Proc. R. Soc. Lond. B. **204**, 301–328 (1979)
Printed in Great Britain

A computational theory of human stereo vision†

BY D. MARR‡ AND T. POGGIO§

‡ *M.I.T. Psychology Department*, 79 *Amherst Street*, *Cambridge Ma* 02139, *U.S.A.*

§ *Max-Planck-Institut für Biologische Kybernetik*, 7400 *Tübingen*, *Spemannstrasse* 38, *Germany*

(*Communicated by S. Brenner, F.R.S. – Received* 26 *January* 1978)

An algorithm is proposed for solving the stereoscopic matching problem. The algorithm consists of five steps: (1) Each image is filtered at different orientations with bar masks of four sizes that increase with eccentricity; the equivalent filters are one or two octaves wide. (2) Zero-crossings in the filtered images, which roughly correspond to edges, are localized. Positions of the ends of lines and edges are also found. (3) For each mask orientation and 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 stored in 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 matching process.

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

† A preliminary and lengthier version of this theory is available from the M.I.T. A.I. Laboratory as Memo 451 (1977).

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 (1960) 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.

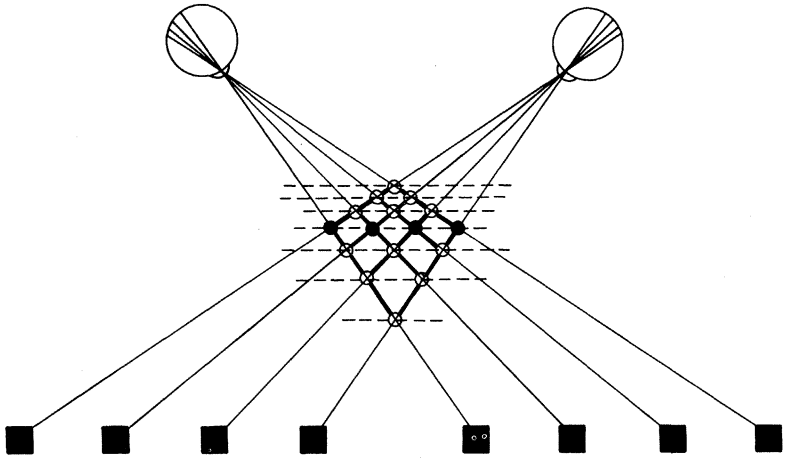


FIGURE 1. Ambiguity in the correspondence between the two retinal projections. In this figure, each of the four points in one eye's view could match any of the four projections in the other eye's view. Of the 16 possible matchings only four are correct (filled circles), while the remaining 12 are 'false 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, fig. 4.5–1).

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 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:

(*R*1) *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 unique physical position.

(*R*2) *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 practice, 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.

In a recent article, Marr & Poggio (1976) derived a cooperative algorithm which implements these rules (see figure 2), showing that it successfully solves the false targets problem and extracts disparity information from random dot stereograms (see also Marr, Palm & Poggio 1978).

The biological evidence

Apart from AUTOMAP (Julesz 1963) and Sperling (1970), all of the current stereo algorithms proposed as models for human stereopsis are based on Julesz's (1971) proposal that stereo matching is a cooperative process (Julesz 1971, p. 203 ff.; Julesz & Chang 1976; Nelson 1975; Dev 1975; Hirai & Fukushima 1976; Sugie & Suwa 1977; Marr & Poggio 1976). None of them has been shown to work on natural images.

An essential feature of these algorithms is that they are designed to select correct matches in a situation where false targets occur in profusion. They require 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. 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.

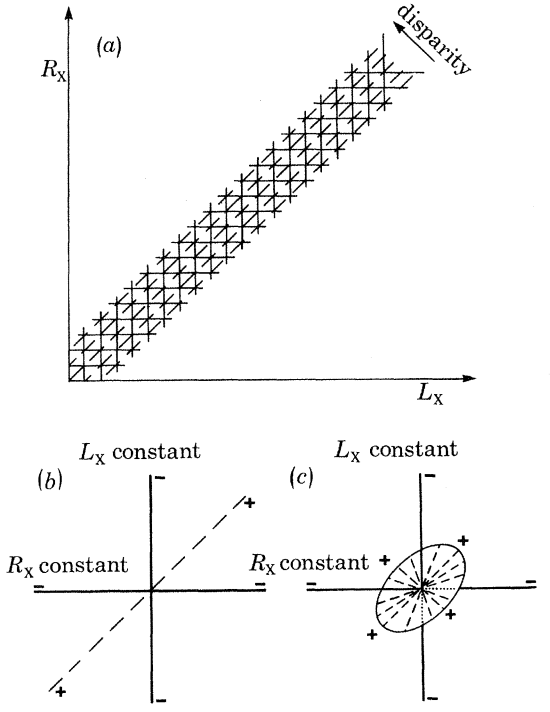


FIGURE 2. The explicit structure of the two rules R1 and R2 for the case of a one dimensional image is represented in (*a*). L_X and R_X represent the positions of descriptive elements in the left and right images. The continuous vertical and horizontal lines represent lines of sight from the left and 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*). The ovals in this figure represent circular two dimensional disks rising out of the plane of the page. Formally, the algorithm represented by this network may be written as the iterative algorithm

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

where $C^{t}_{x,y;d}$ denotes the state of the cell (0 for inactive, 1 for active) corresponding to position (x, y), disparity d and time t in the network of (*a*); $S(x, y, d)$ is a local excitatory neighbourhood confined to the same disparity layer, and (x, y, d) the inhibitory neighbourhood, consists of cells lying on the two lines of sight (*c*). 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 neighbourhoods O and S (*c*). (From Marr & Poggio 1976, fig. 2; copyright by the American Association for the Advancement of Science.)

