MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

Working Paper #131 August 1976

FROM COMPUTATIONAL THEORY TO PSYCHOLOGY AND NEUROPHYSIOLOGY -- a case study from vision

by D. Marr

SUMMARY: The CNS needs to be understood at four nearly independent levels of description: (1) that at which the nature of a computation is expressed; (2) that at which the algorithms that implement a computation are characterised; (3) that at which an algorithm is committed to particular mechanisms; and (4) that at which the mechanisms are realised in hardware. In general, the nature of a computation is determined by the problem to be solved, the mechanisms that are used depend upon the available hardware, and the particular algorithms chosen depend on the problem and on the available mechanisms. Examples are given of theories at each level from current research in vision, and a brief review of the immediate prospects for the field is given.

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

Introduction

Modern neurophysiology has learned much about the operation of the individual neuron, but deceivingly little about the meaning of the circuits they compose. The reason for this can be attributed, at least in part, to a failure to recognise what it means to understand a complex information-processing system.

Complex systems cannot be understood as a simple extrapolation of the properties of their elementary components. One does not formulate a description of thermodynamical effects using a large set of wave equations, one for each of the particles involved. One describes such effects at their own level, and tries to show that in principle, the microscopic and macroscopic descriptions are consistent with one another.

The core of the problem is that a system as complex as a nervous system or a developing embryo must be analyzed and understood at several different levels. For a system that solves an information processing problem, we may distinguish four important levels of description. At the lowest, there is basic component and circuit analysis -- how do transistors, neurons, diodes and synapses work? The second level is the study of particular mechanisms; adders, multipliers, and memories accessed by address or by content. The third level is that of the algorithm, and the top level contains the theory of the overall computation. For example, take the case of Fourier analysis. The computational theory of the Fourier transform is well understood, and is expressed independently of the particular way in which it is computed. One level down, there are several algorithms for implementing a Fourier transform -- the Fast Fourier transform (Cooley & Tukey 1965) which is a serial algorithm; and the parallel "spatial" algorithm that is based on the mechanisms of laser optics. All these algorithms carry out the same computation, and the choice of which one to use depends upon the particular mechanisms that are available. If one has fast digital memory, adders and multipliers, one will use the FFT, and if one has a laser and photographic plates, one will use an "optical" algorithm. In general, mechanisms are strongly determined by hardware, the nature of the computation is determined by the problem, and the algorithms are determined by the computation and the available mechanisms.

Each of these four levels of description has its place in the eventual understanding of perceptual information processing, and it is important to keep them separate. Of course, there are are logical and causal relationships among them, but the important point is that these levels of description are only loosely related. Too often in attempts to relate psychophysical problems to physiology there is confusion about the level at which a problem arises - is it related mainly to biophysics (like after-images) or primarily to information processing (like the ambiguity of the Necker cube)? More disturbingly, although the top level is the most neglected, it is also the most important. This is because the structure of the computations that underly perception depend *more* upon the computational *problems* that have to be solved than on the particular hardware in which their solutions are implemented. There is an analog of this in physics, where a thermodynamical approach represented, at least historically, the first stage in the study of matter. A description in terms of mechanisms or elementary components usually appears afterwards.

The main point then is that the topmost of our four levels, that at which the necessary structure of a computation is defined, is a crucial but neglected one. Its study is separate from the study of particular algorithms, mechanisms or hardware, and the techniques needed to pursue it are new. In the rest of the article, we summarize some examples of vision theories at the different levels we described, and illustrate the types of prediction that can emerge from each.

General remarks about conventional approaches

The problems of visual perception have attracted the curiosity of scientists for many centuries. Important early contributions were made by Newton, who laid the foundations for modern work on color vision, and Helmholtz, whose treatise on physiological optics maintains its interest even today. Early in this century, Wertheimer (1912) noticed that during apparent motion, the correspondence between wholes or "fields" in successive frames seemed to amount to more than correspondence between their constituents. This observation started the Gestalt school of psychology, which was concerned with describing qualities of wholes (thing-quality, solidarity, distinctness), and trying to formulate the laws that governed their creation. The attempt failed for various reasons, and the Gestalt school dissolved into the fog of subjectivism. With the death of the school, many of its early and genuine insights were unfortunately lost to the mainstream of experimental psychology.

The next developments of importance were recent and technical. The advent of electrophysiology in the 1940's and '50's made single cell recording possible, and with Kuffler's (1953) study of retinal ganglion cells a new approach to the problem was born. Its most renowned practitioners are D. H. Hubel and T. N. Wiesel, who since 1959 have conducted an influential series of investigations on single cell responses at various points along the visual pathway in the cat and the monkey.

Hubel & Wiesel (1962) used the notion of a cell's "receptive field" to classify the cell types of primary and secondary visual cortex into simple, complex, and hypercomplex cells. Simple cells are orientation-sensitive and roughly linear: That is, their receptive fields are divided into parallel elongated excitatory and inhibitory parts which summate, and a simple cell's response to a stimulating pattern is roughly predictable from its receptive field geometry. Complex cells are not linear, but apparently respond to edges and bars over a wider range than a simple cell. Hypercomplex cells seem to respond best at points where an edge or bar terminates. How the different types of cell are connected and why they behave as they do is controversial.

Students of the psychology of perception were also affected by a technological advance, the advent of the digital computer. Most notably, it allowed B. Julesz (1971) to construct random dot stereograms, which are image pairs constructed of dot patterns that appear random when viewed monocularly, but which fuse when viewed one through each eye to give a percept of shapes and surfaces with a clear three-dimensional structure. Such percepts are caused solely by the stereo disparity between matching elements in the images presented to each eye.

Very recently, considerable interest has been attracted by a rather

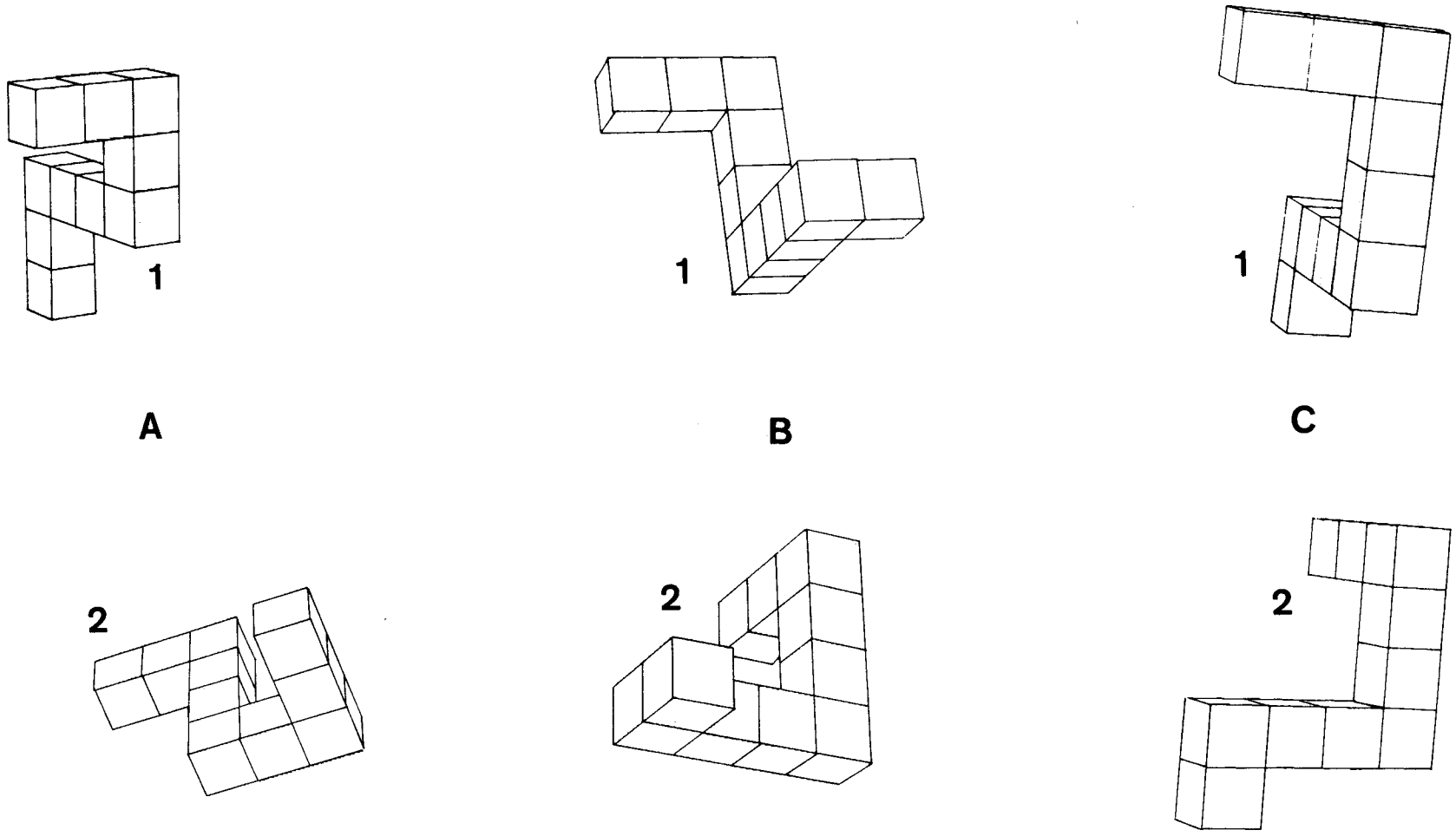


Figure 1. Examples of pairs of perspective line drawings presented to the subjects of Shepard & Metzler's (1971) experiments on mental rotation. (A) A "same" pair, which differs by an 80 degree rotation in the picture plane; (B) a "same" pair which differs by an 80 degree rotation in depth; (C) a "different" pair, which cannot be brought into congruence by any rotation. The time taken to decide whether a pair is the "same" varies linearly with the (3-D) angle by which one must be rotated to be brought into correspondence with the other. (reconstructed from figure 1 of Shepard & Metzler, 1971).

different approach. In 1971, R. N. Shepard and J. Metzler (1971) made line drawings of simple objects, which differed from one another either by a 3-D rotation relative to the viewer, or by a rotation plus a reflection (see figure 1). They asked how long it took to decide whether two depicted objects differed by a rotation and reflection, or merely a rotation. They found that the time taken depended on the 3-D angle of rotation necessary to bring the two objects into correspondence, not the 2-D angle between their images; and that it varied linearly with this angle. Similar findings have been reported in many subsequent investigations, and have led to the resurgence of ideas about mental imagery, and to analogies between visual recognition and computer graphics systems (Shepard 1975).

