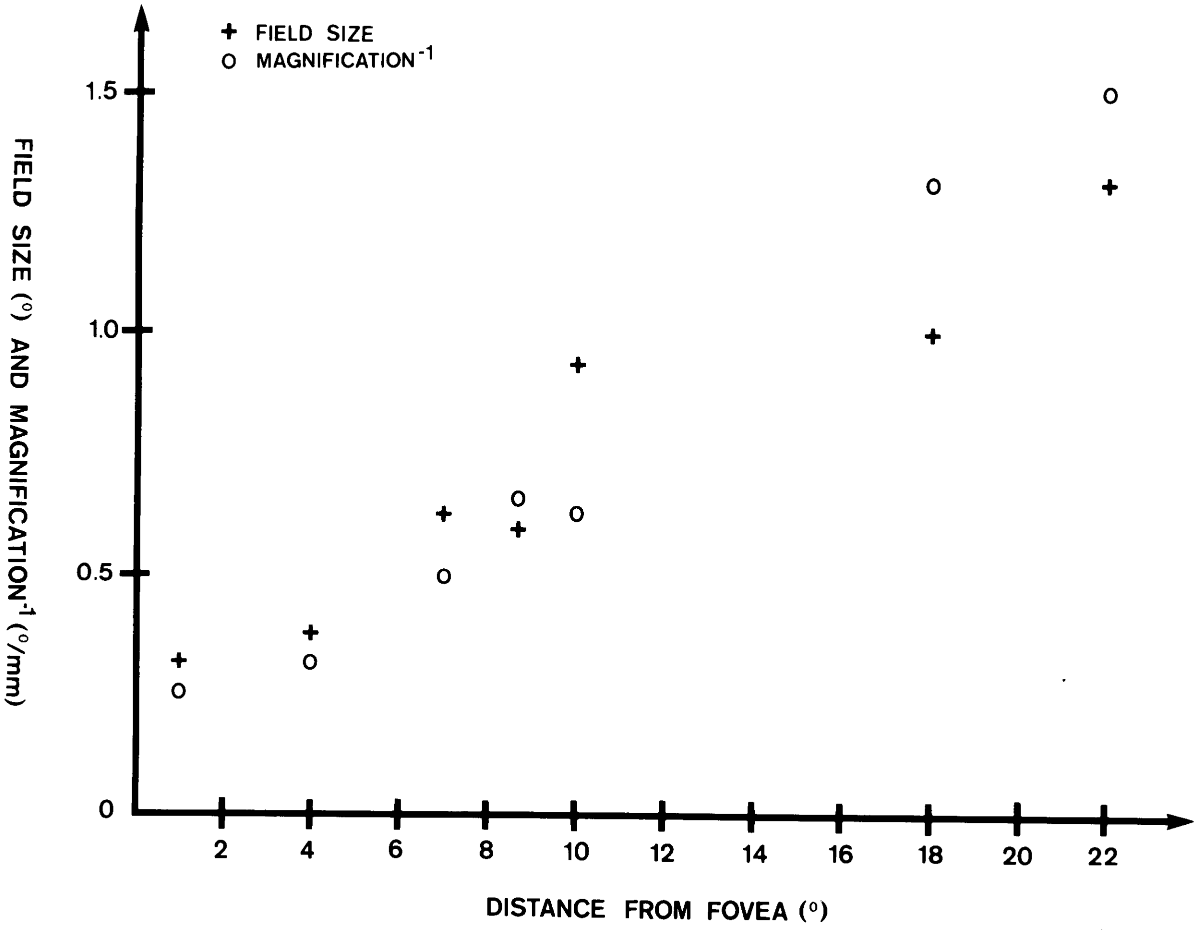


5. (a) Line spread functions measured at two different eccentricities for HRW. The points are fitted using the difference of two Gaussian functions with space constants in the ratio 1.5:1.0. The inhibitory surround exactly balances the excitatory centre so that the area under the curve is zero.

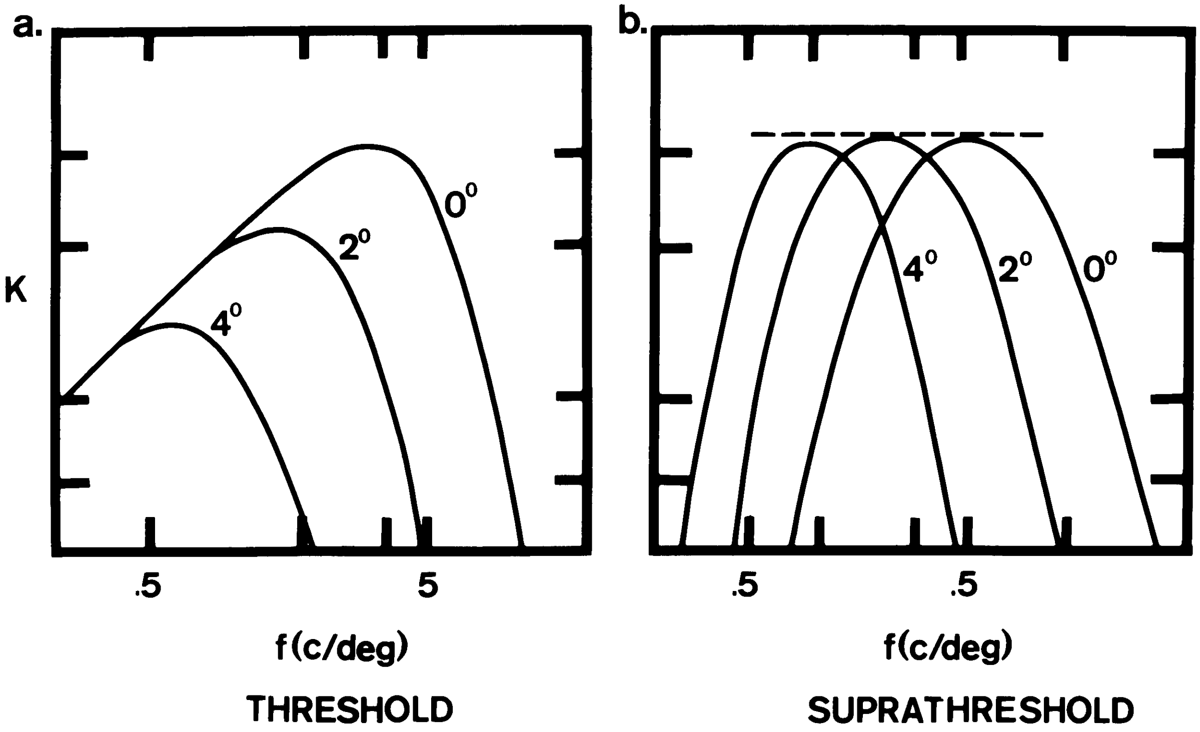
(b) Predictions of local spatial frequency sensitivity from frequency gradient data and from line spread function data. The local frequency sensitivity functions are plotted as solid lines. The dashed lines are the local frequency response predicted by Fourier transforming the line spread functions in (a), which were measured at the appropriate eccentricities. The arrow in the lower graph indicates a translation of the dashed curve by approximately 1.08 log$_{10}$ units. (Redrawn from Wilson & Gieze 1977 figs. 9 & 10).

receptive field sizes are required at each point of the visual field. (3) Average receptive field size increases linearly with eccentricity. In humans at 0 degrees, the mean width w of the central excitatory region of the mask is about 6', (range 3' to 12'); and at 4 degrees eccentricity, w = 12' (range 6' to 24'), (Wilson & Gieze figure 9, Hines 1977 figures 2 & 3). If one assumes that this receptive field is described by the difference of two gaussian functions with space constants in the ratio 1:1.5, the corresponding peak frequency sensitivity of the corresponding channel is given by $1/f = \lambda = 2.3w$. These figures agree quite well with physiological studies in the Macaque. Hubel & Wiesel (1974) reported that the mean width of the receptive field (s) increases linearly with eccentricity e (figure 6) (approximately, $s = 0.05e + 0.25$ degrees, so that at $e = 4$ degrees, s = 27' which gives a value for $w = s/3$ of about 9' as opposed to 12' in humans). The data of Schiller (1977 p. 1347 figures 12 & 14) are in rough agreement with Hubel & Wiesel's. (4) Essentially all of the psychophysical data on the detection of spatial patterns at contrast theshold can be explained by (1), (2) and (3) together with the hypothesis that the detection process is based on a form of spatial probability summation in the channels.

With the characteristic perverseness of the natural world, this happy and concise state of affairs does not provide a precise account of suprathreshold conditions (see figure 7). The known discrepancies can however be explained by introducing two extra hypotheses: (5) Contrast sensitivities of the various channels are adjusted



6. Graph of average receptive field size (crosses) and magnification^{-1} (open circles) against eccentricity, for five cortical locations. Points for 4, 8, 18 and 22 degrees were from one monkey; for 1 degree, from a second. Field size was determined by averaging the fields at each eccentricity, estimating size from (length x width)$^{0.5}$. (Redrawn from Hubel & Wiesel 1974 fig. 6a).



7. Two many-channel schemes, (a) at threshold, in which the receptive-field surrounds are constant, and (b) suprathreshold, in which both centres and surrounds scale linearly, and sensitivity remains constant. (Redrawn from Cowan 1977 fig. 12).

appropriately to the stimulus contrast (Georgeson & Sullivan 1975). The point of this is merely to ensure that bars of the same contrast but different widths actually appear to have the same contrast; (6) Receptive field properties change slightly with contrast, the inhibition being somewhat decreased in low-contrast situations (Cowan 1977 p. 511).

In a more recent article, Wilson & Bergen (1978) have found that the situation at threshold may also be more complicated. They proposed a model consisting of four size-tuned mechanisms centred at each point, the smaller two showing relatively sustained temporal responses, and the larger two being relatively transient. As far as is known, this model accurately accounts for all published threshold sensitivity studies.

The two sustained channels, which Wilson & Bergen call N and S, have *w* values 3.1' and 6.2'; the transient channels, called T and U, have *w*'s of 11.7' and 21'. The sizes of these channels increase with eccentricity in the same way as described above.

The S channel is the most sensitive under both transient and sustained stimulation, and the U channel is the least, having only 1/11 to 1/4 the sensitivity of the S channel. The extent to which the U channel, for example, plays a role in stereopsis is of course unknown.

In what follows, we shall assume that the figures given earlier for the numbers and dimensions of receptive field centres and their scatter hold roughly for suprathreshold conditions. If future experiments confirm that these more recent numbers are relevant for stereopsis, some modification of our quantitative estimates may be necessary.

