REPRESENTING AND COMPUTING VISUAL INFORMATION

DAVID MARR

Vision is the construction of efficient symbolic descriptions from images of the world. An important aspect of vision is the choice of representations for the different kinds of information in a visual scene. In the early stages of the analysis of an image, the representations used depend more on what it is possible to compute from an image than on what is ultimately desirable, but later representations can be more sensitive to the specific needs of recognition. David Marr surveys work in vision at MIT from a perspective in which the representational problems assume a primary importance. An overall framework is suggested for visual information processing which consists of three major levels of representations; the primal sketch, which makes explicit the intensity changes and local two-dimensional geometry of an image; the $2^1/_2$-D sketch, which is a viewer-centered representation of the depth, orientation and discontinuities of the visible surfaces; and the 3-D model representation, which allows an object-centered description of the three-dimensional structure and organization of a viewed shape.

Understanding information processing tasks and vision

Vision is an information processing task, and like any other, it needs understanding at two levels. The first, which I call the computational theory of an information processing task, is concerned with what is being computed and why; and the second level, that at which particular algorithms are designed, with how the computation is to be carried out [Marr and Poggio 1977a]. For example, the theory of the Fourier transform is a level 1 theory, and is expressed independently of ways of obtaining it (algorithms like the Fast Fourier Transform, or the parallel algorithms of coherent optics) that lie at level 2. Chomsky calls level 1 theories competence theories, and level 2 theories performance theories. The theory of a computation must precede the design of algorithms for carrying it out, because one cannot seriously contemplate designing an algorithm or a program until one knows precisely what it is meant to be doing.

I believe this point is worth emphasizing, because it is important to be clear about the level at which one is pursuing one's studies. For example, there has recently been much interest in so-called cooperative algorithms [Marr and Poggio 1976] or relaxation labelling [Rosenfeld, Hummel and Zucker 1976]. The attraction of this technique is that it allows one to write plausible constraints directly into an algorithm, but one must remember that such techniques amount to no more than a style of programming, and they lie at the second of the two levels. They have nothing to do with the theory of vision, whose business it is to derive the constraints and characterize the solutions that are consistent with them.

If one accepts in broad terms this statement of what it means to understand an information processing task, one can go on to ask about the particular theories that one needs to understand vision. Vision can be thought of as a *process*, that produces from images of the external world a description that is useful to the viewer and not cluttered by irrelevant information. These descriptions, in turn, are built or assembled from many

different but fixed representations, each capturing some aspect of the visual scene. In this article, I shall try to present a summary of our work on vision at MIT seen from a perspective in which the representational problems assume a primary importance. I shall include summaries of our present ideas as well as of completed work.

The important point about a representation is that it makes certain information *explicit* (cf. the principle of explicit naming, [Marr 1976]). For example, at some point in the analysis of an image, the intensity changes present there need to be made explicit, so does the geometry -- of the image and of the viewed shape -- and so do other parameters like color, motion, position and binocular disparity. To understand vision thus requires that we first have some idea of which representations to use, and then we can proceed to analyze the computational problems that arise in obtaining and manipulating each representation. Clearly the choice of representation is crucial in any given instance, for an inappropriate choice can lead to unwieldy and inefficient computations. Fortunately, the human visual system offers a good example of an efficient vision processor, and therefore provides important clues to the representations that are most appropriate and likely to yield successful solutions.

This point of view places the nature of the representations at the center of attention, but it is important to remember that the limitations on the processes that create and use these representations are an important factor in determining their structure, because one of the constraints on vision is that the description ultimately produced be derivable from images. In general, the structure of a representation is determined at the lower levels mostly by what it is possible to compute, whereas later on they can afford to be influenced by what it is desirable to compute for the purposes of recognition.

Early processing problems

There are two important kinds of information contained in an intensity array, the intensity changes present there, and the local geometry of the image. The primal sketch [Marr 1976] is a primitive representation that allows this information to be made explicit. Following the clues available from neurophysiology [Hubel and Wiesel 1962], intensity changes are represented by blobs and by oriented elements that specify a position, a contrast, a spatial extent associated with the intensity change, a weak characterization of the type of intensity change involved, and a specification of points at which intensity changes cease (so-called termination points). The representation of local geometry makes explicit two-dimensional geometrical relations between significant items in an image. These include parallel relationships between nearby edges, and the relative positions and orientations of significant places in the image. These significant places are marked by "place-tokens," and they are defined in a variety of ways, by blobs or local patches of different intensity, by small lines, and by the ends of lines or bars. The local geometrical relations between place-tokens are represented by inserting virtual lines that join nearby place-tokens, thus making explicit the existence of a relation between the two tokens, their relative orientation, and the distance between them (figure 1).

The idea of place-tokens and of this way of representing geometrical relations arose from considering the computational problems that are posed by early visual processing, and one of the questions we have been asking is, can one find any psychophysical evidence that the human visual system makes use of a similar representation? We have recently obtained two results related to this point. Stevens [1978] has examined the perception of random-dot interference patterns (figure 2), constructed by superimposing two copies of a random dot pattern where one copy has undergone some composition of expansion, translation, or rotation transformations [Glass 1969]. He found that a simple algorithm suffices to account quantitatively for

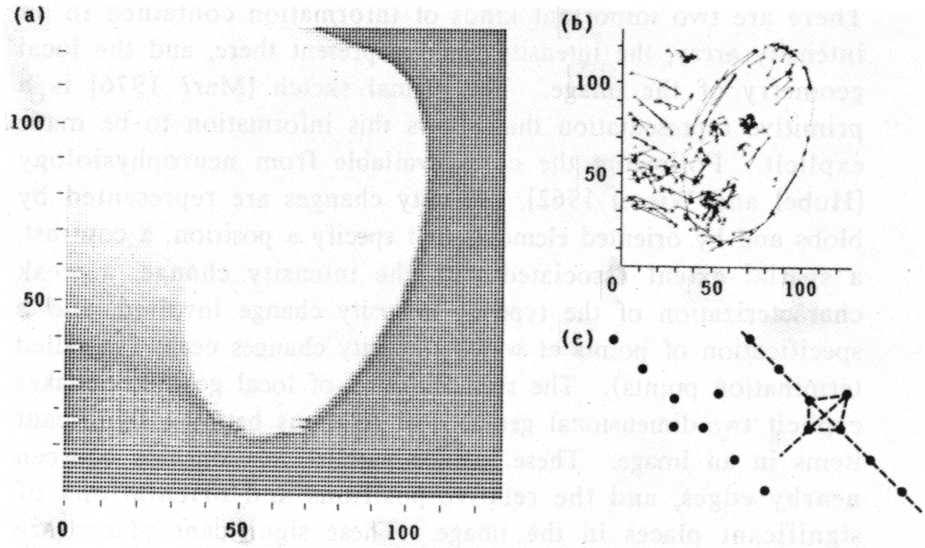


Figure 1. Primal Sketch. The primal sketch makes explicit information held in an intensity array. Changes in intensity are represented by oriented edge, line and bar elements, associated with which is a measure of the contrast and spatial extent of the intensity change. Local two-dimensional geometry of significant places in the image are marked by "place-tokens", which can be defined in a variety of ways, and the geometric relations between them are represented by inserting "virtual lines" between nearby tokens.

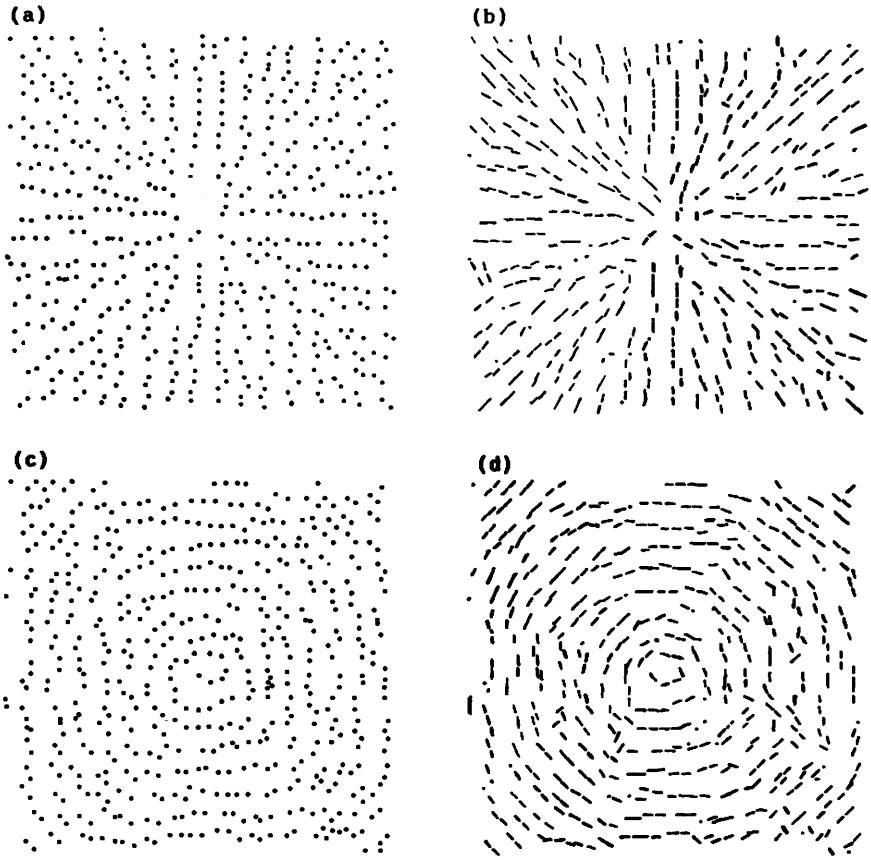


Figure 2. A and C are random-dot interference patterns of the kind described by Glass [1969]. B and D exhibit the results of running the algorithm described in the text and in Figure 3. The neighborhood radius was such that roughly 8 neighborhoods were included [Stevens 1977].

human performance on these patterns. The algorithm consists of three steps:

(1) Each dot defines a place-token. For example, some dots can be replaced by small lines or larger blobs without disrupting the subjective impression of flow.

(2) Virtual lines are inserted between nearby place-tokens, and the neighborhood in which the virtual lines are inserted depends in a predictable way on the density of the dots.

(3) The orientations of the virtual lines attached to all the points in each neighborhood are histogrammed, and locally parallel organization is found by searching for a peak in this histogram. The bucket width that best matches human performance is about 10 degrees.

The details of these steps are set out in figure 3. The interesting features of the algorithm are; (a) It is not iterative. Stevens could find no evidence that human performance rests on a cooperative algorithm, although this type of problem is ideal for that approach. (b) The algorithm is purely local. No global-to-local or top-down interactions are necessary to explain human performance. (c) What the algorithm finds is locally parallel organization. In this case, the organization lies in the virtual lines constructed between nearby dots, but locally parallel organization among the real edges and lines in an image also forms an important part of the structure of an image [Marr 1976].

The second study is one by Schatz [1977] on texture vision discrimination. Marr [1976] suggested that such discriminations could be carried out by first-order discriminations acting on the description in the primal sketch. Marr supposed that certain grouping processes were needed before the discriminations are made in order to account for the full range of human texture discrimination, but in a careful examination of the problem, Schatz found that many of the examples he constructed could be explained by assuming that the discriminations are made only on real edges or on virtual lines inserted between neighboring place-tokens. If this were generally

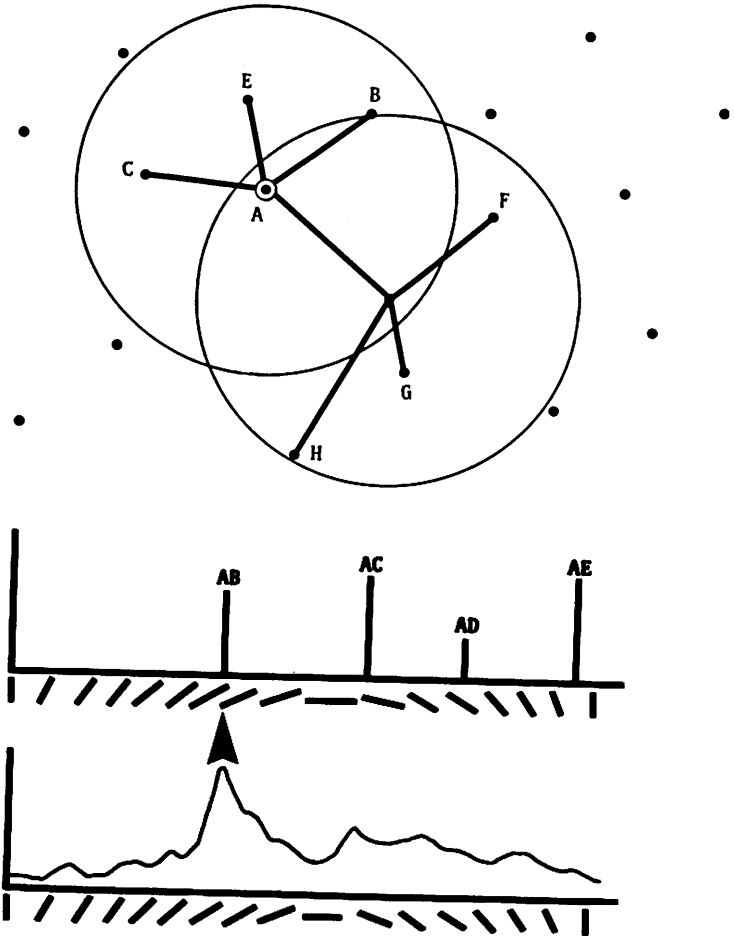
true, it would stand in elegant relation to Julesz's [1975] conjecture, that a necessary condition for the discriminability of two textures is that their dipole statistics differ. This condition is known not to be sufficient, a state of affairs that one can view as implying that we have access to only a proper subset of all dipole statistics. It is possible that this proper subset consists only of real edges and of the virtual lines that join nearby place-tokens.