Interesting and important though these findings are, one must sometimes be allowed the luxury of pausing to reflect upon the overall trends that they represent, in order to take stock of the kind of knowledge that is accessible to these techniques. This proposal is itself an attempt at examining the link between various current approaches, including those of neurophysiology and psychophysics. We would also like to know what are the limitations of these approaches, and how can one compensate for their deficiencies?

Perhaps the most striking feature of these disciplines at present is their phenomenological character. They describe the behavior of cells or of subjects, but do not explain it. What is area 17 actually doing? What are the problems in doing it that need explaining, and at what level of description should such explanations be sought?

In trying to come to grips with these problems, D. Marr and his students at the M. I. T. Artificial Intelligence Laboratory have adopted a point of view that regards visual perception as a problem primarily in information processing. The problem commences with a large, gray-level intensity array, and it culminates in a *description* that depends on that array, and on the purpose that the viewer brings to it. Viewed in this light, a theory of visual information processing will exhibit the four levels of description that, as we saw in the introduction, are attached to any device that solves an information processing problem; and the first task of a theory of vision is to examine the top level. What exactly is the underlying *nature* of the computations being performed during visual perception?

A computational approach to vision

The empirical findings of the last 20 years, together with related anatomical (Allman 1972, 1973, 1974a, b & c, Zeki 1971) and clinical (e.g. Luria 1970, Critchley 1953, Vinken & Bruyn 1969) experience, have strengthened a view for which widespread indirect evidence previously existed, namely that the cerebral cortex is divided into many different areas that are distinguished structurally, functionally and by their anatomical connections. This suggests that, to a first approximation visual information processing can be thought of as having a *modular* structure, a view which is strongly supported by evolutionary arguments. If this is true, the task of a top-level theory of vision is clear; what are the modules, what does each do, and how?

The approach of the M. I. T. Artificial Intelligence Laboratory to the vision problem rests on these assumptions. We believe that the principal problems at present are (a) to formulate the likely modularization, and (b) to understand the computational problems each module presents. Unlike simpler systems like the fly

(Reichardt & Poggio 1976, Poggio & Reichardt 1976), the first step is the most difficult, just as formulating a problem in physics is often more difficult than solving it for a skilled mathematician. Nevertheless a variety of clues is available, from psychophysics and neurophysiology to the wide and interesting range of deficits reported in the literature of clinical neurology. Those cases in which a patient lacks a particular, highly circumscribed faculty are particularly interesting (e.g. Warrington & Taylor 1973, Efron 1968); but more general impairments can also be informative, particularly the agnosias in which higher level analysis and interpretation are damaged while leaving other functions, like texture discrimination and figure-ground separation, relatively unimpaired. Such evidence must be treated with due caution, but it encourages us to examine ways of squeezing the last ounce of information from an image before taking recourse to the descending influence of high-level interpretation on early processing. Computational evidence can also be useful in suggesting that a certain module may exist. For example, Ullman (1976a) showed that fluorescence may often be detected in an image using only local cues, and the method is so simple that one would expect something like it to be incorporated as a computational module in the visual system, even though we are not aware of any supporting evidence, either clinical, physiological, or psychological. The same may be true of other visual qualities like glister and wetness, just as it is generally believed to be true for color, motion and stereopsis.

In order to introduce the reader to our approach, the next few sections present brief summaries of a particular modularization, and the associated theories, that we have been studying over the last two years and which is illustrated by figure 2. We are perfectly aware that the particular decomposition chosen here may not be exactly correct, and even if it is, the separation of modules is certainly not complete. All of the modules described here have been implemented in computer programs which demonstrate that this particular scheme works for a number of natural images. Alternative decompositions that have been tried, in particular those that rely on much more interaction between low-level processing and high-level interpretation of an image, (e.g. Freuder (1975), Shirai (1973)) have not hitherto led to such satisfactory and promising results.

1: The Primal Sketch (Marr 1976a)

It is a commonplace that a scene and a drawing of the scene appear very similar, despite the completely different grey-level images to which they give rise. This suggests that the artist's local symbols correspond in some way to natural symbols that are computed out of the image during the normal course of its interpretation. The first part of our visual information theory therefore asserts that the first operation on an image is to transform it into its raw *primal sketch*, which is a primitive but rich *description* of the intensity changes that are present. Figure 3 shows an example. In order to obtain this description, approximations to the first and second directional derivatives of intensity are measured at several orientations and on several scales everywhere in the image, and these measurements are combined to form local descriptive assertions. The process of computing the primal sketch involves five important steps, the first of which can be compared with the measurements that are apparently made by simple cells in the visual cortex. One prediction made by this part of the theory is that a given intensity change itself determines which