These figures allow us to estimate the minimum sampling density required by each channel, *i.e.* the minimum spatial density of the corresponding receptive fields. From figure 10 of Wilson & Gieze (1977), a channel with peak sensitivity at wavelength λ is band-limited on the high-frequency side by wavelengths of about $2\lambda/3$. This figure is for a threshold criterion of 15-30%, but is rather insensitive to the exact value chosen. Hence by the Sampling Theorem (Papoulis 1968, p. 119), the minimum distance between samples, (*i.e.* receptive fields), in a direction perpendicular to their preferred orientation, is at most $\lambda/3$. Assuming the overall width of the receptive field is about $3\lambda/2$, the minimum number of samples per receptive field width is about 4.5.

An estimate of the minimum longitudinal sampling distance may be obtained as follows. Assume that the receptive field's longitudinal weighting function (see table 1) is gaussian with space-constant σ, thus extending over an effective distance of say 4σ to 6σ. Its fourier transform is also gaussian with space constant in the frequency domain (ω) of $1/\sigma$, and for practical purposes can be assumed to be band-limited with $f_{max} = 3/2\pi\sigma$ to $2/2\pi\sigma$. By the sampling theorem, the corresponding minimum sampling intervals are σ to 1.5σ, *i.e.* about 4 samples per longitudinal receptive field distance. Hence the minimum number of measurements (*i.e.* cells or receptive fields) per receptive field area is about 18. If one assumes that the density of sampled image points is constant over the visual field, it follows that the computational effort required to process the image through a given channel is roughly independent of the receptive field size associated with that channel[3].

This model of the preliminary processing of the image is

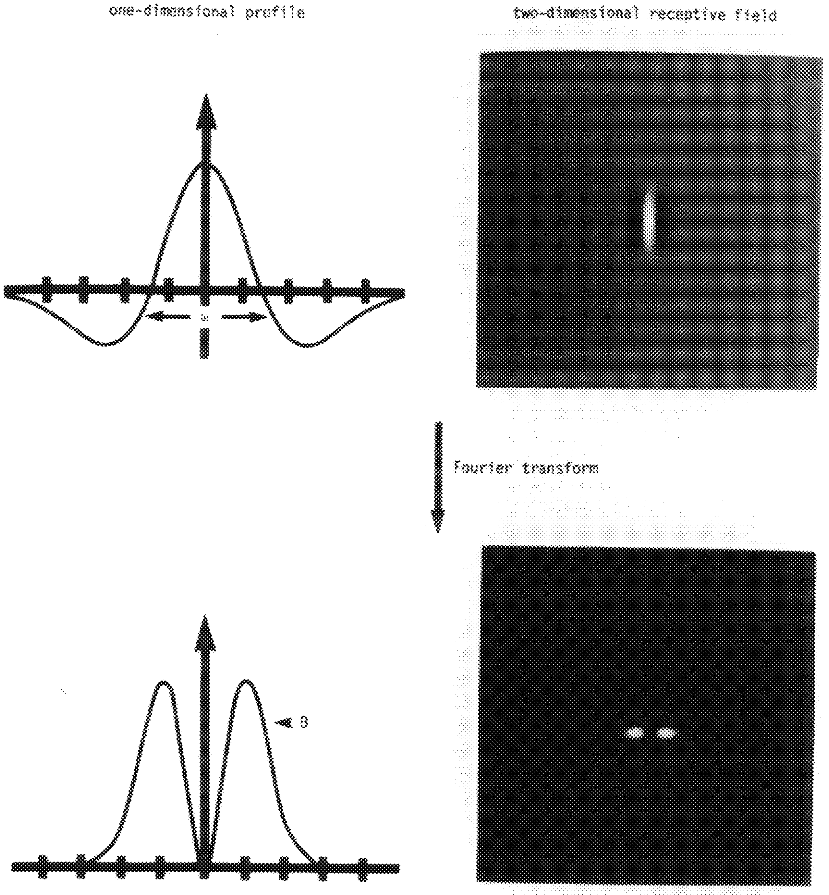
5.3 The domain of the matching function

In view of this information, the first step in our theory consists of filtering the left and right spatial images through four bar masks at each point in the images. We assume that this operation is roughly linear, for a given intensity and contrast. When matching the left and right images, one cannot simply use the raw values measured by this first stage, because they do not correspond directly to physical features on visible surfaces on which matching may be based. One first has to obtain from these measurements some symbol that corresponds with high probability to a physical item with a well-defined spatial position. This observation, which has been verified through computer experiments in the case of stereo vision (Grimson & Marr 1978) formed the starting point for a recent approach to the early processing of visual information (Marr 1974, 1976).

Perhaps the simplest way of obtaining suitable symbols from an image is to find signed peaks in the first (directional) derivative of the intensity array, or alternatively, zero-crossings in the second

TABLE 1. STEP 1 OF THE STEREOPSIS COMPUTATION: SPATIAL FILTERING.

A. At each point in the visual field the image is filtered through receptive fields having these characteristics:



The half-power bandwidth B is about 1 octave.

B. In each position there are four receptive field sizes, the smallest being 1/4 of the largest. The profile R(x) and fourier transform $\hat{R}(\omega)$ of each receptive field is given by

$$R(x) = \frac{1}{\sqrt{2\pi}\,\sigma_e} e^{-\frac{x^2}{2\sigma_e^2}} - \frac{1}{\sqrt{2\pi}\,\sigma_i} e^{\frac{-t^2}{2\sigma_i^2}}$$

$$\hat{R}(\omega) = e^{-\frac{\omega^2}{2\left(\frac{1}{\sigma_e^2}\right)}} - e^{-\frac{\omega^2}{2\left(\frac{1}{\sigma_i^2}\right)}}$$

where σ_e, σ_i are the excitatory and inhibitory space constants, and are in the ratio 1:1.5. The half-power bandwidth spanned by the four receptive field cells at each point is two octaves.

C. w increases with eccentricity: w = 3' - 12' (possibly 20') at 0°
w = 6' - 34' at 4°.

D. Caution notes: receptive field sizes and corresponding spectral sensitivity curves in the suprathreshold condition are different from the threshold values given here. Smaller values of w could be expected under suprathreshold conditions at 0° eccentricity.

E. Formally the output of Step 1 is given by the convolution

$$F_{w,\theta}(x,y) = I * B_{w,\theta}$$

where I(x,y) denotes the light intensity at the image point (x,y), and $B_{w,\theta}(x,y)$ describes the receptive field of a bar-shaped mask at orientation θ, with central region of width w. θ takes the values of 0° and 90° (corresponding to the horizontal and vertical), and w takes four values in the range defined by B and C above.

derivative. The bar-masks of table 1 measure an approximation to the second directional derivative at roughly the resolution of the mask size, so clear signed zero-crossings in the convolution values obtained along a scan line lying perpendicular to the receptive field's longitudinal axis (*cf*. Marr 1976 figure 2) would specify an appropriate location precisely. The fact that the sign of the zero-crossings are important is consistent with experimental data like that of Julesz (1963 figure 2). Since for stereopsis, a precise estimate of only the horizontal coordinate is required, in principle we need to consider only masks having vertically oriented receptive fields. It is known, however, that vertical disparity information is used to help align the two eyes (see *e.g*. Ogle, Martens & Dyer 1967 chapter 11), and so in the human visual system, masks with horizontally oriented receptive fields may be used[4].

In practice, however, it is not enough to use just vertically oriented masks to obtain horizontal disparity information. Julesz (1971, p 80) showed that minute breaks in horizontal lines can lead to fusion of two stereograms even when the breaks lie close to the limit of visual acuity. Such breaks cannot be obtained by simple operations on the the measurements from even the smallest vertical masks. These breaks probably have to be localized by a specialized process for finding terminations by examining the values and positions of rows of zero-crossings obtained from horizontal mask convolutions (cf. Marr 1976 p. 496).

Thus not only zero-crossings but also terminations have to be made explicit, (cf. the principle of explicit naming, Marr 1976 p. 485). The matching process will then operate on descriptions, of the left and right images, that are built of these two symbolic primitives, and which specify their positions, the mask size from which they were obtained, and their signs. This process is summarized in Table 2.

5.4 Matching

At the heart of the matching problem lies the problem of false targets. If false targets arise in profusion, a somewhat sophisticated algorithm must be used to eliminate them (cf. section 2). Computational simplicity can be preserved only if false targets are rare, and the existence of several independent spatial-frequency-tuned channels provides a way of accomplishing this.

We propose that, for each set of masks of a given size, symbols of the same type (zero-crossing or termination) and sign are matched between the two images. If each channel were very narrowly tuned to a wavelength λ, the minimum distance between zero-crossings of the same sign in each image would be about λ. In this case, matching would be unambiguous in a disparity range up to λ. The same argument holds qualitatively for the actual channels, but because they are not so narrowly tuned, the disparity range for unambiguous matching will be smaller and must be estimated. This may be done in the following way. The argument is carried out for zero-crossings, since terminations are

Table 2

Step 2 of the stereopsis computation: Zero-crossings and terminations

(a) The outputs of each of the four filters (for each value of w, evaluated at θ (vertical orientation) are scanned along the horizontal direction (θ = 0), and the positions of positive- and negative-sloped zero-crossings are found.

(b) Step (a) is also carried out at θ = 0, and the spatial distribution of the zero-crossings and amplitudes of the associated gradients are examined for the information they provide about the positions of terminations.

(c) Formally, zero-crossings and terminations may be defined as follows: *Zero-crossing positions* (x^*, y^*) are the non-trivial solutions to (1) $F_{w,90} = 0$ for zero-crossings from the vertical masks

(2) $F_{w,0} = 0$ for zero-crossings from the horizontal masks

The non-trivial solutions to (2) define a set of curves in x and y, each given parametrically by $(x(s), y(s))$. Then ***termination points*** (ξ, η) are the non-trivial solutions to

(3) $d^2(F_{w,0})/ds^2 = 0$, *i.e.* taking the derivatives along each of the curves $(x(s), y(s))$.