If one accepts that texture discrimination relies upon first-order discriminations of this type, it is natural to ask how sensitive are the particular discrimination functions that we can bring to bear on an image. Riley [1977] has found evidence that the available functions are extremely coarse. For example, figure 4 consists of a background in which the line segments have a random orientation, surrounding a square containing lines of only three orientations. Surprisingly, the square cannot be discerned without scrutiny. One interpretation of this and related findings is, that discriminations on orientations other than horizontal and vertical are made on the output of 5 channels, each nearly binary, and with an angular width of about 35 degrees -- in other words, only very little information is available about the distribution of orientations in an image. It appears that our discrimination ability is as poor or poorer for the other stimulus dimensions, for example intensity distribution [Riley 1977].

In another study concerned with what can be extracted from an image, Ullman [1976a] enquired about the possible physical basis for the subjective quality of fluorescence, which is normally associated with the presence of a light source. He noted that at a light source boundary, the ratio of intensity to intensity gradient changes sharply, whereas this is not true at reflectance boundaries unless the surface orientation changes sharply. He showed that, in the mini-world of Mondrians, the discriminant to which this leads predicts human performance satisfactorily.

Ken Forbus [1978] has extended this work to the

Figure 3. The algorithm for computing locally parallel structure has three fundamental steps. In the case of the Moire dot patterns, each dot contributes a place token. A virtual line represents the position, separation, and orientation between a pair of neighboring dots. To favor relatively nearer neighbors, relatively short virtual lines are emphasized. The second step is to histogram the orientations of the virtual lines that were constructed. For example, the neighbor D would contribute orientations AD, DF, DG, and DH to the histogram. The final step (after smoothing the histogram) is to determine the orientation at which the histogram peaks, and to select that virtual line (AB) closest to that orientation as the solution. [Stevens 1977]


E
B
C
A
F
G
H
AB
AC
AD
AE

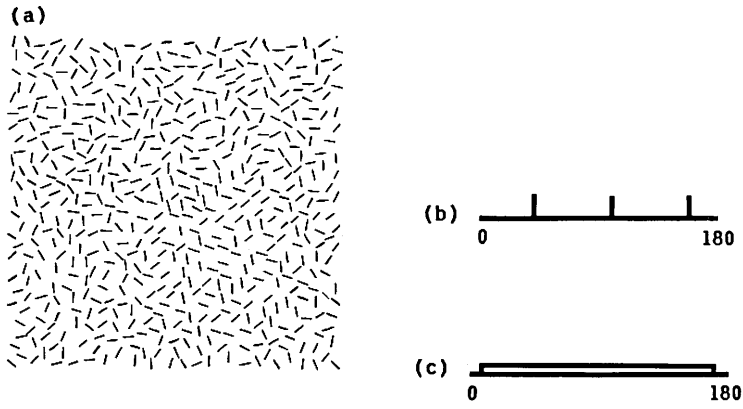


Figure 4. The pattern A contains two regions, one of whose line segments has the orientation distribution shown in B, and the other has the distribution C. Surprisingly, three orientations cannot be distinguished from a random orientation distribution.

detection of surface luster. Since glossiness is due to the specular component of a surface reflectivity function, one can treat the detection of gloss as essentially the detection of light sources that appear reflected in a surface (see [Beck 1974]), and this depends ultimately on the ability to detect light sources. Forbus divided the problem into three categories; (a) in which the specularity is too small to allow gradient measurements, (b) in which both intensity and gradient measurements are available, but the specularity is local (as it is for a curved surface or a point source), and (c) in which the surface is planar and the source is extended. He derived diagnostic criteria for each case.

Delimiting regions from a discriminant

Whenever a region is defined in an image by a predicate, for example by a difference in texture or brightness, one faces the problem of delimiting the region accurately. There are two approaches to designing algorithms for this problem; one is to use the predicate directly, deciding whether a given location lies within or without the region by testing some function of the predicate there. The second approach is to differentiate the predicate, defining the region by its boundaries rather than by properties of its interior.

The difficulties with the problem arise because one is usually ignorant beforehand of the scale at which significant predicate signals may be gathered. For example, suppose one wished to find the boundary between two regions that are distinguished by different densities of dots. Dot density has to be measured by selecting a neighborhood size and counting the number of dots that lie within it. If the neighborhood size is too large, one may not be able to resolve the regions. If it is so small as to contain zero, one or two dots, natural fluctuations may obscure any changes in density.

One solution to this problem is to make the measurements simultaneously at several neighborhood sizes, looking for agreement between the results obtained in those neighborhood sizes that lie just above the size at which random fluctuations appear. This technique can be applied to region finding or to boundary finding, and an example of the results is given in figure 5. The dot density here is not known *a priori.*

This issue is of considerable techical interest, but it is important not to lose sight of the underlying computational problem, which is what kind of boundary is to be found, and why? The techniques of O'Callaghan [1974] for example are designed to find boundaries in dot patterns so accurately that their positions are determined up to the decision about which dots it passes through. The justification for this type of study is that humans can assign boundaries this accurately, but the

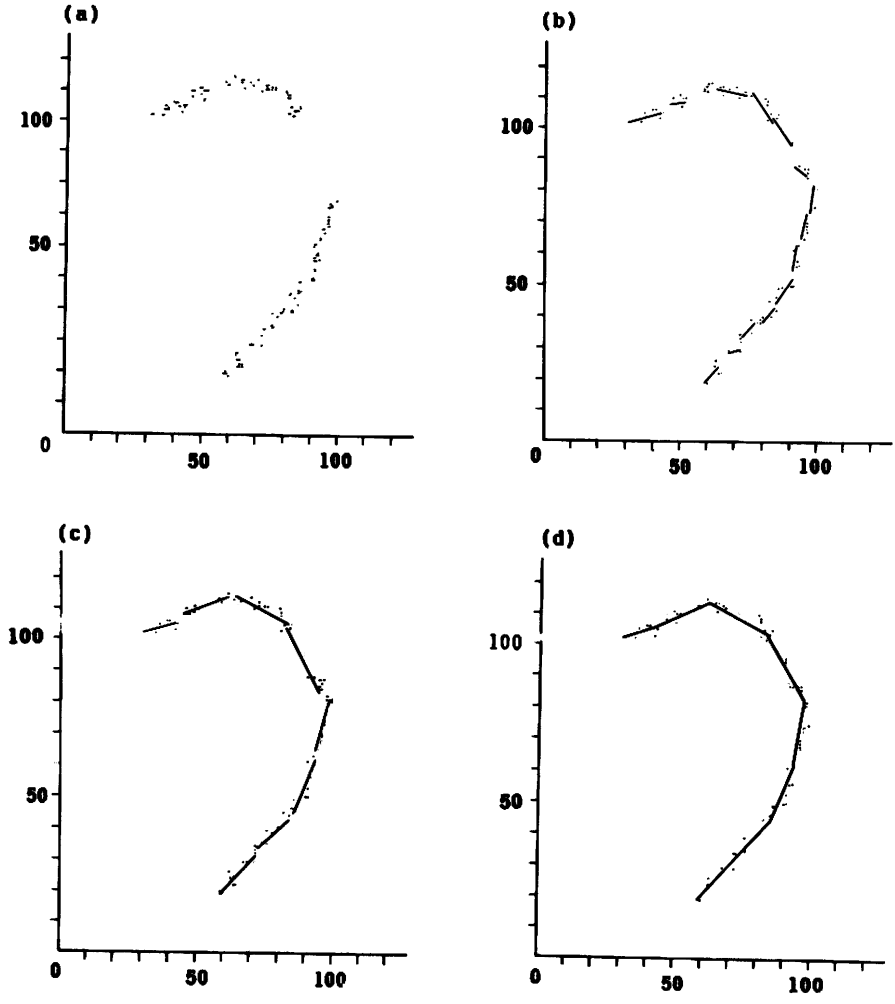


Figure 5. Finding a boundary from dot (or place-token) density changes. Once a rough assignment of boundary points has been made (*a*) local line-fitting (*b*) and grouping (*c* and *d*) techniques can recover a rough specification of the boundary quite easily.

difficulty lies in formulating a reasonable definition of what the boundary is.

This problem is a deep one, touching the heart of the question of what early vision is *for.* I shall return to it later in this essay, but it is perhaps worth remarking here that there seems to be a clear need for being able to do early visual processing roughly and fast as well as more slowly and accurately, which means having ways of handling rough descriptions of regions -- ways of characterizing their approximate extent and shape -- *before* characterizing their precise boundaries. Figure 6 contains one example of a region whose rough extent is clear, but whose exact boundary is not.

The motivation for wanting this is that rough descriptions are very useful during the early stages of building a shape description for recognition [Marr and Nishihara 1977]. For example a man often appears as a roughly vertical rectangle in an image, and this information is useful because it eliminates many other shapes from consideration quite early. Campbell [1977] has suggested that the extraction of rough descriptions from an image may depend on the ability to examine its lower spatial frequencies. Even if this is one of the available mechanisms it is unlikely to be the only one, because sparse line drawings can raise the same problems while having almost no power in their low frequencies. It may be that some notion of rough grouping applied to low resolution place-tokens set up by pieces of contour in the image provides a useful approach to this problem.

Lightness

Ever since Ernst Mach noticed the bands named after him, there has been considerable interest in the problem of computing perceived brightness. Of especial interest is the recent work of Land and McCann [1971] on the retinex theory (see also [Horn 1974]), which is concerned with the quantity they call lightness; and that of Colas-Baudelaire [1973] on the computation of

Figure 6. An example of a region whose rough boundary is clear, but whose exact boundary is not. (Drawing (c) K. Prendergast, 1977).

perceived brightness. Lightness is an approximation to reflectance that is obtained by filtering out slow intensity changes, the underlying idea being that these are usually due to the illuminant, not to changes in reflectance. The problem with this idea is of course that some slow changes in intensity are perceptually important (see [Horn 1977] for an analysis of shape from shading). The linear filter model of Colas-Baudelaire performs well on images in which there are no sharp changes in intensity, but the author found it difficult to extend his model to the more general case. The recent finding of Gilchrist [1977], that perceived depth influences perceived brightness, suggests that some aspects of the problem occur quite late -- in our terms, at the level of the $2^1/_2$-D sketch (see below).

Our own work on the brightness problem is probably not relevant to the perception of brightness, but it is interesting as a demonstration that the primal sketch loses very little information. Woodham and Marr (unpublished program) have written a program that inverts the primal sketch, so that its output is an intensity array. The basic idea is to scan outwards from edges, assigning a constant brightness to points along the scan lines, and arresting the scan when it encounters another edge. Figure 7 exhibits the results of running this program, showing the original image (7a), the primal sketch (7b), and the reconstructed intensity array (7c).

Structure from motion

I said earlier that, especially at the earlier stages of visual information processing, the representations and processes are determined more by what it is possible to compute from an image than by what is desirable. Examples are the problems associated with structure from motion, stereopsis, texture gradients, and shading.

Given a sequence of views of objects in motion, the human visual system is capable of interpreting the changing views in terms of the shapes of the viewed objects, and their motion in

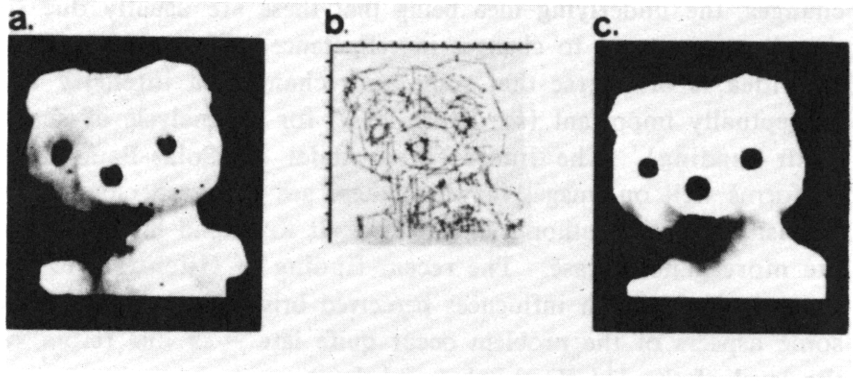


Figure 7. An image (*a*), the spatial components of its primal sketch (*b*), and a reconstruction of the image from the primal sketch (*c*). This shows that our current primal sketch programs lose little of the information in an image.

three-dimensional space. Even if each successive view is unrecognizable, the human observer easily perceives these views in terms of moving objects [Wallach and O'Connell 1953]. To answer the question of how a succession of images yields an interpretation in terms of three-dimensional structure in motion, Ullman [1977] divided the problem into two parts: (1) finding a correspondence between elements in successive views; and (2) determining the three-dimensional structures and their motion from the way corresponding elements move between views.

