PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Visual Information Processing: The Structure and Creation of Visual Representations [and Discussion]

D. Marr, S. Lal and H. B. Barlow

Phil. Trans. R. Soc. Lond. B 1980 **290**, 199-218
doi: 10.1098/rstb.1980.0091

Phil. Trans. R. Soc. Lond. B **290**, 199–218 (1980)
Printed in Great Britain

Visual information processing: the structure and creation of visual representations

By D. Marr
M.I.T. Artificial Intelligence Laboratory and Department of Psychology,
545 Technology Square, Cambridge, Massachusetts 02139, *U.S.A.*

For human vision to be explained by a computational theory, the first question is plain: What are the problems that the brain solves when we see? It is argued that vision is the construction of efficient symbolic descriptions from images of the world. An important aspect of vision is therefore the choice of representations for the different kinds of information in a visual scene. An overall framework is suggested for extracting shape information from images, in which the analysis proceeds through three representations: (1) the primal sketch, which makes explicit the intensity changes and local two-dimensional geometry of an image; (2) the $2\frac{1}{2}$-D sketch, which is a viewer-centred representation of the depth, orientation and discontinuities of the visible surfaces; and (3) the 3-D model representation, which allows an object-centred description of the three-dimensional structure and organization of a viewed shape. The critical act in formulating computational theories for processes capable of constructing these representations is the discovery of valid constraints on the way the world behaves, that provide sufficient additional information to allow recovery of the desired characteristic. 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.

Introduction

Modern neurophysiology has learned much about the operation of the individual nerve cell, but disconcertingly little about the meaning of the circuits that they compose in the brain. The reason for this can be attributed, at least in part, to a failure to recognize what it means to understand a complex information-processing system; for a complex system cannot be understood as a simple extrapolation from the properties of its elementary components. One does not formulate, for example, a description of thermodynamical effects by using a large set of equations one for each of the particles involved. One describes such effects at their own level, that of an enormous collection of particles, 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 analysed and understood at several different levels. Indeed, in a system that solves an information-processing problem, we may distinguish four important levels of description (Marr & Poggio 1977; Marr 1977*a*). At the lowest level, there is basic component and circuit analysis: how do transistors (or neurons) or diodes (or synapses) work? The second level is the study of particular mechanisms: adders, multipliers and memories, these being assemblies made from basic components. The third level is that of the algorithm, the scheme for a computation; and the top level contains the *theory* of the computation. A theory of addition, for example, would encompass the meaning of that operation, quite independent of the representation of the numbers to be added, i.e., whether they are, say arabic or roman. But it would also include the realization that the first of these representations is the more suitable of the two. An algorithm, on the other hand, is a particular method by which to add numbers. It therefore applies to a particular representation, since plainly an algorithm that adds arabic numerals would be useless for roman. At still a further level down, one comes upon a mechanism for addition – say a pocket calculator – which simply implements a particular algorithm. As a second example, take the case of Fourier analysis. Here the computational theory of the Fourier transform – the decomposition of an arbitrary mathematical curve into a sum of sine waves of differing frequencies – is well understood, and is expressed independently of the particular way in which it might be computed. One level down, there are several algorithms for computing a Fourier transform, among them the so-called fast Fourier transform, which comprises a sequence of mathematical operations, and the so-called spatial algorithm, a single, global operation that is based on the mechanisms of laser optics. All such algorithms produce the same result, so the choice of which one to use depends upon the particular mechanisms that are available. If one has fast digitial memory, adders and multipliers, one will use the fast Fourier transform, and if one has a laser and photographic plates, one will use an 'optical' method.

Now each of the four levels of description will have its place in the eventual understanding of perceptual information processing, and of course there are logical and causal relations among them. But the important point is that the four 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, for instance, mainly to the physical mechanisms of vision (like the after-images such as the one seen after staring at a light bulb) or mainly to the computational theory of vision (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 nature of computations that underlie perception depend more upon the computational *problems* that have to be solved than upon the particular hardware in which their solutions are implemented. To phrase the matter another way, an algorithm is likely to be understood more readily by understanding the nature of the problem that it deals with than by examining the mechanism (and the hardware) by which it is embodied. There is, after all, an analogue to all of this in physics, where a thermodynamical approach represented, at least historically, the first stage in the study of matter: it succeeded in producing a theory of gross properties such as temperature. A description in terms of mechanisms or elementary components – in this case atoms and molecules – appeared some decades afterwards.

Our main point, therefore, is that the topmost of our four levels, that at which the necessary structure of 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 this article, I summarize some examples of vision theories at the uppermost level.

Conventional approaches

The problems of visual perception have attracted the curiosity of scientists for many centuries. Important early contributions were made by Newton (1704), who laid the foundations for modern work on colour vision, and Helmholtz (1910), whose treatise on physiological optics maintains its interest even today. Early in this century, Wertheimer (1938) noticed the apparent motion not of individual dots but instead of wholes, or 'fields', in images presented sequentially, as if in a cine film. In much the same way do we perceive the migration across the sky of a

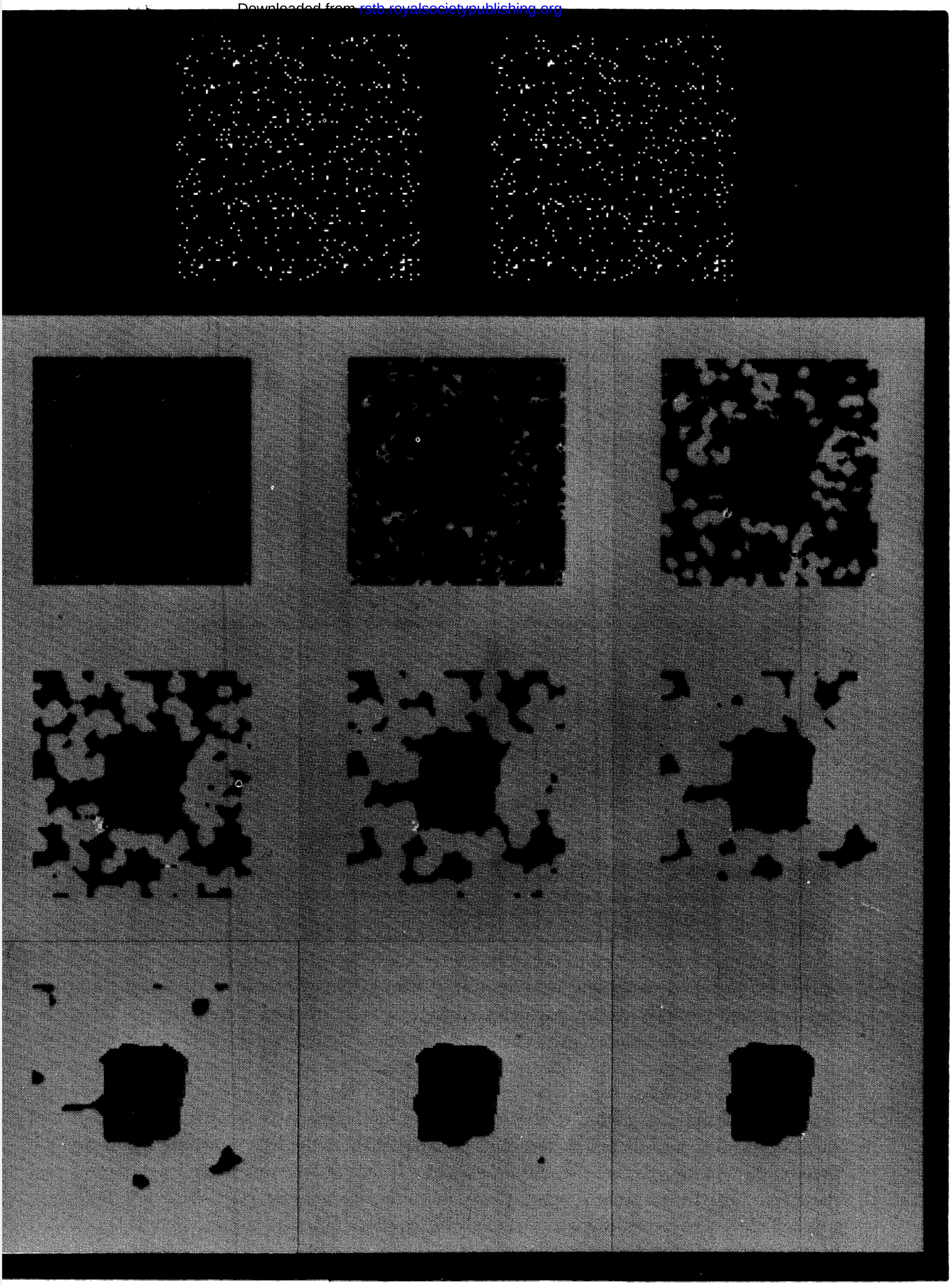

FIGURE 1. A sparse random-dot stereogram (the top two images), and its decoding by Marr & Poggio's (1976) cooperative algorithm. The initial state contains all possible matches within a given disparity range, and the algorithm embodies the constraints of uniqueness and continuity to eliminate false targets. Shades of grey are used to signify matches at different disparities. The figure shows the initial state, and the states after 1, 2, 3, 4, 5, 6, 8 and 14 iterations. The algorithm progressively reveals a square hovering in depth. This algorithm is not the one used by the human visual system.

flock of geese, the flock somehow constituting a single entity, and not individual birds. This observation started the Gestalt school of psychology, which was concerned with describing the qualities of wholes, including solidarity and 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 1940s and 1950s made single-cell recording possible, and with Kuffler's (1953) study of retinal ganglion cells – the neurons of the eye that give rise to the optic nerve – a new approach to the problem was born. Its most renowned practitioners are Hubel & Wiesel (1962, 1968), 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.