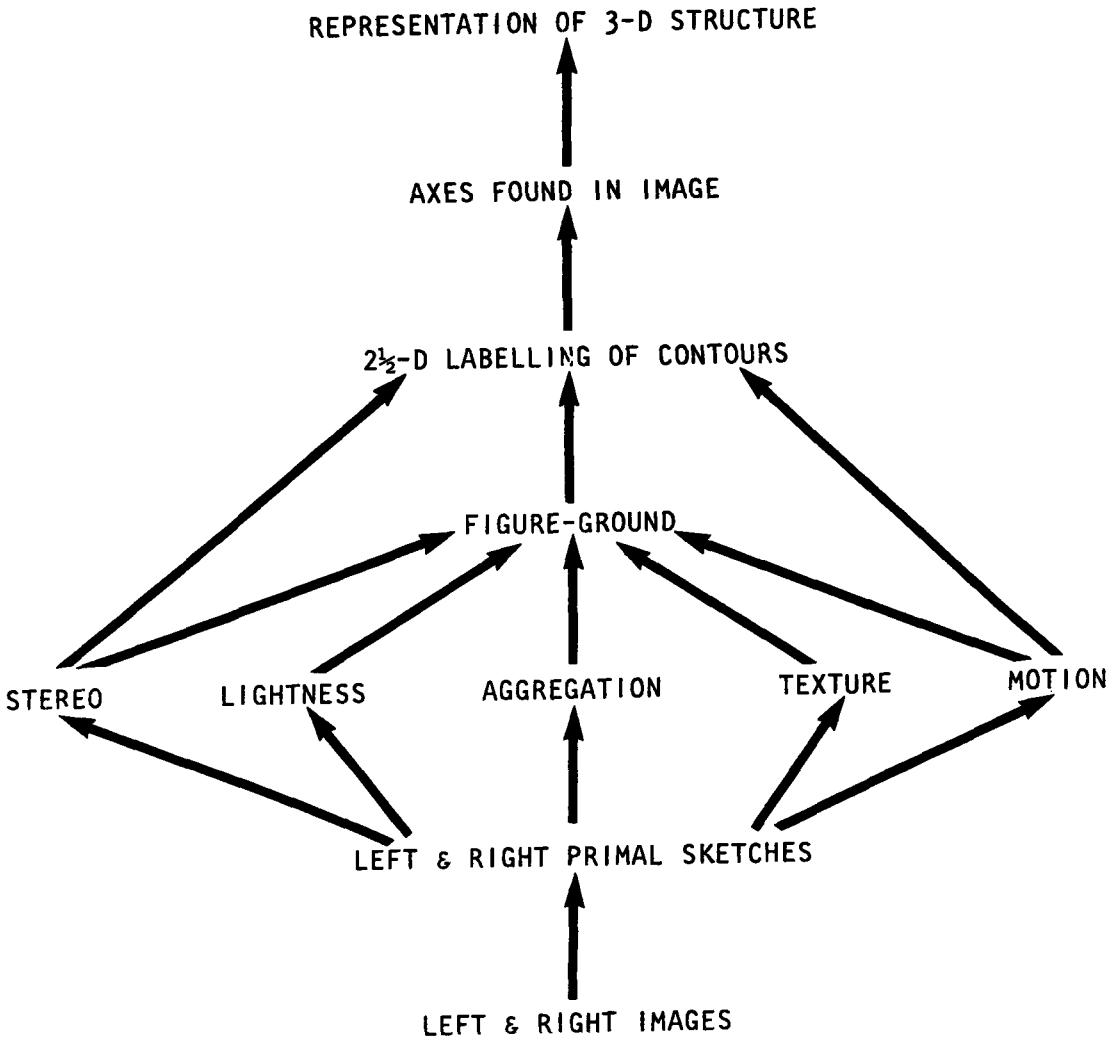
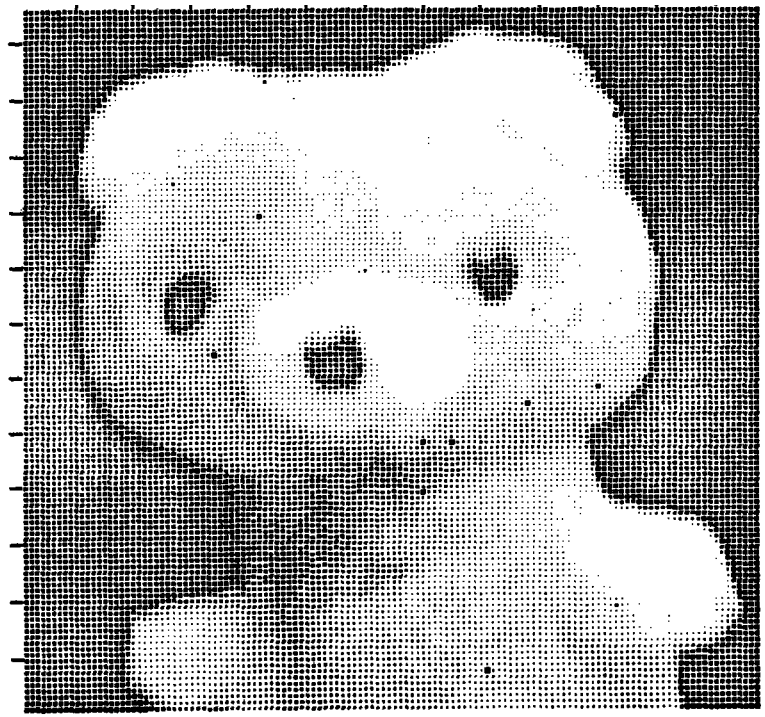
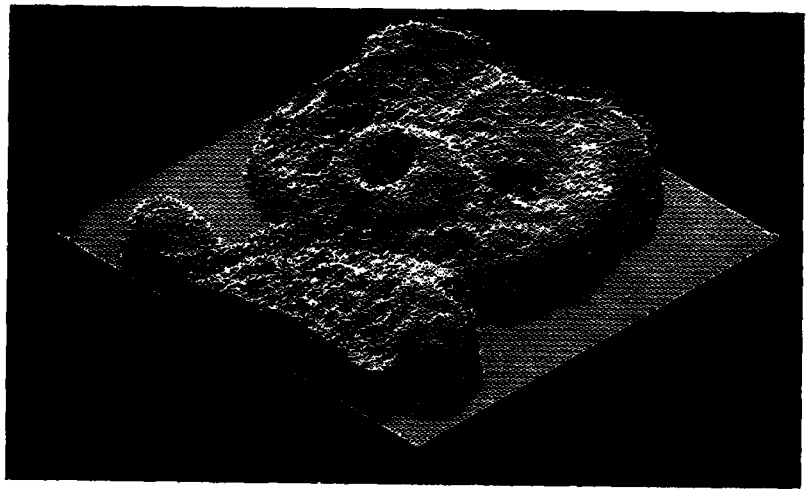
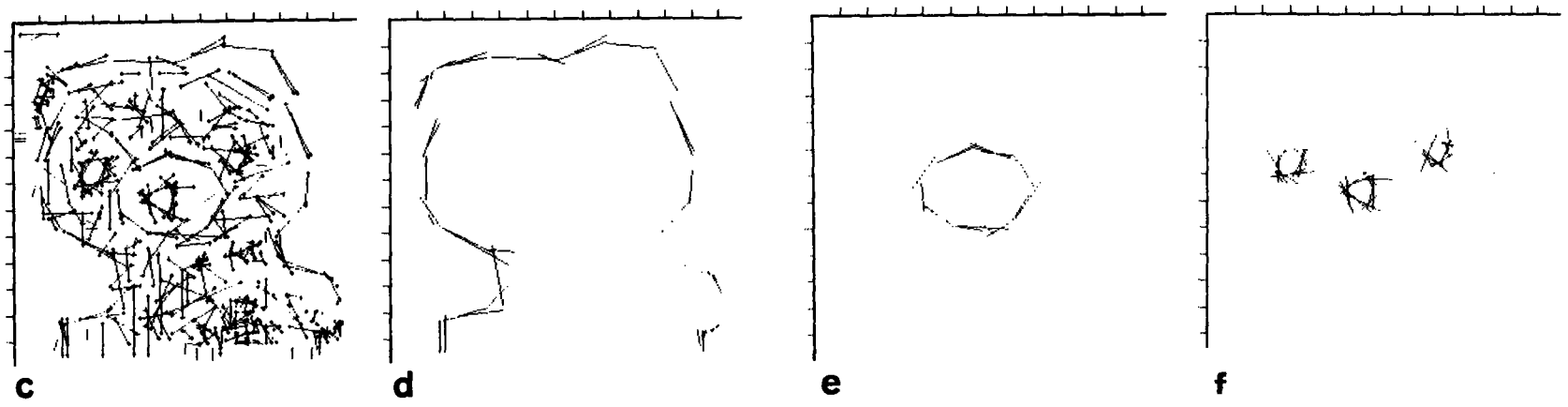


Figure 2. This diagram summarises our overall view of the visual recognition problem, and it embodies several points that our approach takes as assumptions. The first is that the recognition process decomposes to a set of modules that are to a first approximation independent. The simplified subdivision shown here consists of four main stages, each of which may contain several modules. (1) The translation of the image into a primitive description called the *primal sketch* (Marr 1976b); (2) The division of the primal sketch into regions or forms, through the action of various grouping processes ranging in scope from the very local to global predicates like a rough type of connectedness; (3) The assignment of an axis-based description to each form (see figure 4); and (4) The construction of a 3-D model for the viewed shape, based initially on the axes delivered by (3). The relation between the 3-D model representation of a shape and the image of that shape is found and maintained with the help of the image-space processor. Finally, the representation of the geometry of a shape is separate from the representation of the shape's use or purpose (Warrington & Taylor 1973).





a



c d e f

Figure 3. 3a shows the image of a toy bear, printed in a font with 16 grey levels. In 3b, the intensity at each point is represented along the z-axis. 3c illustrates the spatial component of the raw primal sketch as obtained from this image. Associated with each line segment are measures of contrast, type and extent of the intensity change, position and orientation. This image is so simple that purely local grouping processes suffice to extract the major forms from the primal sketch. These forms are exhibited in 3d, e & f.

simple-cell measurements are used to describe it. This is in direct contrast to theories which assert that each simple cell acts as a "feature-detector", whose output is freely available to subsequent processes. If this is true, it requires that a well-defined interaction take place between simple-cell like measurements made at the same orientation and position in the visual field but with different receptive field sizes (see Marr 1976a).

2: Stereopsis (Marr 1974, Marr & Poggio 1976)

Suppose that images of a scene are available, taken from two nearby points at the same horizontal level. In order to compute stereoscopic disparity, the following steps must be carried out: (1) a particular location on a surface in the scene must be chosen from one image; (2) that location must be identified in the other image; (3) the relative positions of the two images of that location must be measured. Notice that methods based on grey-level correlation between images fail to satisfy these conditions because a grey-level measurement does not define a point on a physical surface independently of the optics of the imaging device. The matching must be based on objective markings that lie on a physical surface, and so one has to use predicates that correspond to changes in reflectance. One way of doing this is to obtain a primitive description of the intensity changes that exist in each image, and then to match these descriptions. Line and edge segments, blobs, and edge termination points correspond quite closely to boundaries and reflectance changes on physical surfaces.

The stereo problem may thus be reduced to that of matching two primitive descriptions, one from each eye. One can think of elements of these descriptions as having only position information, like the black points in a random-dot stereogram, although in practise there exist some rules about which matches between descriptive elements are possible, and which are not. There are physical constraints that translate into two rules for how the left and right descriptions are combined:

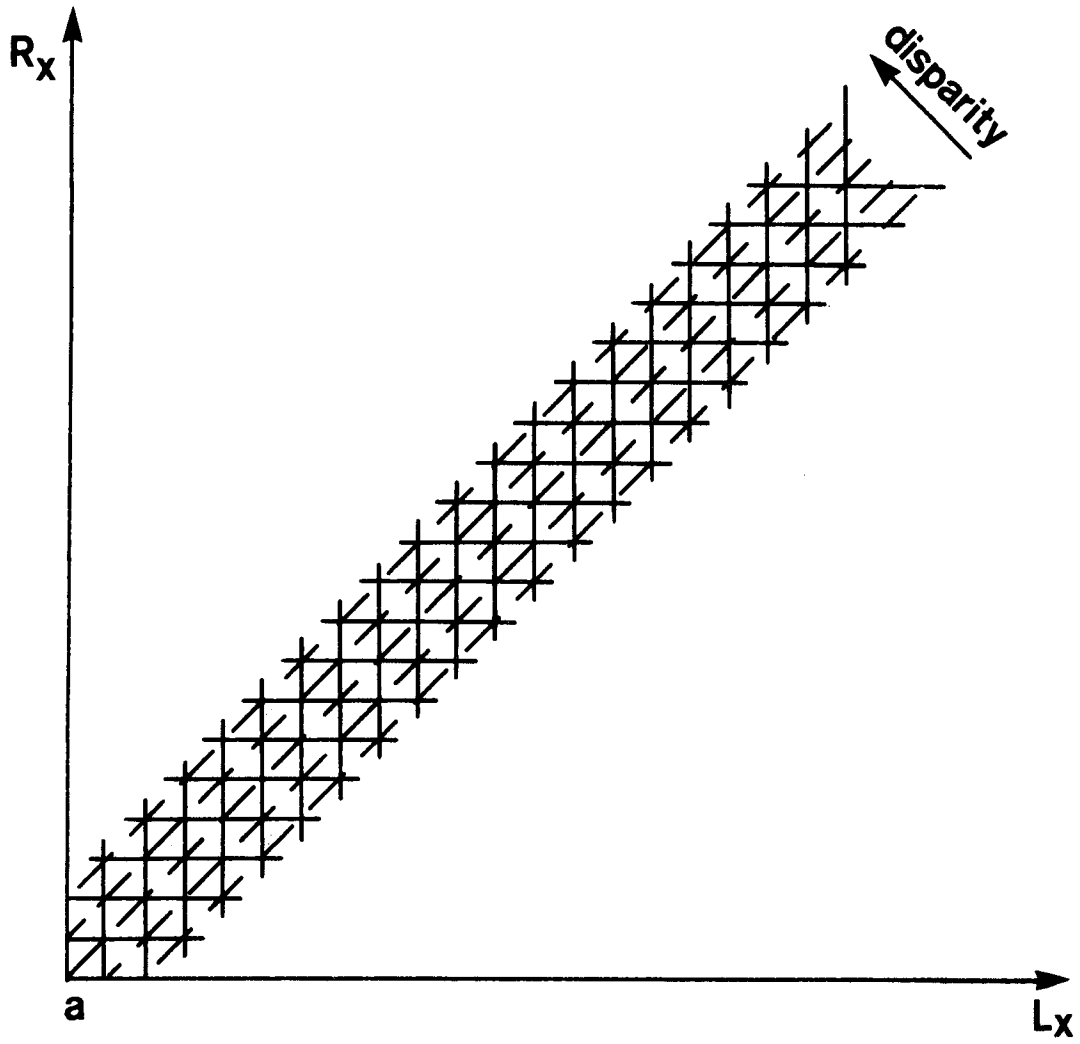
(1) *Uniqueness.* Each item from each image may be assigned at most one disparity value. This condition rests on the premise that the items to be matched have a physical existence, and can be in only one place at a time.