An important preliminary question about the correspondence problem concerns the level at which it takes place. Is it primarily a low-level relation, established between small and simple parts of the scenes and largely independent of

higher-level knowledge and three-dimensional interpretation? Or do higher level influences, like the interpretation of the whole of a shape from one frame, play an important part in determining the correspondence?

Ullman has assembled a considerable amount of evidence that the former view is correct. For example, figure 8 shows two successive frames, one denoted with full lines and the other with dotted lines. If the whole pattern were being analyzed from one frame, the shape of the wheel extracted, and used to match the elements in the next frame, the observer presented with these frames in rapid succession should perceive them as a whole wheel rotating. Notice however that the inner and outer parts of the wheel have their closest neighbors in one direction, whereas the center parts have theirs in the other; because of this, if the matching were done early and locally, the observer should see the center part rotating one way, and the inner and outer rings rotating the other (as shown with arrows in figure 8). When appropriately timed, this is in fact what happens.

Another line of evidence is the following. The most important factor in finding a correspondence between elements is the distance the element moves from one view to the next. But is this distance an objective two-dimensional measurement or an interpreted movement in three-dimensional space? There is some confusion in the literature about this point, since many studies have assumed that correspondence strength is linked to the smoothness of apparent motion [Kolers 1972], and this is apparently more closely related to three- than to two-dimensional distances. Ullman [1977] has however shown that this assumption is false, and that it is the two-dimensional distance alone that determines the correspondence.

The second part of the problem is to determine the three-dimensional structure once the correspondence between successive views has been established. Unless this problem is constrained in some way, it cannot be solved, so one has to search for reasonable assumptions on which to base the design of one's algorithms. (This state of affairs is a common one in the

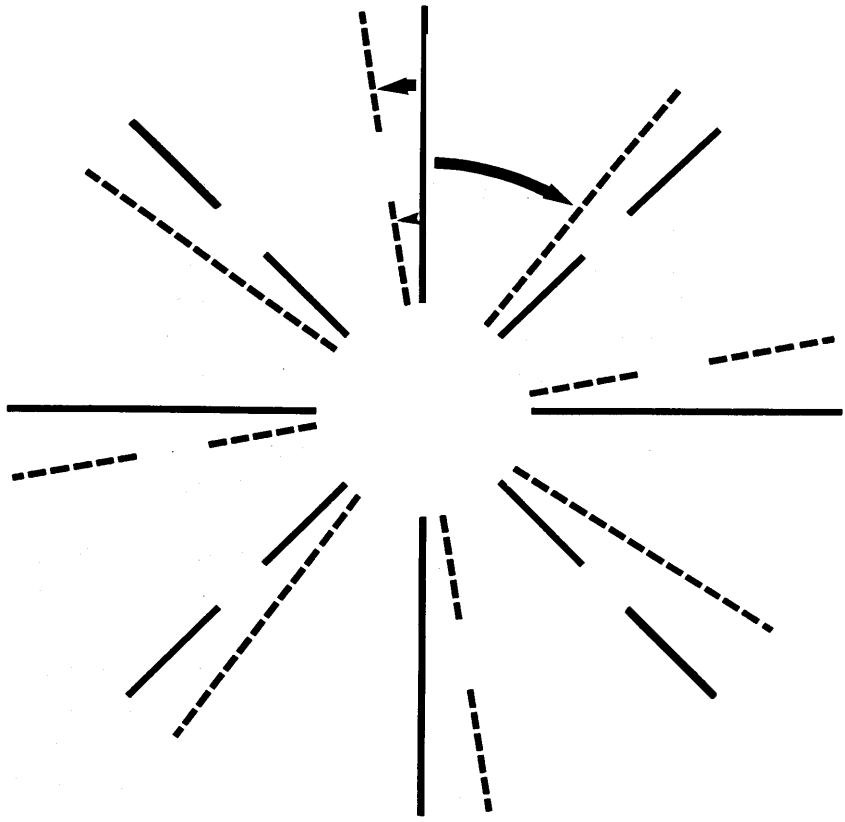


Figure 8. Evidence that the correspondence problem for apparent motion involves matching operations that act at a low level [Ullman 1977].

theory of visual processes, as we shall see when we discuss the problems of stereopsis, and shape from contour). Ullman suggested basing the interpretation on the following assumptions; (1) any two-dimensional transformation that has a unique interpretation as a rigid body moving in space should be interpreted as such an object in motion, and (2) that the imaging process is locally an orthogonal projection. He then showed that under orthogonal projection, three-dimensional shape and motion may be recovered from as little as three views each showing the image of the same five points, no four of which are coplanar. This result leads to algorithms capable of recovering shape and motion from scenes containing arbitrary objects in motion. The final question is whether the algorithms that humans employ to recover shape and motion rely on these same two assumptions, and this question is currently under investigation. The important point here is that for more human-like algorithms, the number of views can be traded off against the accuracy of the computation, decreasing the emphasis on the particular number "three."

Stereopsis

Ever since Julesz [1971] made the first random-dot stereogram, it has been clear that at least to a first approximation stereo vision can be regarded as a modular component of the human visual system. Marr [1974] and Marr and Poggio [1976] formulated the computational theory of the stereo matching problem in the following way:

(R1) Uniqueness. Each item from each image may be assigned at most one disparity value. This condition rests on the premise that the items to be matched correspond to physical marks on a surface, and so can be in only one place at a time.

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

By representing these constraints geometrically, Marr and

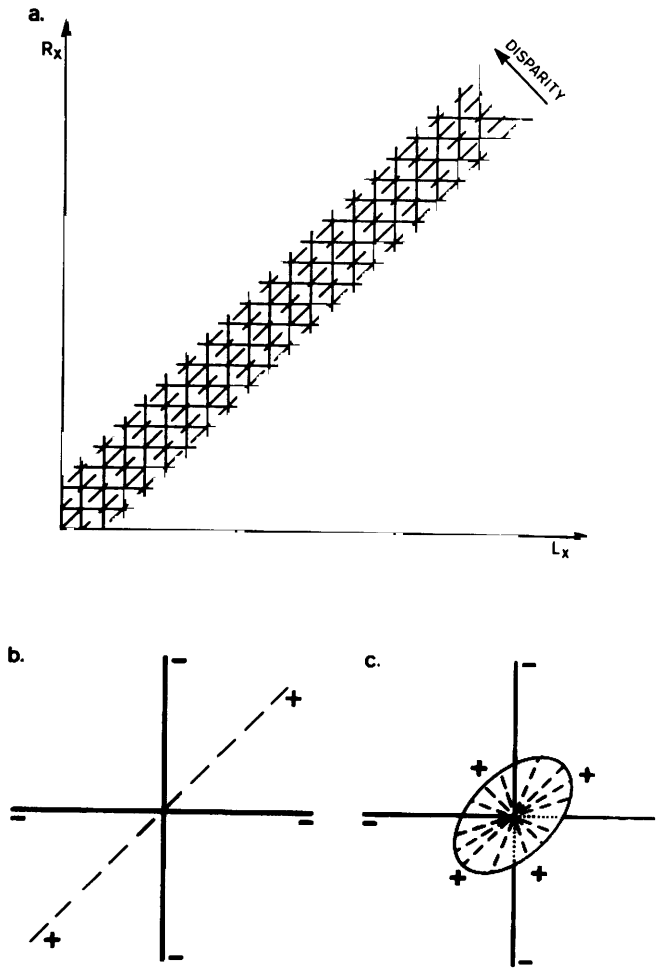


Figure 9. The structure of a network for implementing the algorithm described by equation 1. Such a network was used to solve the stereograms exhibited in figures 10 and 11. [Marr & Poggio 1976]

Poggio [1976] embodied them in a cooperative algorithm. In figure 9, *Lx* and *Rx* represent the positions of descriptive elements from the left and right views, and the horizontal and vertical lines indicate the range of disparity values that can be assigned to left-eye and right-eye elements. The uniqueness condition then corresponds to the assertion that only one disparity value may be "on" along each horizontal or vertical line. The continuity condition states that we seek solutions that tend to spread along the dotted diagonals, which are lines of constant disparity, and between adjacent diagonals. Figure 9b shows how this geometry appears at each intersection point. Figure 9c gives the corresponding local geometry when the images are two-dimensional rather than one. Figure 9a shows the explicit structure of the two rules *R1* and *R2* for the case of a one-dimensional image, and it also represents the structure of a network for implementing the algorithm described by equation 1 Solid lines represent "inhibitory" interactions, and dotted lines represent "excitatory" ones. 9b gives the local structure at each node of the network 9a. This algorithm may be extended to two-dimensional images, in which case each node in the corresponding network has the local structure shown in 9c. Such a network was used to solve the stereograms exhibited in figures 10 and figure 11.

It can be shown [Marr, Poggio and Palm 1977] that, if a network is created with the positive and negative connections shown in figure 9c, states of such a network that satisfy the constraints on the computation are stable, and that given suitable inputs, the network will converge to these stable states for a wide variety of the control parameters. Thus one can think of the network as defining an algorithm that operates on many input elements to produce a global organization *via* local but highly interactive constraints. Formally, the algorithm reads:

$$C_{xyd}^{(n-1)} = u\left\{ \sum_{x'y'd' \in S(xyd)} C_{xyd}^{(n)} - \epsilon \sum_{x'y'd' \in O(xyd)} C_{x'y'd'}^{(n)} + C_{xyd}^{(o)} \right\}$$

where $u(z) = 0$ if $z < \theta$, and $u(z) = 1$ otherwise; S and O are the circular and thick line neighborhoods of the cell C_{xyd} in figure 9c. This is an example of a "cooperative" algorithm [Marr and Poggio 1977a], and it exhibits typical non-linear cooperative phenomena like hysteresis, filling-in, and disorder-order transitions. Figures 10 and 11 illustrate two applications of the algorithm to random-dot stereograms.

In figure 10 the initial state of the network C_{xyd} is defined by the input such that a node takes the value 1 if it occurs at the intersection of a 1 in the left and right eyes (see figure -9), and it has value 0 otherwise. The network iterates on this initial state, and the parameters used here, as suggested by the combinatorial analysis, were θ = 3.0, ϵ = 2.0 and M = 5, where θ is the threshold and M is the diameter of the "excitatory" neighborhood illustrated in figure 9c. The stereograms themselves are labelled LEFT and RIGHT, the initial state of the network as 0, and the state after n iterations is marked as such. To understand how the figures represent states of the network, imagine looking at it from above. The different disparity layers in the network lie in parallel planes spread out horizontally, so that the viewer is looking down through them. In each plane, some nodes are on and some are off. Each of the seven layers in the network has been assigned a different gray level, so that a node that is switched on in the top layer (corresponding to a disparity of +3 pixels) contributes a dark point to the image, and one that is switched on in the lowest layer (disparity = -3) contributes a lighter point. Initially (iteration 0) the network is disorganized, but in the final state, stable order has been achieved (iteration 14), and the inverted wedding-cake structure has been found. The density of this stereogram is 50%.

The algorithm of equation 1 is capable of solving

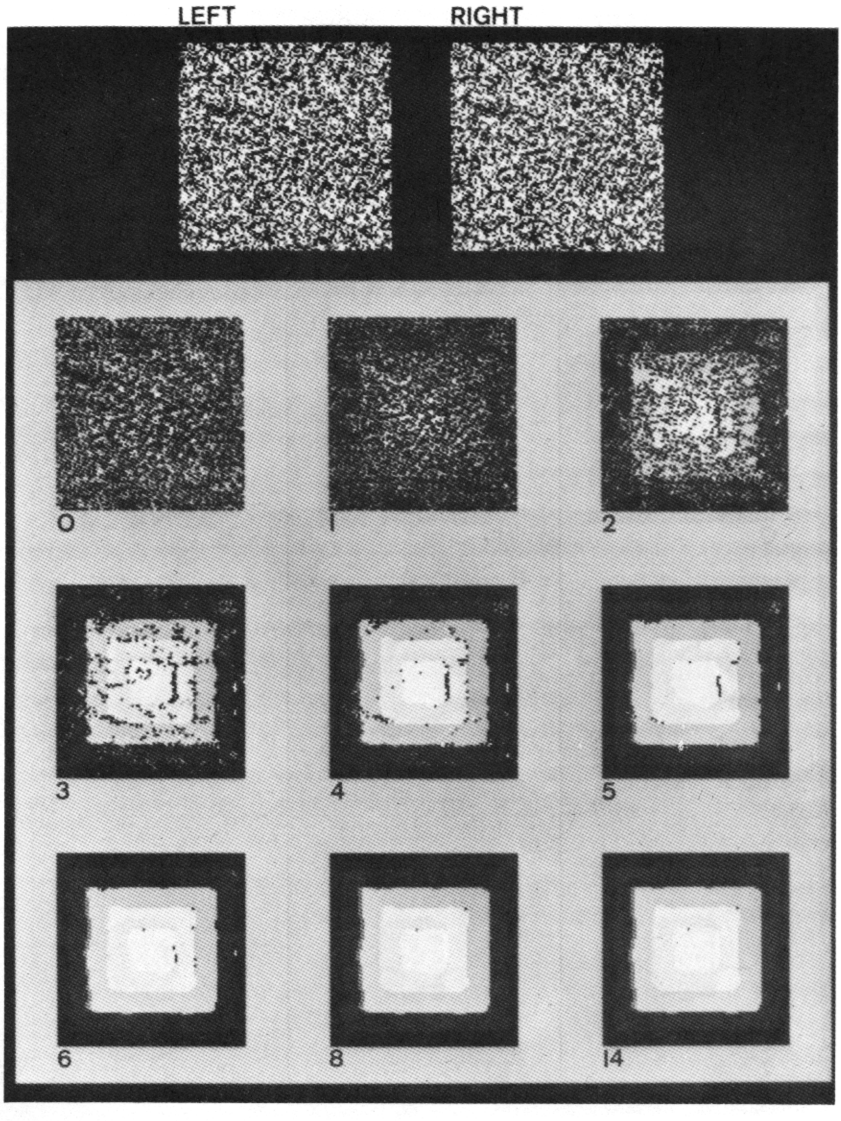


Figure 10. This shows the results of applying the algorithm defined by equation 1 to a random-dot stereograms. The density is 50%. [Marr & Poggio 1976].