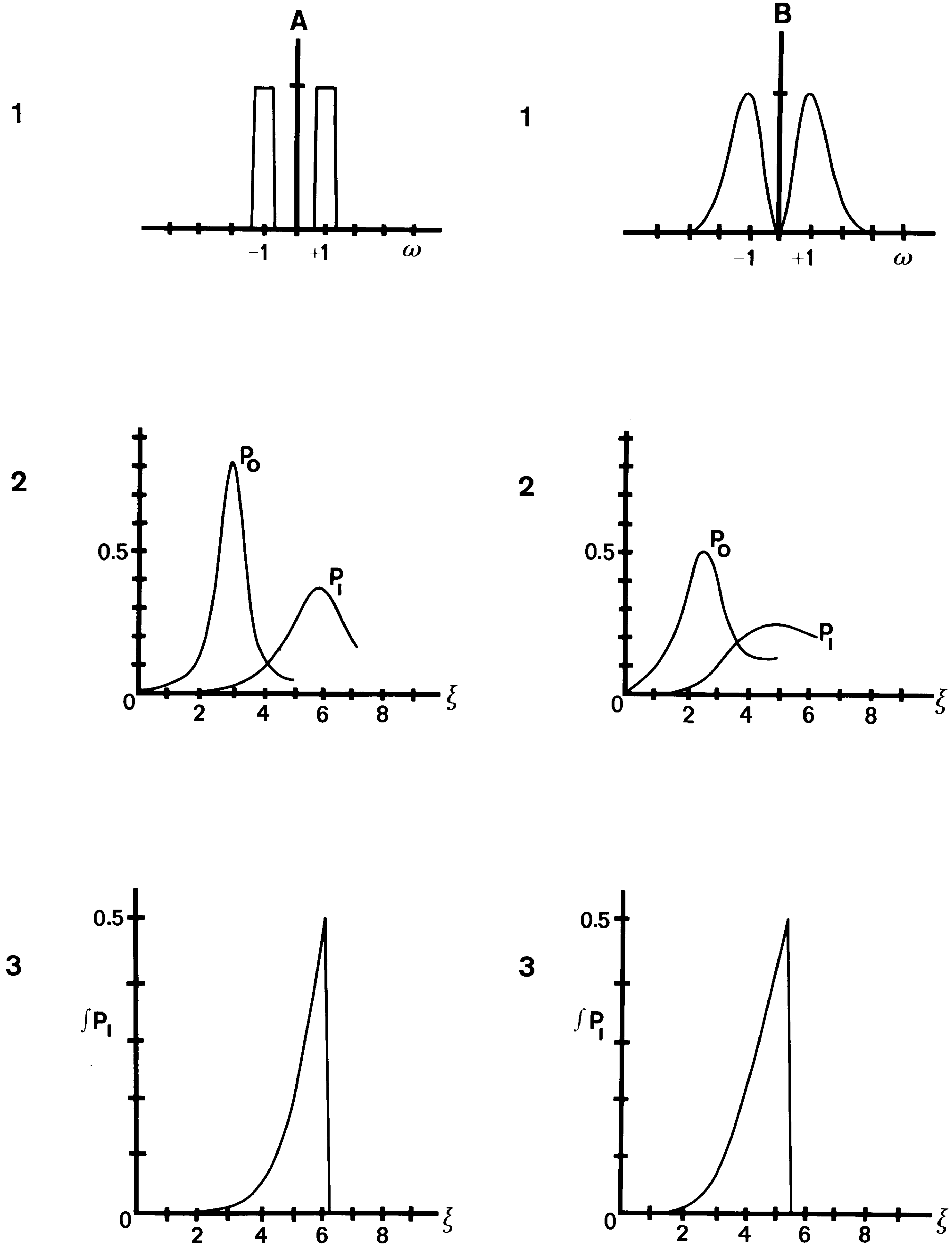
sparser and pose less of a false-target problem.

Statistical analysis of zero-crossings

The quantity of interest is the probability distribution of the interval between adjacent zero-crossings of the same sign in the filtered image. This depends on (a) the image characteristics, and (b) the filter characteristics. For (a), we assume that the input to the masks is gaussian. More precisely, if $I(x, y)$ is the mask input at coordinate (x, y), and $h(y)$ represents the longitudinal weighting function of the mask (see table 1), our assumption is that $f(x) = \int I(x, y)h(y)dy$ is a gaussian process.

For (b), we examine two extreme cases. Since the actual filters have a half-power bandwidth of one octave, the first case we consider is that of an ideal linear band-pass filter of width one octave, as illustrated in figure 8a. The second case (figure 8b) is the receptive field suggested by the threshold experiments of Wilson & Gieze (1977), consisting of excitatory and inhibitory gaussian distributions, with space-constants in the ratio 1:1.5, (see figures 5 and 7a). In both cases, the filtered image is a gaussian zero-mean process. We also take the worst case, *i.e.* that in which the power spectrum of the channel's input is essentially white in the relevant spectral range.

Our problem is now reduced to that of finding the distribution of the intervals between alternate zero-crossings by a stationary normal


A
B
1
−1
+1
ω
1
−1
+1
ω
2
0.5
P_0
P_1
0
2
4
6
8
ξ
2
0.5
P_0
P_1
0
2
4
6
8
ξ
3
0.5
$\int P_1$
0
2
4
6
8
ξ
3
0.5
$\int P_1$
0
2
4
6
8
ξ

8. Interval distribution for zero-crossings. A "white" gaussian process is passed through a filter with the frequency characteristic (transfer function) shown in (1). The interval distribution for the first (P_0) and second (P_1) zero-crossings of the resulting zero-mean gaussian process are approximated in (2). Given a zero-crossing at the origin, the probability of having at least two within a distance ξ is approximated by the integral of P_1 and shown in (3). In (A), these quantities are given for an ideal band-pass filter one octave wide and with centre frequency $\omega = 2\pi/\lambda$; (B) represents the case of the receptive field described by Cowan (1977) and Wilson & Gieze (1977). The corresponding spatial distribution of excitation and inhibition, *i.e.* the inverse fourier transform of (B1) appears, in the same units, in table 1. For case (a) a probability level of $P_1 = 0.001$ occurs at $\xi = 2.3$, and a probability level of 0.5 occurs at $\xi = 6.1$. The corresponding figures for case B are $\xi = 1.5$ and $\xi = 5.4$.

Situation B is derived from channel characteristics at threshold (see figures 8a and 6), and represent the worst case likely from the point of view of stereo matching. Situation A is closer to Cowan's (1977) guess at the suprathreshold condition (figure 8b). In situation B, the ratio of the space constants for excitation and inhibition (table 1 and figure 6) is 1:1.5; the values of P_1 change by not more than 5% if this ratio is 1:1.75 (Wilson 1978b).

process. Many authors have considered this problem, dating from the pioneering work of Rice (1945), (see for example Longuet-Higgins 1962, Leadbetter 1969).

Assume that there is a zero-crossing at the origin, and let $P_0(\xi)$, $P_1(\xi)$ be the probability densities of the distances to the first and second zero-crossings. P_0 and P_1 are approximated by the following formulae (Rice 1945 section 3.4, Longuet-Higgins 1962 eqs. 1.2.1 & 1.2.3):

$$P_o(\xi) = \frac{1}{2\pi}\left[\frac{\Psi(0)}{-\Psi^{n}(0)}\right]^{1/2} \frac{M_{23}(\xi)}{H(\xi)} (\Psi^2(0) - \Psi^2(\xi))[1 + H(\xi)\cot^{-1}(-H(\xi))]$$

$$P_1(\xi) = \frac{1}{2\pi}\left[\frac{\Psi(0)}{-\Psi^{n}(0)}\right]^{1/2} \frac{M_{23}(\xi)}{H(\xi)} (\Psi^2(0) - \Psi^2(\xi))[1 - H(\xi)\cot^{-1}(H(\xi))]$$

where $\psi(\xi)$ is the autocorrelation of the underlying stochastic process, and ' denotes differentiation w.r.t ξ.

$$H(\xi) = M_{23}(\xi)[M_{22}^2(\xi) - M_{23}^2(\xi)]^{-1/2}$$

$$M_{22}(\xi) = -\Psi''(0)(\Psi^2(0) - \Psi^2(\xi)) - \Psi(0)\Psi'^2(\xi)$$

$$M_{23}(\xi) = \Psi''(\xi)(\Psi^2(0) - \Psi^2(\xi)) + \Psi(\xi)\Psi'^2(\xi)$$

These approximations cease to be accurate for large values of ξ, (*i.e.* of order λ), where $2\pi/\lambda$ is the centre frequency of the channel; see Longuet-Higgins (1962) for a discussion of various approximations), where they overestimate P_0 and P_1. $P_1(\xi)$ is the quantity of interest here, since it is the interval distribution between zero-crossings of the same sign.

P_0 and P_1 were computed for the two filters of figures 8a1 & b1, and they are plotted in figures 8a2 & b2. The integrals of the two P_1 curves appear in figures 8a3 & b3. From these graphs, we see for example that the 0.05 probability level for the presence of false targets occurs at $\xi = 4.1$ (approximately $\lambda/1.52$) for the ideal band-pass filter one octave wide, centred on wavelength λ (figure 8a1), and at $\xi = 3.1$ for the receptive field of figure 8b1. In this case, ξ is approximately $\lambda/2$, where λ is the principal wavelength associated with the channel, and $\lambda = 2.2w$, where w is the measured width of the central excitatory area of the receptive field. Thus in this case, the 95% confidence limit occurs at a disparity approximately equal to w (ξ = 3.1, w = 2.8).