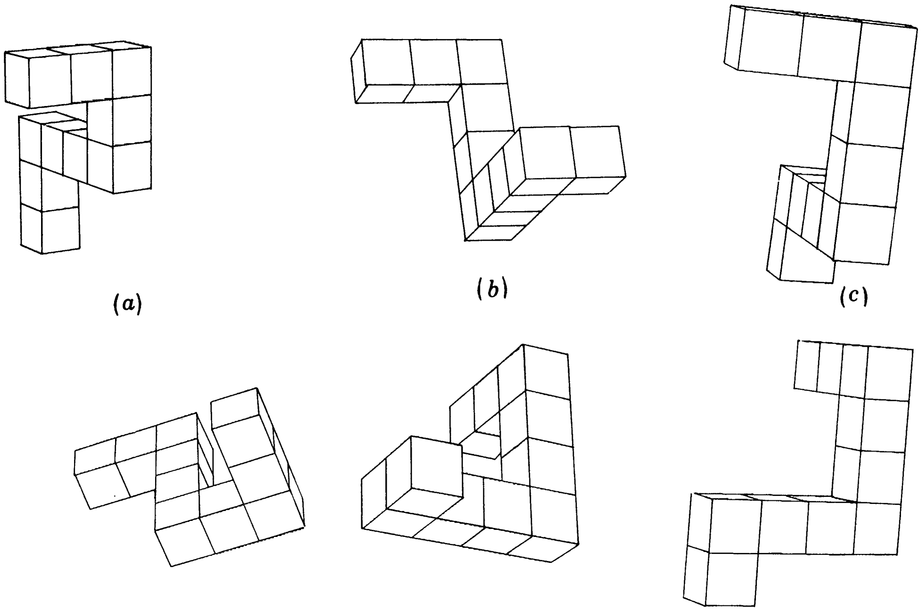

Figure 2. Some drawings similar to those used in Shepard & Metzler's (1971) experiments on mental rotation. Those shown in (*a*) and (*b*) are identical and the relative angle between the two is 80°. Those in (*c*) are not identical, and no rotation will bring them into congruence.

Students of the psychology of perception were also affected by a technological advance, the advent of the digital computer. Most notably, it allowed Bela Julesz in 1959 to devise random-dot stereograms (see Julesz 1971), 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. An example is shown in figure 1. Here the image for the left eye is a matrix of black and white squares generated at random by a computer program. The image for the right is made by copying the left image and then shifting a square-shaped region at its centre slightly to the left, providing a new random pattern to fill in the gap that the shift must create. If each of the eyes sees only one matrix, as if they were both in the same physical place, the result is the sensation of a square floating in space. Plainly such percepts are caused solely by the stereo disparity between matching elements in the images presented to each eye.

More recently, considerable interest has been attracted by a rather different approach. In 1971, Shepard & Metzler made line drawings of simple objects that differed from one another either by a three-dimensional rotation, or by a rotation plus a reflexion (see figure 2). They asked how long it took to decide whether two depicted objects differed by a rotation and a reflexion, 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. Indeed, it varied linearly with this angle. One is led thereby to the notion that a mental rotation of sorts is actually being performed: that a mental description of the first shape in a pair is being adjusted incrementally in orientation until it matches the second, such adjustment requiring greater time when greater angles are involved.

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 through these techniques. For we repeat: perhaps the most striking feature of neurophysiology and psychophysics at present is that they *describe* the behaviour of cells or of subjects, but do not *explain* it. What are the visual areas of the cerebral cortex actually doing? What are the problems in doing it that need explaining, and at what level of description should such explanations be sought?

A Computational approach to vision

In trying to come to grips with these problems, our group at the M.I.T. Artificial Intelligence Laboratory has adopted a point of view that regards visual perception as a problem primarily in information processing. The problem begins with a large, grey-level intensity array, which suffices to approximate an image such as the world might cast upon the retinas of the eyes, and it culminates in a *description* that depends on that array, and on the purpose that the viewer brings to it. Our particular concern in this article will be with the derivation of a description well suited for the recognition of three-dimensional shapes.

The primal sketch

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 symbols correspond in some way to natural symbols that are computed out of the image during the normal course of its interpretation. Our theory therefore asserts that the first operation on an image is to transform it into a primitive but rich description of the way its intensities change over the visual field, as opposed to a description of its particular intensity values in and of themselves. This yields a description of markedly reduced size that still captures the important aspects required for image analysis. We call it a *primal sketch* (Marr 1976). Consider, for example, an intensity array of 1000 by 1000, or 10^6 points in all. Even if the possible intensity at any one point were merely black or white – two different brightnesses – the number of all possible arrays would still be 2^{10^6}. In a real image, however, there tend to be continuities of intensity – areas where brightness varies uniformly – and this tends to eliminate possibilities in which the black and white oscillate wildly. It also tends to simplify the array. Typically, therefore, a primal sketch need not include a set of values for every point in an image. As stored in a computer, it will instead constitute an array with numbers representing the directions, magnitudes, and spatial extents of intensity changes assigned to certain specific points in an image, points that

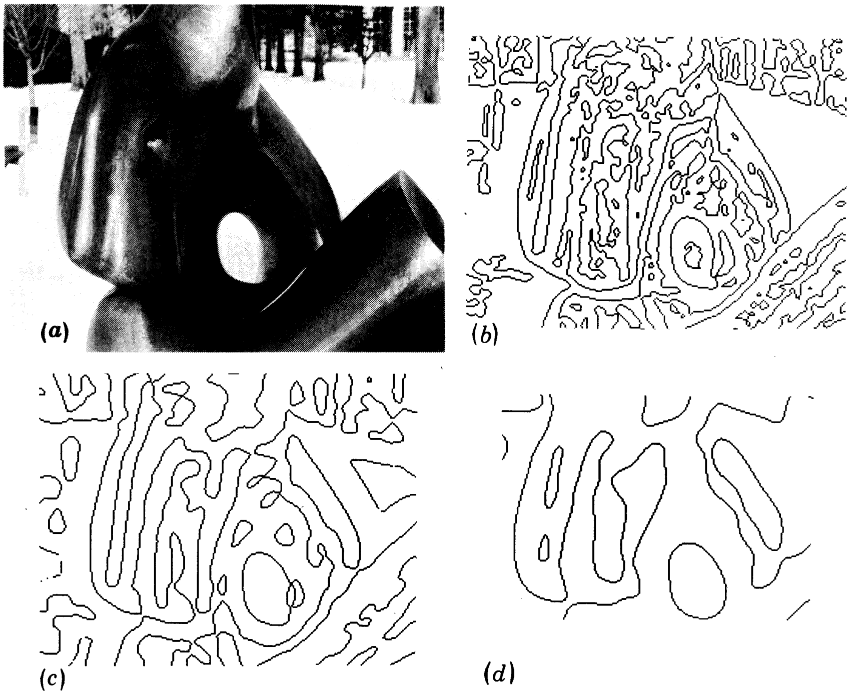

Figure 3. The image (a), which is 320×320 pixels, has been convolved with $\nabla^2 G$, a centre–surround operator with central excitatory region of width $2\sigma = 6$, 12 and 24 pixels. These filters span approximately the range of filters that operate in the human fovea. The zero-crossings of the filtered images are shown in (b), (c) and (d). These are the precursors of the raw primal sketch. (From Marr & Hildreth 1979, figure 6).

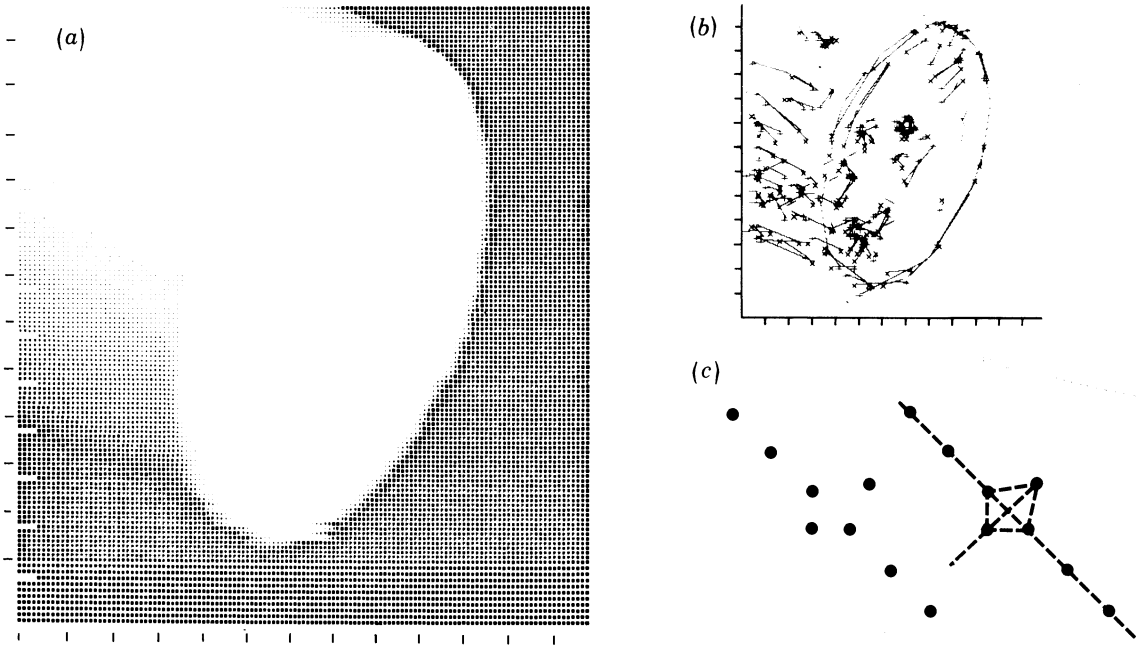

Figure 4. The primal sketch makes explicit information held in an intensity array (a). There are two kinds of information: one (b) concerns changes in intensity, represented by orientated edge, bar and blob primitives, together with associated parameters that measure the contrast and spatial extent of the intensity change; and the other (c) is the local geometry of significant places in the image. Such places are marked by place-tokens, which can be defined in a variety of ways, and the geometric relations between them are represented by virtual lines (Marr 1976, figures 7 and 12a).