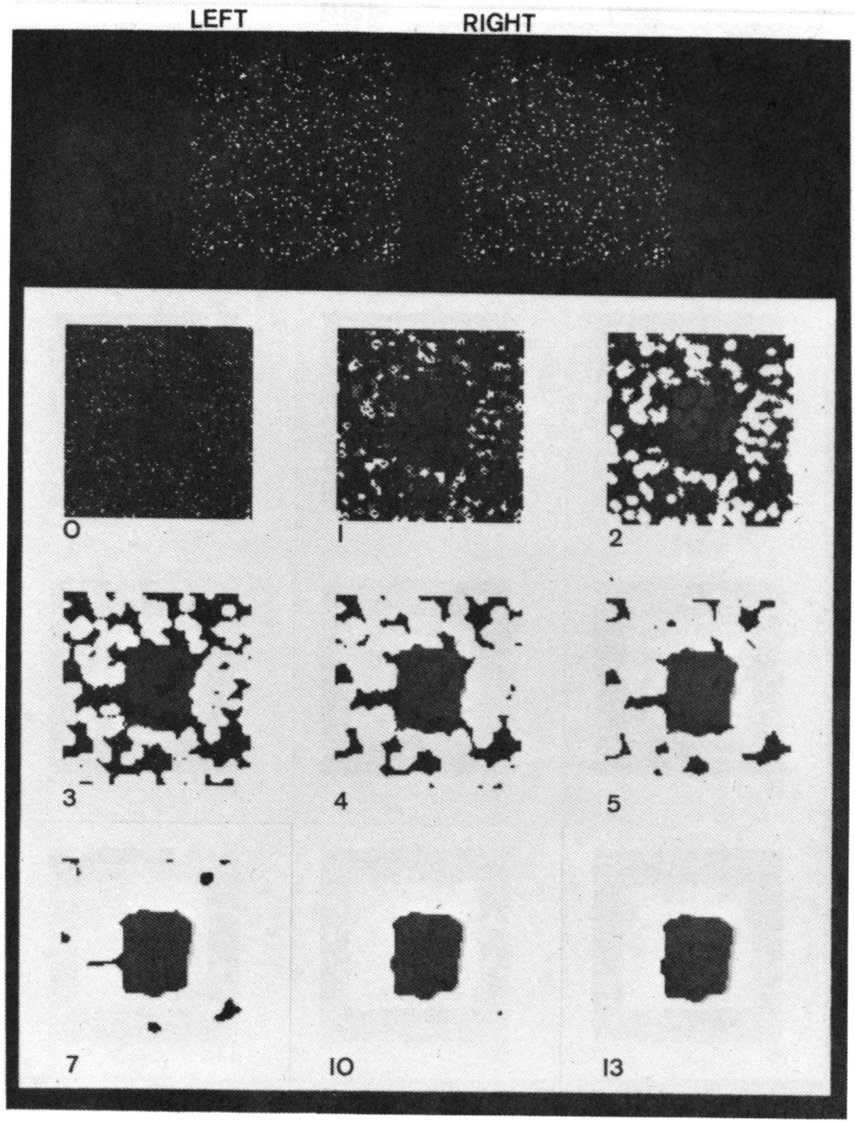


Figure 11. The algorithm of equation 1, with the parameter values used in Figure 10, but with less density. [Marr and Poggio 1976].

random-dot stereograms with densities from 50% down to less than 10%, as shown in figure 11. For this and smaller densities, the algorithm converges increasingly slowly. If a simple homeostatic mechanism is allowed to control the threshold θ as a function of the average activity (number of "on" cells) at each iteration, the algorithm can solve stereograms whose density is very low. In this example, the density is 5% and the central square has a disparity of +2 relative to the background. The algorithm "fills in" those areas where no dots are present, but it takes several more iterations to arrive near the solution than in cases where the density is 50%. When we look at a sparse stereogram, we perceive the shapes in it as cleaner than those found by the algorithm. This seems to be due to subjective contours that arise between dots that lie on shape boundaries.

There are a number of findings that cast doubt on the relevance of this algorithm to the question of how human stereo vision works. The most important of these findings are (a) the apparently crucial role played by eye-movements in human stereo vision (see especially [Richards 1977]); (b) our ability to tolerate up to 15% expansion of one image [Julesz 1971]; (c) our ability to tolerate the severe defocussing of one image [Julesz 1971]; (d) evidence that stereo detectors are organized into three "pools" (convergent, zero disparity, and divergent) and that this organization is important for stereo vision [Richards 1971]; and (e) our ability to perceive depth in rivalrous stereograms [Mayhew and Frisby 1976]. These difficulties led Marr and Poggio [1977b] to formulate a second stereo algorithm, designed specifically as a model for human stereopsis.

Our first stereo theory was inspired by Julesz's belief that stereoscopic fusion is a cooperative process -- a belief based primarily on the observation that it exhibits hysteresis. The main problem with the cooperative algorithm is that it apparently works too well in some ways (it performs better that humans do when eye-movements are eliminated), and not well enough in others (humans see depth in rivalrous stereograms). Our ability to fuse two images when one is blurred, the rivalrous stereogram

results of Mayhew and Frisby [1976], and the recent results of Julesz and Miller [1976] on the existence of independent spatial-frequency-tuned channels in binocular fusion, suggest that several copies of the image, obtained by successively coarser filtering, are used during fusion, perhaps helping one another in a way similar to that in which local regions help each other in our cooperative algorithm.

The second idea was a notion that originated with Marr and Nishihara [1977] and about which I shall have more to say later, which is that one of the things early visual processing does is to construct a "depth map" of the surfaces round a viewer. In this map, each direction away from the viewer is associated with a distance (or some function of distance) and a surface orientation. We have christened the resulting data structure the $2^1/_2$-D sketch.

The important point here is that the $2^1/_2$-D sketch is in some sense a memory. This provided the key idea: Suppose that the hysteresis Julesz observed is not due to a cooperative process at all, 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, and in fact it would not even be necessary for the whole image ever to be fused everywhere provided that a depth map of the viewed surface were built and maintained in this intermediate memory. This idea leads to the following theory. (1) Each image is convolved with bar-shaped masks of various sizes, and matching takes place between peak mask values for disparities up to about twice the panel-width of the mask (see [Felton, Richards and Smith 1972]), for pairs of masks of the same size and polarity. (2) Wide masks can control vergence movements, thus causing small masks to come into correspondence. (3) When a correspondence is achieved, it is held and written down somewhere (e.g. in the $2^1/_2$-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.

This theory leads to many experimental predictions, which are currently being tested.

Intermediate processing problems

We have discussed the types of information that need to be represented early in the processing of visual information, and we have examined the computational structure of some of the processes that can derive and maintain this information. We turn now to the question of what all this information is to be used for.

The current approach to machine vision assumes that the next step in visual processing consists of a process called *segmentation*, whose purpose is to divide the image into regions that are meaningful either in terms of physical objects or for the purpose at hand. Despite considerable efforts over a long period, the theory and practice of segmentation remain primitive, and once again I believe that the main reason lies in the failure to formulate precisely the goals of this stage of the processing. What for example is an object? Is a head one? Is it still one if it is attached to a body? What about a man on horseback?

These questions point to some of the difficulties one has when trying to formulate what should be recovered as a region from early visual processing. Furthermore, however one chooses to answer them, it is usually still impossible to recover the desired regions using only local grouping techniques acting on a representation like the primal sketch. Most images are too complex, and even the simplest images cannot often be segmented entirely at that level [Marr 1976].

Something additional is clearly needed, and one approach to the dilemma has been to invoke specialized knowledge about the nature of the scenes being viewed to aid segmentation of the image into regions that correspond roughly to the objects expected in the scene. Tenenbaum and Barrow [1976], for example, applied knowledge about several different types of scene

to the segmentation of images of landscapes, an office, a room, and a compressor. Freuder [1976] used a similar approach to identify a hammer in a simple scene. If this approach were correct, it would mean that a central problem for vision is arranging for the right piece of specialized knowledge to be made available at the appropriate time during segmentation. Freuder's work, for example, was almost entirely devoted to the design of a heterarchical control system that made this possible. More recently, the constraint relaxation technique of Rosenfeld, Hummel and Zucker [1976] has attracted considerable attention for just this reason, that it appears to offer a technique whereby constraints drawn from disparate sources may be applied to the segmentation problem whilst incurring only minimal penalties in control. It is however difficult to analyze such algorithms rigorously even in very clearly defined situations [Marr, Poggio and Palm 1977], and in the naturally more diffuse circumstances that surround the segmentation problem, it may often be impossible.

Reformulating the problem

The basic problem seems to be how to formulate precisely the next stage of visual processing. Given a representation like the primal sketch, and the many possible boundary-defining processes that are naturally associated with it, which boundaries should one attend to and why? The segmentation approach fails because objects and desirable regions are not visually primitive constructions, and hence cannot be recovered reliably from the primal sketch or similar representation without additional specialized knowledge. If we are to succeed, we must discover precisely what information it is that needs to be made explicit at this stage, what, if any, additional knowledge it is appropriate to apply, and we must design a representation that matches these requirements.

In order to search for clues to a suitable representation, let us return to the physics of the situation. The primal sketch

represents intensity changes and the local two-dimensional geometry of an image. The principle factors that determine these are (1) the illuminant, (2) surface reflectance, (3) the shape of the visible surface, and (4) the vantage point. The first two factors raise the difficult problems of color and brightness, and I shall not discuss them further. The third and fourth factors are independent of the first two (whether two shapes are the same does not depend upon their colors or on the lighting), and so may be treated separately.

I shall argue that, since most early visual processes extract information about the visible surface, it is these surfaces, their shape and disposition relative to the viewer, that need to be made explicit at this point in the processing. Furthermore, because surfaces exist in three-dimensional space, this imposes constraints on them that are general, and not confined to particular objects. It is these constraints that constitute the *a priori* knowledge that it is appropriate to bring to bear next.

One example of the exploitation of fairly general constraints was the work of Waltz [1975], who formulated the constraints that apply to images of polyhedra. The representation on which that work was based was line drawings, but these are not suitable for our needs here because part of the task we wish to carry out is the discovery of physical edges that are only weakly present or even absent in the primal sketch. The approach of Mackworth [1973] was closer to what we want, since it involved a primitive way of representing surfaces.

Part of our task in formulating the problem of intermediate vision is therefore the examination of ways of representing and reasoning about surfaces. We therefore start our enquiry by discussing the general nature of shape representations. What kinds are there, and how may one decide among them? Although it is difficult to formulate a completely general classification of shape representations, Marr and Nishihara [1977] attempted to set out the basic design choices that have to be made when a representation is formulated. They concluded that there are three characteristics of a shape

representation that are largely responsible for determining the information that it makes explicit. The first is the type of *coordinate system* it uses, whether it is defined relative to the viewer or to the object being viewed; the second characteristic concerns the nature of the *shape primitives* used by the representation, that is, the elements whose positions the coordinate system is used to define. Are they two- or three-dimensional, in what sizes do they come, and how detailed are they? And the third characteristic is concerned with the organization a representation imposes on the information in a description; for example is the description modular or does it have little internal structure? We have two sources of information that can help us to formulate the important issues in intermediate visual information processing, firstly the computational problems that arise, and secondly, psychophysics.

Vision provides several sources of information about shape. The most direct are stereo and motion, but texture gradients in a single image are nearly as effective, and the theatrical techniques of facial make-up rely on the sensitivity of perceived shape to shading. It often happens that some parts of a scene are open to inspection by some of these techniques, and other parts by others. Yet different as the techniques are, they have two important characteristics in common. They rely on information from the image rather than on *a priori* knowledge about the shapes of the viewed objects; and the information they specify concerns the depth or surface orientation at arbitrary points in an image, rather than the depth or orientation associated with particular objects.

If one views a stereo pair of a complex surface, like a crumpled newspaper or the "leaves" cube of Ittelson (1960), one can easily state the surface orientation of any piece of the surface, and whether one piece is nearer to or further from the viewer than its neighbors. Nevertheless one's memory for the shape of the surface is poor, despite the vividness of its surface orientation during perception. Furthermore, if the surface contains elements nearly parallel to the line of sight, their

apparent surface orientation when viewed monocularly can differ from the apparent surface orientation when viewed binocularly.

From these observations, one can perhaps draw some simple inferences.

- There is at least one internal representation of the depth, or surface orientation, or both, associated with each surface point in a scene.

- Because surface orientation can be associated with unfamiliar shapes, its representation probably precedes the decomposition of the scene into objects. (This point is particularly relevant to our discussion of intermediate visual information processing.)