Eye movements seem, however, to be important for human stereo vision (Richards 1977; Frisby & Clatworthy 1975; Saye & Frisby 1975). Other findings these algorithms fail to explain include (*a*) the ability of some subjects to tolerate a 15% expansion of one image (Julesz 1971, fig. 2.8–8), (*b*) the findings about independent spatial-frequency-tuned channels in binocular fusion, of which our tolerance to severe defocusing of one image is a striking demonstration (Julesz 1971, fig. 3.10–3), (*c*) the physiological, clinical, and psychophysical evidence about Richards' two pools hypothesis (Richards 1970, 1971; Richards & Regan 1973); and (*d*) the size of Panum's fusional area (6′–18′, Fender & Julesz 1967; Julesz & Chang 1976) which seems surprisingly small to have to resort to cooperative mechanisms for the elimination of false targets.

Taken together, these findings indicate that a rather different approach is necessary. In this article, 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 and formulate the predictions to which it leads.

For a more comprehensive review of the relevant psychophysics and neurophysiology see Marr & Poggio (1977*a*).

An outline of the theory

The basic computational problem in binocular fusion is the elimination of false targets, and for any given monocular features 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 (Kaufman 1964; Julesz 1971, §§ 3.9 and 3.10; Julesz & Miller 1975; Mayhew & Frisby 1976) 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 also attractive in view of the extremely high degree of precision with which they may be controlled (Riggs & Niehl 1960; Rashbass & Westheimer 1961*a*).

These observations suggest a scheme for solving the fusion problem in the following way (Marr & Poggio 1977*a*, *b*): (1) Each image is analysed 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 correspondence.

This scheme contains no hysteresis, and therefore does not account for the hysteresis observed by Fender & Julesz (1967). 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 an 'orientation and depth map' of the visible surfaces round a viewer (Marr & Nishihara 1978, fig. 2; Marr 1977, § 3). In this map, information is combined from a number of different and probably independent processes that interpret disparity, motion, shading, texture, and contour information. These ideas are illustrated by the representation shown in figure 3, which Marr & Nishihara called the $2\frac{1}{2}$-D sketch.

Suppose now that the hysteresis Fender & Julesz observed is not due to a co-operative process during matching, but is in fact the result of using a memory buffer, like the $2\frac{1}{2}$-D sketch, in which to store the depth map of the image as it is discovered. Then, the matching 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 matched 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 in the $2\frac{1}{2}$-D sketch; (4) there is a backwards relation between the memory and the masks, acting 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.

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.

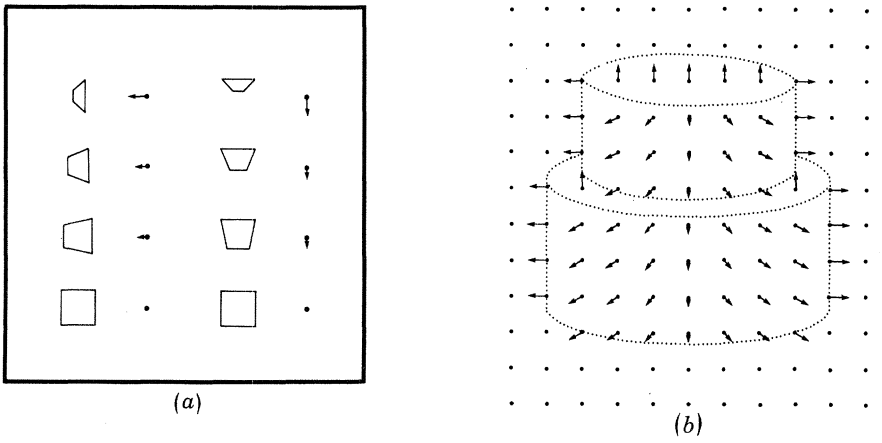


FIGURE 3. Illustration of the $2\frac{1}{2}$-D sketch. In (*a*) the perspective views of small squares placed at various orientations to the viewer are shown. The dots with arrows show a way of representing the orientations of such surfaces symbolically. In (*b*), this representation is used to show the surface orientations of two cylindrical surfaces in front of a background orthogonal to the viewer. The full $2\frac{1}{2}$-D sketch would include rough distances to the surfaces as well as their orientations, contours where surface orientation changes sharply, and contours where depth is discontinuous (subjective contours). A considerable amount of computation is required to maintain these quantities in states that are consistent with one another and with the structure of the outside world (see Marr 1977, § 3). (From Marr & Nishihara 1978, fig. 2.)

The idea that there may be a range of different sized 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 & Giese (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; and (*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 4*a*), whose tuning curves have the form of figure 4*b*, and which have a half power bandwidth of between one and two octaves. (2) The half power bandwidth of the local sensitivity function at each eccentricity

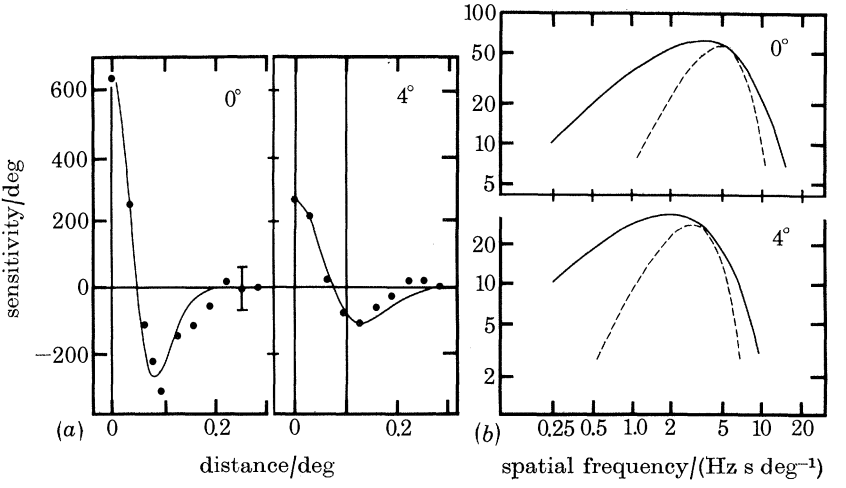


FIGURE 4. (*a*) Line spread functions measured psychophysically at threshold at two different eccentricities. The points are fitted using the difference of two Gaussian functions with space constants in the ratio 1.5 : 1.0. The inhibitory surround exactly balances the excitatory centre so that the area 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. (Redrawn from Wilson & Giese 1977, fig. 9 and 10.)

