Visual cognition: An introduction*

STEVEN PINKER
Massachusetts Institute of Technology

Abstract

This article is a tutorial overview of a sample of central issues in visual cognition, focusing on the recognition of shapes and the representation of objects and spatial relations in perception and imagery. Brief reviews of the state of the art are presented, followed by more extensive presentations of contemporary theories, findings, and open issues. I discuss various theories of shape recognition, such as template, feature, Fourier, structural description, Marr–Nishihara, and massively parallel models, and issues such as the reference frames, primitives, top-down processing, and computational architectures used in spatial cognition. This is followed by a discussion of mental imagery, including conceptual issues in imagery research, theories of imagery, imagery and perception, image transformations, computational complexities of image processing, neuropsychological issues, and possible functions of imagery. Connections between theories of recognition and of imagery, and the relevance of the papers contained in this issue to the topics discussed, are emphasized throughout.

Recognizing and reasoning about the visual environment is something that people do extraordinarily well; it is often said that in these abilities an average three-year old makes the most sophisticated computer vision system look embarrassingly inept. Our hominid ancestors fabricated and used tools for millions of years before our species emerged, and the selection pressures brought about by tool use may have resulted in the development of sophisticated faculties allowing us to recognize objects and their physical properties, to bring complex knowledge to bear on familiar objects and scenes, to

*Preparation of this paper was supported by NSF grants BNS 82-16546 and 82-09540, by NIH grant 1R01HD18381-01, and by a grant from the Sloan Foundation awarded to the MIT Center for Cognitive Science. I thank Donald Hoffman, Stephen Kosslyn, Jacques Mehler, Larry Parsons, Whitman Richards, and Ed Smith for their detailed comments on an earlier draft, and Kathleen Murphy and Rosemary Krawczyk for assistance in preparing the manuscript. Reprint requests should be sent to Steven Pinker, Psychology Department, M.I.T., E10-018, Cambridge, MA 02139, U.S.A.

0010-0277/84/$19.40 Elsevier Sequoia/Printed in The Netherlands

negotiate environments skillfully, and to reason about the possible physical interactions among objects present and absent. Thus visual cognition, no less than language or logic, may be a talent that is central to our understanding of human intelligence (Jackendoff, 1983; Johnson-Laird, 1983; Shepard and Cooper, 1982).

Within the last 10 years there has been a great increase in our understanding of visual cognitive abilities. We have seen not only new empirical demonstrations, but also genuinely new theoretical proposals and a new degree of explicitness and sophistication brought about by the use of computational modeling of visual and memory processes. Visual cognition, however, occupies a curious place within cognitive psychology and within the cognitive psychology curriculum. Virtually without exception, the material on shape recognition found in introductory textbooks in cognitive psychology would be entirely familiar to a researcher or graduate student of 20 or 25 years ago. Moreover, the theoretical discussions of visual imagery are cast in the same loose metaphorical vocabulary that had earned the concept a bad name in psychology and philosophy for much of this century. I also have the impression that much of the writing pertaining to visual cognition among researchers who are not directly in this area, for example, in neuropsychology, in 'ividual differences research, developmental psychology, psychophysics, and information processing psychology, is informed by the somewhat antiquated and imprecise discussions of visual cognition found in the textbooks.

The purpose of this special issue of *Cognition* is to highlight a sample of theoretical and empirical work that is on the cutting edge of research on visual cognition. The papers in this issue, though by no means a representative sample, illustrate some of the questions, techniques, and types of theory that characterize the modern study of visual cognition. The purpose of this introductory paper is to introduce students and researchers in neighboring disciplines to a selection of issues and theories in the study of visual cognition that provide a backdrop to the particular papers contained herein. It is meant to bridge the gap between the discussions of visual cognition found in textbooks and the level of discussion found in contemporary work.

Visual cognition can be conveniently divided into two subtopics. The first is the representation of information concerning the visual world currently before a person. When we behave in certain ways or change our knowledge about the world in response to visual input, what guides our behavior or thought is rarely some simple physical property of the input such as overall brightness or contrast. Rather, vision guides us because it lets us know that we are in the presence of a particular configuration of three-dimensional shapes and particular objects and scenes that we know to have predictable properties. 'Visual recognition' is the process that allows us to determine on the basis of retinal input that particular shapes, configurations of shapes, objects, scenes, and their properties are before us.

The second subtopic is the process of remembering or reasoning about shapes or objects that are not currently before us but must be retrieved from memory or constructed from a description. This is usually associated with the topic of 'visual imagery'. This tutorial paper is divided into two major sections, devoted to the representation and recognition of shape, and to visual imagery. Each section is in turn subdivided into sections discussing the background to each topic, some theories on the relevant processes, and some of the more important open issues that will be foci of research during the coming years.

Visual recognition

Shape recognition is a difficult problem because the immediate input to the visual system (the spatial distribution of intensity and wavelength across the retinas—hereafter, the "retinal array") is related to particular objects in highly variable ways. The retinal image projected by an object—say, a notebook—is displaced, dilated or contracted, or rotated on the retina when we move our eyes, ourselves, or the book; if the motion has a component in depth, then the retinal shape of the image changes and parts disappear and emerge as well. If we are not focusing on the book or looking directly at it, the edges of the retinal image become blurred and many of its finer details are lost. If the book is in a complex visual context, parts may be occluded, and the edges of the book may not be physically distinguishable from the edges and surface details of surrounding objects, nor from the scratches, surface markings, shadows, and reflections on the book itself.