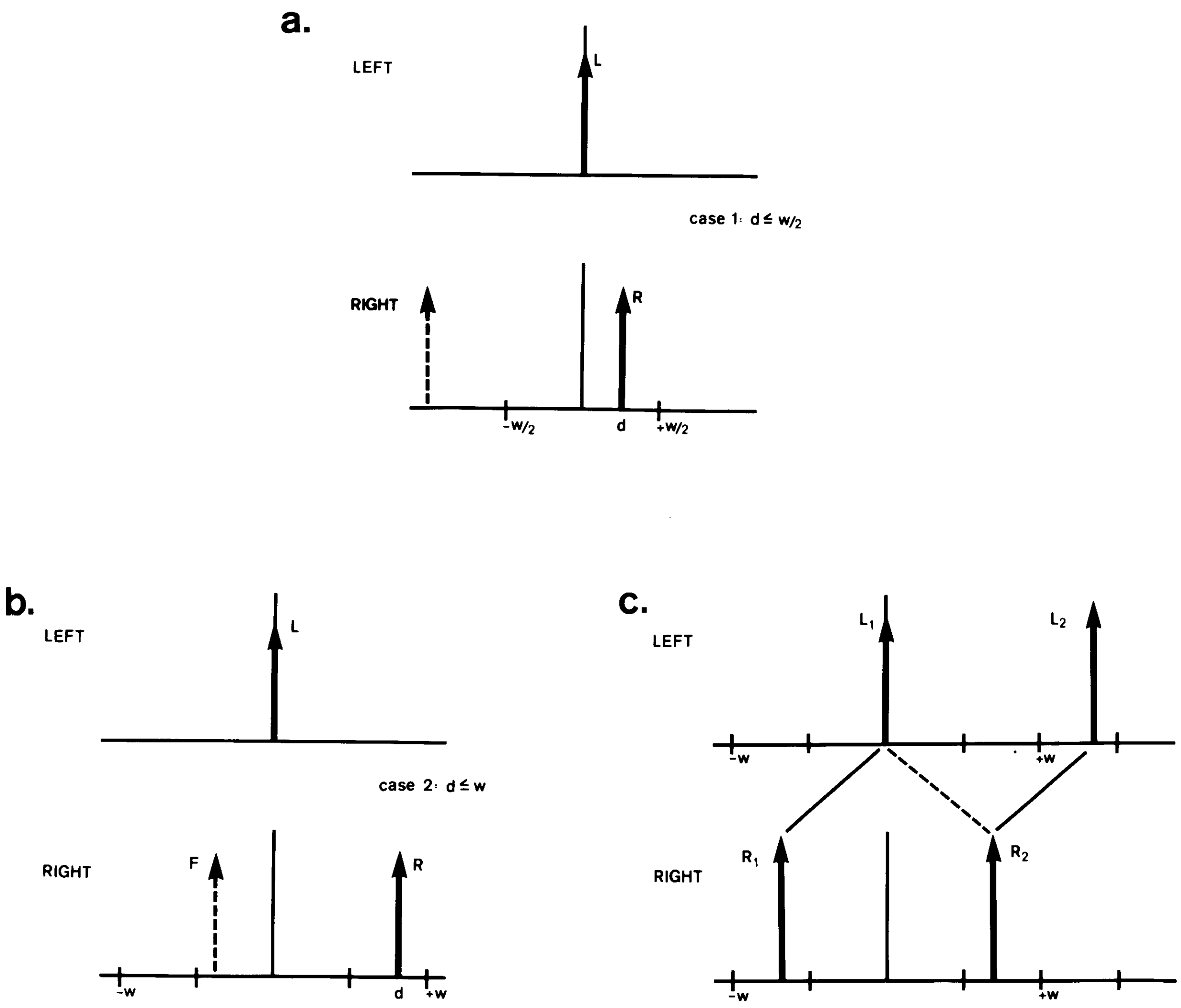
At the 0.001 probability level, the ideal band-pass filter is 50% better (the corresponding ξ is larger) than the receptive field filter with the same centre frequency; at the 0.05 probability level it is 30% better; and at the 0.5 probability level, it is 13% better. The legend to figure 8 provides more details about these results.

We have made a similar comparison between the sustained and transient channels of Wilson (1978a) and of Wilson & Berger (1978). If the sustained channels correspond to the case of figure 8b, the transient channels have a larger ratio of the space constants for inhibition and excitation, a somewhat larger excitatory space-constant, and an excitatory area larger than the inhibitory. Even under these conditions, the values change only slightly.

The matching process

We now apply the results of these calculations to the matching process, and show that within a given channel there are essentially two possible ways of dealing with false targets. If one wishes to avoid false targets altogether, the disparity range over which a match is sought must be restricted to $\pm w/2$ (see figure 9a). For suppose zero-crossing L in the left image matches zero-crossing R in the right image. The above calculations assure us that the probability of another zero-crossing of the same sign within w of R in the right image is less than 0.05. Hence if the disparity between the images is less than $w/2$, a search for matches in the range $\pm w/2$ will yield only the correct match R (with probability 0.95). Such a low error rate can be accomodated without resorting to sophisticated algorithms. For example, two reasonable ways to increase the matching reliability are (a) to demand rough agreement between the slopes of the matched zero-crossings, and (b) to fail to accept an isolated match all of whose neighbours give different disparity values. Of course if the disparity between the images exceeds $w/2$, this procedure will fail, a circumstance that we discuss later.

There is, however, an alternative strategy, that allows one to



9. The matching process. A zero-crossing L in the left image matches one R displaced by disparity *d* in the right image. The probability of a false target within *w* of R is small, so provided that *d* < *w*/2, (case A), almost no false targets will arise in the disparity range ±*w*/2. This gives the first possible algorithm. Alternatively (case B), all matches within the range ±*w* may be considered. Here, false targets (F) can arise in about 50% of the cases, but the correct solution is also present. If the correct match is convergent, the false target will with high probability be divergent. Therefore in the second algorithm, unique matches are accepted as correct, and the remainder as ambiguous and subject to the "pulling effect", illustrated in case C. Here, L_1 could match R_1 or R_2, but L_2 can match only R_2. Because of this, and because the two matches have the same disparity, L_1 is assigned to R_1.

deal with the matching problem over a larger disparity range. Let us consider the possible situations if the disparity between the images is d, where $|d| < w$ (figure 9b). Observe firstly that if $d > 0$, the correct match is almost certainly ($p < 0.05$) the only convergent candidate in the range $(0, w)$. Secondly, the probability of a (divergent) false target is at most 0.5. Therefore, 50% of all possible matches will be unambiguous and correct, and the remainder will be ambiguous, mostly consisting of two alternatives, one convergent and one divergent, one of which is always the correct one. In the ambiguous cases, selection of the correct alternative can be based simply on the sign of neighbouring unambiguous matches. This algorithm will fail for image disparities that significantly exceed $\pm w$, since the percentage of unambiguous matches will be too low (roughly 0.2 for $\pm 1.5w$).

Sparse images like an isolated line or bar, that yield few or no false targets, pose a different problem. They often give rise to unique matches, and may therefore be relied upon over quite a large disparity range. Hence if the above strategy fails to disclose candidate matches in its disparity range, the search for possible matches may proceed outwards, ceasing as soon as one is found.

In summary then (see table 3) there are two immediate candidates for matching algorithms. The simpler is restricted to a disparity range of $\pm w/2$, and in its most straightforward form will fail to assign 5% of the matches. The second involves some straightforward comparisons between neighbouring matches, but even before these comparisons, the 50% unambiguous matches could be used to drive eye-movements, and provide a rough sensation of depth.

The implementation of the first of these algorithms is straightforward. The second one can be implemented most economically using two "pools", one sensitive in a graded way to convergent and the other to divergent disparities (see figure 10). Notice that, in this sense, the first algorithm requires only one "pool", that is, a single unit sensitive in a graded way to the disparity range $\pm w/2$.

In the second algorithm, matches that are unambiguous or already assigned can "pull" neighbouring ambiguous matches to whichever alternative has the same sign. This may be related to the "pulling effect" described in psychophysical experiments by Julesz & Chang (1976). Notice however that this algorithm requires the existence of pulling only across pools and not within pools (in the terminology of Julesz & Chang p. 119).

Disparities larger than w can be examined in very sparse images. If, for example, both primary pools (covering a disparity range of $\pm w$) are silent, detectors operating outside this range, possibly with a broad tuning curve, may be consulted. In a biologically plausible implementation, these detectors should be inhibited by activity in the primary pools (see figure 10). It is tempting to suggest that detectors for these outlying disparities (*i.e.* exceeding about $\pm w$) may give rise to depth sensations and eye-movement control in diplopic conditions.

If the image is not sparse, and the disparity exceeds the

Table 3

Step 3 of the stereopsis computation: Matching (algorithm 2)

(a) For each zero-crossing or termination of a given sign in one image, matches are sought in the other in the range ±*w*. If a unique match is found, it is read by the memory. In no more than 50% of the cases the match will be ambiguous,, involving usually one convergent and one divergent candidate, one of which is correct. The signs of neighbouring matches that are unambiguous or already assigned determines the choice in these cases. The disparity of the chosen pool (either divergent or convergent) is read into the memory.

(b) It may happen that no matches can be found in this range. If failure occurs for a significant pecentage of the zero-crossings within a small neighbourhood of the fixation point, then it is assumed that disparities outside the range ±*w* occur there.

operating range, both algorithms will fail. Can the failure be recognized simply at this low level?

For the first algorithm, no correct match will be possible in the range ±w/2. The probability of a random match in this range is about 0.4, *i.e.* significantly less than 1.0. When the disparity between the two images lies in the range ±w/2, there will ***always*** be at least one match. It is therefore relatively easy to discriminate between these two situations.

For the second algorithm, an analogous argument applies: in this case the probability of no candidate match is about 0.3 for image disparities lying outside the range ±w, and zero for disparities lying within it. Again, it is relatively easy to discriminate between the situations.

Implications for psychophysical measurements of Panum's fusional area

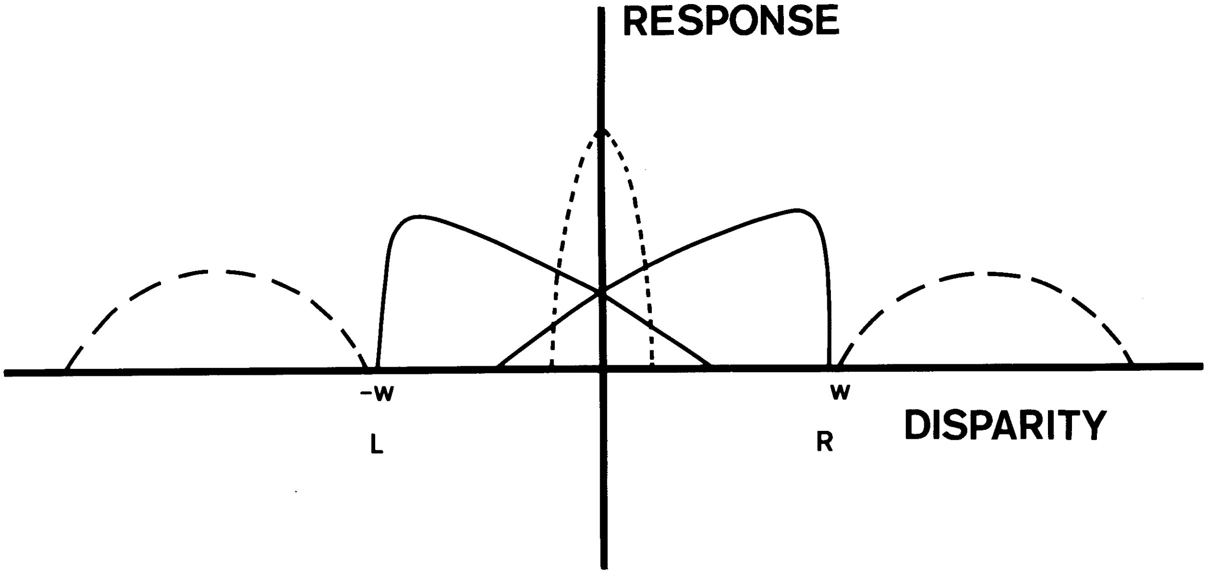
Using the second of the above algorithms, matches may be assigned correctly for a disparity range ±w. The precision of the disparity values thus obtained should be quite high, and a roughly constant proportion of w (which one can estimate from stereoacuity results at about w/20). For foveal channels, this means ±3' disparity with resolution 10" for the smallest, and ±12' (perhaps up to ±20' if Wilson & Bergen (1978) holds for stereopsis) with resolution 40" for the largest ones. At 4 degrees eccentricity, the range is ±5.3' to about

±34'. We assume that this range corresponds to stereoscopic fusion, and that outside it one enters diplopic conditions, in which disparity can be estimated only for relatively sparse images.