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 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° the mean width w of the central excitatory region of the mask is about 6′ (range 3′–12′); and at 4° eccentricity, $w = 12'$ (range 6′–24′) (Wilson & Giese 1977, fig. 9; Hines 1976, figs 2 and 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 channel is given by $1/f = \lambda = 2.2w$. These figures agree quite well with physiological studies in the Macaque. Hubel & Wiesel (1974, fig. 6*a*) reported that the mean width of the receptive field (s) increases linearly with eccentricity e (approximately, $s = 0.05e + 0.25°$, so that at

$e = 4°$, $s = 27'$ which gives a value for $w = \frac{1}{3}s$ of about $9'$ as opposed to the figure of $12'$ assumed here for humans). The data of Schiller, Finlay & Volman (1977, p. 1347, figs. 12 and 14) are in rough agreement with Hubel & Wiesel's. (4) Essentially all of the psychophysical data on the detection of spatial patterns at contrast threshold 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. The known discrepancies can however be explained by introducing two extra hypotheses: (5) contrast sensitivities of the various channels are adjusted 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 when contrast is low (Cowan 1977, p. 511).

In a more recent article, Wilson & Bergen (1979) 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 equal to $11.7'$ and $21'$. The sizes of these channels increase with eccentricity in the same way as described above.

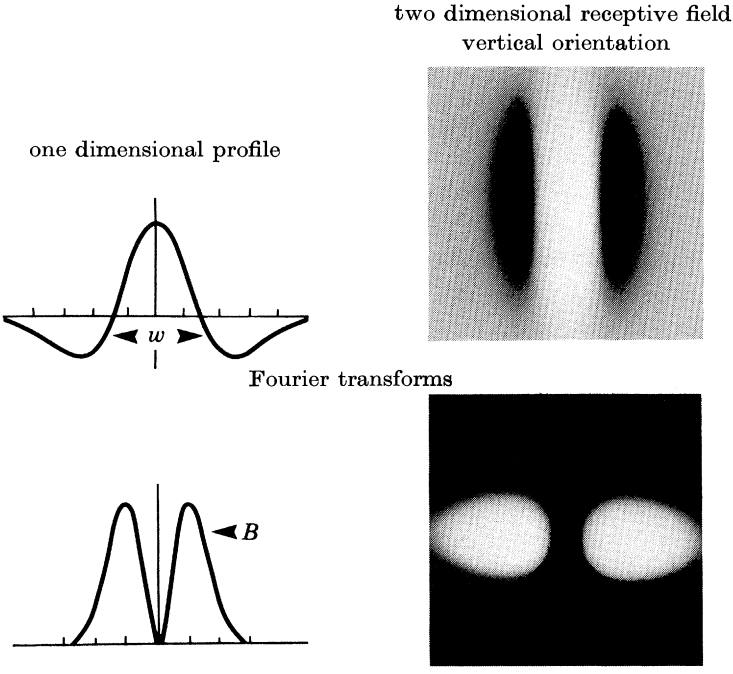
The S channel is the most sensitive under both transient and sustained stimulation, and the U channel is the least, having only $\frac{1}{11}$ to $\frac{1}{4}$ the sensitivity of the S channel. The extent to which the U channel, for example, plays a role in stereopsis is of course unknown.

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

Wilson & Giese's 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 fig. 10 of Wilson & Giese (1977), a channel with peak sensitivity at wavelength λ is band-limited on the high frequency side by wavelengths of about $\frac{2}{3}\lambda$, and $\lambda = 2.2w$. 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

TABLE 1. SPATIAL FILTERING: SUMMARY OF PSYCHOPHYSICAL EVIDENCE

(*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*) For each position and orientation there are four receptive field sizes, the smallest being $\frac{1}{4}$ of the largest. The profile $R(x)$ and Fourier transform $\hat{R}(\omega)$ of each receptive field are given by:

$$R(x) = (2\pi)^{-\frac{1}{2}}\{\sigma_e^{-1}\exp[-x^2/2\sigma_e^2] - \sigma_i^{-1}\exp[-x^2/2\sigma_i^2]\},$$
$$\hat{R}(\omega) = \exp[-\tfrac{1}{2}\omega^2\sigma_e^2] - \exp[-\tfrac{1}{2}\omega^2\sigma_i^2],$$

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 $0°$, and $w = 6' - 34'$ at $4°$. *Note.* receptive field sizes and corresponding spectral sensitivity curves in the suprathreshold condition may be different from the values given here, which were measured at threshold.

(*d*) 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 in suitable units 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 covers the range 0–180° with 12 values about 15° apart and w takes four values in the range defined by (*c*) above.

(*e*) In practice, cells with on-centre receptive fields will signal positive values of the filtered signal, and cells with off-centre receptive fields will signal negative values.

most $\frac{1}{3}\lambda$. Assuming the overall width of the receptive field is about $\frac{3}{2}\lambda$, 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–6σ. (A value of $\sigma = w$ will give an approximately square receptive field.) 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, that is about four 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. It follows that the number of multiplications required to process the image through a given channel is roughly independent of the receptive field size associated with that channel. Not too much weight should be attached to the estimate of 18, although we feel that the sampling density cannot be significantly lower. In the biological situation, total sampling density will decrease as eccentricity increases.