Most theories of shape recognition deal with the indirect and ambiguous mapping between object and retinal image in the following way. In long-term memory there is a set of representations of objects that have associated with them information about their shapes. The information does not consist of a replica of a pattern of retinal stimulation, but a canonical representation of the object's shape that captures some invariant properties of the object in all its guises. During recognition, the retinal image is converted into the same format as is used in long-term memory, and the memory representation that matches the input the closest is selected. Different theories of shape recognition make different assumptions about the long-term *memory representations* involved, in particular, how many representations a single object will have, which class of objects will be mapped onto a single representation, and what the format of the representation is (i.e. which primitive symbols can be found in a representation, and what kinds of relations among them can be specified). They will differ in regards to which sports of *preprocessing* are done to the retinal image (e.g., filtering, contrast enhancement, detection of edges) prior to matching, and in terms of how the retinal input or memory representations are *transformed* to bring them into closer correspondence. And they differ in terms of the metric of *goodness of fit* that determines which memory representation fits the input best when none of them fits it exactly.

Traditional theories of shape recognition

Cognitive psychology textbooks almost invariably describe the same three or so models in their chapters on pattern recognition. Each of these models is fundamentally inadequate. However, they are not always inadequate in the ways the textbooks describe, and at times they *are* inadequate in ways that the textbooks do not point out. An excellent introduction to three of these models—templates, features, and structural descriptions—can be found in Lindsay and Norman (1977); introductions to Fourier analysis in vision, which forms the basis of the fourth model, can be found in Cornsweet (1980) and Weisstein (1980). In this section I will review these models extremely briefly, and concentrate on exactly why they do not work, because a catalogue of their deficits sets the stage for a discussion of contemporary theories and issues in shape recognition.

Template matching

This is the simplest class of models for pattern recognition. The long term memory representation of a shape is a replica of a pattern of retinal stimulation projected by that shape. The input array would be simultaneously superimposed with all the templates in memory, and the one with the closest above-threshold match (e.g., the largest ratio of matching to nonmatching points in corresponding locations in the input array) would indicate the pattern that is present.

Usually this model is presented not as a serious theory of shape recognition, but as a straw man whose destruction illustrates the inherent difficulty of the shape recognition process. The problems are legion: partial matches could yield false alarms (e.g., a 'P' in an 'R' template); changes in distance, location, and orientation of a familiar object will cause this model to fail to detect it, as will occlusion of part of the pattern, a depiction of it with wiggly or cross-hatched lines instead of straight ones, strong shadows, and many other distortions that we as perceivers take in stride.

There are, nonetheless, ways of patching template models. For example, multiple templates of a pattern, corresponding to each of its possible displacements, rotations, sizes, and combinations thereof, could be stored. Or, the input pattern could be rotated, displaced, and scaled to a canonical set of values before matching against the templates. The textbooks usually dismiss these possibilities: it is said that the product of all combinations of transformations and shapes would require more templates than the brain could store, and that in advance of recognizing a pattern, one cannot in general determine which transformations should be applied to the input. However, it is easy to show that these dismissals are made too quickly. For example, Arnold Trehub (1977) has devised a neural model of recognition and imagery, based on templates, that addresses these problems (this is an example of a 'massively parallel' model of recognition, a class of models I will return to later). Contour extraction preprocesses feed the matching process with an array of symbols indicating the presence of edges, rather than with a raw array of intensity levels. Each template could be stored in a single cell, rather than in a space-consuming replica of the entire retina: such a cell would synapse with many retinal inputs, and the shape would be encoded in the pattern of strengths of those synapses. The input could be matched in parallel against all the stored memory templates, which would mutually inhibit one another so that partial matches such as 'P' for 'R' would be eliminated by being inhibited by better matches. Simple neural networks could center the input pattern and quickly generate rotated and scaled versions of it at a variety of sizes and orientations, or at a canonical size and orientation (e.g., with the shape's axis of elongation vertical); these transformed patterns could be matched in parallel against the stored templates.

Nonetheless, there are reasons to doubt that even the most sophisticated versions of template models would work when faced with realistic visual inputs. First, it is unlikely that template models can deal adequately with the third dimension. Rotations about any axis other than the line of sight cause distortions in the projected shape of an object that cannot be inverted by any simple operation on retina-like arrays. For example, an arbitrary edge might move a large or a small amount across the array depending on the axis and phase of rotation and the depth from the viewer. 3-D rotation causes some surfaces to disappear entirely and new ones to come into view. These problems occur even if one assumes that the arrays are constructed subsequent to stereopsis and hence are three-dimensional (for example, rear surfaces are still not represented, there are a bewildering number of possible directions of translation and axes of rotation, each requiring a different type of retinal transformation).

Second, template models work only for isolated objects, such as a letter presented at the center of a blank piece of paper: the process would get nowhere if it operated, say, on three-fifths of a book plus a bit of the edge of the table that it is lying on plus the bookmark in the book plus the end of the pencil near it, or other collections of contours that might be found in a circumscribed region of the retina. One could posit some figure–ground segregation preprocess occurring before template matching, but this has problems of its own. Not only would such a process be highly complex (for example, it would have to distinguish intensity changes in the image resulting from differences in depth and material from those resulting from differences in orientation, pigmentation, shadows, surface scratches, and specular (glossy) reflections), but it probably interacts with the recognition process and hence could not precede it. For example, the figure–ground segregation process involves carving up a set of surfaces into parts, each of which can then be matched against stored templates. This process is unlikely to be distinct from the process of carving up a single object into its parts. But as Hoffman and Richards (1984) argue in this issue, a representation of how an object is decomposed into its parts may be the first representation used in accessing memory during recognition, and the subsequent matching of particular parts, template-style or not, may be less important in determining how to classify a shape.