Under these assumptions, our predicted values apparently correspond quite well to available measures of the fusional limits without eye movements. Mitchell (1966) used small flashed line targets and found, in keeping with earlier studies, that the maximum amout of convergent or divergent disparity without diplopia is 10-14' in the fovea, and about 30' at 5 degrees eccentricity. The extent of the so-called Panum fusional area is therefore twice this[5].

Under stabilized image conditions, Fender & Julesz (1967) found that fusion occurred between line targets (13' by 1 degree high) at a maximum disparity of 40'. This value probably represents the whole extent of Panum's fusional area. Using the same technique on a random-dot stereogram, Fender & Julesz arrived at a figure of 14' (6' displacement and 8' disparity within the stereogram). Since the dot size was only 2', one may expect more energy in the high frequency channels than in the low, which would tend to reduce the fusional area. Julesz & Chang (1976), using a 6' dot size over a visual angle of 5 degrees, routinely achieved fusion up to ±18' disparity. Taking all factors into account, these figures seem to be consistent with our expectations.

The matching process is summarized in table 3.



10. An implementation of the second matching algorithm. For each mask size of central width *w*, there are two pools of disparity detectors, signalling crossed or uncrossed disparities and spanning a range of ±*w*. There may be an additional pool of detectors finely tuned to zero disparity. Additional diplopic disparities probably exist beyond this range. They are vetoed by detectors of smaller absolute disparity.

5.5 Dynamic memory storage: the $2\frac{1}{2}$-D sketch

According to our theory, once matches have been obtained using masks of a given size, they are represented in a temporary buffer. These matches also control vergence movements of the two eyes, thus allowing information from large masks to bring small masks into their range of correspondence. We postpone a detailed discussion of this process until the next section.

Why a memory?

The reasons for postulating the existence of a memory are of two kinds, those arising from general considerations about early visual processing, and those concerning the specific problem of stereopsis. As we have seen, a memory like the $2\frac{1}{2}$-D sketch (see figure 5), is computationally desirable because it makes explicit information about the image in a form that is closely matched to what early visual processes can deliver (Marr 1977 section 3.6 and Table 1). It is possible and reasonable to synthesize the outputs of various early processes in such a representation because the information they extract from images has a well-defined physical interpretation, namely the shape of the visible surfaces. The $2\frac{1}{2}$-D sketch describes these processes' results by representing the relative depth and surface-orientation associated with each viewing direction, together with contours of discontinuity in depth or surface-orientation.

The more particular reason associated specifically with stereopsis is the computational simplicity of the matching process, which requires a buffer in which to preserve its results as (1) disjunctive eye movements change the plane of fixation, and (2) objects move in the visual field. In this way, the 2-$\frac{1}{2}$-D sketch becomes the place where "global" stereopsis is actually achieved, combining the matches provided independently by the different channels and making the resulting disparity map available to other visual processes.

The nature of the memory

The 2-$\frac{1}{2}$-D sketch is a dynamic memory with considerable intrinsic computing power. It is perhaps worth stressing that it belongs to early visual processing, and cannot be influenced directly from higher levels, for example *via* verbal instructions, *a priori* knowledge or even previous visual experience (Ono & Nakamazio 1977, Frisby & Clatworthy 1975, and remarks of Marr 1977 section 3 about Ittelson 1960).

Although we have little direct evidence about the memory, one would expect a number of constraints derived from the physical world to be embedded in its internal structure. For example, the rule R2 of section 1, that disparity changes smoothly almost everywhere, might be implemented in the 2-$\frac{1}{2}$-D sketch by connexions similar to those that implement it in Marr & Poggio's (1976) cooperative algorithm (figure

2c). This active rule in the memory may be responsible for the sensation of a continuous surface to which even a sparse stereogram can give rise (Julesz 1971 figure 4.5-5, Grimson, Marr & Nishihara 1978).

We would expect other constraints to be embedded there in a similar way, for example the continuity of discontinuities in the visible surfaces, which we believe underlies the phenomenon of subjective contours (Marr 1977 section 3.6). It is possible that even more complicated consistency relations, concerning the possible arrangements of surfaces in three-dimensional space, are realized by computations in the memory, (e.g. constraints in the spirit of those made explicit by Waltz 1975). Such constraints may eventually form the basis for an understanding of phenomena like the Necker-cube reversal.

From this point of view, it is natural that many illusions concerning the interpretation of three-dimensional structure (the Necker cube, subjective contours, the Muller-Lyer figure, the Poggendorff figure, etc.) should take place after stereoscopic fusion (see Julesz 1971, Blomfield 1973).

The information represented in the $2\frac{1}{2}$-D sketch

The $2\frac{1}{2}$-D sketch represents the surface orientation, relative depth, and contours of discontinuities of these quantities in a scene. The exact form of the representation remains an open question, both from the computational and biological points of view, (Marr 1977 section 3). Because of the variety of purposes for which it has to be used, one would expect it to be easy to obtain any of the zeroth, first or second derivatives of depth in any direction in the visual field. The representational question here is, are all these quantities stored directly, or are some obtained from the others on demand?

For the purposes of this article, however, we can imagine that the memory records the matches obtained by binocular fusion in the form $(x_L, y_L; x_R, y_R; d)$, that is, the matching coordinate on the left and right images together with specification of the corresponding depth d relative to some reference point in the scene.

Dynamic management of stored information

According to this theory, the memory preserves depth (or disparity) information during the scanning of a scene with disjunctive eye-movements, and during movement of viewed objects. Information management will have limitations both in depth and in time, and the main questions here are over what range of disparities can the $2\frac{1}{2}$-D sketch maintain a record of a match in the presence of incoming information, and how long can it do this in its absence. The temporal question is less interesting because the purpose of the buffer is to organize incoming perceptual information, not to preserve it when there is none. In fact, Fender & Julesz's (1967) occluded one image of a random-dot stereogram, and found that fusion was destroyed for occlusions times longer than about 200 msec (see their figure 10).

The spatial aspects of the $2\frac{1}{2}$-D sketch raise a number of interesting questions. Firstly, are the maximal disparities that are preserved by the memory in stabilized image conditions the same as the maximum range of disparities that are simultaneously visible in a random-dot stereogram under normal viewing conditions? Secondly, does the distribution of the disparities that are present in a scene affect the range that the memory can store? For example, is the range greater for a stereogram of a spiral, in which disparity changes smoothly, than in a simple square-and-surround stereogram of similar overall disparity?

For the first question, the available evidence seems to indicate that the range is the same in the two cases. According to Fender & Julesz (1967), the range is about 2 degrees for a random-dot stereogram. When the complex stereograms given by Julesz (1971 e.g. 4.5-3) are viewed from about 20 cms, they give rise to disparities of about the same order. If this were true, it would imply that the maximal range of simultaneously perceivable disparities is a property of the $2\frac{1}{2}$-D sketch alone, and is independent of eye movements.

Fender & Julesz (1967) reported that under their experimental conditions, the maximum range for line stimuli (13' wide) was less (about 70') than the two degree range for random-dot stereograms. It may well be that "cooperative" effects in the $2\frac{1}{2}$-D sketch, that arise from the implementation of rule R2 ("filling-in" phenomena), may increase the maximum storable disparity range for textured surfaces.

Foley, Applebaum & Richards (1975), however, obtained a figure of 2 degrees for 18' wide line stimuli flashed for 40 msec. This discrepancy may be due in part to contrast and luminance effects.

In addition to these restrictions on the overall disparity range of the $2\frac{1}{2}$-D sketch, there may also be limitations on the degree of steepness in depth of a surface that can be represented. At some point, too steep a gradient in a surface will be represented instead as a discontinuity in depth. The data of Tyler (1974) are quite suggestive in this respect (see his figure 2), and may help to characterize the filling-in properties of the memory, (rule R2 of section 1). One may also expect the critical steepness, at which sujective contours may arise, to depend in part on the eccentricity of the surface in the visual field.

With regard to the second question, it seems at present unlikely that the maximum range of simultaneously perceivable disparities is much affected by their distribution. It can be shown that the figure of about 2 degrees, which holds for stabilized image conditions and for freely viewed stereograms with continuously varying disparities, also applies to stereograms with a single disparity.

Perception times do however depend on the distribution of disparities in a scene (Frisby & Clatworthy 1975, Saye & Frisby 1975). A stereogram of a spiral staircase ascending towards the viewer did not produce the long perception times associated with a two-planar stereogram of similar disparity range. This is to be expected, within the framework of our theory, because of the way in which we propose vergence movements are controlled. We now turn to this topic.

5.6 Vergence Movements

Disjunctive eye movements, which change the plane of fixation of the two eyes, are independent of conjunctive eye-movements (Rashbass & Westheimer 1961b), are smooth rather than saccadic, have a reaction time of about 160 msec, and follow a rather simple control strategy. The (asymptotic) velocity of eye vergence depends linearly on the amplitude of the disparity, the constant of proportionality being about 8 degrees/sec per degree of disparity (Rashbass & Westheimer 1961a). Vergence movements are accurate to within about 2' (Riggs & Niehl 1960), and voluntary binocular saccades preserve vergence nearly exactly (Williams & Fender 1977).

These data strongly suggest that the control of vergence movements is continuous rather than ballistic. Furthermore, Westheimer & Mitchell (1969) found that tachistoscopic presentation of disparate images led to the initiation of an appropriate vergence movement, but not to its completion.