This model of the preliminary processing of the image is summarized in table 1. There are in fact more efficient ways of implementing it (see Marr & Hildreth 1979).

The domain of the matching function

In view of this information, the first step in our theory can be thought of as filtering the left and right spatial images through bar masks of four sizes and about twelve orientations 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 1979) 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 derivative. The bar masks of table 1 measure an approximation to the second directional derivative at roughly the resolution of the mask size, provided that the image does not vary locally along the orientation of the mask (Marr & Hildreth 1979). If this is so, clear signed zero-crossings in the convolution values obtained along a line lying perpendicular to the receptive field's longitudinal axis (cf. Marr 1976, fig. 2) would specify

accurately the position of an edge in the image.† Edges whose orientations lie near the vertical will of course play a dominant role in stereopsis.

In practice, however, it is not enough to use just oriented edges 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, and such breaks cannot be obtained by simple operations on 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 (cf. Marr 1976, p. 496).

Thus zero-crossings and (less importantly) terminations have both to be made explicit (cf. Marr 1976, p. 485). The matching process will then operate on descriptions, of the left and right images, that are built of these two kinds of symbolic primitives, and which specify their precise positions, the mask size and orientation from which they were obtained, and their signs.

Matching

At the heart of the matching problem lies the problem of false targets. 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. We have done this only for zero-crossings, since terminations are sparser and pose less of a false-target problem.

Let us consider a two dimensional image filtered through a vertically oriented mask. Matching will take place between zero-crossings of the same sign along corresponding horizontal lines in the two images. If two such zero-crossings lie very close together in one image, the danger of false targets will arise. Hence a critical parameter in our analysis will be the distance between adjacent zero-crossings of the same sign along each of these lines.

This problem is now one dimensional, and we approach it by estimating the probability distribution of the interval between adjacent zero-crossings of the same sign. This depends on (*a*) the image characteristics, and (*b*) the filter (or mask) characteristics. For (*a*) we take the worst case, that in which the power spectrum of the input to the filter is white (within the filter's spectral range). We also assume, for computational convenience, that the filtered output is a

† It is perhaps worth noting that this rather direct way of locating sharp intensity changes in the image is not the only nor necessarily the best method from the point of view of an actual implementation. It is shown elsewhere (Marr & Hildreth 1979) that under certain conditions, the zeroes in an image filtered through a Laplacian operator (like an X-type retinal ganglion cell) provide an equivalent way of locating edges, whose orientation must then be determined.

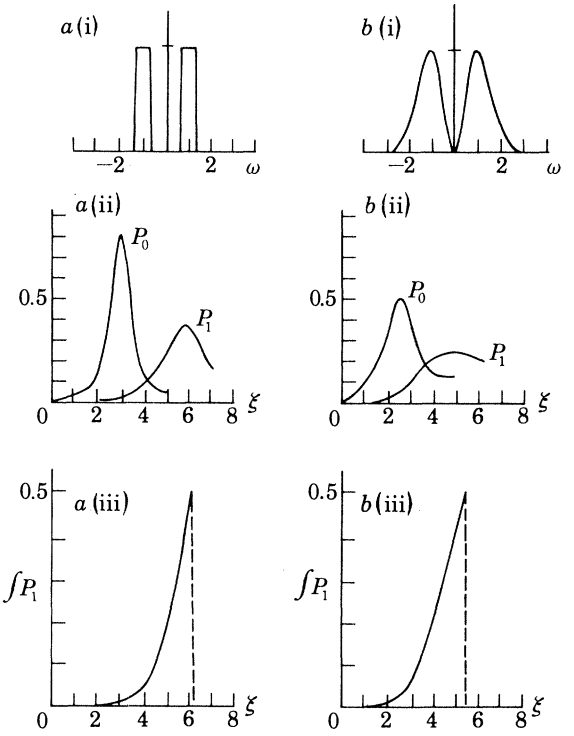


FIGURE 5. Interval distributions for zero-crossings. A 'white' Gaussian random process is passed through a filter with the frequency characteristic (transfer function) shown in (i). The approximate interval distribution for the first (P_0) and second (P_1) zero-crossings of the resulting zero-mean Gaussian process is shown in (ii). Given a positive zero-crossing at the origin, the probability of having another within a distance ξ is approximated by the integral of P_1 and shown in (iii). 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 & Giese (1977). The corresponding spatial distribution of excitation and inhibition, i.e. the inverse Fourier transform of (*b*i) appears, in the same units, in table 1. The ratio of space constants for excitation and inhibition is 1 : 1.5. The width w of the central excitatory portion of the receptive field is 2.8 in the units in which ξ is plotted.

For case (*a*) a probability level of $\int P_1 = 0.001$ 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 $\xi = 5.4$. If the space constant ratio is 1 : 1.75 (Wilson 1978*b*) the values of $\int P_1$ change by not more than 5 %.

Gaussian (zero-mean) process. This hypothesis is quite realistic (E. Hildreth, personal communication).

For (*b*), we examine two cases. Since the actual filters have a half-power bandwidth of around one octave, the first case we consider is that of an ideal linear bandpass filter of width one octave, as illustrated in figure 5*a*(i). The second case (figure 5*b*(i)) is the receptive field suggested by the threshold experiments of Wilson & Giese (1977), consisting of excitatory and inhibitory Gaussian distributions, with space constants in the ratio 1 : 1.5 (see figure 4).