Feature models

This class of models is based on the early "Pandemonium" model of shape recognition (Selfridge, 1959; Selfridge and Neisser, 1960). In these models, there are no templates for entire shapes; rather, there are mini-templates or 'feature detectors' for simple geometric features such as vertical and horizontal lines, curves, angles, 'T'-junctions, etc. There are detectors for every feature at every location in the input array, and these detectors send out a graded signal encoding the degree of match between the target feature and the part of the input array they are 'looking at'. For every feature (e.g., an open curve), the levels of activation of all its detectors across the input array are summed, or the number of occurrences of the feature are counted (see e.g., Lindsay and Norman, 1977), so the output of this first stage is a set of numbers, one for each feature.

The stored representation of a shape consists of a list of the features composing the shape, in the form of a vector of weights for the different features, a list of how many tokens of each feature are present in the shape, or both. For example, the representation of the shape of the letter 'A' might specify high weights for (1) a horizontal segment, (2) right-leaning diagonal segment, (3) a left-leaning diagonal segment, (4) an upward-pointing acute angle, and so on, and low or negative weights for curved and vertical segments. The intent is to use feature weights or counts to give each shape a characterization that is invariant across transformations of it. For example, since the features are all independent of location, any feature specification will be invariant across translations and scale changes; and if features referring to orientation (e.g. "left-leaning diagonal segment") are eliminated, and only features distinguishing straight segments from curves from angles are retained, then the description will be invariant across frontal plane rotations.

The match between input and memory would consist of some comparison of the levels of activation of feature detectors in the input with the weights of the corresponding features in each of the stored shape representations, for example, the product of those two vectors, or the number of matching features minus the number of mismatching features. The shape that exhibits the highest degree of match to the input is the shape recognized.

The principal problem with feature analysis models of recognition is that no one has ever been able to show how a *natural* shape can be defined in terms of a vector of feature weights. Consider how one would define the shape of a horse. Naturally, one could define it by giving high weights to features like 'mane', 'hooves', 'horse's head', and so on, but then detecting these features would be no less difficult than detecting the horse itself. Or, one could try to define the shape in terms of easily detected features such as vertical lines and curved segments, but horses and other natural shapes are composed of so many vertical lines and curved segments (just think of the nose alone, or the patterns in the horse's hide) that it is hard to believe that there is a feature vector for a horse's shape that would consistently beat out feature vectors for other shapes across different views of the horse. One could propose that there is a hierarchy of features, intermediate ones like 'eye' being built out of lower ones like 'line segment' or 'circle', and higher ones like 'head' being built out of intermediate ones like 'eye' and 'ear' (Selfridge, for example, posited "computational demons" that detect Boolean combinations of features), but no one has shown how this can be done for complex natural shapes.

Another, equally serious problem is that in the original feature models the spatial relationships among features—how they are located and oriented with respect to one another—are generally not specified; only which ones are present in a shape and perhaps how many times. This raises serious problems in distinguishing among shapes consisting of the same features arranged in different ways, such as an asymmetrical letter and its mirror image. For the same reason, simple feature models can turn reading into an anagram problem, and can be shown formally to be incapable of detecting certain pattern distinctions such as that between open and closed curves (see Minsky and Papert, 1972).

One of the reasons that these problems are not often raised against feature models is that the models are almost always illustrated and referred to in connection with recognizing letters of the alphabet or schematic line drawings. This can lead to misleading conclusions because the computational problems posed by the recognition of two-dimensional stimuli composed of a small number of one-dimensional segments may be different in kind from the problems posed by the recognition of three-dimensional stimuli composed of a large number of two-dimensional surfaces (e.g., the latter involves compensating for perspective and occlusion across changes in the viewer's vantage point and describing the complex geometry of curved surfaces). Furthermore, when shapes are chosen from a small finite set, it is possible to choose a feature inventory that exploits the minimal contrasts among the particular members of the set and hence successfully discriminates among those members, but that could be fooled by the addition of new members to the set. Finally, letters or line drawings consisting of dark figures presented against a blank background with no other objects occluding or touching them avoids the many difficult problems concerning the effects on edge detection of occlusion, illumination, shadows, and so on.

Fourier models

Kabrisky (1966), Ginsburg (1971, 1973), and Persoon and Fu (1974; see also Ballard and Brown, 1982) have proposed a class of pattern recognition models that that many researchers in psychophysics and visual physiology adopt implicitly as the most likely candidate for shape recognition in humans. In these models, the two-dimensional input intensity array is subjected to a spatial trigonometric Fourier analysis. In such an analysis, the array is decomposed into a set of components, each component specific to a sinusoidal change in intensity along a single orientation at a specific spatial frequency. That is, one component might specify the degree to which the image gets brighter and darker and brighter and darker, etc., at intervals of 3° of visual angle going from top right to bottom left in the image (averaging over changes in brightness along the orthogonal direction). Each component can be conceived of as a grid consisting of parallel black-and-white stripes of a particular width oriented in a particular direction, with the black and white stripes fading gradually into one another. In a full set of such grating-like components, there is one component for each stripe width or spatial frequency (in cycles per degree) at each orientation (more precisely, there would be a continuum of components across frequencies and orientations).