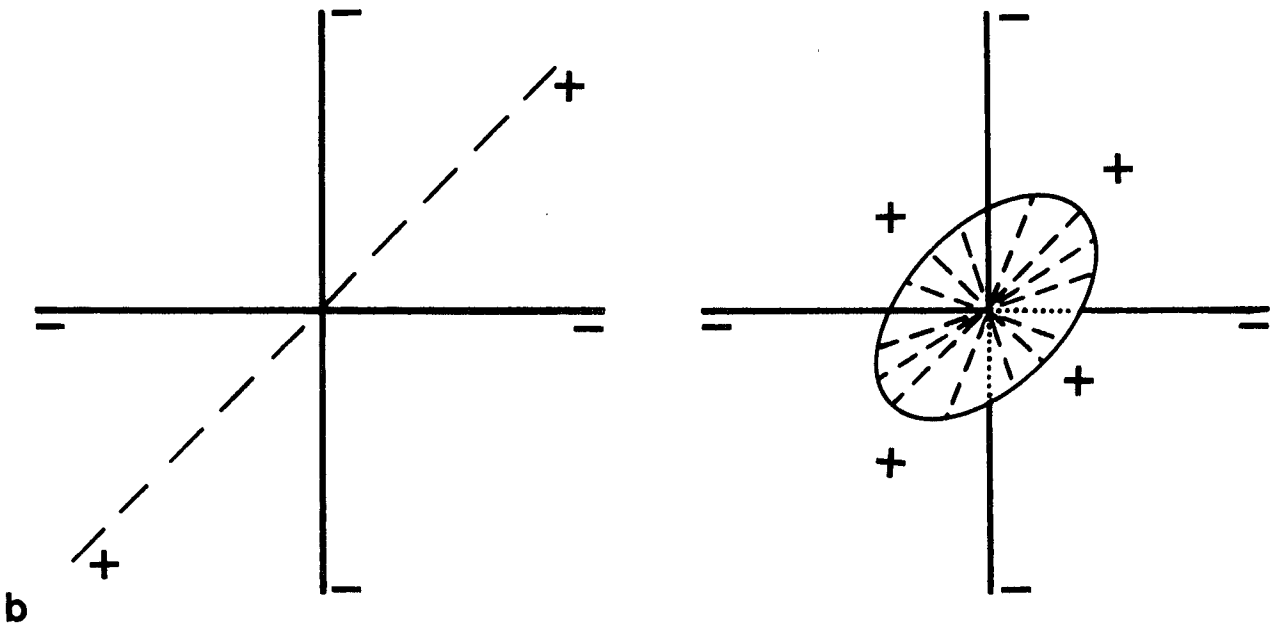
(2) *Continuity.* Disparity varies smoothly almost everywhere. This condition is a consequence of the cohesiveness of matter, and it states that only a relatively small fraction of the area of an image is composed of boundaries.

These conditions on the computation are represented geometrically in figure 4a. Later in the article, we exhibit a network that implements these conditions, and we illustrate how it solves random-dot stereograms.

In this case the computational problem is rather well-defined, essentially because of Julesz's (1971) demonstration that random-dot stereograms, containing no monocular information, yield stereopsis. It is not yet completely clear however what mechanisms are actually available for implementing this computation (for instance do eye movements play a critical role?). As a consequence, it is an open question whether the cooperative algorithm introduced later is used, or whether simpler "serial" scanning algorithms may actually be implementing the stereopsis computation (see Marr & Poggio 1976).

Figure 4. The geometry of constraints on the computation of binocular disparity. 4a illustrates the constraints for the case of a one-dimensional image. Lx and Ly 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 states that only one disparity value may be assigned to each descriptive element. That is, only one disparity value may be "on" along each horizontal or vertical line. The continuity condition states that we seek solutions in which disparity values vary smoothly almost everywhere. That is, solutions tend to spread along the dotted diagonals, which are lines of constant disparity, and between adjacent diagonals. 4b shows how this geometry appears at each intersection point. The constraints may be implemented by a network with positive and negative interactions that obey this geometry, because the stable states of such a network are precisely the states that satisfy the constraints on the computation. 4c shows the constraint geometry for a 2-dimensional image. The negative interactions remain essentially unchanged, but the positive ones now extend over a small 2-dimensional neighbourhood. A network with this geometry was used to perform the computation exhibited in figure 8.


R_x
disparity
a
L_x


b

3: Structure from apparent motion

It has long been known that as an object moves relative to the viewer, the way its appearance changes provides information about its shape, and we are able to use that information to infer shape. This problem decomposes into two parts: (a) matching the elements that occur in consecutive frames; and (b) deriving shape information from measurements of the changes in position between successive frames. It presently looks as though (a) will yield to a combination of the local preference measures that Ullman (1976b) has measured psychophysically, and a method similar to that of Marr & Poggio (described above) for the stereo matching problem; and Ullman (1976b) has solved problem (b). The idea is this. In general, nothing can be inferred about the shape of an object given only a set of views of it. Some extra assumptions have to be made. Ullman made the assumption that the viewed objects are *rigid*, and derived a method for computing shape from successive views based on this. The method gives results that are quantitatively superior to the ability of humans to determine shape from motion, and which fail in qualitatively similar circumstances. He has also devised a set of simple parallel algorithms by which the method may be implemented.

4: Grouping and texture vision (Marr 1976a)

The primal sketch of an image is in general a large and unwieldy collection of data. This is an unavoidable consequence of the irregularity and complexity of natural images. The next important computational problem is how to decode the primal sketch. For most images, it appears unnecessary to invoke specific hypotheses about what is there until considerably later in the processing. The theory next applies a number of quite general selection and grouping processes to elements in the primal sketch. The purpose of these processes is to organise the local descriptive elements into *forms* and *regions*, which are closed contour groups that are obtained in various ways. Regions may be defined by their boundaries, which have been formed by grouping together some set of edge, line, or place-tokens; or they may be defined by a first-order predicate operating on the primal sketch elements within it. This second method corresponds to the definition of a region by a texture, and it leads to a theory of the processes on which texture discrimination is based.

It is important to realize that the descriptive items that may be grouped here can be very abstract - like tokens for the end of a line, a blob, or a constructed line that joins two blobs. Tokens are created for each new group, and these tokens themselves become subject to the operation of the same or similar grouping processes as operated on elements of the raw primal sketch. The grouping processes are very conservative. They satisfy a principle that seems to have general application to recognition problems, called the *principle of least commitment*, which states that nothing should be done that may later have to be undone. Only "obvious" groupings are made, and where there is doubt between two possible groupings, both are constructed and held pending subsequent selection. Figure 3 illustrates some results of applying these grouping processes.

5: Representation and recognition of 3-dimensional shape

The last two components of the theory concern the representation of

three-dimensional shapes. One component deals with the nature of the representation system that is used, and the other with how to obtain it from the types of description that can be delivered from the primal sketch. The key ingredients of the representation system are:

(a) The deep structure of the three-dimensional representation of an object consists of a stick figure, where in formal terms each stick represents one or more axes in the object's generalized cone representation, as illustrated in figure 5. In fact, a hierarchy of stick figures exists, that allows one to describe an object on various scales with varying degrees of detail.

(b) Each stick figure is defined by a propositional database called a *3-D model*. The geometrical structure of a 3-D model is specified by storing the relative orientations of pairs of connecting axes. This specification is local rather than global, and it contrasts with schemes in which the position of each axis is specified in isolation, using some circumscribing frame of reference. (See legend to figure 5).

(c) When a 3-D model is being used to interpret an image, the geometrical relationships in the model are interpreted by a computationally simple mechanism called the *image-space processor*, which may be thought of as a device for representing the positions of two vectors in 3-space, and for computing their projections onto the image.

(d) During recognition, a sophisticated interaction takes place between the image, the 3-D model, and the image-space processor. This interaction gradually relaxes the stored 3-D model onto the axes computed from the image. Some facets of this process resemble the computation of a 3-D rotation, but a simple computer graphics metaphor is misleading. In fact, the rotations take place on abstract vectors (the axes) that are not even present in the original image; and at any moment, only two such vectors are explicitly represented.