Our problem is now transformed into one that many authors have considered, dating from the pioneering work of Rice (1945), and the appendix sets out the formulae in detail. The results we need are all contained in figure 5. P_0 is the probability distribution of the interval between adjacent zero-crossings (which perforce have opposite signs), and P_1 the distribution of the interval to the second zero-crossing. Since alternate zero-crossings have the same sign, P_1 is the quantity of interest, and its integral $\int P_1$ is also given in figure 5.

$\int P_1$ can be understood in the following way. Suppose a positive zero-crossing occurs at the point O. Then $\int P_1$ represents the probability that at least one other positive zero-crossing will occur within a distance ξ of O. (In figure 5*b*(iii), the width w of the central part of the receptive field associated with the filter is equal to 2.8 on the ξ scale.)

From the graphs in figure 5, we see for example that the 0.05 probability level for $\int P_1$ occurs at $\xi = 4.1$ (approximately $\lambda/1.52$) for the ideal band pass filter one octave wide, centred on wavelength λ (figure 5*a*(i)), and at $\xi = 3.1$ for the receptive field of fig. 5*b*(i). In the second case, ξ is approximately $\frac{1}{2}\lambda$, 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 approximately w ($\xi = 3.1$, $w = 2.8$).

At the 0.001 probability level, the ideal bandpass filter is 50% better (the corresponding ξ is 50% 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 5 provides more details about these results.

We have made a similar comparison between the sustained and transient channels of Wilson (1978*a*) and of Wilson & Bergen (1979). If the sustained channels correspond to the case of figure 5*b*, 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. Our analysis applies directly to channels with vertical orientation, and is roughly valid for channels with orientation near the vertical.

Within a channel of given size, there are in practice two possible ways of dealing with false targets. If one wishes essentially to avoid them altogether, the disparity range over which a match is sought must be restricted to $\pm\frac{1}{2}w$ (see figure 6*a*). 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 $\frac{1}{2}w$, a search for matches in the range $\pm\frac{1}{2}w$ will yield only the correct match R (with probability

0.95). Such a low error rate can be accommodated without resorting to sophisticated algorithms. For example, two reasonable ways of increasing 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 $\frac{1}{2}w$, this procedure will fail, a circumstance that we discuss later.

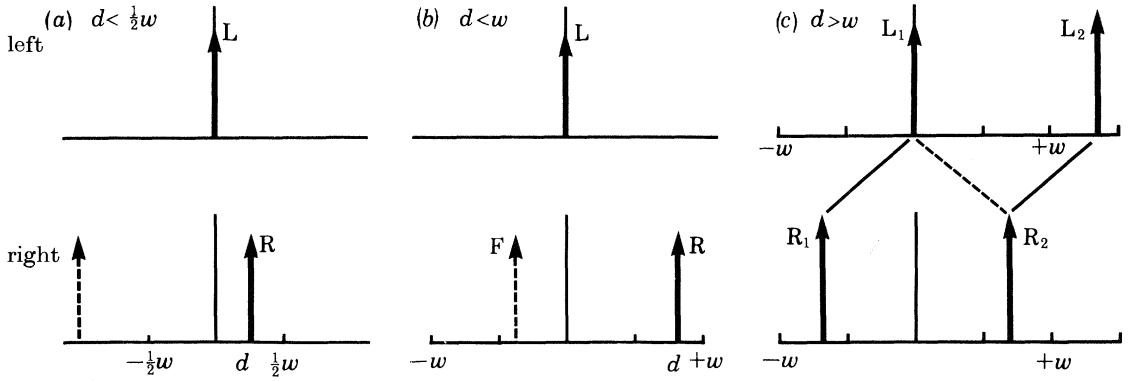


FIGURE 6. The matching process driven from the left image. A zero-crossing L in the left image matches one R displaced by disparity d in the right image. The probability of a false target within w of R is small, so provided that $d < \frac{1}{2}w$ (case *a*), almost no false targets will arise in the disparity range $\pm\frac{1}{2}w$. This gives the first possible algorithm. Alternatively (case *b*), all matches within the range $\pm w$ 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 from either image 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.

There is, however, an alternative strategy, that allows one to 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 6*b*). 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 $\pm 1.5w$). Notice that if there is a match near zero disparity, it is likely ($p > 0.9$) to be the only candidate.

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 there-

fore 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 there are two immediate candidates for matching algorithms. The simpler is restricted to a disparity range of $\pm \frac{1}{2}w$ 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.

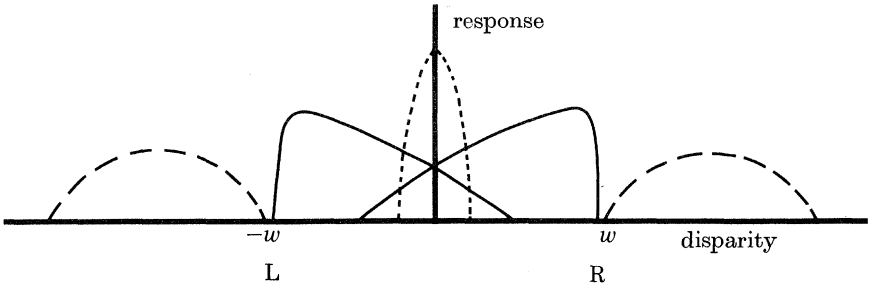


FIGURE 7. An implementation of the second matching algorithm. For each mask size of central width w, there are two pools of disparity detectors, signalling crossed or uncrossed disparities and spanning a range of $\pm w$. There may be additional detectors finely tuned to near-zero disparities. Additional diplopic disparities probably exist beyond this range. They are vetoed by detectors of smaller absolute disparity.

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 7). (In this sense, the first algorithm requires only one 'pool', that is, a single unit sensitive in a graded way to the disparity range $\pm \frac{1}{2}w$.) Candidate matches near zero disparity are likely to be correct, and this fact can be used to improve performance. One way is to add, to the two basic pools, high resolution units tuned to near-zero disparities.