tend to be places of locally high or low intensity. The positions of these points, particularly their arrangement among their immediate neighbours – that is to say, the local geometry of the image – must also be made explicit in the primal sketch, as it would otherwise be lost. (It was implicit, of course, in the 1000 by 1000 array, but we are no longer retaining data for each of those 10^6 places.) One way to do this is to specify 'virtual lines' – directions and distances – between neighbouring points of interest in the sketch.

The process of computing the primal sketch involves several steps. The first is the derivation of the *raw primal sketch* (see Marr & Hildreth 1979), which involves detecting and representing the intensity changes in the image. First, the image is filtered through a set of medium bandpass second differential operators $\nabla^2 G$, (where ∇^2 is the Laplacian and G is a Gaussian distribution), and the zero-crossings in the filtered images are found (see figure 3). This representation of the intensity changes is probably complete (Marr *et al.* 1979).

Although in general there is no reason why the zero-crossings found by the different channels should be related, in practice they will be. The reason is that most intensity changes in an image arise from physical phenomena that are spatially localized. This constraint allowed Marr & Hildreth to formulate the *spatial coincidence assumption* which states: If a zero-crossing is present in a set of independent $\nabla^2 G$ channels over a contiguous range of sizes, and it has the same position and orientation in each channel, then the set of such zero-crossings may be taken to indicate the presence of an intensity change in the image that is due to a single physical phenomenon (a change in reflectance, illumination, depth or surface orientation).

This assumption allows one to combine the zero-crossings from different channels into edge-segment descriptors, bars and blobs (see figure 4*b*), which constitute the raw primal sketch. To obtain the full primal sketch, these primitive elements are grouped, perhaps hierarchically, into units called place-tokens, which associate properties like length, width, brightness and so forth with positions in the image (Marr 1976). Virtual lines may then be used to represent the local geometry of these place-tokens (see figure 4*c* and Stevens 1978).

Recently, Marr & Ullman (1979) have extended the work of Marr & Hildreth to include the detection and use of directional selectivity. They have proposed specific roles for the X and Y channels found originally by Enroth-Cugell & Robson (1966), and in an explicit model for one class of cortical simple cell, they showed how to combine X and Y information to form a directionally selective unit.

Modules of early visual processing

The primal sketch of an image is typically a large and unwieldy collection of data, even despite its simplification relative to a grey-level array; for this is the unavoidable consequence of the irregularity and complexity of natural images. The next computational problem is thus its decoding. Now the traditional approach to machine vision assumes that the essence of such a decoding is a process called *segmentation*, whose purpose is to divide a primal sketch, or more generally an image, into regions that are meaningful, perhaps as physical objects. Tenenbaum & 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 (1975) used a similar approach to identify a hammer in a simple scene. Upon finding a blob, his computer program would tentatively label it as the head of a hammer, and begin a search for confirmation in the form of an appended shaft. 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 in the segmentation of an image. Freuder's work, for example, was almost entirely devoted to the design of a system that made this possible. But despite considerable efforts over a long period, the theory and practice of segmentation remain rather primitive, and here again we believe that the main reason lies in the failure to formulate precisely the goals of this stage of the processing – a failure, in other words, to work at the topmost level of visual theory. What, for example, is an object? Is a head an object? Is it still an object if it is attached to a body? What about a man on horseback?

Marr (1978) argued that the early stages of visual information processing ought instead to squeeze the last possible ounce of information from an image before taking recourse to the descending influence of 'high-level' knowledge about objects in the world. Let us turn, then, to a brief examination of the physics of the situation. As noted earlier, the visual process begins with arrays of intensities projected upon the retinas of the eyes. The principal factors that determine these intensities are (1) the illuminant, (2) the surface reflectance properties of the objects viewed, (3) the shapes of the visible surfaces of these objects, and (4) the vantage point of the viewer. Thus if the analysis of the input intensity arrays is to operate autonomously, at least in its early stages, it can only be expected to extract information about these four factors. In short, early visual processing must be limited to the recovery of localized physical properties of the visible *surfaces* of a viewed object, particularly local surface dispositions (orientation and depth) and surface material properties (colour, texture, shininess, and so on). More abstract matters such as a description of overall three-dimensional shape must come after this more basic analysis is complete.

An example of early processing is stereopsis. Imagine that images of a scene are available from two nearby points at the same horizontal level – the analogue of the images that play upon the retinas of your left and right eyes. The images are somewhat different, of course, in consequence of the slight difference in vantage point. Imagine further that a particular location on a surface in the scene is chosen from one image; that the corresponding location is identified in the other image; and that the relative positions of the two versions of that location are measured. This information will suffice for the calculation of depth – the distance of that location from the viewer. Notice that methods based on grey-level correlation between the pair of images fail to be suitable because a mere grey-level measurement does not reliably define a point on a physical surface. To put the matter plainly, numerous points in a surface might fortuitously be the same shade of grey, and differences in the vantage points of the observer's eyes could change the shade as well. The matching must evidently be based instead on objective markings that lie upon the surface, and so one had to use changes in reflectance. One way of doing this is to obtain a primitive description of the intensity changes that exist in each image (such as a primal sketch), and then to match these descriptions. After all, the line segments, edge segments, blobs, and edge termination points included in such a description correspond quite closely to boundaries and reflectance changes on physical surfaces. The stereo problem – the determination of depth given a stereo pair of images – may thus be reduced to that of matching two primitive descriptions, one from each eye; and to help in this task there are physical constraints that translate into two rules for how the left and right descriptions are combined.

Uniqueness

Each item from each image may be assigned at most one disparity value, that is to say, a unique position relative to its counterpart in the stereo pair. 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.

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 discontinuities in depth.

In the case of random-dot stereograms, the computational problem is rather well defined, essentially because of Julesz's demonstration that random-dot stereograms, containing no monocular information, still yield stereopsis. In 1976, Marr & Poggio developed a method for computing local disparities in a pair of random-dot stereograms by an iterative, parallel procedure known technically as cooperative algorithm (see figure 1 and Marr *et al.* 1977). This sort of algorithm has the property that it can be defined completely in terms of simple local interactions because at each of its iterations, each point is affected only by a calculation performed on its immediate neighbourhood. Yet all points are so affected during each successive iteration, so the transformations take on a complex global nature. Subsequent comparison of the algorithm's performance with psychophysical data showed that it did not hold up well as a model for human stereopsis. To be sure, it performed better than people do on the standard stereograms like that shown in figure 1; but it did not explain people's ability to see stereograms in which one of the two images is defocused slightly or enlarged slightly relative to the other. These observations led Marr & Poggio (1979) to devise another algorithm, based on the human use of spatial-frequency-tuned channels and vergence eye movements. This algorithm is consistent with all of the currently known psychophysical data.

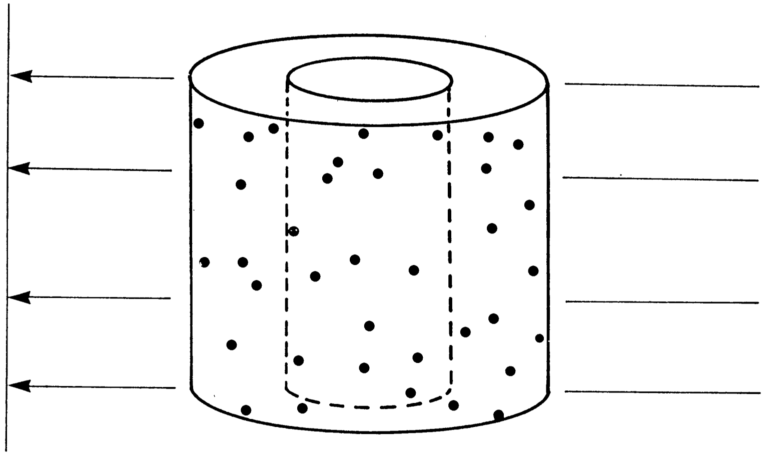

FIGURE 5. The motion analogue of the random-dot stereogram. Two transparent, concentric cylinders are rotated in opposite directions. Each has dots scattered on its surface. A cine camera photographs the scene from the side, and each frame contains only a pattern of random dots. When a human watches the film, however, he immediately perceives the two counter-rotating cylinders. (From Ullman 1979*b*.)