A Fourier transform of the intensity array would consist of two numbers for each of these components. The first number would specify the degree of contrast in the image corresponding to that frequency at that orientation (that is, the degree of difference in brightness between the bright areas and the dark areas of that image for that frequency in that orientation), or, roughly, the degree to which the image 'contains' that set of stripes. The full set of these numbers is the *amplitude spectrum* corresponding to the image. The second number would specify where in the image the peaks and troughs of the intensity change defined by that component lie. The full set of these numbers of the *phase spectrum* corresponding to the image. The amplitude spectrum and the phase spectrum together define the *Fourier transform* of the image, and the transform contains all the information in the original image. (This is a very crude introduction to the complex subject of Fourier analysis. See Weisstein (1980) and Cornsweet (1970) for excellent nontechnical tutorials).

One can then imagine pattern recognition working as follows. In long-term memory, each shape would be stored in terms of its Fourier transform. The Fourier transform of the image would be matched against the long-term memory transforms, and the memory transform with the best fit to the image transform would specify the shape that is recognized.[1]

How does matching transforms differ from matching templates in the original space domain? When there is an exact match between the image and one of the stored templates, there are neither advantages nor disadvantages to doing the match in the transform domain, because no information is lost in the transformation. But when there is no exact match, it is possible to define metrics of goodness of fit in the transform domain that might capture some of the invariances in the family of retinal images corresponding to a shape. For example, to a first approximation the amplitude spectrum corresponding to a shape is the same regardless of where in the visual field the object is located. Therefore if the matching process could focus on the amplitude spectra of shape and input, ignoring the phase spectrum, then a shape could be recognized across all its possible translations. Furthermore, a shape and its mirror image have the same amplitude spectrum, affording recognition of a shape across reflections of it. Changes in orientation and scale of an object result in corresponding changes in orientation and scale in the transform, but in some models the transform can easily be normalized so that it is invariant with rotation and scaling. Periodic patterns and textures, such as a brick wall, are easily recognized because they give rise to peaks in their transforms corresponding to the period of repetition of the pattern. But most important, the Fourier transform segregates information about sharp edges and small

[1] In Persoon and Fu's model (1974), it is not the transform of brightness as a function of visual field position that is computed and matched, but the transform of the tangent angle of the boundary of an object as a function of position along the boundary. This model shares many of the advantages and disadvantages of Fourier analysis of brightness in shape recognition.

details from information about gross overall shape. The latter is specified primarily by the lower spatial-frequency components of the transform (i.e., fat gratings), the former, by the higher spatial-frequency components (i.e. thin gratings). Thus if the pattern matcher could selectively ignore the higher end of the amplitude spectrum when comparing input and memory transforms, a shape could be recognized even if its boundaries are blurred, encrusted with junk, or defined by wiggly lines, dots or dashes, thick bands, and so on. Another advantage of Fourier transforms is that, given certain assumptions about neural hardware, they can be extracted quickly and matched in parallel against all the stored templates (see e.g., Pribram, 1971).

Upon closer examination, however, matching in the transform domain begins to lose some of its appeal. The chief problem is that the invariances listed above hold only for entire scenes or for objects presented in isolation. In a scene with more than one object, minor rearrangements such as moving an object from one end of a desk to another, adding a new object to the desk top, removing a part, or bending the object, can cause drastic changes in the transform. Furthermore the transform cannot be partitioned or selectively processed in such a way that one part of the transform corresponds to one object in the scene, and another part to another object, nor can this be done within the transform of a single object to pick out its parts (see Hoffman and Richards (1984) for arguments that shape representations must explicitly define the decomposition of an object into its parts). The result of these facts is that it is difficult or impossible to recognize familiar objects in novel scenes or backgrounds by matching transforms of the input against transforms of the familiar objects. Furthermore, there is no straightforward way of linking the shape information implicit in the amplitude spectrum with the position information implicit in the phase spectrum so that the perceiver can tell where objects are as well as what they are. Third, changes in the three-dimensional orientation of an object do not result in any simple cancelable change in its transform, even it we assume that the visual system computes three-dimensional transforms (e.g., using components specific to periodic changes in binocular disparity).

The appeal of Fourier analysis in discussions of shape recognition comes in part from the body of elegant psychophysical research (e.g., Campbell and Robson, 1968) suggesting that the visual system partitions the information in the retinal image into a set of channels each specific to a certain range of spatial frequencies (this is equivalent to sending the retinal information through a set of bandpass filters and keeping the outputs of those filters separate). This gives the impression that early visual processing passes on to the shape recognition process not the original array but something like a Fourier transform of the array. However, *filtering* the image according to its spatial frequency components is not the same as *transforming* the image into its spectra. The psychophysical evidence for channels is consistent with the notion that the recognition system operates in the space domain, but rather than processing a single array, it processes a family of arrays, each one containing information about intensity changes over a different scale (or, roughly, each one bandpass-filtered at a different center frequency). By processing several bandpass-filtered images separately, one obtains some of the advantages of Fourier analysis (segregation of gross shape from fine detail) without the disadvantages of processing the Fourier transform itself (i.e. the utter lack of correspondence between the parts of the representation and the parts of the scene).