In the second algorithm, matches that are unambiguous or already assigned can 'pull' neighbouring ambiguous matches to whichever alternative has the same sign. This is a form of cooperativity, and 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 1976, 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 7). 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 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 $\pm \frac{1}{2}w$. 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 $\pm \frac{1}{2}w$, 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 $\pm w$, and zero for disparities lying within it. Again, it is relatively easy to discriminate between the situations.

Finally, W. E. L. Grimson (personal communication) has pointed out that matching can be carried out from either image or from both. Observe for example in figure 6*c*, that if matching is initiated from the left image, the match for L_1 is ambiguous, but for L_2 it is unambiguous. Similarly from the right image.

It seems most sensible to initiate matching simultaneously from both images. Then, before any 'pulling', there are three possible outcomes. (1) The matching of an element starting from both images is unambiguous, in which case the two must agree. (2) Matching from one image is ambiguous, but from the other it is not. In this case, the unambiguous match should be chosen. (3) Matching from both images is ambiguous, in which case they must be resolved by pulling from unambiguous neighbours.

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 $\pm w$. 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 $\frac{1}{20}w$). For foveal channels, this means $\pm 3'$ disparity with resolution $10''$ for the smallest, and $\pm 12'$ (perhaps up to $\pm 20'$ if Wilson & Bergen (1979) holds for stereopsis) with resolution $40''$ for the largest ones. At 4° eccentricity, the range is $\pm 5.3'$ to about $\pm 34'$. 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 (see Mitchell 1966; Fender & Julesz 1967; Julesz & Chang 1976; and predictions 3–6 below).

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.

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. A memory like the $2\frac{1}{2}$-D sketch (see figure 3) is computationally desirable on general grounds, because it provides a representation in which information obtained from several early visual processes can be combined (Marr 1977; § 3.6 and table 1). 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 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.

One would however expect a number of constraints derived from the physical world to be embedded in its internal structure. For example, the rule R2 stated early in this article, that disparity changes smoothly almost everywhere, might be implemented in the $2\frac{1}{2}$-D sketch by connections similar to those that implement it in Marr & Poggio's (1976) cooperative algorithm (figure 2*c*). 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; fig. 4.4–5).

Another constraint is, for example, the continuity of discontinuities in the visible surfaces, which we believe underlies the phenomenon of subjective contours (Marr 1977, § 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., Julesz 1971, Blomfield 1973) should take place after stereoscopic fusion.

According to this theory, the memory roughly 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.

The spatial aspects of the $2\frac{1}{2}$-D sketch raise a number of interesting questions. First, 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° for a random dot stereogram. When the complex stereograms given by Julesz (1971, e.g. 4.5–3) are viewed from about 20 cm, 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.

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°, 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.

Vergence movements

Disjunctive eye movements, which change the plane of fixation of the two eyes, are independent of conjunctive eye movements (Rashbass & Westheimer 1961*b*), are smooth rather than saccadic, have a reaction time of about 160 ms, 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°/s per degree of disparity (Rashbass & Westheimer 1961*a*). Vergence

movements are accurate to within about 2′ (Riggs & Niehl 1960), and voluntary binocular saccades preserve vergence nearly exactly (Williams & Fender 1977). 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. These data strongly suggest that the control of vergence movements is continuous rather than ballistic.

Our hypothesis is, that vergence movements are accurately controlled by matches obtained through the various channels, acting either directly or indirectly through the $2\frac{1}{2}$-D sketch. This hypothesis is consistent with the observed strategy and precision of vergence control, and 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 stereograms 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-like search strategy by the vergence control system. In other words, guidance of vergence movements is a simple continuous closed loop process (cf. Richards 1975) which is usually inaccessible from higher levels.

There may 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.

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 two categories according to whether their results are critical and are already available (A), or are critical and not available and therefore amount to predictions (P). 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.

Computation

The algorithm we have described has been implemented, and is apparently reliable at solving the matching problem for stereo pairs of natural images (Grimson & Marr 1979). It depends on the uniqueness and continuity conditions formulated at the beginning of this article, and it is perhaps of some interest to see exactly how.

The continuity assumption is used in two ways. First, vergence movements

driven by the larger masks are assumed to bring the smaller masks into register over a *neighbourhood* of the match obtained through the larger masks. Secondly, local matching ambiguities are resolved by consulting the sign of *nearby* unambiguous matches.

The uniqueness assumption is used in quite a strong way. If a match found from one image is unique, it is assigned without further checking. This is permissible only because the uniqueness assumption is based on true properties of the physical world. If the algorithm is presented with a stereo pair in which the uniqueness assumption is violated, as it is in Panum's limiting case, the algorithm will assign a match that is unique from one image but not from the other (O. J. Braddick, in preparation).

Psychophysics

1 (A, P**). Independent spatial-frequency-tuned channels are known to exist in binocular fusion and rivalry. The theory identifies these with the channels described from monocular experiments (Julesz & Miller 1975; Mayhew & Frisby 1976; Frisby & Mayhew 1979; Wilson & Giese 1977; Cowan 1977; Wilson 1978*a*, *b*; Wilson, Phillips, Rentschler & Hilz 1979; Wilson & Bergen 1979; and Felton, Richards & Smith 1972).

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

3 (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. Stereo acuity should scale with the width w of the smallest active matched channels (i.e. about 10″ for the smallest and 40″ for the largest foveal channels).

4 (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.

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

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

Comments. (1) Mitchell (1966) used small flashed line targets and found, in keeping with earlier studies, that the maximum amount of convergent or divergent disparity without diplopia is 10–14′ in the fovea, and about 30′ at 5° eccentricity. The extent of the so-called Panum fusional area is therefore twice this.