A second example of early visual processing concerns the derivation of structure from motion. It has long been known that as an object moves relative to the viewer, the way its appearance changes provides information that we can use to determine its shape (Wallach & O'Connell 1953). The motion analogue of a random-dot stereogram is illustrated in figure 5, and as expected, humans can easily perceive shape from a succession of frames, each of which on its own is merely a set of random-dots. In various papers and a forthcoming book on the subject, Ullman (1979*a*, *b*) decomposed the problem into two parts: matching the elements that occur in consecutive images, and deriving shape information from measurements of their changes in position. Ullman then showed that these problems can be solved mathematically. His basic idea is that, in general, nothing can be inferred about the shape of an object given only a set of sequential views of it; some extra assumptions have to be made. Accordingly, he formulates an assumption of rigidity, which states that if a set of moving points has a *unique* interpretation as a rigid body in motion, that interpretation is correct. (The assumption is based on a theorem, which he proves, stating that three distinct views of four non-coplanar points on a rigid body are sufficient to determine uniquely their three-dimensional arrangement in space.) From this he derives a method for computing structure from motion. 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. Ullman has also devised a set of simple algorithms by which the method may be implemented.

The $2\frac{1}{2}$-D sketch

Both of the techniques of image analysis discussed in the preceding paragraphs provide information about the relative distances to various places in an image. In stereopsis, it is the matching of points in a stereo pair that leads to such information. In structure from motion, it is the matching of points in successive images. More generally, however, we know that vision provides several sources of information about shapes in the visual world. The most direct, perhaps, are the aforementioned stereo and motion, but texture gradients in a single image are nearly as effective. Furthermore, the theatrical techniques of facial make-up reveal the sensitivity of perceived shapes to shading (see Horn 1975), and colour sometimes suggests the manner in which a surface reflects light. It often happens that different parts of a scene are open to inspection by different techniques. Yet different as the techniques are, they all have two important characteristics in common: they rely on information from the image rather than *a priori* knowledge about the shapes of the viewed objects, and the information that they specify concerns the depth or surface orientation at arbitrary points in an image, rather than the depth of orientation associated with particular objects (see table 1).

Table 1. The form in which various early visual processes deliver information about the changes in a scene

(r is depth; δr is small, local change in depth; Δr is large changes in depth; $\boldsymbol{s}$ is local surface orientation.)

information source	natural parameter
stereo	disparity, hence especially δr and Δr
motion	r, hence δr, Δr
shading	$\boldsymbol{s}$
texture gradients	$\boldsymbol{s}$
perspective cues	$\boldsymbol{s}$
occlusion	Δr
contour	$\boldsymbol{s}$

To make the most efficient use of different and often complementary channels of information deriving from stereopsis, from motion, from contours, from texture, from colour, from shading, they need to be combined in some way. The computational question that now arises is thus how best to do this, and the natural answer is to seek some representation of the visual scene that makes explicit just the information that these processes can deliver. We seek, in other words, a representation of surfaces in an image that makes explicit their shapes and orientations, much as the arabic rerpesentation of a number makes explicit its composition by powers of ten. It might be contrasted with the representation of a surface as a mathematical expression, in which the orientation is only implicit, and not at all apparent. We call such a representation the $2\frac{1}{2}$D sketch (Marr & Nishihara 1978; Marr 1978), and in the particular candidate for it shown in figure 6, surface orientation is 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 the point, and increasing lengths denote surfaces that dip increasingly away from the viewer. The orientation of each needle defines the local direction of dip.

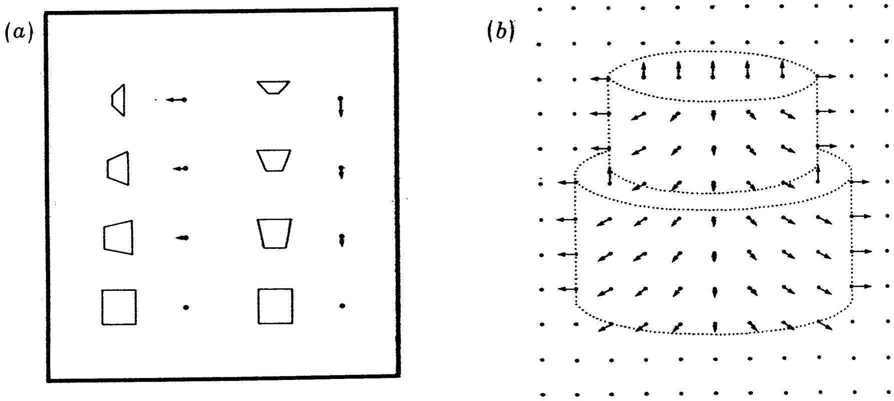

FIGURE 6. Illustration of the $2\frac{1}{2}$-dimensional sketch. In (*a*), the perspective views of small squares placed at various orientations to the viewer are shown. The dots with arrows show a way of representing the orientations of such surfaces symbolically. In (*b*), this representation is used to show the surface orientations of two cylindrical surfaces in front of a background orthogonal to the viewer. The full $2\frac{1}{2}$-dimensional sketch would include rough distances to the surfaces as well as their orientations, contours where surface orientation changes sharply, and contours where depth is discontinuous (subjective contours). A considerable amount of computation is required to maintain these quantities in states that are consistent with one another and with the structure of the outside world (see Marr 1978, §3). (From Marr & Nishihara 1978, figure 2.)

Our argument is that the $2\frac{1}{2}$-D sketch is useful because it makes explicit information about the image in a form that is closely matched to what image analysis can deliver. To put it another way, we can formulate the goals of this stage of 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 $2\frac{1}{2}$-D sketch, and which contours are missing in the primal sketch and need to be inserted into the $2\frac{1}{2}$-D sketch to bring it into a state that is consistent with the nature of three-dimensional space. This formulation avoids the difficulties associated with the terms 'region' and 'object' – the difficulties inherent in the image segmentation approach; for the grey level intensity array, the primal sketch, the various modules of early visual processing, and finally the $2\frac{1}{2}$-D sketch itself deal only with discovering the properties of *surfaces* in an image. One is pleased about that, for we know of ourselves as perceivers that surface orientation can be associated with unfamiliar shapes, so its representation probably precedes the decomposition of the scene into objects. One is thus free to ask precise questions about the computational structure of the $2\frac{1}{2}$-D sketch and of processes to create and maintain it. We are currently much occupied with these matters.

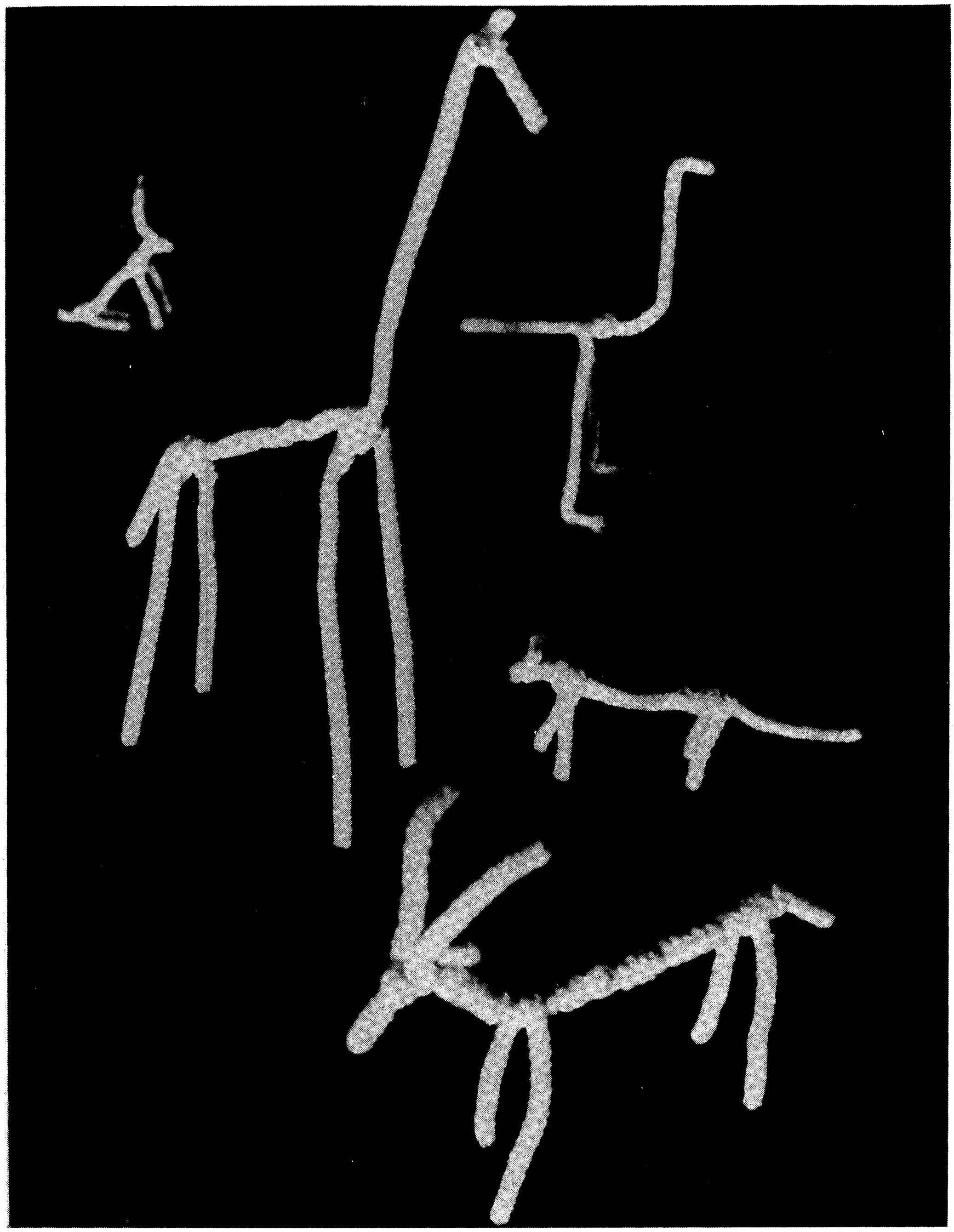

FIGURE 7. The portrayal of animals by a small number of pipe-cleaners serves to show that the representation of a three-dimensional shape need not make explicit its surface to describe it so well that it can easily be recognized. The success of the representation is due, one suspects, in large measure to the correspondence between the pipe-cleaners and the axes of the volumes that they stand for. (From Marr & Nishihara 1978, figure 1.)