Thus our hypothesis, that vergence movements are accurately controlled through matches obtained by the various channels, is consistent with the observed strategy and precision of vergence control. The hypothesis also accounts for the findings of Saye & Frisby (1975). Scenes like the spiral staircase, in which disparity changes smoothly, allow vergence movements to scan a large disparity range under the continuous control of the outputs of even the smallest masks. On the other hand, two-planar sterograms with the same disparity range require a large vergence shift, but provide no accurate information for its continuous control. The long perception times for such stereograms may therefore be explained in terms of a random-walk search strategy by the vergence control system. Furthermore, Saye & Frisby (1975) concluded from other evidence that "merely knowing where to direct eye movements is not sufficient to shorten stereopsis perception times, whereas monocularly conspicuous features may be sufficient." In other words, "on-line" guidance of vergence movements seems to be required. The process is a simple continuous closed-loop system which is usually inaccessible from higher levels.

By analogy with the fixation system of the fly (Reichardt & Poggio 1976), it may be possible to describe quantitatively the control of vergence by disparity (Richards 1975 figure 9). Richards' suggestion, that the relation between initial vergence velocity and disparity should be proportional to the relation between perceived depth and disparity, is attractive but far from being proved (compare Richards 1977 figure 1 and Rashbass & Westheimer 1961a figure 22). If it were true, however, it would be consistent with the direct control of vergence movements by the $2\frac{1}{2}$-D sketch. This would imply that the "diplopic disparity detectors" that we mentioned earlier (see figure 9) achieve their control of vergence movements not directly, but *via* the signals they provoke in the $2\frac{1}{2}$-D sketch, and which correspond to a real sensation of depth.

There may however exist some simple learning ability in the vergence control system. There is some evidence that an observer can learn to make an efficient series of vergence movements (Frisby & Clatworthy 1975). This learning effect seems however to be confined to the type of information used by the closed-loop vergence control system. *A priori*, verbal or high-level cues about the stereogram are ineffective.

5.7 Open Questions

There are several questions about the 2-$\frac{1}{2}$-D sketch that relate specifically to stereo. They will have to be examined through further psychophysical and computational studies. Some of the most immediate questions are:

(1) How many matches over what area are sufficient to cause information to be written into the memory?

(2) What is the relationship between the spatial structure of the information written in the memory and the scanning strategy of disjunctive and conjunctive eye movements?

(3) What are the rules that govern when filling-in takes place, and what is the three-dimensional shape of the filled-in surface?

(4) Is information moved around in the 2-$\frac{1}{2}$-D sketch during disjunctive or conjunctive eye movements, and if so, how? For example,

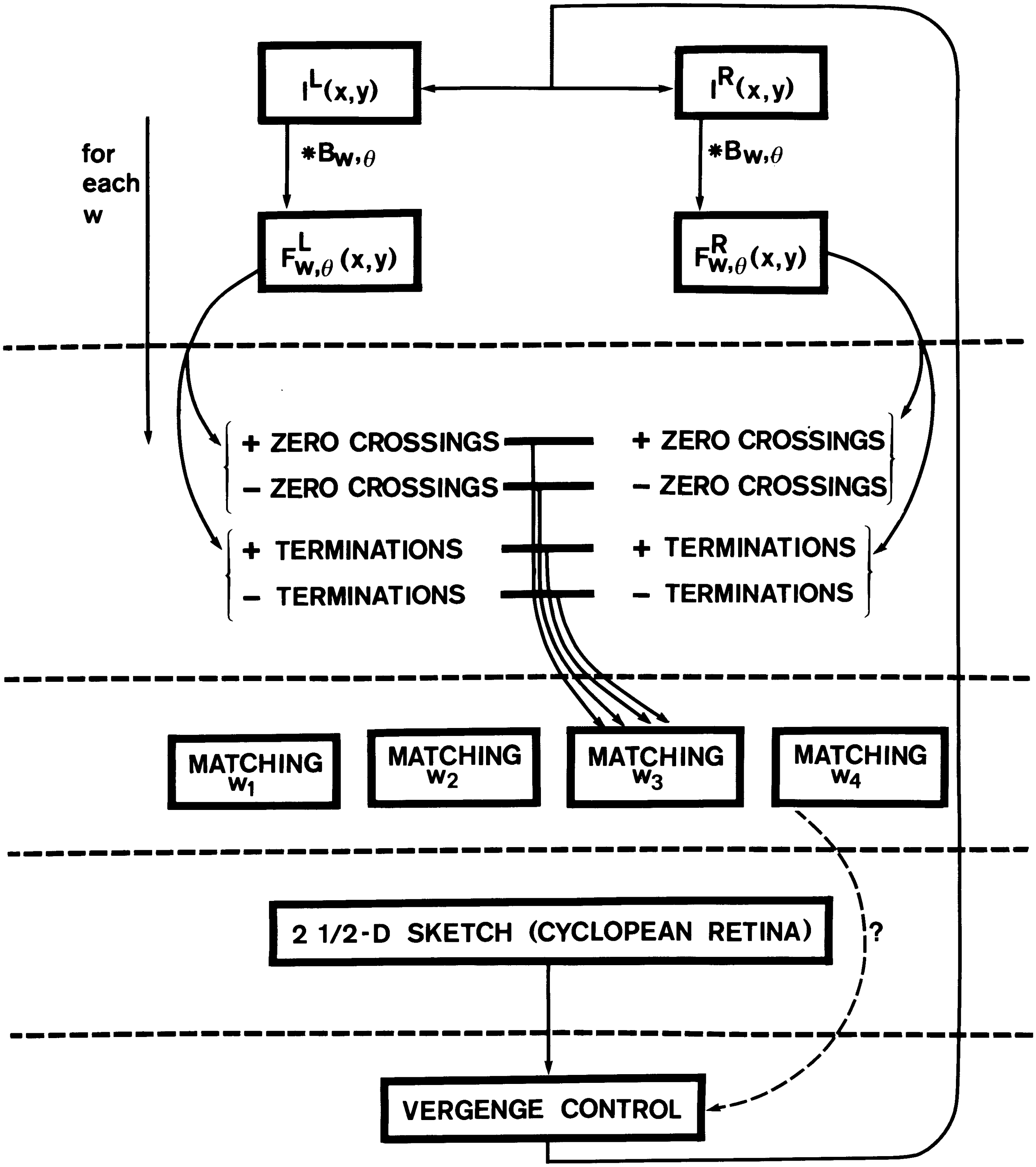
does the current fixation point always correspond to the same point in the $2\frac{1}{2}$-D sketch? If so, this implies that information is being moved almost continuously, both laterally and in depth. If not, there must be distinguished moments at which information is moved or the memory is cleared. To some extent one can of course simulate the movement of information in the memory by modifying the way it is addressed, but beyond a certain point this implies unacceptably inefficient use of the memory's representation capacity.

(5) What precisely is stored in the $2\frac{1}{2}$-D sketch, a function of depth or of its derivatives? As we have seen, the available evidence suggests that the overall range of depths that can be represented in the memory corresponds to about 2 degrees of disparity. If this is right, it implies that the $2\frac{1}{2}$-D sketch represents some function of depth explicitly, rather than implicitly through one or more of its derivatives.

5.7 Summary

The basic structure of the theory is summarized in table 4.

TABLE 4. FLOW DIAGRAM OF THE STEREOPSIS COMPUTATION.



6 Experiments

In this section, we summarize the experiments that are important for the theory. We separate psychophysical experiments from neurophysiological ones, and divide the experiments themselves into four categories according to whether their results are critical and are already available (A), are critical and not available and therefore amount to predictions (P), are neither critical nor available, but are of interest (I), and available results left unexplained by the theory (W). In the case of experimental predictions, we make explicit their importance to the theory by a system of stars; three stars indicates a prediction which, if falsified, would disprove the theory. One star indicates a prediction whose disproof remnants of the theory could survive.

6.1 Computation

The theory is capable of solving the matching problem for stereo vision of natural images (Grimson & Marr 1978). Solution of the overall stereo vision problem will require more detailed information about the $2\frac{1}{2}$-D sketch.

6.2 Psychophysics

1(A) The existence of independent spatial-frequency-tuned channels in binocular fusion and rivalry (Julesz & Miller 1975, Mayhew & Frisby 1976). These channels are independent up to the level of the $2\frac{1}{2}$-D sketch, where they are combined to form a "global" stereo percept (cf. Frisby & Mayhew 1978, finding (a)).

2(A) The nature and characteristics of the monocular channels are as described by Wilson & Gieze (1977) and Cowan (1977). (See also Wilson 1978a, b, Wilson & Phillips 1978, and Wilson & Berger 1978).

3(P)** The channels referred to in (1) and (2) are the same. Evidence consistent with this is provided by Felton, Richards & Smith (1972), who concluded that disparity mechanisms make bar-by-bar correlations, as opposed to edge-by-edge ones.

4(P)*** Terminations and signed zero-crossings in the filtered image are used as the input to the matching process.

5(P)** In the absence of eye movements, discrimination between two disparities in a random-dot stereogram is only possible within the range $\pm w$, where w is the width associated with the largest active channel. Using filtered stereograms, Frisby & Mayhew (1978) found that discrimination was possible without eye movements in the range 0-8'. If the fixation point was actually at the zero disparity position, this range is significantly higher than our theory would predict. Their experimental procedure does not eliminate the possibility that fixation point lay in the middle of the disparity range.

6(P)*** In the absence of eye movements, the magnitude of perceived depth in non-diplopic conditions is limited by the lowest spatial frequency channel stimulated.

7(P)*** In the absence of eye movements, the minimum fusable disparity range (Panum's fusional area) is ±3.1' in the fovea, and ±5.3' at 4 degrees eccentricity. This requires that only the smallest channels be active.

8(P)*** In the absence of eye movements, the maximum fusable disparity range is ±12' (possibly up to ±20') in the fovea, and about ±34' at 4 degrees eccentricity. This requires that the largest channels be active, for example by using bars or other large bandwidth stimuli.

9(P)** In the absence of eye-movements, the perception of rivalrous random-dot stereograms is subject to certain limitations. For example, for images of sufficiently high quality, figure 2b of Mayhew & Frisby (1976) should give rise to depth sensations, but figure 2c should not. In the presence of eye movements, figure 2c gives a sensation of depth. This could be explained if vergence eye movements can be driven by the relative imbalance between the numbers of unambiguous matches in the crossed and uncrossed pools over a small neighbourhood of the fixation point.