The essence of this part of the theory is a method for representing the spatial disposition of the parts of an object and their relation to the viewer.

6: 2 1/2 - dimensional analysis of an image (Marr 1976c, Marr & Vatan in preparation)

In simple images, the forms delivered from the primal sketch correspond to the contours of physical objects. Finally therefore, we need to bridge the gap between such forms and the beginning of the 3-D analysis described in the previous paragraph. We call this 2 1/2 - dimensional analysis, and it consists largely of assigning to contours labels, that reflect aspects of their 3-dimensional configuration, before that configuration has been made explicit. The most powerful single idea here is the distinction between convex and concave edges and contour segments. One can show that these distinctions are preserved by orthogonal projections, and can be made the basis of a segmenting technique that decomposes a figure into 2-D regions that correspond to the appropriate 3-D decomposition for a wide range of viewing angles (see figure 6). Marr (1976c) has proved that the assumptions, that are implicit in the use of the convex-concave distinction to analyze a contour, are equivalent to assuming that the viewed shapes are composed of generalized cones. This adds additional support for using the stick-figure scheme based on generalized

Figure 5. Examples of 3-D models, and their arrangement into the 3-D model representation of a human shape. A 3-D model consists of a model axis (a) and component axes (b) that consist of a principal axis (the torso) and several auxiliary axes (the head and limbs) whose positions are described relative to the principal axis. The complete human 3-D model is enclosed in a rectangle (c). The 3-D model representation is obtained by concatenating 3-D models for different parts at different levels of detail. This is achieved by allowing a component axis of one 3-D model to be the model axis of another. Here, for example, the arm auxiliary axis in the human 3-D model acts as the model axis for the arm 3-D model, which itself has two component axes, the upper and lower arms. The figure shows how this scheme extends downwards as far as the fingers.

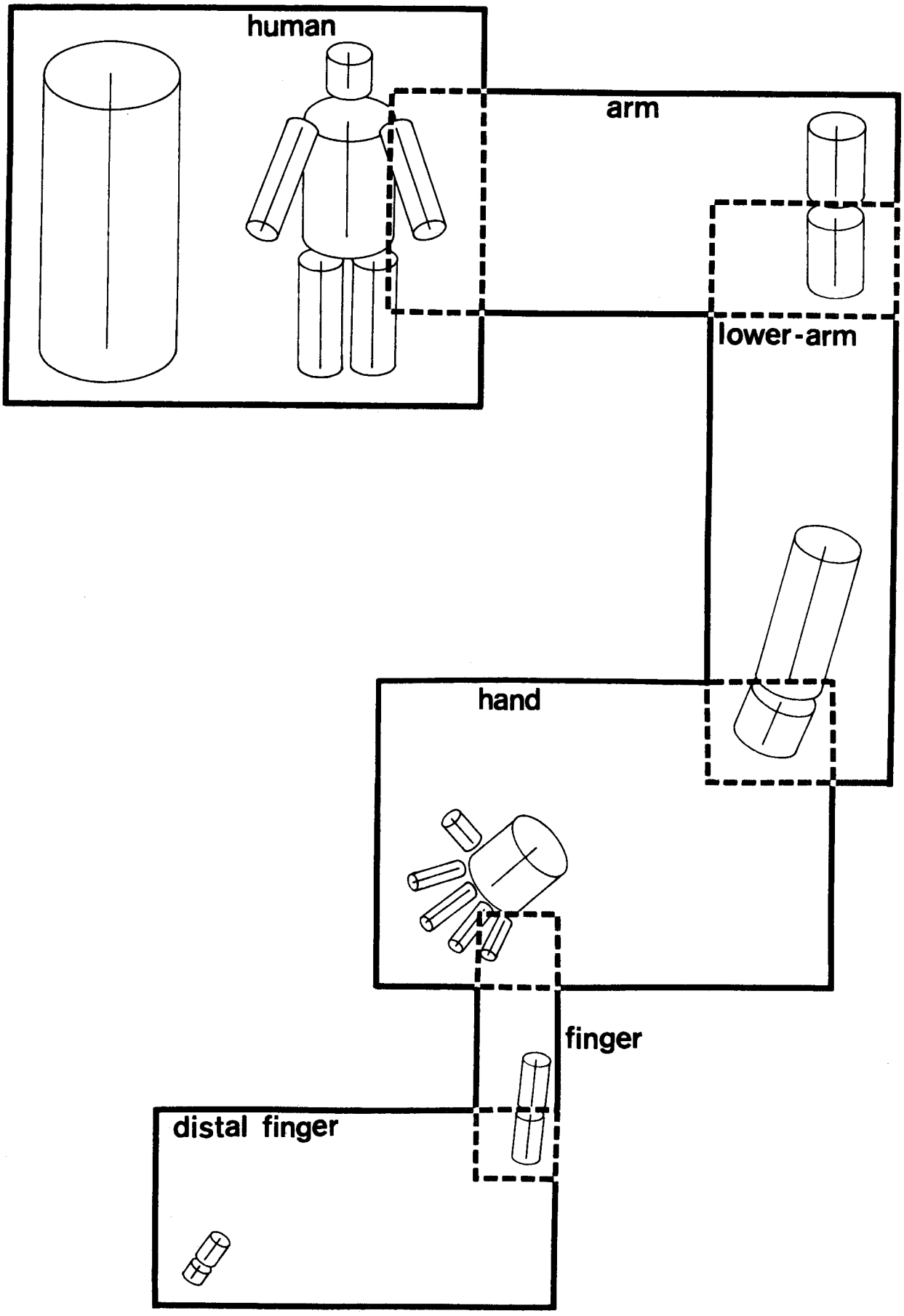
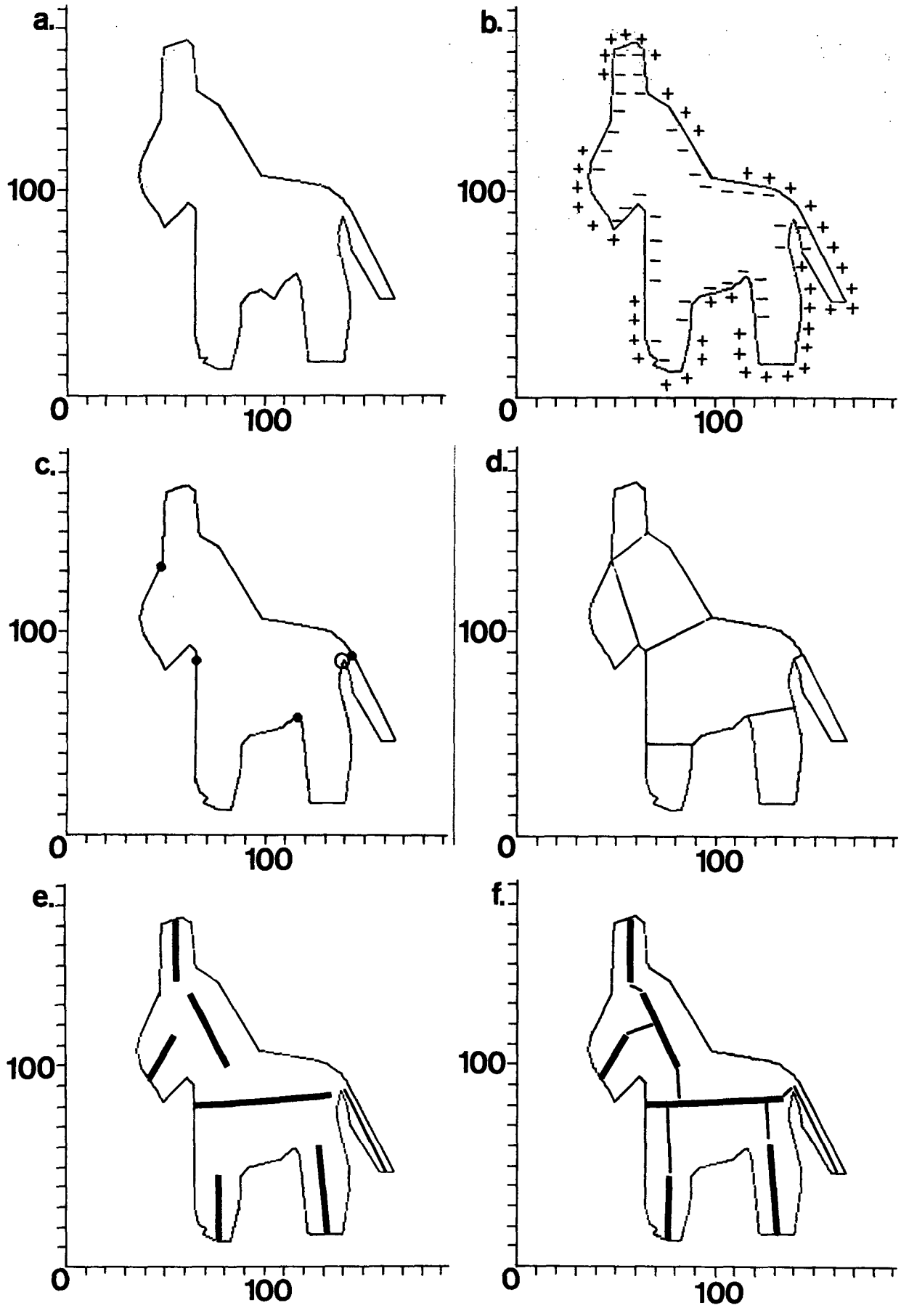
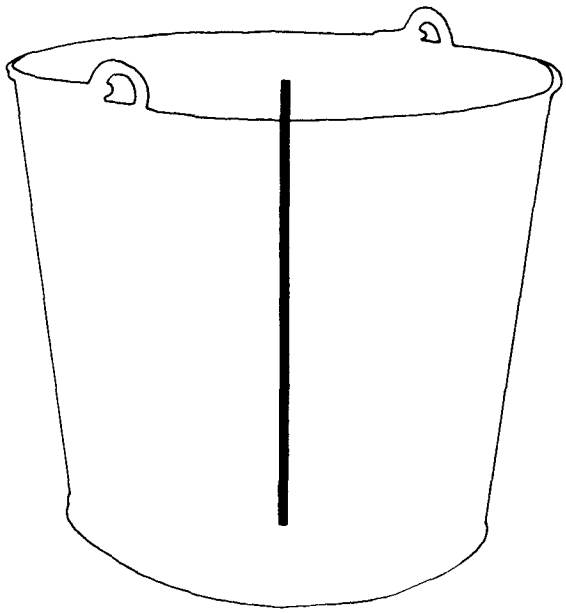