Structural descriptions

A fourth class of theories about the format in which visual input is matched against memory holds that shapes are represented *symbolically*, as *structural descriptions* (see Minsky, 1975; Palmer, 1975a; Winston, 1975). A structural description is a data structure that can be thought of as a list of propositions whose arguments correspond to parts and whose predicates correspond to properties of the parts and to spatial relationships among them. Often these propositions are depicted as a graph whose nodes correspond to the parts or to properties, and whose edges linking the nodes correspond to the spatial relations (an example of a structural description can be found in the upper left portion of Fig. 6). The explicit representation of spatial relations is one aspect of these models that distinguishes them from feature models and allows them to escape from some of the problems pointed out by Minsky and Papert (1972).

One of the chief advantages of structural descriptions is that they can factor apart the information in a scene without necessarily losing information in it. It is not sufficient for the recognition system simply to supply a list of labels for the objects that are recognized, for we need to know not only what things are but also how they are oriented and where they are with respect to us and each other, for example, when we are reaching for an object or driving. We also need to know about the visibility of objects: whether we should get closer, turn up the lights, or remove intervening objects in order to recognize an object with more confidence. Thus the recognition process in general must not boil away or destroy the information that is not diagnostic of particular objects (location, size, orientation, visibility, and surface properties) until it ends up with a residue of invariant information; it must *factor apart* or *decouple* this information from information about shape, so that different cognitive processes (e.g., shape recognition *versus* reaching) can access the information relevant to their particular tasks without becoming overloaded, distracted, or misled by the irrelevant information that the retina conflates with the relevant information. Thus one of the advantages of a structural description is that the shape of an object can be specified by one set of propositions, and its location in the visual field, orientation, size, and relation to other objects can be specified in different propositions, each bearing labels that processing operations can use for selective access to the information relevant to them.

Among the other advantages of structural descriptions are the following. By representing the different parts of an object as separate elements in the representation, these models break up the recognition process into simpler subprocesses, and more important, are well-suited to model our visual system's reliance on decomposition into parts during recognition and its ability to recognize novel rearrangements of parts such as the various configurations of a hand (see Hoffman and Richards (1984)). Second, by mixing logical and spatial relational terms in a representation, structural descriptions can differentiate among parts that must be present in a shape (e.g., the tail of the letter 'Q'), parts that may be present with various probabilities (e.g., the horizontal cap on the letter 'J'), and parts that must not be present (e.g., a tail on the letter 'O') (see Winston, 1975). Third, structural descriptions represent information in a form that is useful for subsequent visual reasoning, since the units in the representation correspond to objects, parts of objects, and spatial relations among them. Nonvisual information about objects or parts (e.g., categories they belong to, their uses, the situations that they are typically found in) can easily be associated with parts of structural descriptions, especially since many theories hold that nonvisual knowledge is stored in a propositional format that is similar to structural descriptions (e.g., Minsky, 1975; Norman and Rumelhart, 1975). Thus visual recognition can easily invoke knowledge about what is recognized that may be relevant to visual cognition in general, and that knowledge in turn can be used to aid in the recognition process (see the discussion of top-down approaches to recognition below).

The main problem with the structural description theory is that it is not really a full theory of shape recognition. It specifies the format of the representation used in matching the visual input against memory, but by itself it does not specify what types of entities and relations each of the units belonging to a structural description corresponds to (e.g., 'line' *versus* 'eye' *versus* 'sphere'; 'next-to' *versus* 'to-the-right-of' *versus* '37-degrees-with-respect-to'), nor how the units are created in response to the appropriate patterns of retinal stimulation (see the discussion of feature models above). Although most researchers in shape recognition would not disagree with the claim that the matching process deals with something like structural descriptions, a genuine theory of shape recognition based on structural descriptions must specify these components and justify why they are appropiate. In the next section, I discuss a theory proposed by David Marr and H. Keith Nishihara which makes specific proposals about each of these aspects of structural descriptions.

Two fundamental problems with the traditional approaches

There are two things wrong with the textbook approaches to visual representation and recognition. First, none of the theories specifies where perception ends and where cognition begins. This is a problem because there is a natural factoring part of the process that extracts information about the geometry of the visible world and the process that recognizes familiar objects. Take the recognition of a square. We can recognize a square whether its contours are defined by straight black lines printed on a white page, by smooth rows and columns of arbitrary small objects (Kohler, 1947; Koffka, 1935), by differences in lightness or in hue between the square and its background, by differences in binocular disparity (in a random-dot stereogram), by differences in the orientation or size of randomly scattered elements defining visual textures (Julesz, 1971), by differences in the directions of motion of randomly placed dots (Ullman, 1982; Marr, 1982), and so on. The square can be recognized as being a square regardless of how the boundaries are found; for example, we do not have to learn the shape of a square separately for boundaries defined by disparity in random-dot stereograms, by strings of asterisks, etc., nor must we learn the shapes of other figures separately for each type of edge once we have learned how to do so for a square. Conversely, it can be demonstrated that the ultimate recognition of the shape is not necessary for any of these processes to find the boundaries (the boundaries can be seen even if the shape they define is an unfamiliar blob, and expecting to see a square is neither necessary nor sufficient for the perceiver to see the boundaries; see Gibson, 1966; Marr, 1982; Julesz, 1971). Thus the process that recognizes a shape does not care about how its boundaries were found, and the processes that find the boundaries do not care how they will be used. It makes sense to separate the process of finding boundaries, degree of curvature, depth, and so on, from the process of recognizing particular shapes (and from other processes such as reasoning that can take their input from vision).