Under stabilized image conditions, Fender & Julesz (1967) found that fusion occurred between line targets (13′ by 1° 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°, routinely achieved fusion up to $\pm 18'$ disparity. Taking all factors into account, these figures seem to be consistent with our expectations.

(2) Prediction 6 should hold for dynamic stereograms with the following caveats. First, motion cues must be eliminated. Secondly nonlinear temporal summation between frames at the receptor level may introduce unwanted low spatial-frequency components in the two images.

7 (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, fig. 2*b* of Mayhew & Frisby (1976) should give rise to depth sensations, but fig. 2*c* should not. In the presence of eye movements, fig. 2*c* 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.

8 (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, λ is approximately $2.2w$ where λ is the centre frequency of the channel.)

9 (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 (figures 6 and 7).

10 (P***). In the absence of eye movements, the perception of tilt in stereoscopically viewed grating pairs of different spatial frequencies is limited by 4, 5, and 6 above.

11 (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).

12 (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.

13 (P***). For a novel two planar stereogram, 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.

14 (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.

15 (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 (13). Once found, their representation in the memory should be labile if an important aspect of the representation there consists of local disparity differences.

Neurophysiology

16 (*partly* A). At each point in the visual field, the scatter of bar mask receptive field sizes is about 4 : 1 (Hubel & Wiesel 1974, figs. 1 and 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.

17 (P**). There exist binocularly driven cells sensitive to 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, size and rough orientation.

18 (P**). Each of the populations defined by (17) is divided into at least two 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 7). Being sensitive to pure disparity, these cells are sensitive to changes in disparity induced by vergence movements. In addition, there may be units quite sharply tuned to near-zero disparities.

19 (P*). In addition to the basic disparity pools of (18), there may exist cells tuned to more outlying (diplopic) disparities (compare figure 7). These cells should be inhibited by any activity in the basic pools (cf. Foley, Applebaum & Richards 1975).

20 (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° 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° disparity range currently being represented in the $2\frac{1}{2}$-D sketch.

21 (P*). The diplopic disparity cells of (20) are especially concerned with the control of disjunctive eye movements.

Comments. 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 different ways, our physiological predictions are more speculative than our psychophysical ones. They should perhaps be regarded as guidelines for investigation rather than as necessary consequences of the theory.

Unfortunately, the technical problems associated with the neuro-

physiology of stereopsis are considerable, and rather little quantitative data is currently available. Since Barlow, Blakemore & Pettigrew's (1967) original paper, relatively few examples of disparity tuning curve have been published (see for example, Pettigrew, Nikara & Bishop 1968; Bishop, Henry & Smith 1971; Nelson, Kato & Bishop 1977). Recently however, Poggio & Fischer (1978, in the monkey), and von der Heydt, Adorjani, Hanny & Baumgartner (1978, in the cat) have published properly controlled disparity tuning curves. On the whole, these studies (see also Clarke, Donaldson & Whitteridge 1976) favour the pools idea (see prediction 18).

Discussion

Perhaps one of the most striking features of our theory is the way it returns to Fender & Julesz's (1967) 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 1979) 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 the predictions 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 figure 6. 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, especially concerning the $2\frac{1}{2}$-D sketch. For example:

(1) What is its exact structure, and how are the various constraints implemented there?

(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) 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?

Finally, we feel that an important feature of this theory is that it grew from an analysis of the computational problems that underlie 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.

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, Eric Grimson, David Hubel, Bela Julesz, John Mayhew and Hugh Wilson, and to Werner Reichardt and the Max Planck Society for their kind hospitality in Tübingen. Karen Prendergast prepared the illustrations. The Royal Society kindly gave permission for reproduction of figure 3, and *Science* and the American Association for the Advancement of Science for figure 2. This work was conducted at the Max-Planck-Institut für Biologische Kybernetik in Tübingen, 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. D.M. was partly supported by NSF contract number 77-07569-MCS.

References

Barlow, H. B., Blakemore, C. & Pettigrew, J. D. 1967 The neural mechanism of binocular depth discrimination. *J. Physiol., Lond.* **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., Lond.* **216**, 39–68.

Blakemore, C. & Campbell, 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., Lond.* **203**, 237–260.

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

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

Clarke, P. G. H., Donaldson, I. M. L. & Whitteridge, D. 1976 Binocular visual mechanisms in cortical areas I and II of the sheep. *J. Physiol., Lond.* **256**, 509–526.

Cowan, J. D. 1977 Some remarks on channel bandwidths for visual contrast detection. *Neurosci. Res. Progr. Bull.* **15**, 492–517.

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

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

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. & Mayhew, J. E. W. 1979 Spatial frequency selective masking and stereopsis. (In preparation.)

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. 1979 A computer implementation of a theory of human stereo vision. (In preparation.)

von der Heydt, R., Adorjani, Cs., Hanny, P. & Baumgartner, G. 1978 Disparity sensitivity and receptive field incongruity of units in the cat striate cortex. *Exp. Brain Res.* **31**, 523–545.

Hines, M. 1976 Line spread function variation near the fovea. *Vision Res.* **16**, 567–572.

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. 1974 Sequence regularity and geometry of orientation columns in monkey striate cortex. *J. comp. Neurol.* **158**, 267–294.

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

Julesz, B. 1971 *Foundations of cyclopean perception*. 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. *Am. 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.

Longuet-Higgins, M. S. 1962 The distribution of intervals between zeros of a stationary random function. *Phil. Trans. R. Soc. Lond.* 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. Lond.* B **275**, 483–524.