human
arm
lower-arm
hand
finger
distal finger

Figure 6. Analysis of a contour from Vatan and Marr (1976). The outline (a) was obtained by applying local grouping operations to a primal sketch, as in figure 4. It is then smoothed, and divided into convex and concave components (b). The outline is searched for deeply concave points or components, which correspond to strong segmentation points. One such point is marked with an open circle in (c). There are usually several possible matching points for each strong segmentation point, and the candidates for the marked point are shown here by filled circles (c). The correct mates for each segmentation point can usually be found by eliminating relatively poor candidates. The result of doing this here is the segmentation shown in (d). Once these segments have been defined, their corresponding axes (thick lines) are easy to obtain (e). They do not usually connect, but may be related to one another by intermediate lines which are called *embedding relations* (thin lines in f). According to the 3-D representation theory, the resulting stick figure (f) is the deep structure on which interpretation of this image is based.


a.
100
0
100
b.
100
0
100
c.
100
0
100
d.
100
0
100
e.
100
0
100
f.
100
0
100



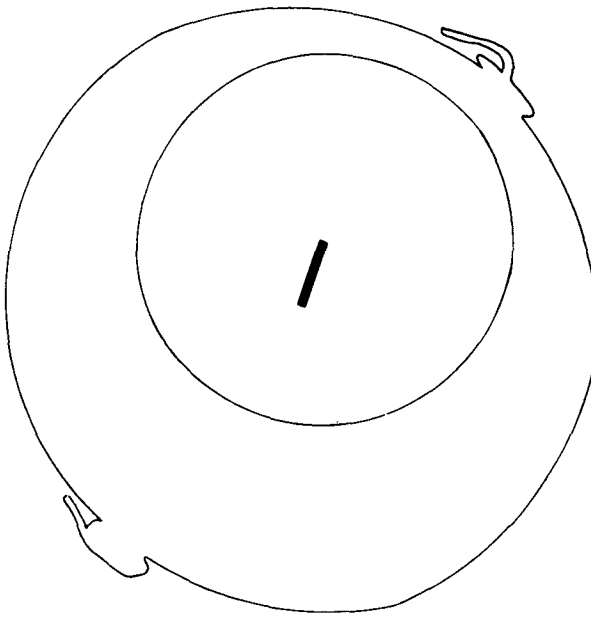


Figure 7. (a) and (b) show two views of a water-pail. Warrington & Taylor's (1973) patients are impaired on (b), but not on (a). This is consistent with the 3-D model representation, for reasons that are clear from (c) and (d). The outlines of the original figures are shown as thin lines, and the axis is shown as a thick one. This axis is directly recoverable from image (a), but not from (b) where it is severely foreshortened. Since the 3-D model representation relies on an explicit representation of this axis, the successful recognition of views like (b) requires considerable extra computation.

cones to represent 3-D shapes. The theory assigns many alternating figure effects like the Necker cube to the existence of alternative self-consistent labellings computed at this stage.

It is perhaps worth mentioning one interesting point that has emerged from this way of recognising and representing 3-D shapes. Warrington & Taylor (1973) described patients with right parietal lesions who had difficulty in recognising objects seen in "unconventional" views - like the view of a water pail seen from above (see figure 7). They did not attempt to define what makes a view unconventional. According to our theory, the most troublesome views of an object will be those in which its stick-figure axes cannot easily be recovered from the image. The theory therefore predicts that unconventional views in the Warrington & Taylor sense will correspond to those views in which an important axis in the object's generalised cylinder representation is foreshortened. Such views are by no means uncommon - if a 35mm camera is directed towards you, you are seeing an unconventional view of it, since the axis of its lens is foreshortened.

Examples of algorithms and mechanisms

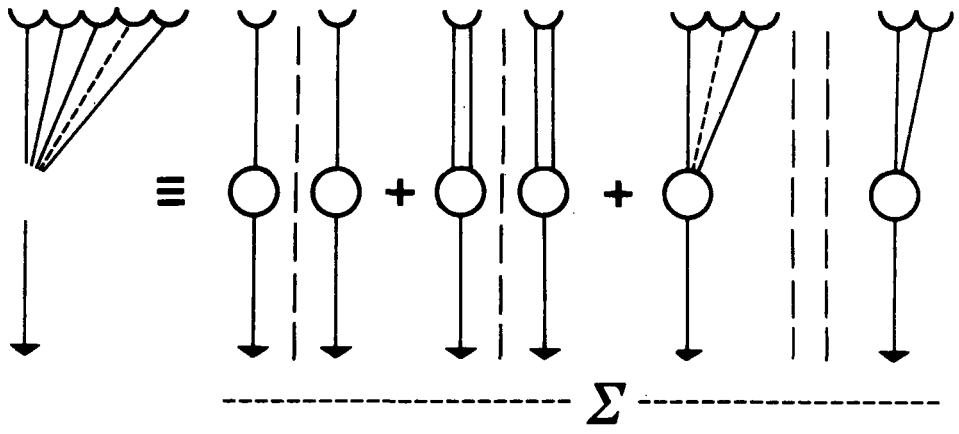
Between the top and bottom of our four levels lie descriptions of algorithms and descriptions of mechanisms. The distinction between these two levels is rather subtle, since they are often closely related. The form of a specific algorithm can impose strong constraints on the mechanisms, and conversely. Let us consider three examples.

1: "Simple" algorithms

An algorithm operates on some kind of input and yields a corresponding output. In formal terms, an algorithm can be thought of as a mapping between the input and the output space. Perhaps the simplest of all nonlinear operators on a linear space are the so-called polynomial operators. They encompass a broad spectrum of applications including all linear problems, and they approximate all sufficiently smooth, nonlinear operators. For this particular class of "simple" algorithms (i.e. representable through a "smooth" operator) polynomial representations provide a canonical decomposition in a series of simpler, multilinear operators. Figure 8 shows this decomposition in terms of interactions or "graphs" of various orders: in this way an algorithm, or its network implementation, may be decomposed into an additive sequence of simple, canonical terms, just as in another context, a function can be conveniently characterized by its various Fourier terms. Moreover, functional and computational properties can be associated with interactions of a given order and type.

Poggio & Reichardt (1976) used the polynomial representation of functionals to classify the algorithms underlying movement, position and figure-ground computation in the fly's visual system. The idea was to identify which terms, among the diversity of the possible ones, are implied by the experimental data. Figure 8 shows the graphs that play a significant role in the fly's control of flight and, in this sense, characterize the algorithms involved. The notion that seems to capture best the "computational complexity" of these simple, smooth mappings is the notion of p-order (perceptron-order, see Poggio and Reichardt, 1976). Movement computation in the fly is of

a



b

Separation of the three types of interactions in the fly

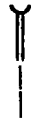
Movement computation	Position ("attractiveness") computation	
		
Corresponding to r^{ψ}	Corresponding to $D(\psi)$	Correction to superposition rule
Homogeneously distributed in the eye (no strong dependence on ψ and ν)	Mostly in the lower part of the eye ($D(\nu)$ and $L(\vartheta)$ dependence)	Mostly in the lower part of the eye
No "age" dependence	(?)	"Age" dependence
Light intensity threshold at about 10^{-4} candel/m^2 (Eckert, 1973)	Light intensity threshold (of fixation!) at about 10^{-2} cd/m^2 (Reichardt, 1973; Wehrhahn, 1976	?
Present in the Drosophila mutant S 129 (Heisenberg, pers. comm.)	Disturbed in the Drosophila mutant S 129 (Heisenberg, pers. comm.)	?

Figure 8. Graphical representation (a) of the decomposition of a "simple" nonlinear, n-input "algorithm" into a sum of interactions of various order. The functional representation

$$S\{\ldots x(t)\ldots\} = L^{(0)} + \sum_i L_i^{(1)}\{x_i(t)\} + \sum_{ij} L_{ij}^{(2)}\{x_i(t), x_j(t)\} + \ldots$$

where L is an n-linear mapping, can be read from an appropriate sequence of such elementary graphs. Fig. 8b shows the graphs that implement the fly's orientation behavior, studied by Reichardt and Poggio. Several findings suggest that they may correspond to separate physiological modules. Characteristic functional and computational properties can be associated to each interaction type. (From Poggio and Reichardt, 1976).

order 2, and figure-ground discrimination in the simple case of relative motion depends on fourth-order graphs, but possibly with p-order 2. A closed or Type 1 (Marr 1976b) theory of this kind may be a useful way of characterizing preprocessing operations in nervous systems. The approach has a rather limited validity however, since it does not apply to the large and important class of "non-smooth" algorithms, where cooperative effects, decisions and symbols play an essential role. While an arbitrary number of mechanisms and circuits may implement these "smooth" algorithms, it is clear that "forward" interactions between neurons are the most natural candidates.