- Because the apparent orientation of a surface element can change, depending on whether it is viewed binocularly or monocularly, the representation of surface orientation is probably driven almost entirely by perceptual processes, and is influenced only slightly by specific knowledge of what the surface orientation actually is. Our ability to "perceive" the surface much better than we can "memorize" it may also be connected with this point.

In addition, it seems likely that the different sources of information can influence the *same* representation of surface orientation.

The computational problem

In order to make the most efficient use of these different and often complementary sources of information, they need to be combined in some way. The computational question is, how best to do this? The natural answer is to seek some representation of the visual scene that makes explicit just the information these processes can deliver.

Fortunately, the physical interpretation of the representation we seek is clear. All these processes deliver information about the depth or surface orientation associated with surfaces in an image, and these are well-defined physical quantities. We therefore seek a way of making this information explicit, of maintaining it in a consistent state, and perhaps also of incorporating into the representation any physical constraints that hold for the values that depth and surface orientation take over the kinds of surface that occur in the real world. Table 1 lists the type of information that the different early processes can extract from images. The interesting point here is that although processes like stereo and motion are in principle capable of delivering depth information directly, they are in practice more likely to deliver information about local *changes* in depth, for example by measuring local changes in disparity. Texture gradients and shading provide more direct information about surface orientation. In addition, occlusion, brightness, and size clues can deliver information about discontinuities in depth. It is for example amazing how clear an impression of depth can be obtained from a monocular image containing bright or dim rectangles of different sizes against a dark background. The main function of the representation we seek is therefore not only to make explicit information about depth, local surface orientation, and discontinuities in these quantities, but also to create and maintain a global representation of depth that is consistent with the local cues that these sources provide. We call such a representation the $2^1/_2$-D sketch, and the next section describes a particular candidate for it.

A possible form for the $2^1/_2$-D sketch

The example I give for the $2^1/_2$-D sketch is a viewer-centered representation, which uses surface primitives of one (small) size. It includes a representation of contours of surface discontinuity, and it has enough internal computational structure to maintain its descriptions of depth, surface orientation and surface

Table 1

The form in which various early visual processes deliver information about the changes in a scene.

r = depth
δr = small, local changes in depth
Δr = large changes in depth
$\underset{\sim}{s}$ = local surface orientation

Information source	Natural parameter
Stereo	Disparity, hence especially δr and Δr
Motion	r, hence δr, Δr
Shading	$\underset{\sim}{s}$
Texture gradients	$\underset{\sim}{s}$
Perspective cues	$\underset{\sim}{s}$
Occlusion	Δr

discontinuity in a consistent state. The representation itself has no additional internal structure.

Depth may be represented by a scalar quantity *r*, the distance from the viewer of a point on a surface. Surface discontinuities may be represented by oriented line elements. Surface orientation may be represented by a unit vector *(x, y, z)* in three-dimensional space. Following those who have used gradient space ([Huffman 1971] [Horn 1977]) we can rewrite this as *(p, q, 1)*, which can be represented as a vector *(p, q)* in two-dimensional space. In other words, surface orientation may be represented by covering an image with needles. The length of each needle defines the dip of the surface at that point, so that zero length corresponds to a surface that is perpendicular to the vector from the viewer to that point, and the length increases as the surface tilts away from the viewer. The orientation of the needle defines the direction of the surface's dip. Figure 12 illustrates this representation.

In principle, the relation between depth and surface orientation is straightforward -- one is simply the integral of the other, taken over regions bounded by surface discontinuities. It is therefore possible to devise a representation with intrinsic computational facilities that can maintain the two variables, of depth and surface orientation, in a consistent state. But note that, in any such scheme, *surface discontinuities* acquire a special status (as curves across which integration stops). Furthermore, if the representation is an active one, maintaining consistency through largely local operations, curves that mark surface discontinuities (e.g. contours that arise from occluding contours in the image) must be "filled in" completely, so that at no point along an object boundary can the integration leak across it. It is interesting that subjective contours have this property, and that they are closely related to subjective changes in brightness that are often associated with changes in perceived depth. If the human visual processor contains a representation that resembles the $2^1/_2$-D sketch, it would therefore be interesting to ask whether subjective contours occur within it. (See [Ullman 1976b]

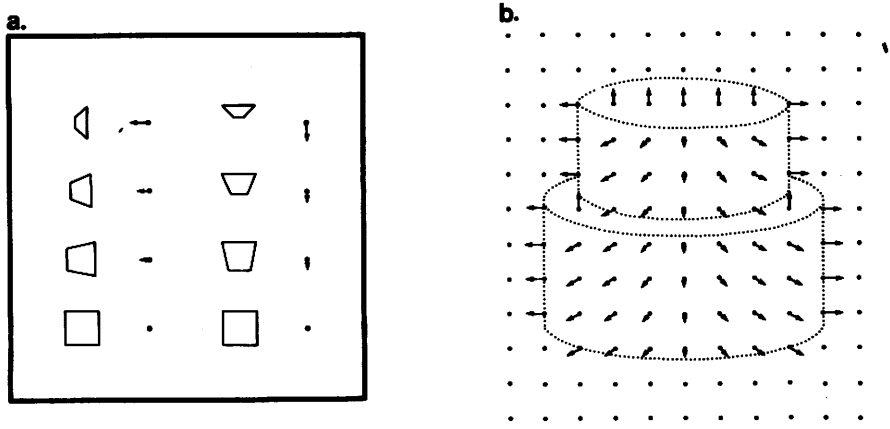


Figure 12. The $2^1/_2$-D sketch represents depth, contours of surface discontinuity, and the orientation of visible surfaces. A convenient representation of surface orientation is described in the text and illustrated here. The orientation of the needles is determined by the projection of the surface normal on the image plane, and the length of the needles represents the dip out of that plane (*a*). A typical $2^1/_2$-D sketch appears in *b*, although depth information is not represented in the figure.

for an analysis of the shape of curved subjective contours).

In summary, my argument is that the $2^1/_2$-D sketch is useful because it makes explicit information about the image in a form that is closely matched to what early visual processes can deliver. We can formulate the goals of intermediate visual processing as being primarily the construction of this representation, discovering for example what are the surface orientations in a scene, which of the contours in the primal sketch correspond to surface discontinuities and should therefore be represented in the $2^1/_2$-D sketch, and which contours are missing in the primal sketch and need to be inserted into the

$2^1/_2$-D sketch in order to bring it into a state that is consistent with the structure of three-dimensional space. This formulation avoids the difficulties associated with the terms "region" and "object," and allows one to ask precise questions about the computational structure of the $2^1/_2$-D sketch and of processes to create and maintain it. We are currently much occupied with these problems.

Later processing problems

The $2^1/_2$-D sketch is a poor representation for the purposes of recognition because it is unstable (in the sense of [Marr and Nishihara 1977]), it depends on the vantage point, and it fails to make explicit pieces of a shape (like an arm) that are larger that the primitive size. Except for the simplest of purposes, it is an inadequate vehicle for a visual system to convey information about shape to other processes, and so I turn now to representations that are more suitable for recognition tasks.

If one were to design a shape representation to suit the problems of recognition, one would naturally base it on an object-centered coordinate system. In addition, one would have to include shape primitives of many different sizes, so as to be able to make explicit shape characteristics that can range from a wart to an elephant. Marr and Nishihara [1977] discuss these questions in detail, and I shall not repeat their observations here. The deepest issues are those raised by having to define an object-based coordinate system. Since they are central to the problem of defining representations for use in later processing of visual information, I shall spend the remainder of the essay discussing this topic.

Marr and Nishihara [1977] pointed out that there are two types of object-centered coordinate system that one might attempt to define precisely. One refers all locations on an object to a single coordinate frame that embraces the entire object, and the other distributes the coordinate system, making it local to each articulated component or individual shape characteristic. Marr

and Nishihara concluded that the second of these schemes is the more desirable, and they gave as an example the representation illustrated in figure 13. But with a representation of this kind, the most difficult questions begin after its internal structure has been defined. How can one define canonically the coordinate scheme for an arbitrary shape, and even more difficult, how can such a thing be found from an image *before* a description of the viewed shape has been computed? Some kind of answers to these questions must be found if the representation is to be used for recognition.

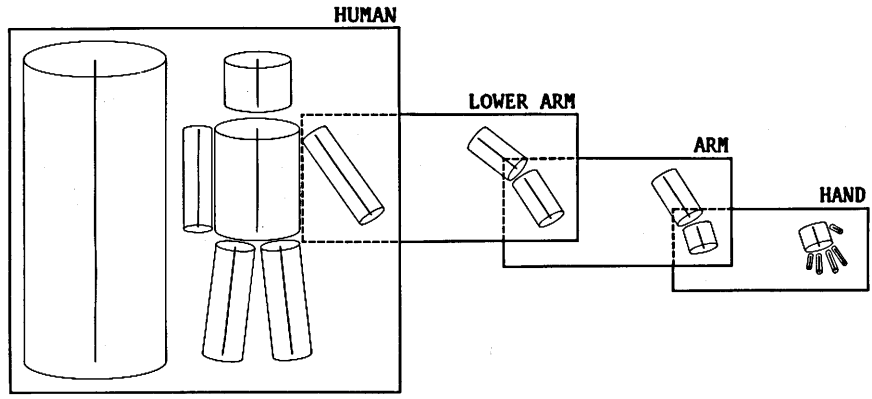


Figure 13. Organization of shape information in a 3-D model description. Each box corresponds to a 3-D model. Its model axis is on the left side, and the arrangement of its component axes are shown on the right side.

If the coordinate system used for a given shape is to be canonical, its definition must take advantage of any salient geometrical characteristics that the shape possesses. For example,

if a shape has natural axes, distinguished by length or by symmetry, then they should be used. The coordinate system for a sausage should take advantage of its major axis, and for a face, of its axis of symmetry.

Highly symmetrical objects, like a sphere, square, or circular disc, will inevitably lead to ambiguities in the choice of coordinate systems. For a shape as regular as a sphere this poses no great problem, because its description in all reasonable systems is the same. One can even allow other factors, like the direction of motion or of spin, to influence the choice of coordinate frame. For other shapes, the existence of more than one possible choice probably means that one has to represent the object in several ways. This is acceptable provided that the number of ways is small. For example, there are four possible axes on which one might wish to base the coordinate system for representing a door, the midlines along its length, its width, its thickness, and to represent how the door opens, the axis of its hinges. For a typewriter, there are two choices at the top level; an axis parallel to its width, because that is usually its largest dimension, and the axis about which a typewriter is roughly symmetrical.

In general, if an axis can be distinguished in a shape, it can be used as the basis for a local coordinate system. One approach to the problem of defining object-centered coordinate systems is therefore to examine the class of shapes having an axis as an integral part of their structure. One such is the class of *generalized cones.* (A generalized cone is the surface swept out by moving a cross section of constant shape but smoothly varying size along an axis, as in figure 14).

Binford [1971] drew attention to this class of surfaces, suggesting that it might provide a convenient way of describing three-dimensional surfaces for the purposes of computer vision. I regard it as an important class not because the shapes themselves are easily decribable, but because the presence of an axis allows one to define a canonical local coordinate system. Fortunately many objects, especially those whose shape was achieved by