A failure to separate these processes has tripped up the traditional approaches in the following ways. First, any theory that derives canonical shape representations directly from the retinal arrays (e.g., templates, features) will have to solve all the problems associated with finding edges (see the previous paragraph) at the same time as solving the problem of recognizing particular shapes—an unlikely prospect. On the other hand, any theory that simply assumes that there is some perceptual processing done before the shape match but it does not specify what it is is in danger of explaining very little since the putative preprocessing could solve the most important part of the recognition process that the theory is supposed to address (e.g., a claim that a feature like 'head' is supplied to the recognition process). When assumptions about perceptual preprocessing *are* explicit, but are also incorrect or unmotivated, the claims of the recognition theory itself could be seriously undermined: the theory could require that some property of the world is supplied to the recognition process when there is no physical basis for the perceptual system to extract that property (e.g., Marr (1982) has argued that it is impossible for early visual processes to segment a scene into objects).

The second problem with traditional approaches is that they do not pay serious attention to what in general the shape recognition process has to do, or, put another way, what problem it is designed to solve (see Marr, 1982). This requires examining the input and desired output of the recognition process carefully: on the one hand, how the laws of optics, projective geometry, materials science, and so on, relate the retinal image to the external world, and on the other, what the recognition process must supply the rest of cognition with. Ignoring either of these questions results in descriptions of recognition mechanisms that are unrealizable, useless, or both.

The Marr–Nishihara theory

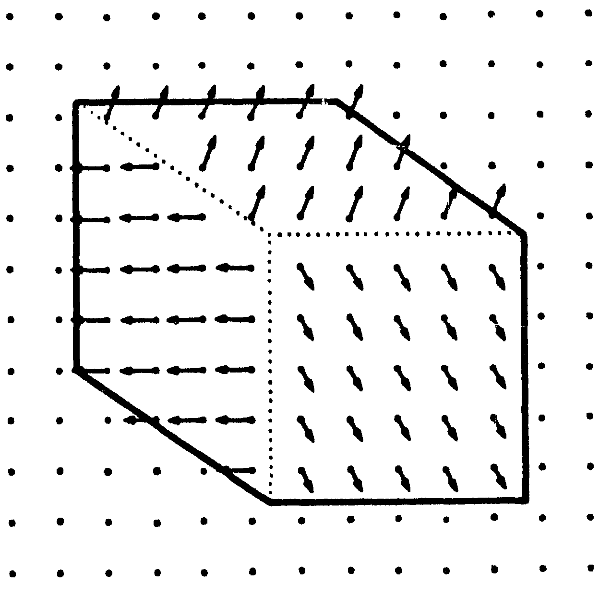
The work of David Marr represents the most concerted effort to examine the nature of the recognition problem, to separate early vision from recognition and visual cognition in general, and to outline an explicit theory of three-dimensional shape recognition built on such foundations. In this section, I will briefly describe Marr's theory. Though Marr's shape recognition model is not without its difficulties, there is a consensus that it addresses the most important problems facing this class of theories, and that its shortcomings define many of the chief issues that researchers in shape recognition must face.

The 2½-D sketch

The core of Marr's theory is a claim about the interface between perception and cognition, about what early, bottom-up visual processes supply to the recognition process and to visual cognition in general. Marr, in collaboration with H. Keith Nishihara, proposed that early visual processing culminates in the construction of a representation called the *2½-D sketch*. The 2½-D sketch is an array of cells, each cell dedicated to a particular line of sight from the viewer's vantage point. Each cell in the array is filled with a set of symbols indicating the depth of the local patch of surface lying on that line of sight, the orientation of that patch in terms of the degree and direction in which it dips away from the viewer in depth, and whether an edge (specifically, a discontinuity in depth) or a ridge (specifically, a discontinuity in orientation) is present at that line of sight (see Fig. 1). In other words, it is a representation of the surfaces that are visible when looking in a particular direction from a single vantage point. The 2½-D sketch is intended to gather together in one representation the richest information that early visual processes can deliver. Marr claims that no top-down processing goes into the construction of the 2½-D sketch, and that it does not contain any global information about shape (e.g., angles between lines, types of shapes, object or part boundaries), only depths and orientations of local pieces of surface.

The division between the early visual processes that culminate in the 2½-D sketch and visual recognition has an expository as well as a theoretical advantage: since the early processes are said not to be a part of visual cognition

Figure 1 *Schematic drawing of Marr and Nishihara's 2½-D sketch. Arrows represent surface orientation of patches relative to the viewer (the heavy dots are foreshortened arrows). The dotted line represents locations where orientation changes discontinuously (ridges). The solid line represents locations where depth changes discontinuously (edges). The depths of patches relative to the viewer are also specified in the 2½-D sketch but are not shown in this figure. From Marr (1982).*