Although the various levels of description are only loosely related, knowledge of the computation and of the algorithm may sometimes admit inferences at the lowest level of anatomy and physiology. The description of the visual system of the fly at the computational and functional level suggests, for instance, that different, separate neural structures may correspond to the different computations. Recent data support this conjecture. Movement computation seems to depend mainly on receptor system 1-6, while the position computation seems dependent on receptor system 7-8 (Wehrhahn, 1976 and in preparation). Mutants of Drosophila, normal with respect to the movement algorithm, are apparently disturbed in the position algorithm (Heisenberg, in preparation).

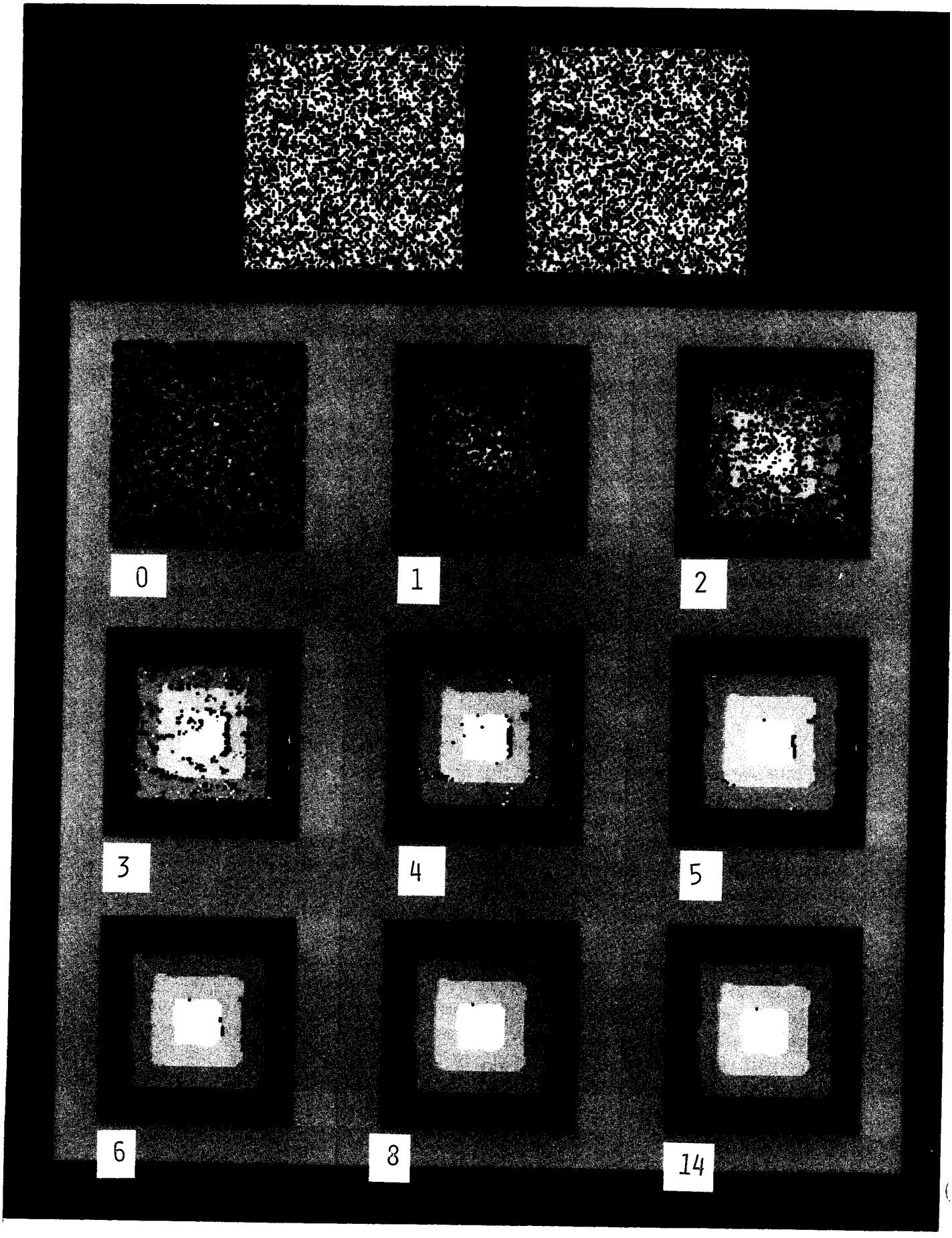
2: "Cooperative" algorithms

A more general and not precisely definable class of algorithms includes what one might call cooperative algorithms. Such algorithms may describe bifurcations and phase transitions in dynamical systems. An essential feature of a cooperative algorithm is that it operates on many "input" elements and reaches a global organization *via* local but highly interactive constraints. An apparently cooperative algorithm plays a major role in binocular depth perception (Julesz 1971). The stereopsis computation defined by figure 4a applies many local constraints to many local inputs to yield a final state consistent with these constraints. Various mechanisms could implement this type of algorithms. Parallel, recurrent, nonlinear interactions, both excitatory and inhibitory, seem to represent a natural implementation. In the stereopsis case such a mechanism is illustrated in the rest of figure 4. This mechanism may be realized through many different components and circuitries. In the nervous system, however, there are certain very obvious candidates, which allow some definite predictions. For instance, one is led to conjecture the existence of disparity columns (actually layers) of cells with reciprocal excitatory short-range interactions on each layer and long-range inhibitory interactions between layers with the characteristic "orthogonal" geometry of figure 4. Figure 9 shows that this algorithm succesfully extracts depth information from random-dot stereograms. The algorithm exhibits typical cooperative phenomena, like hysteresis and disorder-order transitions. It is important to stress that it is the computational problem which determines the structure of the excitatory and inhibitory interactions, and not "hardware" considerations about neurons or synapses. The apparent success of this cooperative algorithm in tackling the stereo problem suggests that other perceptual computations may be easy to implement in similar ways. Likely candidates are "filling-in" phenomena, subjective contours, figural reinforcement, some kinds of perceptual grouping and associative retrieval. In fact the associative retrieval network described by

Figure 9. A pair of random dot stereograms (left and right), the initial state of a network that implements the algorithm illustrated in figure 4, and various iterations of the network operating on this stereo pair. To understand how the figures represent states of the network, imagine looking down on it from above. The different disparity layers in the network are in parallel planes spread out horizontally, and the viewer is looking down through them. In each plane, some nodes are on and some are off. Each layer in the network has been assigned a different gray level, so that a node that is switched on in the lowest layer contributes a light point to the image, and one that is switched on in the top layer contributes a darker point. Initially (iteration 0) the the network is disorganised, but in the final state order has been achieved (iteration 14). The central squares have a divergent disparity relative to the background, and they therefore appears lighter. The density of the original random dot stereogram was 50%, but the algorithm succeeds in extracting disparity values at densities down to less than 5%. Let C_{xyd} denote the state of a cell (either 0 or 1) in the 3-D array of fig. 4b in position (x, y) and disparity d at the n-th iteration. Then the algorithm used here reads

$$C_{xyd}^{(n+1)} = u_\theta \{ \sum_{S(x,y,d)} C_{x'y'd'}^{(n)} - \epsilon \sum_{O(x,y,d)} C_{x'y'd'}^{(n)} + C_{xyd}^{(o)} \},$$

where $u_\theta(z) = 0$ if $z < \theta$, and $u(z) = 1$ otherwise; S(xyd) is a neighborhood of cell (xyd) on the same disparity layer; O(xyd) represents the neighborhood of cell (xyd) defined by the "orthogonal" directions shown in Fig. 4b. Excitation between parallel disparity layers may also be present.


0
1
2
3
4
5
6
8
14

Marr (1971), in connection with a theory of the hippocampal cortex, implements a cooperative algorithm.

3: Procedural algorithms

Still another and larger class of algorithms is represented by the specification of procedures, and the construction and manipulation of explicit symbolic descriptions. For example, the 3-D representation theory described in part 5 of the previous section explains how the stick figure representation of a viewed object may be obtained from an image, and manipulated during recognition. The detailed specification of the algorithms involved here is carried out by defining the datastructures that are created to represent the situation, and by specifying procedures that operate on these datastructures in accordance with the information currently being delivered from the image, and that available from stored models.

This way of specifying an algorithm is very general and powerful, although unlike the two other ways that we discussed, it is a far cry from the circuitry level of description at which neurophysiological experiments are carried out. In a digital computer, one does not try to bridge the gap between these two levels in one step. Instead, a basic instruction set, an assembler, a high level language (LISP, ALGOL) and a compiler are interposed to ease the burden of passing from the description of a computation down to the specification of a particular pattern of current flow.

We may eventually expect a similar intermediate vocabulary to be developed for describing the central nervous system. Hitherto, only one non-trivial "machine-code" operation has been studied in the context of neural hardware, namely simple storage and retrieval functions (Marr 1969 & 1971, and Brindley 1969).

Discussion

The prospects for this approach to the problems of visual information processing look very bright. In one case, that of stereopsis, we have been able to carry out a complete analysis of the problem, starting with the structure of the computation, and following it right down to psychophysical and neurophysiological predictions. As far as we are aware, this is the first time that this has been done. Although the cerebellar theory of Marr (1969) provided a theory of a particular mechanism (of associative memory) and its neurophysiological predictions, it did not show how this mechanism could be used to execute motor skills. The stereopsis theory also contains what we believe is the first use of a cooperative algorithm to solve an information processing problem. As a psychological model the theory may of course be shown to be false, and part of its value is that it can be. Even if it is, we feel that it exhibits the correct *form* for a theory in this subject, and is valuable if only because of that. Another point of particular interest is the way that clues from clinical neurology, and predictions about such findings, are often combined by a computational theory with information about neurophysiology and psychology (e.g. the 3-D theory described above). We feel that this aspect of the research, together with the extensive use of psychophysical techniques, will become increasingly important as the theories' details are worked out.