Later processing problems

The final components of our visual processing theory concern the application of visually derived surface information for the representation of three-dimensional shapes in a way that is suitable specifically for recognition (Marr & Nishihara 1978). By this we mean the ability to recognize a shape as being the same as a shape seen earlier, and this in essence depends on being able to describe shapes consistently each time they are seen, whatever the circumstances of their positions relative to the viewer. The problem with local surface representations such as the $2\frac{1}{2}$-D sketch is that the description depends as much on the viewpoint of the observer as it does on the structure of the shape. In order to factor out a description of a shape that depends on its structure alone, the representation must be based on readily identifible geometric features of the overall shape, and the dispositions of these features must de specified relative to the shape in itself. In brief, the coordinate system must be 'object-centred', not 'viewer-centred'. One aspect of this deals with the nature of the representation scheme that is to be used, and another with how to obtain it from the $2\frac{1}{2}$-D sketch. We begin by discussing the first, and will then move on to the second.

The 3-D model representation

The most basic geometric properties of the volume occupied by a shape are (1) its average location (or centre of mass), (2) its overall size, as exemplified, for example, by its mean diameter or volume, and (3) its principal axis of elongation or symmetry, if one exists. A description based on these qualities would certainly be inadequate for an application such as shape recognition; after all, one can tell little about the three-dimensional structure of a shape given only its position, size and orientation. But if a shape itself has a natural decomposition into components that can be so described, this volumetric scheme is an effective means for describing the relative spatial arrangment of those components. The illustration of figure 7 shows a familiar version of this type of description, the stick figure (see Blum 1973). The recognizability of the animal shapes depicted in the illustration is surprising considering the simplicity of representation used to describe them.

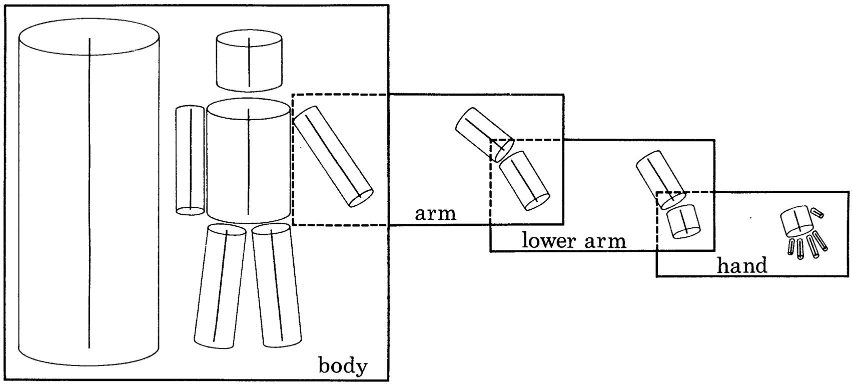

FIGURE 8. The arrangement of 3-D models into the representation of a human shape. First the overall form – the 'body' – is given an axis. This yields an object-centred coordinate system which can then be used to specify the arrangement of the 'arms', 'legs', 'torso' and 'head'. The position of each of these is specified by an axe is of its own, which in turn serves to define a coordinate system for specifying the arrangement of further subsidiary parts. This gives us a hierarchy of 3-D models, shown here extending downwards as far as the fingers. The shapes in the figure are drawn as if they were cylindrical, but that is purely for illustrative convenience. (From Marr & Nishihara 1978, figure 3.)

The reason that such a description works so well lies, we think, in (1) the volumetric (as opposed to surface-based) definition of the primitive elements – the sticks – used by the representation, (2) the relatively small number of elements used, and (3) the relation of elements to each other rather than to the viewer. In short, this type of shape representation is volumetric, modular, and can be based on object-centred coordinates. Figure 8 illustrates the scheme of representation that was developed from these ideas. Here the description of a shape is composed of a hierarchy of stick-figure specifications that we call 3-D models. In the simplest, a single axis element is used to specify the location, size and orientation of the entire shape; the human body displayed in the illustration will serve as an instance, This element is also used to define a coordinate system that will specify the dispositions of subsidiary axes, each of these specifying in turn a coordinate system for 3-D models of 'arm', 'hand', and so on. This hierarchical structure makes it possible to treat any component of a shape as a shape in itself. It also provides flexibility in the detail of a description.

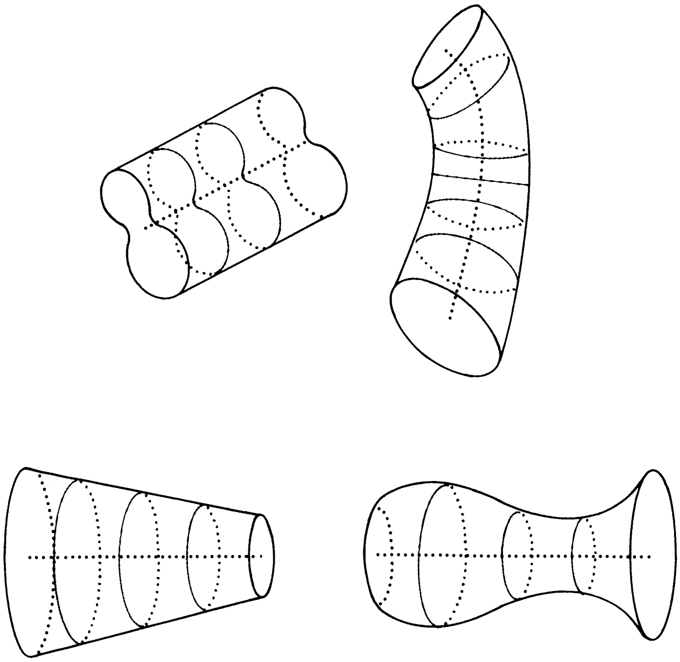

FIGURE 9. The definition of a generalized cone. It is the surface created by moving a cross-section along a given smooth axis. The cross section may vary smoothly in size, but its shape remains constant. Several examples are shown here. In each, the cross section is shown at several positions along the trajectory that spins out the construction.

Shapes admitting 3-D model descriptions

If the scheme for a given shape is to be uniquely defined and stable over unimportant variations such as viewpoint – if, in a word it is to be canonical – its definition must take advantage of any salient geometrical characteristics that the shape possesses inherently. If a shape has natural axes, then those 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, a square, or a circular disk, 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 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, but this is acceptable provided that their number is small. For example, there are four possible axes on which one might wish to base the coordinate system for representing a door, namely the midlines along its length, its width, and its thickness, and also the axis of its hinges. (This last would be especially useful to represent how the door opens.) For a typewriter, there are two reasonable choices, an axis parallel to its width, because that is usually its largest dimension, and the axis about which a typewriter is roughly symmetrical.

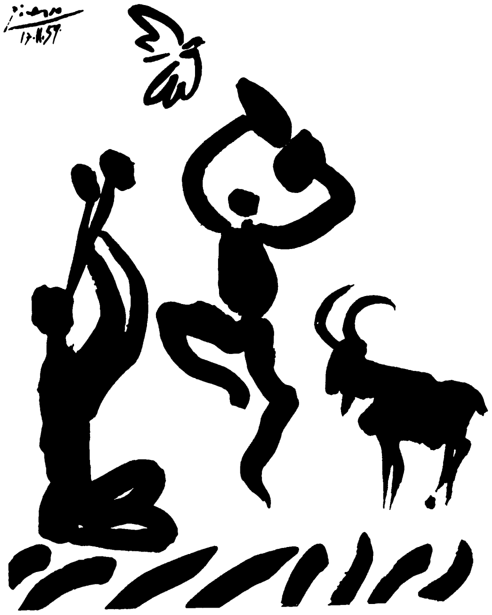

FIGURE 10. 'Rites of Spring' by Pablo Picasso. We immediately interpret such silhouettes in terms of particular three-dimensional surfaces – this despite the paucity of information in the image itself. In order to do this, we plainly must invoke certain *a priori* assumptions and constraints about the nature of the shapes.

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-centred coordinates is therefore to examine the class of shapes having an axis as an integral part of their structure. Consider, accordingly, the class of so-called *generalized cones*, each of these being the surface swept out by moving a cross section of constant shape but smoothly varying size along an axis, as shown in figure 9. Binford (1971) has drawn attention to this class of constructions, suggesting that it might provide a convenient way of describing three-dimensional surfaces for the purposes of computer vision (see also Agin 1972; Nevatia 1974). We regard it as an important class not because the shapes themselves are easily describable, 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 growth, are described quite naturally in terms of one or more generalized cones. The animal shapes of figure 7 provide some examples; the individual sticks are simply the axes of generalized cones that approximate the shapes of parts of these creatures.