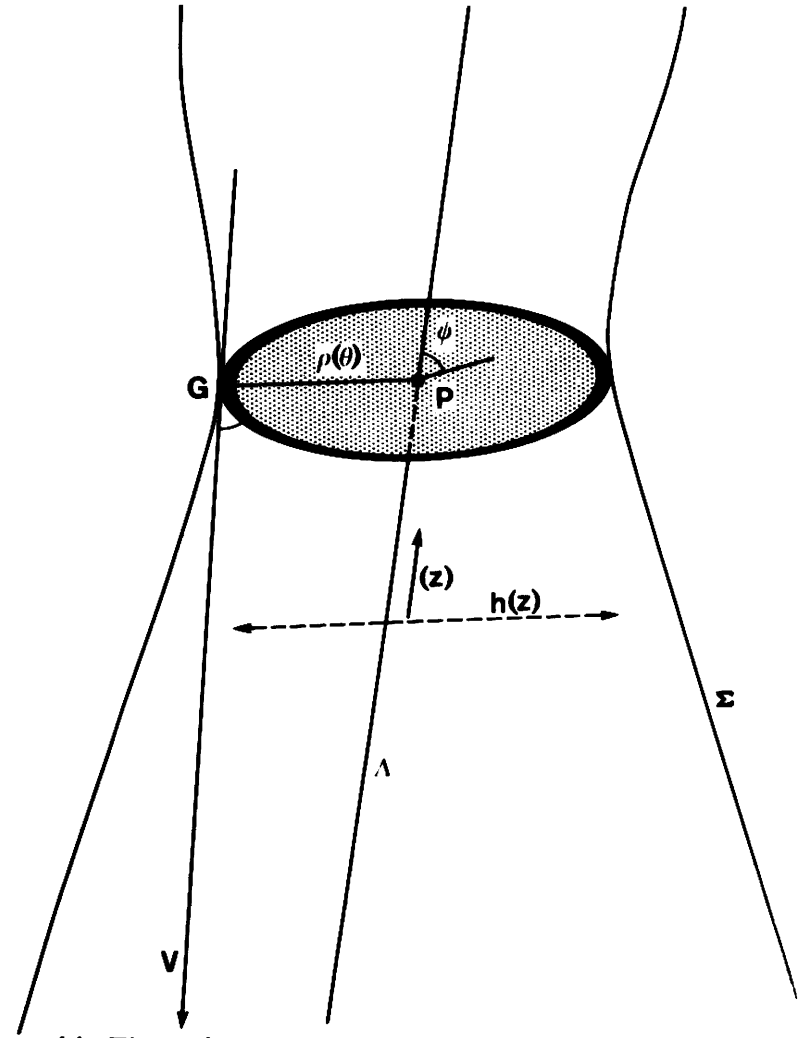
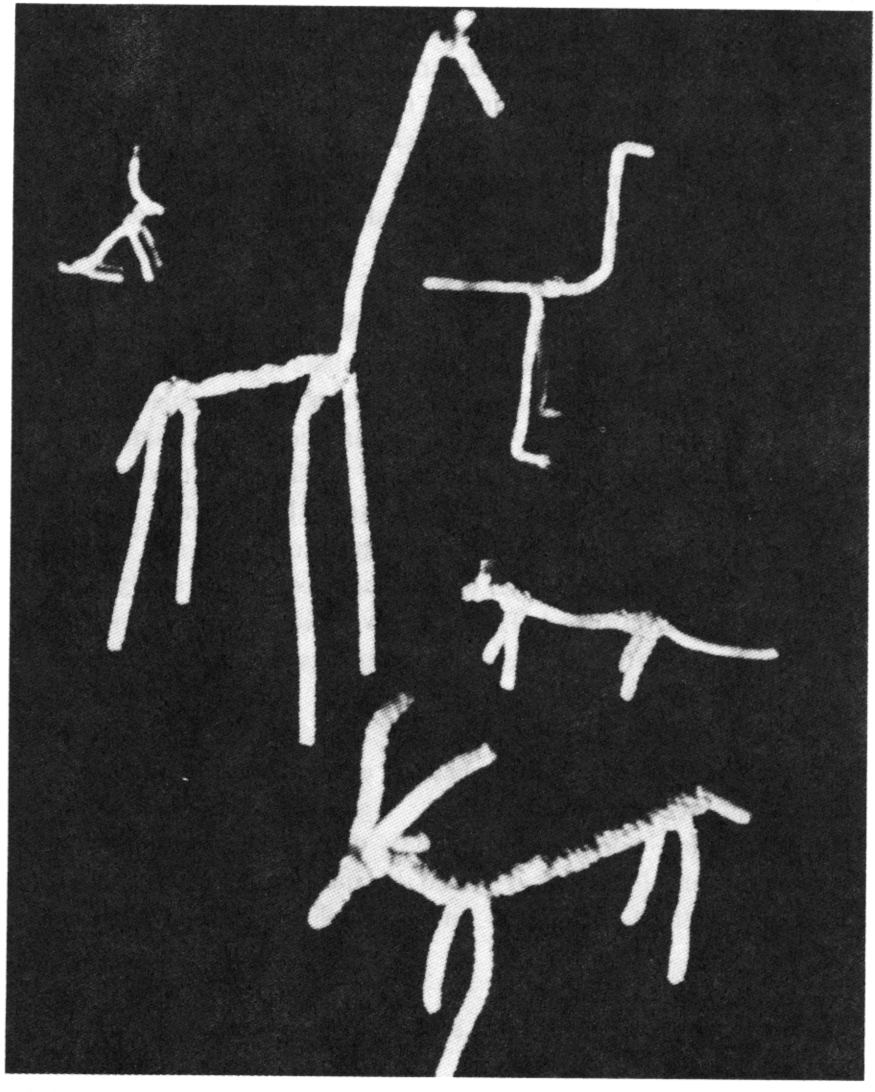


Figure 14. The **definition** of a generalized cone. A generalized cone is the surface generated by moving a smooth cross-section ρ along a straight axis Λ. The cross-section may vary smoothly in size (as prescribed by the function $h(z)$), but its shape remains constant. The eccentricity of the cone is the angle ψ between its axis and a plane containing a cross-section.

Figure 15. These pipecleaner figures illustrate the point that a shape representation does not have to reproduce a shape's surface in order to describe it adequately for recognition; as we see here, animal shapes can be portrayed quite effectively by the arrangement and relative sizes of a small number of sticks. The simplicity of these descriptions is due to the correspondence between the sticks shown here and natural or canonical axes of the shapes described. To be useful for recognition, a shape representation must be based on characteristics that are uniquely defined by the shape and which can be derived reliably from images of it. [Marr & Nishihara 1977]



growth, are described quite naturally in terms of one or more generalized cones. The animal shapes in figure 15 provide some examples -- the individual sticks are simply axes of generalized cones that approximate the shapes of parts of these animals. Many artifacts can also be described in this way, like a car (a small box sitting atop and in the middle of a longer one), and a building (a box with a vertical axis).

It is important to remember that there exist surfaces that cannot conveniently be approximated by generalized cones, for example a cake that has been cut at its intersection with some arbitrary plane, or the surface formed by a crumpled newspaper. Cases like the cake can be dealt with by introducing suitable surface primitives that describe the plane of the cut, but the crumpled newspaper poses apparently intractable problems.

Even if a shape possesses a canonical coordinate system, one is still faced with the problem of finding it from an image. Blum [1973], Agin [1972] and Nevatia [1974] have addressed problems that are related to this question. Blum's sym-axis theory is an interesting one, because he specifies precisely what it is that is computed from a two-dimensional outline. Unfortunately, it is not clear that what this theory computes is in fact useful for shape recognition (see e.g. figure 16), and when applied to a three-dimensional shape, the sym-axis is in general a two-dimensional sheet, so it cannot easily be used to define an object-centered coordinate system. Agin's and Nevatia's work, on the other hand, concerns the analysis of a depth map. This is an important problem, and it would be interesting to see a careful analysis of the conditions under which their techniques will succeed.

Finding the natural coordinate system from an image

My own interest in the problem grew from the 3-D representation theory of Marr and Nishihara [1977], in particular from the question of how to interpret the outlines of objects as seen in a two-dimensional image. The rest of this essay

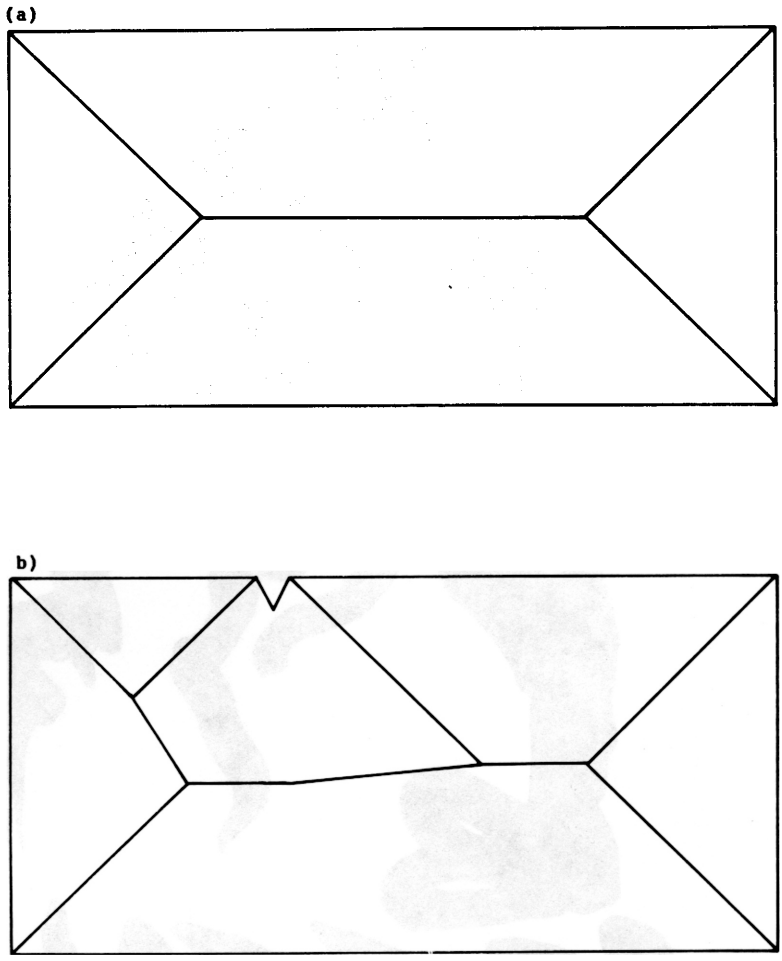


Figure 16. Blum's [1973] grassfire technique for recovering an axis from a silhouette is undesirably sensitive to small perturbations in the contour. *a* shows the Blum transform of a rectangle, and *b*, of a rectangle with a notch [Agin 1972].

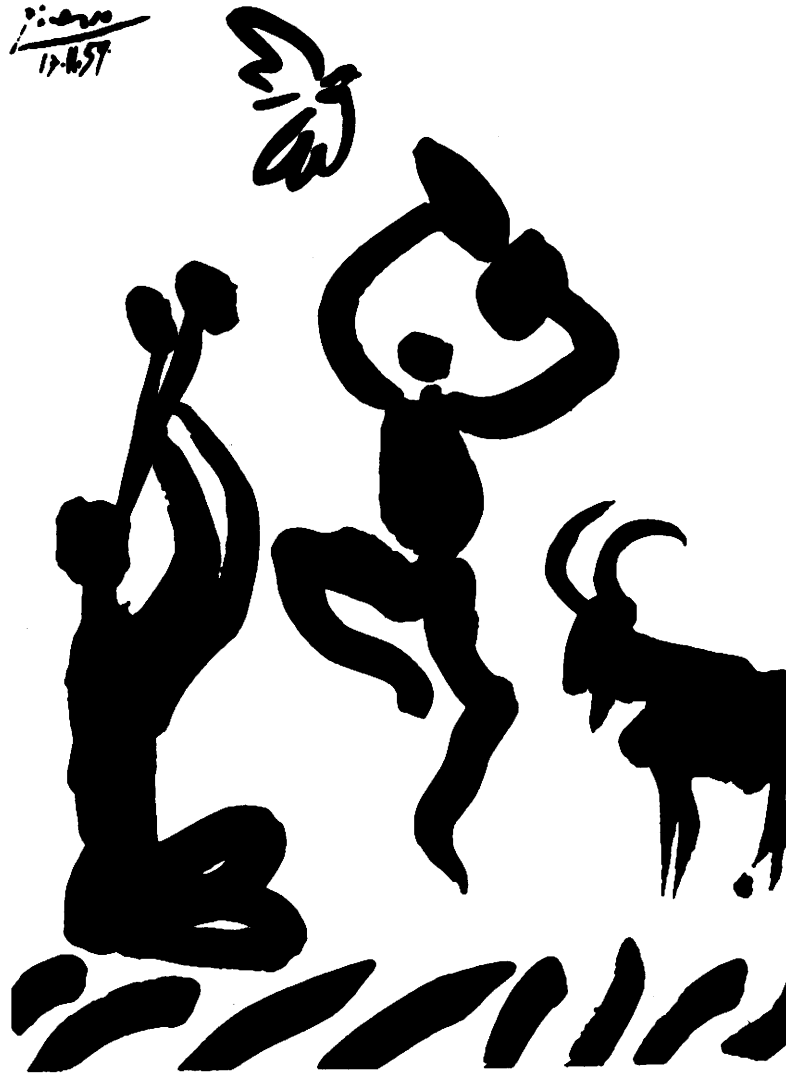


Figure 17. "Rites of spring" by P. Picasso. We immediately interpret the silhouettes in terms of particular 3-D surfaces, despite the paucity of information in the image.

summarizes a recent article by Marr [1977a]. The starting point for this work was the observation that when one looks at the silhouettes in Picasso's work "Rites of Spring" (figure 17), one perceives them in terms of very particular three-dimensional shapes, some familiar, some less so. This is quite remarkable, because the silhouettes could in theory have been generated by an infinite variety of shapes which, from other viewpoints, have no discernable similarities to the shapes we perceive. One can perhaps attribute part of the phenomenon to a familiarity with the depicted shapes; but not all of it, because one can use the medium of a silhouette to convey a new shape, and because even with considerable effort it is difficult to imagine the more bizarre three-dimensional surfaces that could have given rise to the same silhouettes. The paradox is, that the bounding contours in figure 17 apparently tell us more than they should about the shape of the dark figures. For example, neighboring points on such a contour could in general arise from widely separated points on the original surface, but our perceptual interpretation usually ignores this possibility.

The first observation to be made here is that the occluding contours that bound these silhouettes are contours of surface discontinuity, that is precisely the contours with which the $2^1/_2$-D sketch is concerned. Second, because we can interpret the contours as three-dimensional shapes, implicit in the way we interpret them must lie some *a priori* assumptions that allow us to infer a shape from an outline. If a surface violates these assumptions, our analysis will be wrong, in the sense that the shape we assign to the contours will differ from the shape that actually caused them. An everyday example of this phenomenon is the shadowgraph, where the appropriate arrangement of one's hands can, to the surprise and delight of a child, produce the shadow of an apparently quite different shape, like a duck or a rabbit.