10(A) As measured by disparity-specific adaptation effects, the optimum stimulus for a small disparity is a high spatial frequency grating, whereas for large disparities, the most effective stimulus is a low spatial frequency grating. Furthermore, the adaptation effect specific to disparity is greatest for gratings whose periods are twice the disparity (Felton, Richards & Smith 1972). (In our terms, in fact, $2w$ is approximately 2.2λ, where λ is the centre frequency of the channel).

11(A) Evidence for the two pools hypothesis (Richards 1970, 1971, Richards & Regan 1973) is consistent with the minimal requirement for the second of the matching algorithms we described (section 5.3).

12(W) Stereoscopically viewed grating pairs of identical frequency but different contrast are reported to produce a sensation of tilt (Fiorentini & Maffei 1971).

13(P)*** In the absence of eye-movements, the perception of tilt in stereoscopically viewed grating pairs of different spatial frequencies is limited by (6, 7 & 8) above.

14(I) In the presence of eye-movements, is the range of simultaneously fused disparities independent of their distribution, (see section 5.5 on the dynamic management of stored information). For example, is the fusable range the same for a spiral and for a single square in depth?

15(I) Tyler (1974) found a limit in the rate of change of perceived disparity across the retina. Does this limit vary with eccentricity?

16(I) What are the critical parameters (density, distribution, correlation, etc.) for the perception in depth of a "solid" surface in a random-dot stereogram? (See Julesz 1971 pp 150, 79, figure 8.1-2, and White 1962). What are the rules for filling a surface in in depth?

17(A) Individuals impaired in one of the two disparity pools show corresponding reductions in depth sensations accompanied by a loss of vergence movements in the corresponding direction, (Jones 1972).

18(P)* Outside Panum's area, the dependence of depth sensation on disparity should be roughly proportional to the initial vergence velocity under the same conditions.

19(A) Perception times for novel two-planar stereograms are much longer than peception times for stereograms with smoothly varying disparities of the same (large) overall range.

20(P)*** In the two-planar case of (19), vergence movements should exhibit a random-search-like structure. The three star status holds when the disparity range exceeds the size of the largest masks activated by the pattern.

21(P)*** The range of vergence movements made during the successful and precise interpretation of complex, high-frequency, multi-layer, random-dot stereograms should span the range of disparities.

22(I) What is the relationship between scanning strategy and the three-dimensional spatial structure of a stereo image pair?

23(P)* Perception times for a random-dot stereogram portraying two small planar targets separated laterally and in depth, against an uncorrelated background, should be longer than the two-planar case (20). Once found, their representation in the memory should be labile if an important aspect of the representation there consists of local disparity differences.

6.3 Neurophysiology

24(partly A) At each point in the visual field, the scatter of bar mask receptive field sizes is about 4:1 (Hubel & Weisel 1974 figs. 1 & 4). [Wilson & Gieze 1977 p. 27]. More data are however needed on this point. This range is spanned by four populations of receptive field size.

25(P)* For each receptive field size, the local density of receptive field centers should be (at least) 18 per receptive field area.

26(P)** For a given intensity and contrast, these cells perform a nearly linear convolution of the image with a bar-shaped receptive field of medium bandwidth in the spatial frequency domain. The representation of positive and negative values probably involves different cells.

27(P)* A different class of cells may represent the peak and termination positions and signs in the monocularly filtered images.

28(P)** There exist binocularly driven cells sensitive to the disparity. A given cell signals a match between either a zero-crossing pair or a termination pair, both items in its pair having the same sign.

29(P)** For each sign (±) and type (zero-crossing or termination) of match at each point in the visual field, there should exist four populations of matching cells (28), fed independently by the four populations in (24).

30(P)** Each of the sixteen populations defined by (29) is divided into at least two (and possibly three) main disparity pools, tuned to crossed and uncrossed disparities respectively, with sensitivity curves extending outwards to a disparity of about the width of its corresponding receptive field centre (see figure 9). Being sensitive to pure disparity, these cells are sensitive to changes in disparity induced by vergence movements. In addition, there may be one pool quite sharply tuned to zero disparity.

31(P)* In addition to the two (or three) basic disparity pools of (30), there may exist cells tuned to more outlying (diplopic) disparities (compare figure 9). These cells should be inhibited by any activity in the basic pools.

32(P)** There exists a neural representation of the $2\frac{1}{2}$-D sketch. This includes cells that are highly specific for some monotonic function of depth and disparity, and which span a depth range corresponding to about 2 degrees of disparity. Within a certain range, these cells may not be sensitive to disjunctive eye movements. This corresponds to the notion that the plane of fixation can be moved around within the 2 degree disparity range currently being represented in the $2\frac{1}{2}$-D sketch.

33(P)* The diplopic disparity cells of (32) are especially concerned with the control of disjunctive eye movements.

6.3 Cautionary remarks

Because of the computational nature of this approach, we have been able to be quite precise about the nature of the processes that are involved in this theory. Since a process may in general be implemented in several diffferent ways, our physiological predictions are more speculative than our psychophysical ones. They should perhaps be regarded more as guidelines for investigations rather than as necessary consequences of the theory.

A number of other general remarks are in order here.

(1) The first concerns our hypothesis of the near linearity of the filtering operation for a given intensity and contrast. This hypothesis may not be strictly correct, but small deviations from it should not greatly affect our theory.

(2) The second remark concerns our quantitative estimates of the channel characteristics. In suprathreshold conditions, the receptive field may change slightly. Inhibition may be more prominent, corresponding to a narrowing of the channel's bandwidth in the spatial frequency domain (cf. figures 8a & b). It is worth noting that such a change can be implemented easily in a system that separates the positive ("on-centre") from the negative ("off-centre") parts of the signal (De Valois 1977, Burton, Nagshineh & Ruddock 1978). The existence of a natural rectification of the convolution allows a

narrowing of the channel without adding excitatory side-lobes to the receptive field of the cell simply by increasing the strength of its inhibitory surround. Narrowing the channels in this way could move the filter characteristics in the direction of the ideal band-pass filter of figure 8a (cf. figure 8b), thus increasing estimates of Panum's fusional area by up to 30%.

7 Discussion

Perhaps one of the most striking features of our theory is the way it returns to Fender & Julesz's original suggestion, of a cortical memory that accounts for the hysteresis and which is distinct from the matching process. Consequently fusion does not need to be cooperative, and our theory and its implementation (Grimson & Marr 1978) demonstrate that the computational problem of stereoscopic matching can be solved without cooperativity. These arguments do *not* however forbid its presence. Critical for this question are predictions (5) - (7) about the exact extent of Panum's fusional area for each channel. If the empirical data indicate a fusable disparity range significantly larger than $\pm w$, false targets will pose a problem not easily overcome using straightforward matching techniques like algorithm (2) of section 5.3. In these circumstances, the matching problem could be solved by an algorithm like Marr & Poggio's (1976) operating within each channel, to eliminate possible false targets arising as a result of an extended disparity sensitivity range.

As it stands, there are a number of points on which the theory is indefinite, for example, the exact structure of the $2\frac{1}{2}$-D sketch and the way the various constraints are implemented there, the dynamic management of the representation and its dependence on eye movements, and the details of the strategy by which eye movements are controlled.

On the other hand, the theory is precise enough to be implemented as a computer program that deals with the stereo matching problem for raw natural images. It assimilates a large amount of empirical data, and admits of a number of experimental predictions concerning each of the four main parts of the theory, the preprocessing through four independent channels, matching of zero-crossings and terminations, the $2\frac{1}{2}$-D sketch, and the control of vergence movements.

Finally, we feel that an important feature of this theory is that it grew from an analysis of the computational problems that underly stereopsis, and is devoted to a characterization of the processes capable of solving it without specific reference to the machinery in which they run. The elucidation of the precise neural mechanisms that implement these processes, obfuscated as they must inevitably be by the vagaries of natural evolution, poses a fascinating challenge to classical techniques in the brain sciences.

Acknowledgements. We are deeply indebted to Whitman Richards for many remarks that we understand only in retrospect. We are especially grateful to Jack Cowan, John Frisby, David Hubel, Bela Julesz, John Mayhew and Hugh Wilson, and to Werner Reichardt and the Max Planck Society for their kind hospitality in Tubingen. Karen Prendergast prepared the illustrations. This work was conducted the Max-Planck-Institut fur Biologische Kybernetik in Tubingen, and at the Artificial Intelligence Laboratory, a Massachusetts Institute of Technology research program supported in part by the Advanced Research Projects

Agency of the Department of Defense, and monitored by the Office of Naval Research under contract number N00014-75-C-0643.

Footnotes

1: In addition, a near-optimal linear algorithm of the type of Dev's equations (1) & (2) will suffer from stability problems.

2: Owing presumably to a printing error, the left-hand image of their figure 1 has been rotated 90 degrees.

3: Not too much weight should be attached to the estimate of 18, although we feel that the sampling density cannot be significantly lower.

4: The role played in our theory by the channels and the zero-crossings may represent a deep property of the visual process. It has recently been shown that the information carried by a band-limited function is contained in a specification of its zero-crossings, provided that some non-trivial conditions are satisfied (Logan 1977). We conjecture that there may be a relationship between these theorems and the use of zero-crossings summarized in table 1 (Marr, Poggio & Ullmann, in preparation).

5: These values may be compatible with Wilson & Bergen (1978), because of the extremely low sensitivity of the U channel.

References

Barlow, H. B., Blakemore, C. & Pettigrew, J. D. 1967 The neural mechanism of binocular depth discrimination. ***J. Physiol. 193,*** 327-342.

Bishop, P. Q., Henry, G. H. & Smith, C. J. 1971 Binocular interaction fields of single units in the cat striate cortex. ***J. Physiol. 216,*** 39-68.

Blakemore, C. & Campell, F. W. 1969 On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images. ***J. Physiol. 203***, 237-260.

Blomfield, S. 1973 Implicit features and stereoscopy. ***Nature 245NB***, 256.

Burton, G. J., Nagshineh, S. & Ruddock, K. H. 1978 Processing by the human visual system of the light and dark contrast components of the retinal image. ***Biol. Cybernetics*** (in the press).

Campbell, F. W. & Robson, J. 1968 Application of Fourier analysis to the visibility of gratings. ***J. Physiol. 197,*** 551-566.

Cowan, J. D. 1977 Some remarks on channel bandwidths for visual contrast detection. ***Neurosciences Res. Prog. Bull. 15,*** 492-517.

De Valois, K. K. 1977 Independence of black and white: phase-specific adaptation. ***Vision Res. 17,*** 209-215.