Many artefacts can also be described in this way – say a car (a small box sitting atop a longer one) or a building (a box with a vertical axis.)

It is important to remember, however, that there exist surfaces that cannot conveniently be approximated by generalized cones, for example a cake that has been transected at some arbitrary plane, or the surface formed by a crumpled newspaper. Cases like the cake could be dealt with by introducing a suitable surface primitive for describing the plane of the cut, in much the same way as an axis in the 3-D model representation is a primitive that describes a volumetric element. But the crumpled newspaper poses apparently intractable problems.

Finding the natural coordinate system

Even if a shape possesses a canonical coordinate frame, one still is faced with the problem of finding it from an image. Our own interest in this problem grew from the question of how to interpret the *outlines* of objects as seen in a two-dimensional image (Marr 1977*b*), and our starting point was the observation that when one looks at the silhouettes in Picasso's 'Rites of Spring' (reproduced here in figure 10), 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 three-dimensional shapes which, from other viewpoints, would have no descernible 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

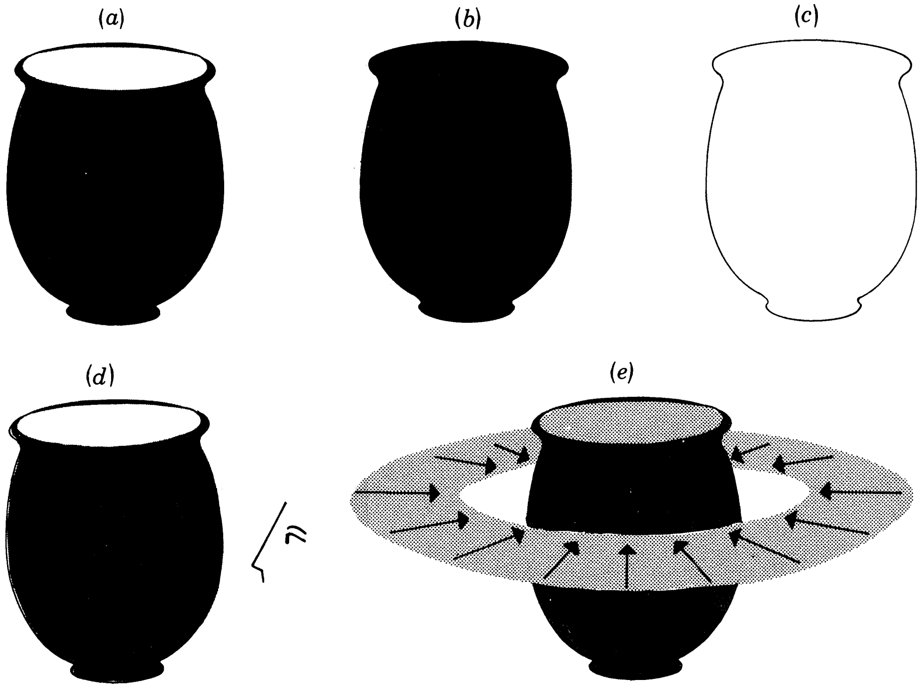

Figure 11. Four structures of importance in studying the *a priori* conditions mentioned in figure 10. (*a*) A three-dimensional surface Σ; (*b*) its silhouette, S_V, as seen from viewpoint V; (*c*) the contour C_V of S_V; (*d*) the set of points Γ_V on Σ that project onto the contour. Finally, (*e*) illustrates schematically the meaning of the phrase 'all distant viewing directions that lie in a plane'.

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, then, is that the bounding contours in Picasso's 'Rites' apparently tell us more than they should about the shape of the figures. For example, neighbouring 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 is that the contours that bound these silhouettes are contours of surface discontinuity, which are precisely the contours with which the $2\frac{1}{2}$-D sketch is concerned. Secondly, because we can interpret the silhouettes as three-dimensional shapes, then 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 that we assign to the contours will differ from the shape that actually caused them. An everyday example is the shadowgraph, where the appropriate arrangement of one's hands can, to the surprise and delight of a child, produce the shadow of a duck or a rabbit.

What assumptions is it reasonable to suppose that we make? To explain them, we need to define the four constructions that appear in figure 11. These are (1) a three-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 the *contour generator* of C_V, and we shall denote it by Γ_V.

Observe that the contour C_V, like the contours in the work of Picasso, imparts very little information about the three-dimensional surface that caused it. Indeed, the only obvious feature available in the contour is the distinction between convex and concave places – that is to say, the presence of inflexion points. In order that these inflections be 'reliable', one needs to make some assumptions about the way in which the contour was generated, and we choose the following restrictions (Marr 1977).

(1) Each point on the contour generator Γ_V projects to a different point on the contour C_V.

(2) Nearby points on the contour C_V arise from nearby points on the contour generator Γ_V.

(3) The contour generator Γ_V lies wholly in a single plane.

The first and second restrictions say that each point on the contour of 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 third restriction is simply the demand that the difference between convex and concave contour segments reflects properties of the surface, rather than of the imaging process.

It turns out to be a theorem that if the surface is smooth (for our purposes, if it is twice differentiable with continuous second derivative) and if restrictions 1–3 hold for all distant viewing positions in any one plane (as illustrated in figure 11), then the viewed surface is a generalized cone with straight axis. (The converse is also true: if the surface is a generalized cone with straight axis, then conditions 1–3 will be found to be true.)

This means that if the convexities and concavities of a bounding contour in an image are actual properties of a surface, then that surface is a generalized cone or is composed of several such cones. In brief, the theorem says that a natural link exists betwen generalized cones and the imaging process itself. The combination of these two must mean, we think, that generalized cones will play an initimate role in the development of vision theory.

Discussion

I have tried in this survey of visual information processing to make two principal points. The first is methodological: namely, that it is important to be very clear about the nature of the understanding that we seek. The results that we try to achieve should be precise ones, at the level of what we call a computational theory. The critical act in formulating computational theories turns out to be the discovery of valid constraints on the way the world is structured – constraints that provide sufficient information to allow the processing to succeed. Consider stereopsis, which presupposes continuity and uniqueness in the world, or structure from visual motion, which presupposes rigidity, or shape from contour, which presupposes the three restrictions just discussed, or even edge detection, which presupposes the assumption of spatial coincidence. The discovery of constraints that are valid and universal leads to results about vision that have the same quality of permanence as results in other branches of science.

The second point is that the critical issues for vision seem to me to revolve around the nature of the representations and the nature of the processes that create, maintain and eventually interpret them. I have suggested an overall framework for visual information processing (summarized in table 2) that includes three categories of representation upon which the processing is to operate. The first encompasses representations of intensity variations and their local geometry in the input to the visual system. One among these, the primal sketch, is expressly intended to be an efficient description of these variations which captures just that information required by the image analysis to follow. The second category encompasses the representations of visible surfaces – the descriptions, in other words, of the physical properties of the surfaces that caused the images in the first place. The nature of these representations, the $2\frac{1}{2}$-dimensional sketch in particular, is determined primarily by what information can be extracted by modules of image analysis such as stereopsis and structure from motion. Like the primal sketch of the previous category, the $2\frac{1}{2}$-dimensional sketch is intended to be a final or output representation: this is where the separate contributions from the various image-analysis modules can be combined into a unified description. The third category encompasses all representations that are subsequently constructed from information contained in the $2\frac{1}{2}$-D sketch. The designs of these tertiary representations are determined largely by the use to which they are to be put, as for the 3-D model representation, to be used for shape recognition. If one had wanted instead, for example, to represent a shape simply for later *reproduction*, say by the milling of a block of metal, then the $2\frac{1}{2}$-D sketch would itself have been sufficient, as the milling process depends explicitly on information about local depth and orientation, such as that sketch can provide.

Table 2. A framework for the derivation of shape information from images

image(s)	
↓	
primal sketch(es)	Describes the intensity changes present in an image, labels distinguished locations like termination points, and makes explicit local two-dimensional geometrical relations
↓	
$2\frac{1}{2}$-D sketch	Represents contours of surface discontinuity, and depth and orientation of visible surface elements, in a coordinate frame that is centred on the viewer
↓	
3-D model representation	Shape description that includes volumetric shape primitives of a variety of sizes, whose positions are defined by using an object-centred coordinate system. This representation imposes considerable modular organization on its descriptions

Finally, a remark of a rather different nature. As we have seen, some aspects of human early visual processing, like stereopsis, have apparently been understood well enough to implement them in machines (Marr & Poggio 1979; Marr & Grimson 1979). The computational power required by these early processes is prohibitive, and until recently the prospects for real-time implementation of human-like early vision were remote. It now appears, however, that the emerging VLSI and CCD technologies will be able to supply the necessary processing power. This could make the next two decades very interesting.

I thank K. Nishihara and M. Feiertag for help in assembling this survey. I thank the Royal Society for permission to reproduce figures 3, 4, 6, 7 and 8, and the M.I.T. Press for figure 5. This work was conducted at the Artificial Intelligence Laboratory, a Massachusetts Institute of Technology research programme supported in part by the Advanced Research Projects Agency of the Department of Defense, and monitored by the Office of Naval Research under contract number N00014-75-C-0643, and in part by NSF contract number 77-07569-MCS.