What assumptions is it reasonable to suppose that we make? In order to explain them, I need to define the four structures that appear in figure 18. These are (1) some three

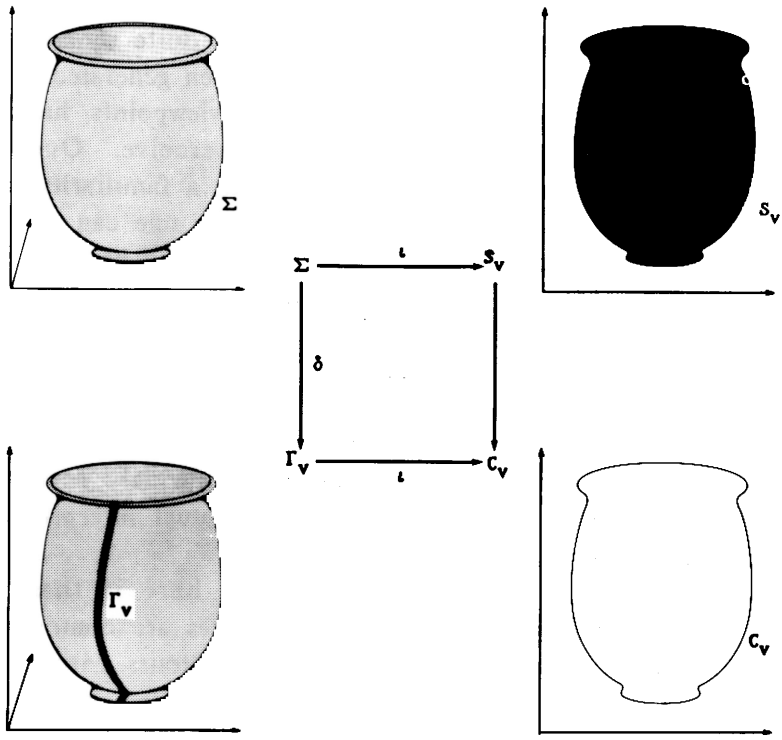


Figure 18. From viewpoint V, the three-dimensional surface Σ forms the silhouette S_V in the image *via* the imaging process ι. The boundary of S_V, obtained by the boundary operator ∂ is denoted by C_V and we call it the contour of Σ. The set of points on Σ that ι maps onto C_V we call the contour generator of C_V, and it is denoted by Γ_V. The map from Σ to Γ_V induced by ∂ is denoted by δ. [Marr 1977].

dimensional surface Σ; (2) its image or silhouette S_V as seen from a viewpoint V; (3) the bounding contour C_V of S_V; and (4) the set of points on the surface Σ that project onto the contour C_V. We shall call this last set the *contour generator* of C_V, and we shall denote it by Γ_V.

If one is presented with a contour in an image, without any knowledge of the surface or perspective that caused it, there is very little information on which one can base one's analysis. The only obvious feature available is the distinction between convex and concave pieces of contour -- that is, the presence of inflection points. In order that inflection points be "reliable," one needs to make some assumptions about the way the contour was generated, and I chose the following restrictions:

R1: The surface Σ is smooth.

R2: Each point on the contour generator Γ_V projects to a different point on the contour C_V.

R3: Nearby points on the contour C_V arise from nearby points on the contour generator Γ_V.

R4: The contour generator Γ_V of C_V is planar.

The first restriction is only a technical one. The second and third say that each point on the contour in the image comes from one point on the surface (which is an assumption that facilitates the analysis but is not of fundamental importance), and that where the surface looks continuous in the image, it really is continuous in three dimensions. The fourth condition, together with the constraint that the imaging process be an orthogonal projection, is simply a necessary and sufficient condition that the difference between convex and concave contour segments reflects properties of the surface, rather than characteristics of the imaging process.

It turns out that the following theorem is true, and it is

a result that I found very surprising.

> *Theorem.* If *R1* is true, and *R2* - *R4* hold for all distant viewing directions that lie in some plane, then the viewed surface is a generalized cone.

This means that if, for distant viewpoints whose viewing directions lie parallel to some plane, a surface's shape can successfully be inferred using only the convexities and concavities of its bounding contours in an image, then that surface is a generalized cone or is composed of several such cones. The interesting thing about this result is that it implicates generalized cones. We have already seen that the important thing about these cones is that an axis forms an integral part of their structure. But this is a feature of their three-dimensional organization, and ought in some sense to be independent of the issues raised by vision. What the theorem says is that there is a natural link between generalized cones and the imaging process itself. The combination of these two must mean, I think, that generalized cones will play an intimate role in the development of vision theory.

Interpreting the image of a single generalized cone

If we take this result at face value, we can now ask an obvious question. Let us assume that our data consist of contours of surface discontinuity in the image of a generalized cone, since without this assumption we can deduce nothing. How may such contours be interpreted? To specify a generalized cone, we have to specify its axis $\wedge$, cross-section $\rho(\theta)$, and axial scaling function $h(z)$ (figure 14); how can we discover these from an image?

The answer to this question is based on the notion of the *skeleton* of a generalized cone. The skeleton is not a difficult idea, since it is very like the set of lines a cartoonist draws to convey the shape of a curved object. It consists of three classes of contour: (a) the contours that occur in a generalized cone's

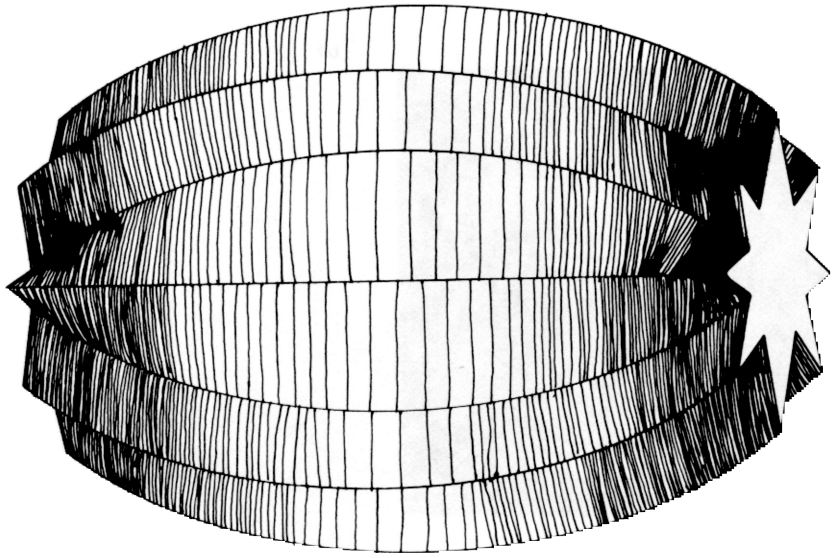


Figure 19. A sketch of a generalized cone showing its silhouette (the circumscribing contour), and its fluting (the contours spanning its length). The radial extremities of a generalized cone are illustrated in Figure 20.

silhouette; (b) the contours that arise from maxima and minima in a cone's axial scaling function (called the cone's *radial extremities)*; and (c) contours that arise from maxima and minima in the cone's cross-section (its *fluting*). These categories are illustrated in figure 19.

The reason why the skeleton is a useful construct for recognition is that one can detect its presence in an image by the many relationships that exist among its parts. For example, radial extremities are all parallel to each other, and the silhouette

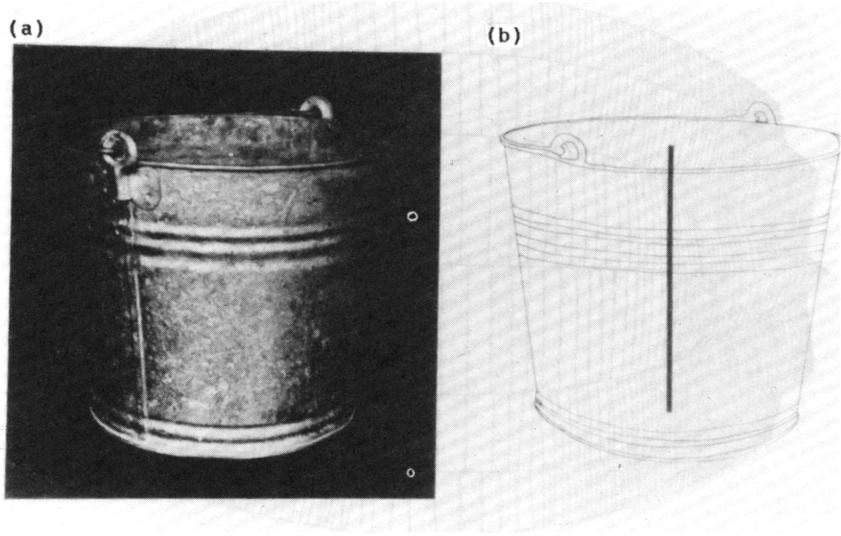


Figure 20. Methods based on the theory described here suffice to solve this image of a bucket. An axial symetry is established by its sides about the bucket's axis (shown thickened), and a parallel relationship holds between components of its radial extremity.

and fluting have a kind of symmetry about the image of the cone's axis. It turns out that one can use these relationships to set up constraints on a set of contours such that, if those constraints are all satisfied by a unique interpretation of the contours in the image, one can be reasonably certain that a skeleton has been found, and hence that the contours can be interpreted as arising from a generalized cone whose axis is then determined. The practical importance of this result is illustrated in figure 20, where one can see that the image of the "sides" is

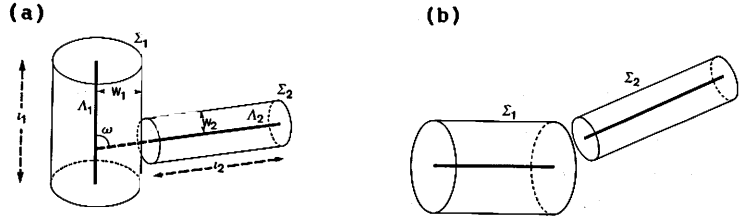


Figure 21. The two main types of joins between two generalized cones. *a* shows a side-to-end join, and *b* shows an end-to-end join.

symmetrical about the bucket's axis, and there is a clear parallel relationship between the image of the bucket's top, the corrugations in its side, and the visible part of its base (the bucket's radial extremities). These relations, of symmetries and parallelism, are preserved by an orthogonal projection. Hence provided that the contours are formed along a viewing direction that is not too close to the axis of the cone, these relations will still be present in the image. If the viewing direction lies so close to the cone's axis that its image is substantially foreshortened, these relationships will no longer be present, but it is part of the overall theory that such views have to be handled differently [Marr and Nishihara 1977].

Real-life objects are often approximately composed of several different cones, joined together in various ways (see figure 13), and we therefore have to study ways of decomposing a multiple cone into its components -- for example, a human body into arms, legs, torso and head. Marr [1977a] analyzed the two types of join shown in figure 21, giving criteria that define segmentation points on the contour produced by two joined

cones. Figure 22 exhibits the segmentation points P and Q for the case in which two short cones are joined side-to-end. P. Vatan has written a computer program that can carry out this segmentation, and an example of its operation is illustrated in figure 23. The initial outline in (a) was obtained by applying local grouping processes to the *primal sketch* of the image of a toy donkey [Marr 1976]. This outline was then smoothed and divided into convex and concave sections to get (b). Next, strong segmentation points, like the deep concavity circled in (c), are identified and a set of heuristic rules are used to connect them with other points on the contour to get the segmentation shown in (d). The component axes shown in (e) are then derived from these. The resulting segments are checked to see that they obey the rules for images of generalized cones. The boundaries must for example be symmetric about the axes, and in the case of side-to-end joins, the axis of the cone that is attached by its end must intersect the segmentation points that separate the two cones' contours. In this example, most of the symmetry relations have degenerated into parallelism. The thin lines in (f) indicate the position of the head, leg, and tail components along the torso axis, and the snout and ear components along the head axis. (This algorithm is due to P. Vatan).

Some comments on the limitations of this theory

The results of this theory are limited in their scope to a particular class of views and surfaces, but on the other hand, they use only a limited kind of visual information, little more than occluding contours that are formed in an image by rays that graze a smooth surface. Interestingly, these particular contours are unsuitable for use in stereopsis or structure-from-motion computations, because they are not formed from markings that define precise locations on the viewed surface. Creases and folds on a surface also give rise to contours in an image, and these have yet to be studied in detail. Information about shape from shading, texture, stereo or motion information has not yet been

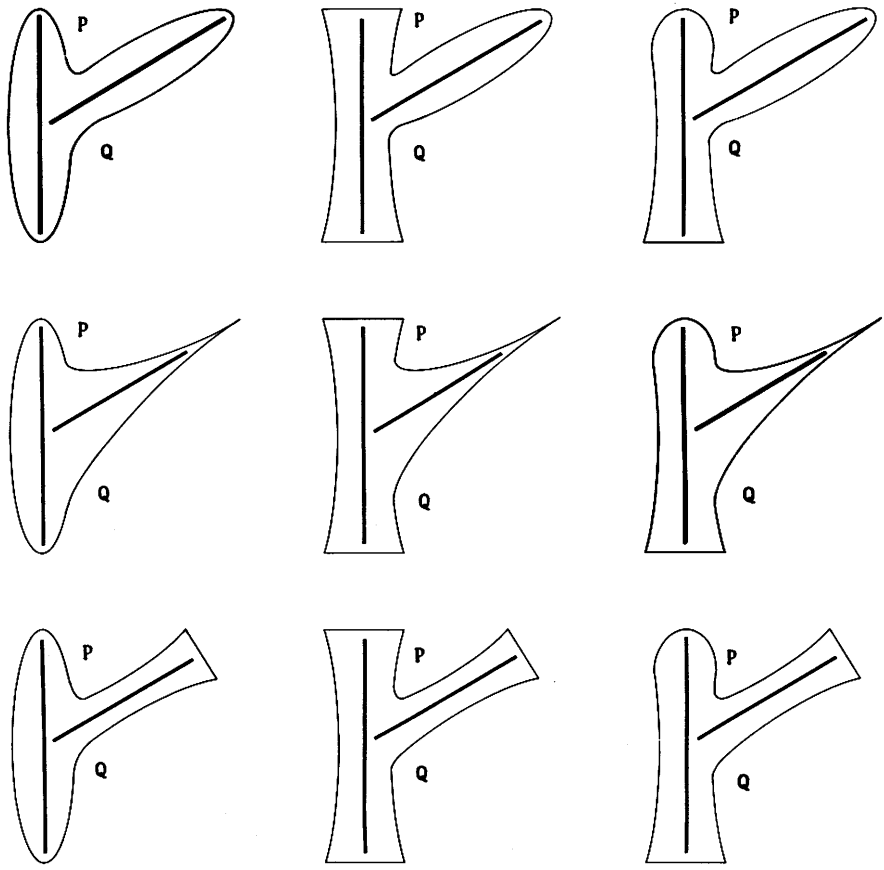


Figure 22. This figure illustrates the types of side-to-end join that can occur between two short generalized cones. In the first column, the left-hand cone is convex; in the center it is concave, and in the right it is convex on one side of the join and concave on the other. The other cone is convex in the top row, and concave in the other two. Segmentation depends upon finding the points *P* and *Q*, which are defined by theorem 7 of Marr [1977] and illustrated here for each case.