Marr, D. 1977 Representing visual information. *AAAS 143rd Annual Meeting. Symposium on Some Mathematical Questions in Biology*, February. Published in *Lectures on mathematics in the life sciences* **10**, 101–180 (1978). Also available as *M.I.T. A.I. Lab. Memo* 415.

Marr, D. & Hildreth, E. 1979 Theory of edge detection. (In preparation.)

Marr, D. & Nishihara, H. K. 1978 Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond.* B **200**, 269–294.

Marr, D., Palm, G. & Poggio, T. 1978 Analysis of a cooperative stereo algorithm. *Biol. Cybernetics* **28**, 223–229.

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

Marr, D. & Poggio, T. 1977*a* A theory of human stereo vision. *M.I.T. A.I. Lab. Memo* 451.

Marr, D. & Poggio, T. 1977*b* Theory of human stereopsis. *J. opt. Soc. Am.* **67**, 1400.

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

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.

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**, 311–410.

Poggio, G. F. & Fischer, B. 1978 Binocular interaction and depth sensitivity of striate and prestriate cortical neurons of the behaving rhesus monkey. *J. Neurophysiol.* **40**, 1392–1405.

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

Rashbass, C. & Westheimer, G. 1961*b* Independence of conjunctive and disjunctive eye movements. *J. Physiol., Lond.* **159**, 361–364.
Rice, S. O. 1945 Mathematical analysis of random noise. *Bell Syst. Tech. J.* **24**, 46–156.
Richards, W. 1970 Stereopsis and steroblindness. *Exp. Brain Res.* **10**, 380–388.
Richards, W. 1971 Anomalous stereoscopic depth perception. *J. opt. Soc. Am.* **61**, 410–414.
Richards, W. 1975 Visual space perception. In *Handbook of Perception*, vol. 5, *Seeing*, ch. 10, pp. 351–386 (ed E. C. Carterette & M. D. Freidman). New York: Academic Press.
Richards, W. A. 1977 Stereopsis with and without monocular cues. *Vision Res.* **17**, 967–969.
Richards, W. A. & Regan, D. 1973 A stereo field map with implications for disparity processing. *Invest. Ophthal.* **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.
Schiller, P. H., Finlay, B. L. & Volman, S. F. 1977 Quantitative studies of single-cell properties in monkey striate cortex. III. Spatial frequency. *J. Neurophysiol.* **39**, 1334–1351.
Sperling, G. 1970 Binocular vision: a physical and a neural theory. *Am. J. Psychol.* **83**, 461–534.
Sugie, N. & Suwa, M. 1977 A scheme for binocular depth perception suggested by neurophysiological evidence. *Biol. Cybernetics* **26**, 1–15.
Waltz, D. 1975 Understanding line drawings of scenes with shadows. In *The psychology of computer vision* (ed. P. H. Winston), pp. 19–91. New York: McGraw-Hill.
Westheimer, G. & Mitchell, D. E. 1969 The sensory stimulus for disjunctive eye movements. *Vision Res.* **9**, 749–755.
Williams, R. H. & Fender, D. H. 1977 The synchrony of binocular saccadic eye movements. *Vision Res.* **17**, 303–306.
Wilson, H. R. 1978*a* Quantitative characterization of two types of line spread function near the fovea. *Vision Res.* **18**, 971–981.
Wilson, H. R. 1978*b* Quantitative prediction of line spread function measurements: implications for channel bandwidths. *Vision Res.* **18**, 493–496.
Wilson, H. R. & Bergen, J. R. 1979 A four mechanism model for spatial vision. *Vision Res.* (in the press).
Wilson, H. R. & Giese, S. C. 1977 Threshold visibility of frequency gradient patterns. *Vision Res.* **17**, 1177–1190.
Wilson, H. R., Phillips, G., Rentschler, I. & Hilz, R. 1979 Spatial probability summation and disinhibition in psychophysically measured line spread functions. *Vision Res.* (in the press).

Appendix. Statistical analysis of zero-crossings

We assume that $f(x) = \int I(x,y)\, h(y)\, \mathrm{d}y$, where $I(x,y)$ is the image intensity and $h(y)$ represents the longitudinal weighting function of the mask, is a white Gaussian process. Our problem is that of finding the distribution of the intervals between alternate zero-crossings by the stationary normal process obtained by filtering $f(x)$ through a linear (bandpass) filter.

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, § 3.4; Longuet-Higgins 1962, eqns 1.2.1 and 1.2.3; Leadbetter 1969):

$$P_0(\xi) = \frac{1}{2\pi}\left[\frac{\psi(0)}{-\psi''(0)}\right]^{\frac{1}{2}} \frac{M_{23}(\xi)}{H(\xi)} (\psi^2(0) - \psi^2(\xi)) [1 + H(\xi) \operatorname{arccot}(-H(\xi))],$$

$$P_1(\xi) = \frac{1}{2\pi}\left[\frac{\psi(0)}{-\psi''(0)}\right]^{\frac{1}{2}} \frac{M_{23}(\xi)}{H(\xi)} (\psi^2(0) - \psi^2(\xi)) [1 - H(\xi) \operatorname{arccot}(H(\xi))],$$

where $\psi(\xi)$ is the autocorrelation of the underlying stochastic process, a prime denotes differentiation with respect to ξ, and also

$$H(\xi) = M_{23}(\xi)\,[M_{22}(\xi) - M_{23}(\xi)]^{-\frac{1}{2}},$$
$$M_{22}(\xi) = -\psi''(0)\,(\psi^2(0) - \psi^2(\xi)) - \psi(0)\,\psi'^2(\xi),$$
$$M_{23}(\xi) = \psi''(\xi)\,(\psi^2(0) - \psi^2(\xi)) + \psi(\xi)\,\psi'^2(\xi).$$

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. The autocorrelation $\psi(\xi)$ can be easily computed analytically for the two filters of figure 5.