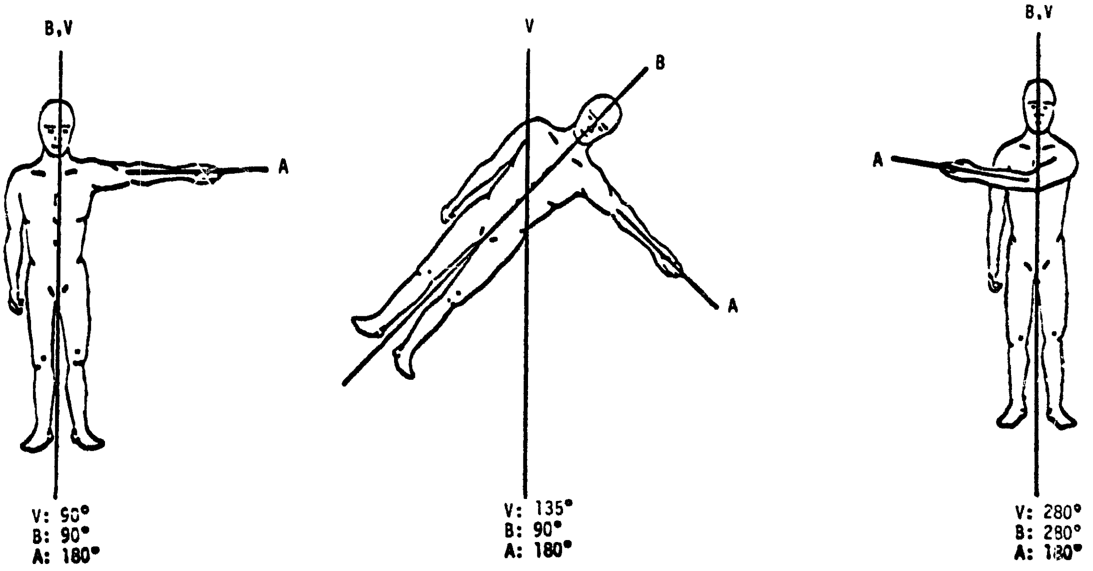
(i.e., not affected by a person's knowledge or intentions), I will discuss them only in bare outline, referring the reader to Marr (1982) and Poggio (1984) for details. The 2½-D sketch arises from a chain of processing that begins with mechanisms that convert the intensity array into a representation in which the locations of edges and other surface details are made explicit. In this 'primal sketch', array cells contain symbols that indicate the presence of edges, corners, bars, and blobs of various sizes and orientations at that location. Many of these elements can remain invariant over changes in overall illumination, contrast, and focus, and will tend to coincide in a relatively stable manner with patches of a single surface in the world. Thus they are useful in subsequent processes that must examine similarities and differences among neighboring parts of a surface, such as gradients of density, size, or shape of texture elements, or (possibly) processes that look for corresponding parts of the world in two images, such as stereopsis and the use of motion to reconstruct shape.

A crucial property of this representation is that the edges and other features are extracted separately at a variety of scales. This is done by looking for points where intensity changes most rapidly across the image using detectors of different sizes that, in effect, look at replicas of the image filtered at different ranges of spatial frequencies. By comparing the locations of intensity changes in each of the (roughly) bandpass-filtered images, one can create families of edge symbols in the primal sketch, some indicating the boundaries of the larger blobs in the image, others indicating the boundaries of finer details. This segregation of edge symbols into classes specific to different scales preserves some of the advantages of the Fourier models discussed above: shapes can be represented in an invariant manner across changes in image clarity and surface detail (e.g., a person wearing tweeds *versus* polyester).

The primal sketch is still two-dimensional, however, and the next stage of processing in the Marr and Nishihara model adds the third dimension to arrive at the 2½-D sketch. The processes involved at this stage compute the depths and orientations of local patches of surfaces using the binocular disparity of corresponding features in the retinal images from the two eyes (e.g., Marr and Poggio, 1977), the relative degrees of movement of features in successive views (e.g., Ullman, 1979), changes in shading (e.g., Horn, 1975), the size and shape of texture elements across the retina (Cutting and Millard, 1984; Stevens, 1981), the shapes of surface contours, and so on. These processes cannot indicate explicitly the overall three-dimensional shape of an object, such as whether it is a sphere or a cylinder; their immediate output is simply a set of values for each patch of a surface indicating its relative distance from the viewer, orientation with respect to the line of sight, and whether either depth or orientation changes discontinuously at that patch (i.e., whether an edge or ridge is present).

The 2½-D sketch itself is ill-suited to matching inputs against stored shape representations for several reasons. First, only the visible surfaces of shapes are represented; for obvious reasons, bottom-up processing of the visual input can provide no information about the back sides of opaque objects. Second, the 2½-D sketch is viewpoint-specific; the distances and orientations of patches of surfaces are specified with respect to the perceiver's viewing position and viewing direction, that is, in part of a spherical coordinate system centered on the viewer's vantage point. That means that as the viewer or the object moves with respect to one another, the internal representation of the object in the 2½-D sketch changes and hence does not allow a successful match against any single stored replica of a past 2½-D representation of the object (see Fig. 2a). Furthermore, objects and their parts are not explicitly demarcated.

Figure 2. *The orientation of a hand with respect to the retinal vertical* V *(a viewer-centered reference frame), the axis of the body* B *(a global object-centered reference frame), and the axis of the lower arm* A *(a local object-centered reference frame). The retinal angle of the hand changes with rotation of the whole body (middle panel); its angle with respect to the body changes with movement of the elbow and shoulder (right panel). Only its angle with respect to the arm remains constant across these transformations.*



Shape recognition and 3-D models

Marr and Nishihara (1978) have proposed that the shape recognition process (a) defines a coordinate system that is centered on the as-yet unrecognized object, (b) characterizes the arrangement of the object's parts with respect to that coordinate system, and (c) matches such characterizations against canonical characterizations of objects' shapes stored in a similar format in memory. The object os described with respect to a coordinate system that is centered on the object (e.g., its origin lies on some standard point on the object and one or more of its axes are aligned with standard parts of the object), rather than with respect to the viewer-centered coordinate system of the 2½-D sketch, because even though the locations of the object's parts with respect to the viewer change as the object as a whole is moved, the locations of its parts with respect to the object itself do not change (see Fig. 2b). A structural description representing an object's shape in terms of the arrangement of its parts, using parameters whose meaning is determined by a coordinate system centered upon that object, is called the *3-D model description* in Marr and Nishihara's theory.