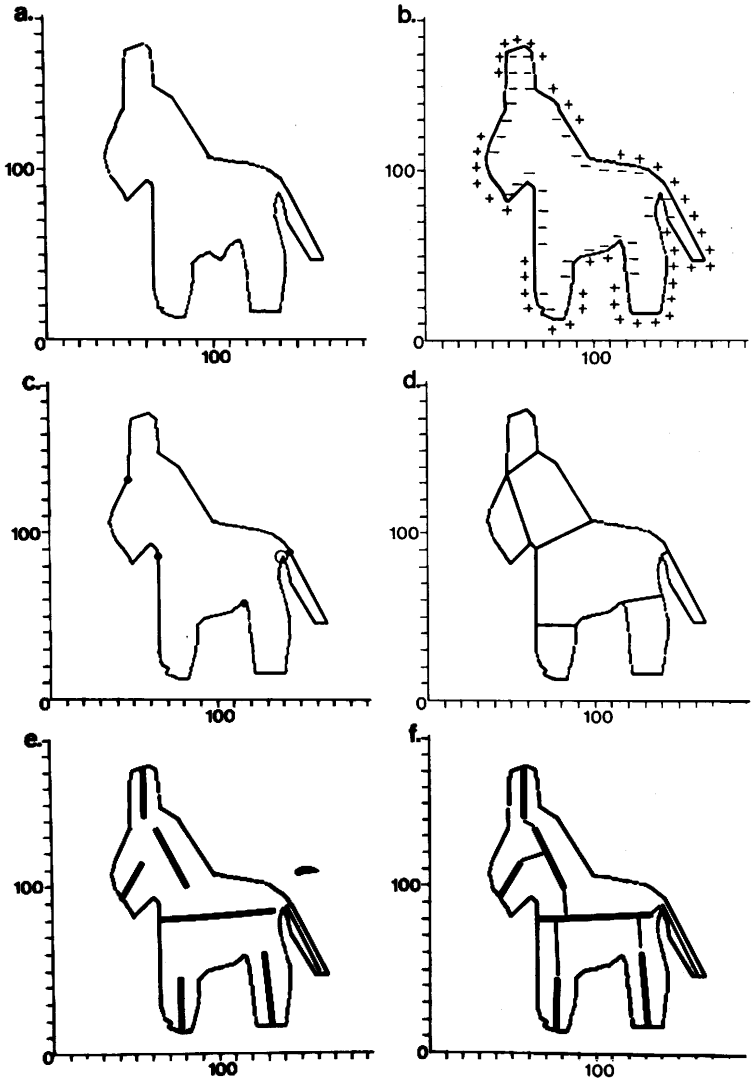
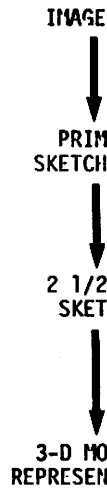


Figure 23. The occluding contours in an image can be used to locate the images of the natural axes of a shape composed of generalized cones [Marr 1977].

TABLE 2

A framework for the derivation of shape information from images.



Describes the intensity changes present in an image, labels distinguished locations like termination points, and makes explicit local two-dimensional geometrical relations.

Represents contours of surface discontinuity, and depth and orientation of visible surface elements, in a coordinate frame that is centered on the viewer.

Shape descriptions that include volumetric shape primitives of a variety of sizes, whose positions are defined using an object-centered coordinate system. This representation imposes considerable modular organization on its descriptions.

considered. By adding these other sources of information, I hope that a set of methods can eventually be assembled that together approach a comprehensive treatment of possible image configurations.

Conclusion

I have tried to make three main points. The first is methodological, namely that it is important to be very clear about the nature of the understanding we seek [Marr and Poggio 1977a] [Marr 1977b]. The results we try to achieve should be precise, at the level of what I called a computational theory, and should deal with problems that can confidently be attributed to a real aspect of vision, and not (for example) to an artifact of the limitations of one's current vision program.

The second main point is that the critical issues for vision seem to me to revolve around the nature of the representations used - that is, the particular characteristics of the world that are made explicit - and the nature of the processes that recover these characteristics, create and maintain the representations, and eventually read them. By analyzing the spatial aspects of the problem of vision [Marr and Nishihara 1977], an overall framework for visual information processing is suggested, that consists of three principal representations: (1) the primal sketch, which makes explicit the intensity changes and local two-dimensional geometry of an image; (2) the $2^1/_2$-D sketch, which is a viewer-centered representation of the depth and orientation of the visible surfaces and includes contours of discontinuities in these quantities; and (3) the 3-D model representation, whose important features are (a) that its coordinate system is object-centered, (b) that it includes volumetric primitives, that make explicit the space occupied by an object and not just its visible surfaces, and (c) that primitives of various sizes are included, arranged in a modular, hierarchical organization.

The third main point concerns the study of processes for

recovering the various aspects of the physical characteristics of a scene from images of it. The critical act in formulating computational theories for such processes is the discovery of valid constraints on the way the world behaves that provide sufficient additional information to allow recovery of the desired characteristic. Several examples are already available, including Land and McCann [1971], which rests on the distinction between sharp and shallow intensity changes; stereopsis [Marr 1974] [Marr and Poggio 1976] [Marr and Poggio 1977b] which uses continuity and uniqueness; structure from visual motion [Ullman 1977], which uses rigidity; fluorescence [Ullman 1976a]; and shape from contour [Marr 1977a]. The discovery of constraints that are valid and sufficiently universal leads to results about vision that have the same quality of permanence as results in other branches of science [Marr 1977b].

Finally, once a computational theory for a process has been formulated, algorithms for implementing it may be designed, and their performance compared with that of the human visual processor. This allows two kinds of result. Firstly, if performance is essentially identical, one has good evidence that the constraints of the underlying computational theory are valid and may be implicit in the human processor; and secondly, if a process matches human performance, it is probably sufficiently powerful to form part of a general purpose vision machine.

References

G. J. Agin, *Representation and Description of Curved Objects*, Stanford AI Memo 173, 1972.

J. Beck, *Surface Color Perception*, Cornell University Press, 1974.

T. O. Binford, *Visual Perception by Computer*, presented to the IEEE Conference on Systems and Control, 1971.

H. Blum, "Biological Shape and Visual Science, (part 1)," *J.*

Theor. Biol. , 38, 1973.

F. W. C. Campbell, "Sometimes a Biologist Has to Make a Noise Like a Mathematician," *NRP Bulletin on Neurophysiology and Psychophysics* (in press), 1977.

P. Colas-Baudelaire, *Digital Picture Processing and Psychophysics: a Study of Brightness Perception* Report No. UTEC-CSC-74-025, Department of Computer Science, University of Utah, 1973.

T. B. Felton, W. Richards, and R. A. Smith, Jr., "Disparity Processing of Spatial Frequencies in Man," *J. of Physiology,* 255, 1972.

K. Forbus, *Light Source Effects,* MIT AI Laboratory Memo 422, 1977.

E. C. Freuder, *A Computer Vision System for Visual Recognition Using Active Knowledge,* MIT AI Laboratory Technical Report 345, 1976.

L. Glass, "Moire Effect from Random Dots," *Nature, 243,* 1969.

A. L. Gilchrist, "Perceived Lightness Depends on Perceived Spatial Arrangement," *Science, 195,* 1977.

B. K. P. Horn, "Determining Lightness from an Image," *Computer Graphics and Image Processing, 3,* 1974.

B. K. P. Horn, "Understanding Image Intensities," *Artificial Intelligence,* 1977.

D. H. Hubel and T. N. Wiesel, "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex," *J. Physiol., Lond. 160,* 1962.

D. A. Huffman, "Impossible Objects as Nonsense Sentences," in *Machine Intelligence 6*, R. Meltzer and D. Michie (eds.), The Edinburgh University Press, 1971.

W. H. Ittelson, *Visual Space Perception*, pp. 145-147, Springer, 1960.

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

B. Julesz, "Experiments in the Visual Perception of Texture," *Scientific American 232*, 1975.

B. Julesz and J. E. Miller, "Independent Spatial-Frequency-Tuned Channels in Binocular Fusion and Rivalry," *Perception 4*, 1976.

P. A. Kolers, *Aspects of Motion Perception*, Pergamon Press, 1972.

E. H. Land, and J. J. McCann, "Lightness and Retinex Theory," *J. Opt. Soc. Am. 61*, 1971.

A. K. Mackworth, "Interpreting Pictures of Polyhedral Scenes." *Artificial Intelligence 4*, 1973.

D. Marr, *A Note on the Computation of Binocular Disparity in a Symbolic, Low-Level Visual Processor*, MIT AI Laboratory Memo 327, 1974.

D. Marr, "Early Processing of Visual Information," *Phil. Trans. Roy. Soc. B. 275*, 1976.

D. Marr, "Analysis of Occluding Contour," *Proc. Roy. Soc. B 197*, 1977a.

D. Marr, "Artificial Intelligence - a Personal View," *Artificial*

Intelligence 9, 1977b.

D. Marr and H. K. Nishihara, "Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes," *Proc. Roy. Soc. B. 200*, 1977.

D. Marr and T. Poggio, "Cooperative Computation of Stereo Disparity," *Science 194*, 1976.

D. Marr and T. Poggio, "From Understanding Computation to Understanding Neural Circuitry," *Neurosciences Res. Prog. Bull. 15*, 1977a.

D. Marr and T. Poggio, "A Theory of Human Stereo Vision," MIT AI Laboratory Memo 451, 1977b.

D. Marr, T. Poggio, and G. Palm, "Analysis of a Cooperative Stereo Algorithm," *Biol. Cybernetics 28*, 1977.

J. E. W. Mayhew and J. P. Frisby, "Rivalrous Texture Stereograms," *Nature 264*, 1976.

R. Nevatia, *Structured Descriptions of Complex Curved Objects for Recognition and Visual Memory*, Stanford AI Memo 250, 1974.

J. F. O'Callaghan, "Computing the Perceptual Boundaries of Dot Patterns," *Computer Graphics and Image Processing 3*, 1974.

W. A. Richards, "Anomalous Stereoscopic Depth Perception," *J. Opt. Soc. Amer. 61*, 1971.

W. A. Richards, "Stereopsis With and Without Monocular Cues," *Vision Res.* 17, 1977.

M. Riley, *Discriminant Functions in Early Visual Processing*, (in preparation) 1977.

A. Rosenfeld, R. A. Hummel, and S. W. Zucker, "Scene Labelling by Relaxation Operations," *IEEE Transactions on Systems, Man and Cybernetics, SMC-6,* 420-433.

B. R. Schatz, "Computation of Texture Discrimination," *Proc. 5th Int. Joint Conf. Art. Intelligence,* 1977, also MIT AI Laboratory Memo 426, 1977.

K. A. Stevens, "Computation of Locally Parallel Structure," *Biol. Cybernetics 29* (also available as MIT AI Laboratory Memo 392), 1978.

J. M. Tenenbaum and H. G. Barrow, *Experiments in Interpretation-Guided Segmentation,* Stanford Research Institute Technical Note 123, 1976.

S. Ullman, "On Visual Detection of Light Sources," *Biol. Cybernetics 21,* 1976a.

S. Ullman, "Filling-in the Gaps: The Shape of Subjective Contours and a Model for their Generation," *Biol. Cybernetics 25,* 1976b.

S. Ullman, *The Interpretation of Visual Motion,* MIT PhD Thesis, 1977, to be published by MIT Press, 1978.

H. Wallach and D. N. O'Connell, "The Kinetic Depth Effect," *J. Exp. Psychol. 45,* 1953.

D. Waltz, "Understanding Line Drawings of Scenes with Shadows," in *The Psychology of Computer Vision,* P. H. Winston (ed.), McGraw-Hill, 1975.

The complete version of this paper appears as "Representing Visual Systems" *Lectures on Mathematics in the Life Sciences,*