Dev, P. 1975 Perception of depth surfaces in random-dot stereograms: a neural model. ***Int. J. Man-Machine Studies 7***, 511-528.

Evans, C. R. & Clegg, J. M. 1967 Binocular depth perception of Julesz patterns viewed as perfectly stabilized retinal images. ***Nature 215,*** 893-895.

Felton, T. B., Richards, W. & Smith, R. A. Jr. 1972 Disparity processing of spatial frequencies in man ***J. Physiol. 225,*** 349-362.

Fender, D. & Julesz, B. 1967 Extension of Panum's fusional area in binocularly stabilized vision. ***J. opt. Soc. Am. 57***, 819-830.

Fiorentini, A. & Maffei, L. 1971 Binocular depth perception without geometrical cues. ***Vision Res. 11,*** 1299-1305.

Foley, J. M., Applebaum, T. H. & Richards, W. A. 1975 Stereopsis with large disparities: discrimination and depth magnitude. ***Vision Res. 15,*** 417-422.

Frisby, J. P. & Clatworthy, J. L. 1975 Learning to see complex random-dot stereograms. *Perception 4,* 173-178.

Frisby, J. P. & Julesz, B. 1975 The effect of orientation difference on stereopsis as a function of line length. *Perception 4,* 179-186.

Frisby, J. P. & Mayhew, J. E. W. 1978 Spatial frequency selective processes and stereopsis. *Vision Res.* (submitted for publication).

Georgeson, M. A. & Sullivan, G. D. 1975 Contrast constancy: deblurring in human vision by spatial frequency channels. *J. Physiol. (Lond.) 252*, 627-656.

Grimson, W. E. L. & Marr, D. (in preparation).

Grimson, W. E. L., Marr, D. & Nishihara, H. K. 1978 (in preparation).

Harmon, L. D. & Julesz, B. 1973 Masking in visual recognition: effects of two-dimensional filtered noise. *Science 180*, 1194-1197.

Hirai, Y. & Fukushima, K. 1976 An inference upon the neural network finding binocular correspondence. *Trans. IECE J59-D*, 133-140.

Hubel, D. H. & Wiesel, T. N. 1970 Cells sensitive to binocular depth in area 18 of the Macaque monkey cortex. *Nature 225*, 41-42.

Hubel, D. H. & Wiesel, T. N. 1973 A reexamination of stereoscopic mechanisms in area 17 of the cat. *J. Physiol. 232*, 29-30.

Ittelson, W. H. 1960 *Visual space perception.* New York: Springer-Verlag (pp. 145-147).

Jones, R. 1972 Psychophysical and oculomotor responses of manual and stereoanomalous observers to disparate retinal stimulation. *Doctoral dissertation, Ohio State University, Dissertation Abstract N. 72-20970.*

Julesz, B. 1960 Binocular depth perception of computer-generated patterns. *Bell System Tech. J. 39*, 1125-1162.

Julesz, B. 1962 Towards the automation of binocular depth perception (AUTOMAP-1). *Proceedings of the IFIPS Congres, Munich 1962*, ed. C. M. Popplewell. Amsterdam, North Holland, 1963.

Julesz, B. 1963 Stereopsis and binocular rivalry of contours. *J. Opt. Soc. Amer. 53,* 994-999.

Julesz, B. 1971 *Foundations of Cyclopean Perception.* Chicago: The University of Chicago Press.

Julesz, B. & Chang, J. J. 1976 Interaction between pools of binocular disparity detectors tuned to different disparities. *Biol. Cybernetics*

22, 107-120.

Julesz, B. & Miller, J. E. 1975 Independent spatial-frequency-tuned channels in binocular fusion and rivalry. *Perception 4*, 125-143.

Kaufman, L. 1964 On the nature of binocular disparity. *Amer. J. Psychol. 77*, 393-402.

Leadbetter, M. R. 1969 On the distributions of times between events in a stationary stream of events. *J. R. Statist. Soc. B 31*, 295-302.

Logan, B. F. Jr. 1977 Information in the zero crossings of bandpass signals. *Bell System Tech. J. 56*, 487-510.

Longuet-Higgins, M. S. 1962 The distribution of intervals between zeros of a stationary random function. *Phil. Trans. R. Soc. A 254*, 557-599.

Marr, D. 1974 A note on the computation of binocular disparity in a symbolic, low-level visual processor. *M.I.T. A.I. Lab. Memo 327.*

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

Marr, D. 1977 Representing visual information. *AAAS 143rd Annual Meeting, Symposium on Some Mathematical Questions in Biology*, February, (in the press). Also available as *M.I.T. A.I. Lab. Memo 415.*

Marr, D. & Nishihara, H. K. 1977 Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. B.* (in the press).

Marr, D. & Poggio, T. 1976 Cooperative computation of stereo disparity. *Science 194*, 283-287.

Marr, D., Palm, G. & Poggio, T. 1978 Analysis of a cooperative stereo algorithm. *Biol. Cybernetics*, (submitted for publication).

Mayhew, J. E. W. & Frisby, J. P. 1976 Rivalrous texture stereograms. *Nature 264*, 53-56.

Mayhew, J. E. W. & Frisby, J. P. 1978 Relative depth discriminations in narrow-band-filtered random-dot stereograms. *Vision Res.*, submitted for publication.

Mitchell, D. E. 1966 Retinal disparity and diplopia. *Vision Res. 6*, 441-451.

Nelson, J. I. 1975 Globality and stereoscopic fusion in binocular vision. *J. Theor. Biol. 49*, 1-88.

Nelson, J. I., Kato, H. & Bishop, P. O. 1977 Discrimination of orientation and position disparities by binocularly activated neurons in cat striate cortex. *J. Neurophysiol.* ***40***, 260-283.

Nikara, T., Bishop, O. P. & Pettigrew, J. D. 1968 Analysis of retinal correspondence by studying receptive fields of binocular single units in cat striate cortex. *Exp. Brain. Res.* ***6***, 353-372.

Ogle, K. N., Martens, T. G. & Dyer, J. A. 1967 ***Oculomotor imbalance in binocular vision and fixation disparity.*** Philadelphia: Lea & Febiger.

Ono, H. & Nakamizo, S. 1977 Saccadic eye movements during changes in fixation to stimuli at different distances. *Vision Res.* ***17***, 233-238.

Papoulis, A. 1968 ***Systems and transforms with applications in optics.*** New York: McGraw Hill.

Pettigrew, J. D., Nikara, T. & Bishop, P. O. 1968 Binocular interaction on single units in cat striate cortex: simultaneous stimulation by single moving slit with receptive fields in correspondence. *Exp. Brain Res.* ***6***, 391-410.

Poggio, G. F. & Fischer, B. 1977 Binocular interaction and depth sensitivity of striate and prestriate cortical neurons of the behaving rhesus monkey. *J. Neurophysiol.*, (in the press).

Rashbass, C. & Westheimer, G. 1961a Disjunctive eye movements. *J. Physiol.* ***159***, 339-360.

Rashbass, C & Westheimer, G. 1961b Independence of conjunctive and disjunctive eye movements. *J. Physiol.* ***159***, 361-364.

Reichardt, W. & Poggio, T. 1976 Visual control of orientation behavior in the fly. Part I, a quantitative analysis. ***Quarterly Reviews of Biophysics*** ***9***, 311-375.

Rice, S. O. 1945 Mathematical analysis of random noise. ***Bell Syst. Tech. J.*** ***24***, 46-156.

Richards, W. 1970 Stereopsis and stereoblindness. *Exp. Brain Res.* ***10***, 380-388.

Richards, W. 1971 Anomalous stereoscopic depth perception. *J. opt. Soc. Amer.* ***61***, 410-414.

Richards, W. 1975 Visual space perception. Ch. 10, pp. 351-386 of *Handbook of Perception, Vol. 5, Seeing.* Eds. E. C. Carterette & M. D. Freidman. New York: Academic Press.

Richards, W. A. 1977 Stereopsis with and without monocular cues. *Vision Res*. (in the press).

Richards, W. A. & Marr, D. (In preparation).

Richards, W. A. & Regan, D. 1973 A stereo field map with implications for disparity processing. ***Investigative Ophthalmology 12***, 904-909.

Riggs, L. A. & Niehl, E. W. 1960 Eye movements recorded during convergence and divergence. ***J. opt. Soc. Am. 50***, 913-920.

Saye, A. & Frisby, J. P. 1975 The role of monocularly conspicuous features in facilitating stereopsis from random-dot stereograms. ***Perception 4***, 159-171.

Shapley, R. M. & Tolhurst, D. J. 1973 Edge detectors in human vision. ***J. Physiol. (Lond.) 229***, 165-183.

Sperling, G. 1970 Binocular vision: a physical and a neural theory. ***Am. J. Psychol. 83***, 461-534.

Stromeyer, C. F. III & Julesz, B. 1972 Spatial-frequency masking in vision: critical bands and spread of masking. ***J. opt. Soc. Amer. 62***, 1221-1232.

Sugie, N. & Suwa, M. 1977 A scheme for binocular depth perception suggested by neurophysiological evidence. ***Biol. Cybernetics 26***, 1-15.

Tyler, C. W. 1974 Depth perception in disparity gratings. ***Nature 251***, 140-142.

Waltz, D. 1975 Understanding line drawings of scenes with shadows. In: ***The psychology of computer vision***, Ed. P. H. Winston, pp19-91. New York: McGraw-Hill.

Westheimer, G. & Mitchell, D. E. 1969 The sensory stimulus for disjunctive eye movements. ***Vision Res. 9***, 749-755.

White, B. W. 1962 Stimulus conditions affecting a recently discovered stereoscopic effect. ***Am. J. Psychol. 75***, 411-420.

Williams, R. H. & Fender, D. H. 1977 The synchromy of binocular saccadic eye movements. ***Vision Res. 17***, 303-306.

Wilson, H. R. 1978a Quantitative characterization of two types of line spread function near the fovea. ***Vision Res***. (in the press).

Wilson, H. R. 1978b Quantitative prediction of line spread function measurements: implications for channel bandwidths. ***Vision Res***. (in the press).

Wilson, H. R. & Berger, J. R. 1978 A four mechanism model for spatial vision. (In preparation).

Wilson, H. R. & Gieze, S. C. 1977 Threshold visibility of frequency gradient patterns. *Vision Res. 17*, 1177-1190.

Wilson, H. R. & Phillips, G. 1978 Evidence against disinhibition in psychophysically measured line spread functions. *Vision Res.* (in the press).