The research that needs to be done falls into two main categories, research at level 3 on algorithms, and research at level 4 on the many computational problems that have not yet been examined. An example of the first type is to understand the relation between various types of best-fit associative retrieval algorithm; the cooperative algorithm of Marr (1971); Elia's algorithm and Rivest's extensions of it (see e.g. Rivest 1974); and Poggio's algorithm based on the Moore-Penrose pseudo-inverse of a singular matrix (Poggio 1975). An understanding of the power of these algorithms, and of the relation between them, seems to be an essential prerequisite for studying neural networks for retrieval. This in turn is probably an important problem, because most computations demand that, at some stage, information is retrieved from a store and used.

Examples of the second category are abundant. The modules we have already described need to be cleaned up; there are many problems in figure-ground separation, texture vision, color vision and other early phenomena that need careful study, and later areas of visual information processing are almost untouched. The importance to this research of establishing good communication between the disciplines of computer science and experimental psychology needs no emphasis, since that is how computational theories can ultimately be put to the test. As far as we can tell, an interdisciplinary approach to the vision problem based on a theoretical computational approach of the kind we have described, stands poised to make a significant contribution to our knowledge of visual perception.

ACKNOWLEDGEMENTS: I thank Tomaso Poggio for many discussions and K. Prendergast for preparing the drawings.

References

Allman, J. M., Kaas, J. H., Lane, R. H. & Miezin, F. M. (1972) A representation of the visual field in the inferior nucleus of the pulvinar in the owl monkey (*Aotus trivirgatus*). *Brain Research, 40*, 291-302.

Allman, J. M., Kaas, J. H. & Lane, R. H. (1973) The middle temporal visual area (MT) in the bushbaby, *Galago senegalensis*. *Brain Research, 57*, 197-202.

Allman, J. M. & Kaas, J. H. (1974a) The organization of the second visual area (V-II) in the owl monkey: a second order transformation of the visual hemifield. *Brain Research, 76*, 247-265.

Allman, J. M. & Kaas, J. H. (1974b) A visual area adjoining the second visual area (V-II) on the medial wall of parieto-occipital cortex of the owl monkey (*Aotus trivirgatus*). *Anat. Rec., 178*, 297-8.

Allman, J. M. & Kaas, J. H. (1974c) A crescent-shaped cortical visual area surrounding the middle temporal area (MT) in the owl monkey (*Aotus trivirgatus*). *Brain Research, 81*, 199-213.

Brindley, G. S. (1969). Nerve net models of plausible size that perform many simple learning tasks. *Proc. Roy. Soc. B., 174*, 173-191.

Cooley, J. M. & Tukey, J. W. (1965). An algorithm for the machine computation of complex Fourier series. *Math. Comp., 19*, 297-301.

Critchley, M. (1953) *The parietal lobes*. London: Edward Arnold.

Efron, R. (1968). What is perception? In:*Boston studies in the philosophy of science, IV*, Eds. R. S. Cohen & M. W. Wartovsky. Dordrecht, Holland: Reide.

Freuder, E. C. (1975). A computer vision system for visual recognition using active knowledge. *M. I. T. A. I. Lab. Technical Report 345*.

Geiger, G. & Poggio, T. (1975). The Mueller-Lyer figures and the fly. *Science 190*, 479-480.

Hubel, D. H. & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.), 160*, 106-154.

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

Marr, D. (1969). A theory of cerebellar cortex. *J. Physiol. (Lond.) 202*, 437-470.

Marr, D. (1971). Simple memory: a theory for archicortex. *Phil. Trans. Roy. Soc. B., 252*, 23-81.

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. (1976a). Early processing of visual information. *Phil. Trans. Roy. Soc. B.* (in the press).

Marr, D. (1976b). Artificial Intelligence -- a personal view. *M. I. T. A. I. Lab. Memo 355*.

Marr, D. (1976c). Analysis of contour. *M. I. T. A. I. Lab. Memo 372*.

Marr, D. & Nishihara, H. K. (1976). Representation and recognition of the spatial

organization of three-dimensional shapes. (Submitted to the Royal Society for publication).

Marr, D. & Poggio, T. (1976). Cooperative computation of stereo disparity. *Science*, (submitted for publication).

Poggio, T. (1975). On optimal non-linear associative recall. *Biol. Cybernetics 19*, 201-209.

Poggio, T. & Reichardt, W. (1973). A theory of the pattern induced flight orientation of the fly *Musca Domestica*. *Kybernetik 12*, 185-203.

Poggio, T. & Reichardt, W. (1976). Visual control of the orientation behavior of the fly: towards the underlying neural interactions. *Quarterly Reviews in Biophysics*, (in the press).

Reichardt, W. (1970). The insect eye as a model for analysis of uptake, transduction and processing of optical data in the nervous system. In: *The Neurosciences, Second Study Program.* (ed. F. O. Schmitt), 494-511. New York: The Rockfeller University Press.

Reichardt, W. & Poggio, T. (1975). A theory of pattern induced flight orientation of the fly *Musca Domestica* II. *Kybernetik, 18*, 69-80.

Reichardt, W. & Poggio, T. (1976). Visual control of the orientation behavior of the fly: a quantitative analysis. *Quarterly Reviews in Biophysics*, (in the press).

Rivest, R. L. (1974). On the optimality of Elia's algorithm for performing best-match searches. *Information Processing 74*, 678-681. Amsterdam: North Holland Publishing Co.

Shepard, R. N. (1975). Form, formation, and transformation of internal representations. In: *Information processing and cognition: The Loyola Symposium*, Ed. R. Solso, pp 87-122. Hillsdale, N. J.: Lawrence Erlbaum Assoc.

Shepard, R. N. & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*, 701-703.

Shirai, Y. (1973). A context-sensitive line finder for recognition of polyhedra. *Artificial Intelligence, 4*, 95-120.

Ullman, S. (1976a). On visual detection of light sources. *Biol. Cybernetics 21*, 205-212.

Ullman, S. (1976b). Structure from motion. *M. I. T. Ph. D. Thesis in preparation.*

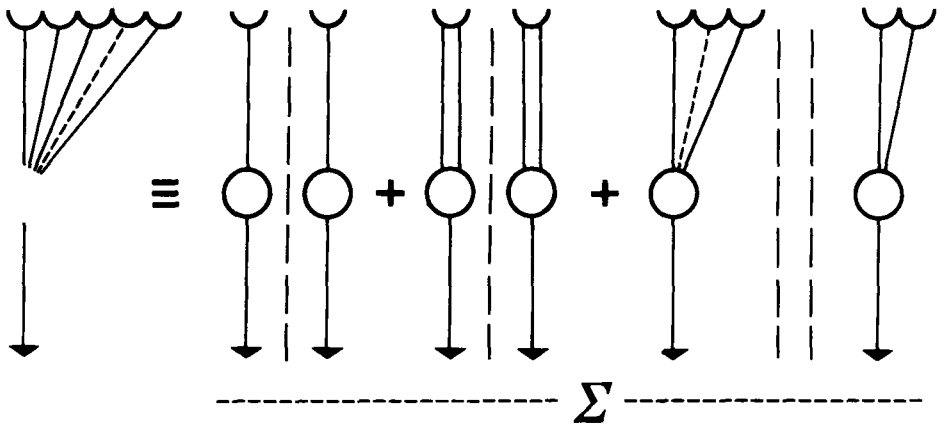
Vinken, P. J. & Bruyn, G. W. (1969) Eds. *Handbook of Clinical Neurology: Vol. 2, Localization in Clinical Neurology.* (In association with A. Biemond). Amsterdam: North Holland Publishing Co.

Warrington, E. K. & Taylor, A. M. (1973). The contribution of the right parietal lobe to object recognition. *Cortex, 9*, 152-164.

Wehrhahn, C. (1976). Evidence for the role of retinal receptors R 7/8 in the orientation behavior of the fly. *Biol. Cybernetics, 21*, 213-220.

Zeki, S. M. (1971) Cortical projections from two prestriate areas in the monkey. *Brain Research, 34*, 19-35.

a



b

Separation of the three types of interactions in the fly

Movement computation	Position ("attractiveness") computation	
		
Corresponding to $r\dot{\psi}$	Corresponding to $D(\psi)$	Correction to superposition rule
Homogeneously distributed in the eye (no strong dependence on ψ and ϑ)	Mostly in the lower part of the eye ($D(\psi)$ and $L(\vartheta)$ dependence)	Mostly in the lower part of the eye
No "age" dependence	(?)	"Age" dependence
Light intensity threshold at about 10^{-4} candel/m^2 (Eckert, 1973)	Light intensity threshold (of fixation!) at about 10^{-2} cd/m^2 (Reichardt, 1973; Wehrhahn, 1976	?
Present in the Drosophila mutant S 129 (Heisenberg, pers. comm.)	Disturbed in the Drosophila mutant S 129 (Heisenberg, pers. comm.)	?

Figure 8. Graphical representation (a) of the decomposition of a "simple" nonlinear, n-input "algorithm" into a sum of interactions of various order. The functional representation

$$S\{\ldots x(t)\ldots\} = L^{(0)} + \sum_i L_i^{(1)}\{x_i(t)\} + \sum_{ij} L_{ij}^{(2)}\{x_i(t), x_j(t)\} + \ldots$$

where L is an n-linear mapping, can be read from an appropriate sequence of such elementary graphs. Fig. 8b shows the graphs that implement the fly's orientation behavior, studied by Reichardt and Poggio. Several findings suggest that they may correspond to separate physiological modules. Characteristic functional and computational properties can be associated to each interaction type. (From Poggio and Reichardt, 1976).