References (Marr)

Agin, G. J. 1972 Representation and description of curved objects. *Stanford Artificial Intelligence Project*, memo AIM-173. Stanford University.

Binford, T. O. 1971 Visual perception by computer. Presented to the I.E.E.E. Conference on Systems and Control, Miami, December.

Blum, H. 1973 Biological shape and visual science, part 1. *J. theor. Biol.* **38**, 205–287.

Enroth-Cugell, C. & Robson, J. D. 1966 The contrast sensitivity of retinal ganglion cells of the cat. *J. Physiol., Lond.* **187**, 517–522.

Freuder, E. C. 1975 A computer vision system for visual recognition using active knowledge. *M.I.T. A.I. Lab. Tech. Rep.* no. 345.

Helmholtz, H. L. F. von 1910 *Treatise on physiological optics* (trans. J. P. Southall, 1925). N.Y.: Dover.

Horn, B. K. P. 1975 Obtaining shape from shading information. In *The psychology of computer vision* (ed. P. H. Winston), pp. 115–155. New York: McGraw-Hill.

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.

Hubel, D. H. & Wiesel, T. N. 1968 Receptive fields and functional architecture of monkey striate cortex. *J. Physiol., Lond.* **195**, 215–243.

Julesz, B. 1971 *Foundations of cyclopean perception.* Chicago: University of Chicago Press.

Kuffler, S. W. 1953 Discharge patterns and functional organization of mammalian retina. *J. Neurophysiol.* **16**, 37–68.

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

Marr, D. 1977*a* Artificial intelligence – a personal view. *Artificial Intelligence* **9**, 37–48.

Marr, D. 1977*b* Analysis of occluding contour. *Proc. R. Soc. Lond.* B **197**, 441–475.

Marr, D. 1978 Representing visual information. *Lectures on mathematics in the life sciences*, volume 10: *Some mathematical Questions in Biology*, pp. 101–180.

Marr, D. & Hildreth, E. 1979 Theory of edge detection. *Proc. R. Soc. Lond.* B. (In the press.)

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

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

Marr, D. & Poggio, T. 1977 From understanding computation to understanding neural circuitry. *Neurosci. Res. Prog. Bull.* **15**, 470–488.

Marr, D. & Poggio, T. 1979 A computational theory of human stereo vision. *Proc. R. Soc. Lond.* B **204**, 301–328.

Marr, D., Poggio, T. & Palm, G. 1977 Analysis of a cooperative stereo algorithm. *Biol. Cybernet.* **28**, 223–239.

Marr, D., Poggio, T. & Ullman, S. 1979 Bandpass channels, zero-crossings and early visual information processing. *J. opt. Soc. Am.* **69**, 914–916.

Marr, D. & Ullman, S. 1979 Directional selectivity and its use in early visual processing. (In preparation.)

Nevatia, R. 1974 Structured descriptions of complex curved objects for recognition and visual memory. *Stanford Artificial Intelligence Project*, memo AIM-250. Stanford University.

Newton, I. 1704 *Optics*. London.

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

Stevens, K. A. 1978 Computation of locally parallel structure. *Biol. Cybernet.* **29**, 19–28.

Tenenbaum, J. M. & Barrow, H. G. 1976 Experiments in interpretation-guided segmentation. *Stanford Res. Inst. Tech. Note* no. 123.

Ullman, S. 1979*a* The interpretation of structure from motion. *Proc. R. Soc. Lond.* B **203**, 405–426.

Ullman, S. 1979*b* *The interpretation of visual motion.* M.I.T. Press.

Wallach, H. & O'Connell, D. N. 1953 The kinetic depth effect. *J. exp. Psychol.* **45**, 205–217.

Wertheimer, M. 1938 Principles of perceptual organizations. In *Source book of Gestalt psychology* (ed. W. H. Ellis), pp. 71–88. New York: Routledge Kegan Paul.

Discussion

S. Lal (*Department of Physiology, Chelsea College, London SW*3 6*LX, U.K.*). There seem to me to be two objections to the theory of early visual processing presented. The first relates to the well-known result that any differentiation process will magnify noise, i.e. it will decrease the signal: noise ratio. Since we know that random fluctuations of neural signals are common, it would be most disadvantageous (at least at low signal levels) to use differentiating operations to extract information about a visual scene. The second objection relates to the impossibility of deducing from local features alone the global characteristics of the intensity function. Yet some global description and analysis is a prerequisite for extracting relevant information from a visual scene.

D. Marr.

1. Noise is probably not so important under conditions of normal vision (to which my remarks apply). In low illumination after sufficient adaptation, of course, retinal receptive field properties do change – away from acting as differentiators and towards acting as photon collectors.

2. The general idea is that one starts with rather local descriptive elements and gradually computes more global constructs, but things can go the other way. In stereopsis, for example, the coarse results obtained from matching the larger channels are then used to control disjunctive eye-movements, which bring signals from the smaller channels into the 'relevant' disparity range (see Marr & Poggio 1979).

H. B. Barlow, F.R.S. (*Physiological Laboratory, Cambridge CB*2 3*EG, U.K.*). I am afraid that this is more commentary than question. For many years the kind of problems of early processing that Marr has been talking about were thought, at least by some, to be trivially easy. According to this view the problem of finding the position of an edge was not the sort of thing to which one should devote valuable computer time, nor the even more valuable artificial intelligencer's intelligence. That view was wrong, for it is a difficult task, and it is a crucial first step for further progress to understanding the nature of these difficulties. It is crucial not only to improve edge-finders in programmes, but even more to enable physiologists to understand neural mechanisms. I am sure many of us must come away from a talk like this with new ideas about what to look for when recording from cortical neurons and analysing how they achieve their selectivity.

Understanding the nature of the difficulties is also necessary when one wants to test whether a computer method of doing something is what really happens. The natural test to apply is to measure *how well* the computer method overcomes the difficulties and compare this with the human observer's performance under various conditions. This is what I was trying to do with symmetry detection, where the first step was to identify (I hope correctly) the natural difficulty in symmetry detection as the occurrence of spurious symmetry. The point here is that one can only begin to make such comparisons when the natural difficulty has been identified, and it seems to be the artificial intelligence approach that leads most directly to this information.

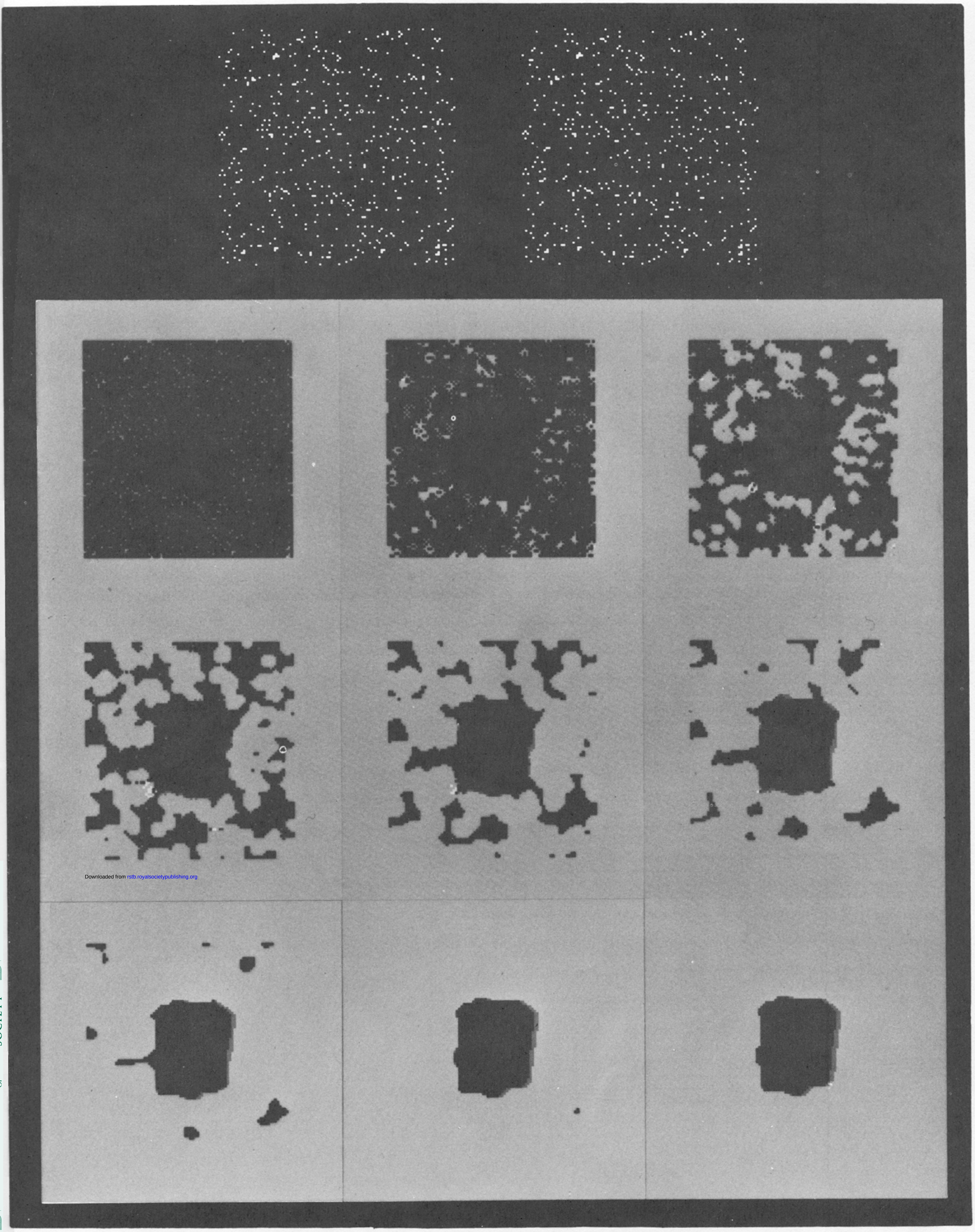

FIGURE 1. A sparse random-dot stereogram (the top two images), and its decoding by Marr & Poggio's (1976) cooperative algorithm. The initial state contains all possible matches within a given disparity range, and the algorithm embodies the constraints of uniqueness and continuity to eliminate false targets. Shades of grey are used to signify matches at different disparities. The figure shows the initial state, and the states after 1, 2, 3, 4, 5, 6, 8 and 14 iterations. The algorithm progressively reveals a square hovering in depth. This algorithm is not the one used by the human visual system.

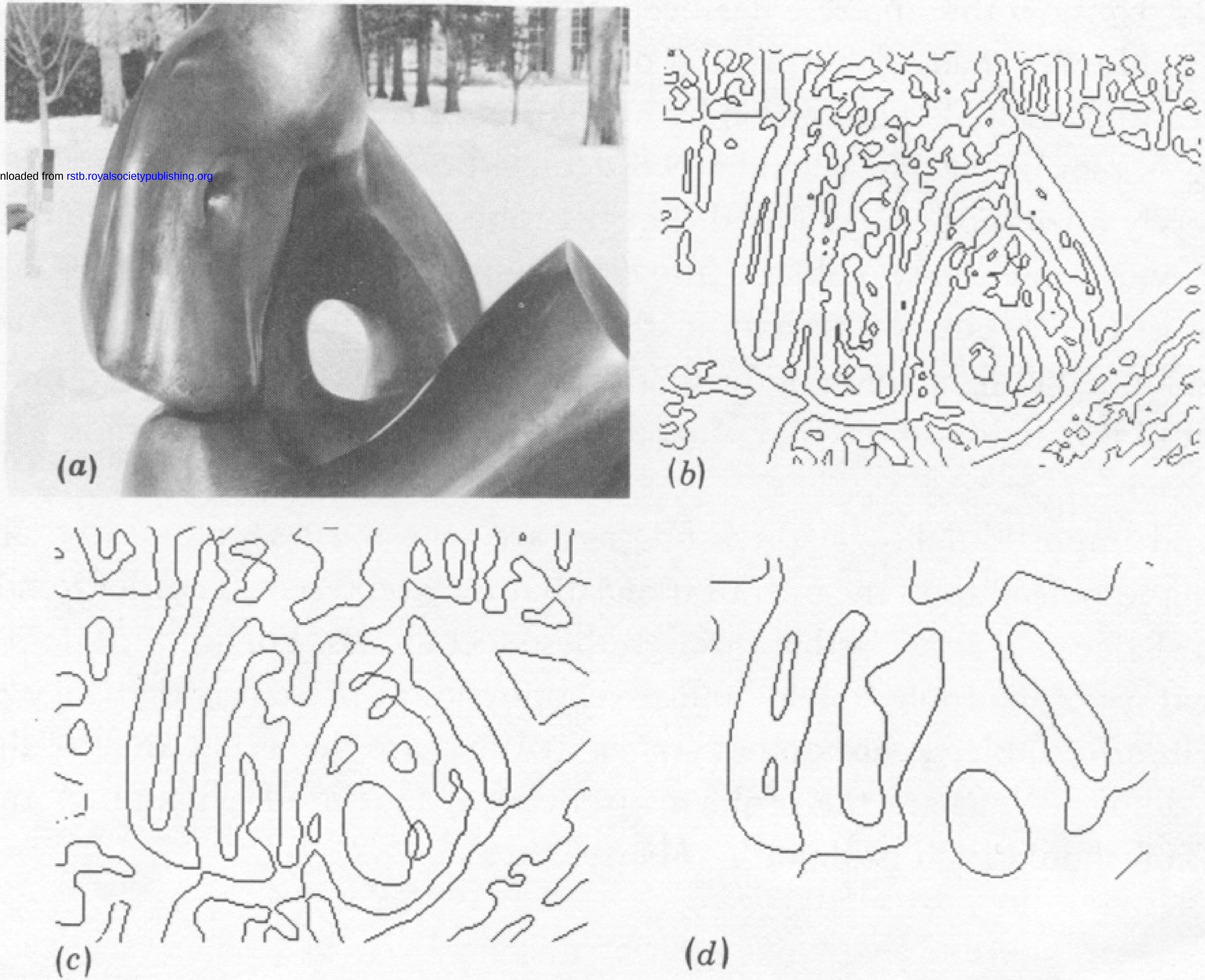

FIGURE 3. The image (*a*), which is 320 × 320 pixels, has been convolved with $\nabla^2 G$, a centre–surround operator with central excitatory region of width $2\sigma = 6$, 12 and 24 pixels. These filters span approximately the range of filters that operate in the human fovea. The zero-crossings of the filtered images are shown in (*b*), (*c*) and (*d*). These are the precursors of the raw primal sketch. (From Marr & Hildreth 1979, figure 6).

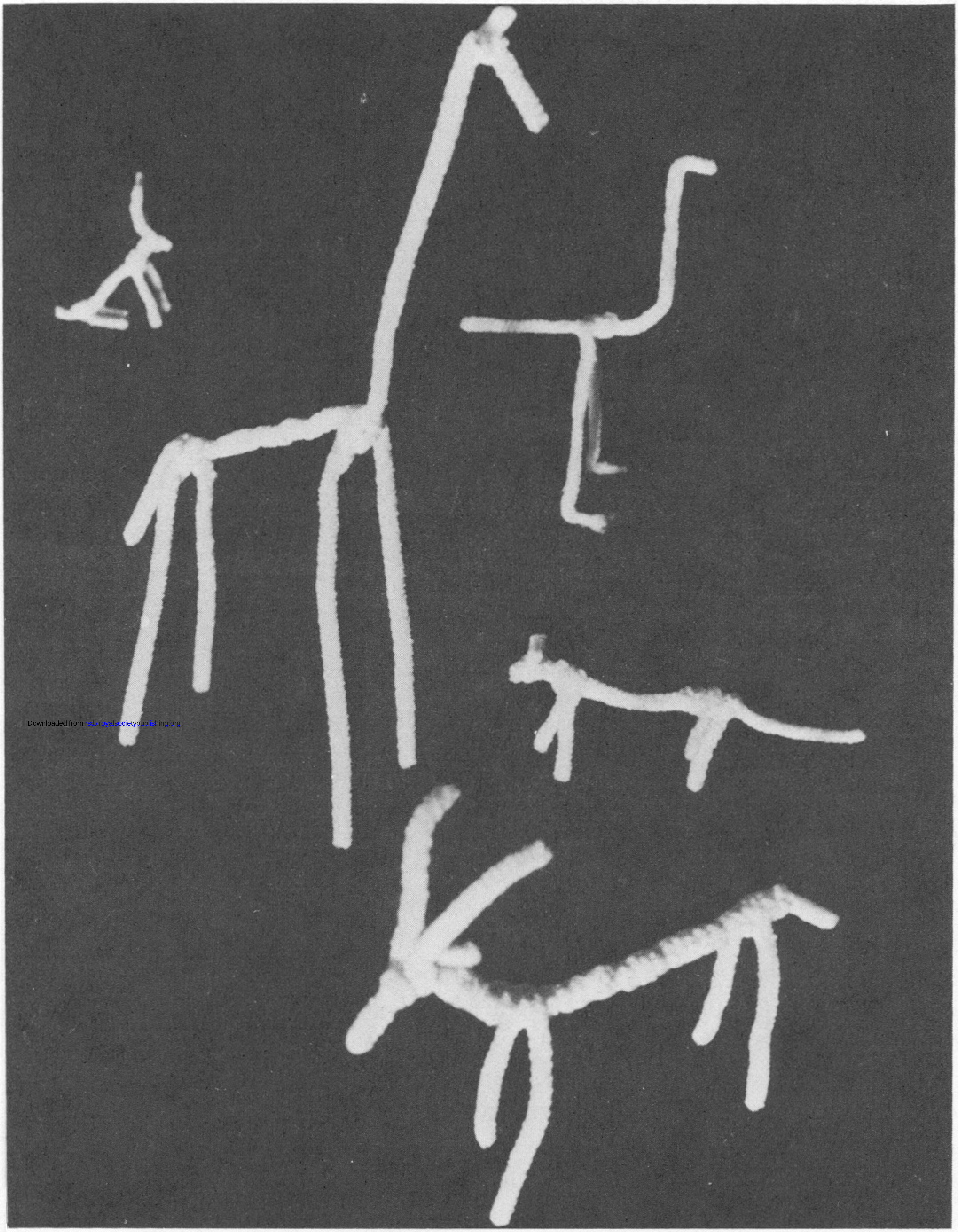

FIGURE 7. The portrayal of animals by a small number of pipe-cleaners serves to show that the representation of a three-dimensional shape need not make explicit its surface to describe it so well that it can easily be recognized. The success of the representation is due, one suspects, in large measure to the correspondence between the pipe-cleaners and the axes of the volumes that they stand for. (From Marr & Nishihara 1978, figure 1.)