Centering a coordinate system on the object to be represented solves only some of the problems inherent in shape recognition. A single object-centered description of a shape would still fail to match an input object when the object bends at its joints (see Fig. 2c), when it bears extra small parts (e.g., a horse with a bump on its back), or when there is a range of variation among objects within a class. Marr and Nishihara address this *stability* problem by proposing that information about the shape of an object is stored not in a single model with a global coordinate system but in a hierarchy of models each representing parts of different sizes and each with its own coordinate system. Each of these local coordinate systems is centered on a part of the shape represented in the model, aligned with its axis of elongation, symmetry, or (for movable parts) rotation.

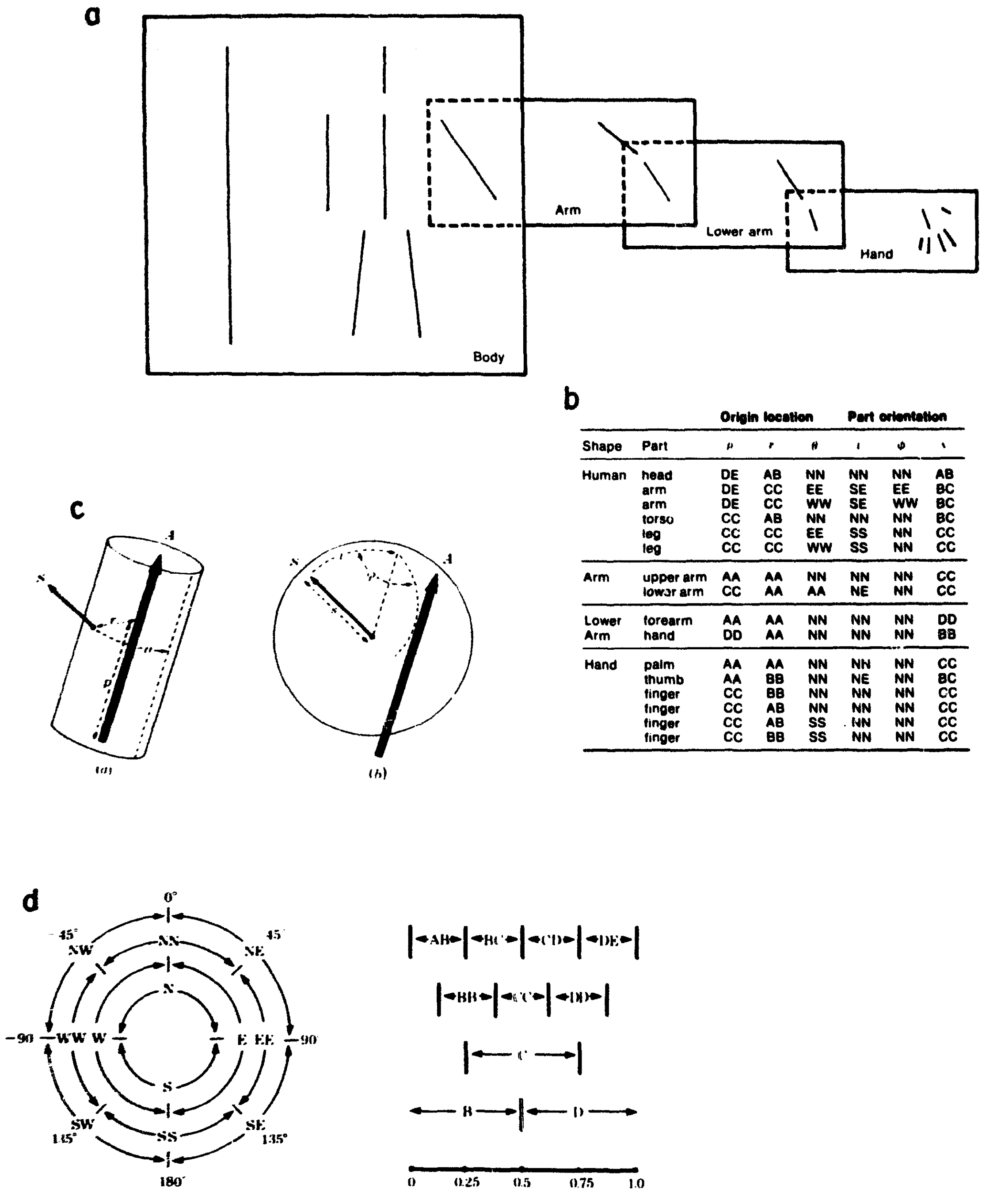
For example, to represent the shape of a horse, there would be a top-level model with a coordinate system centered on the horse's torso. That coordinate system would be used to specify the locations, lengths, and angles of the main parts of the horse: the head, limbs, and tail. Then subordinate models are defined for each of those parts: one for the head, one for the front right leg, etc. Each of those models would contain a coordinate system centered on the part that the model as a whole represents, or on a part subordinate to that part (e.g., the thigh for the leg subsystem). The coordinate system for that model would be used to specify the positions, orientations, and lengths of the subordinate parts that comprise the part in question. Thus, within the head model, there would be a specification of the locations and angles of the neck axis and of the head axis, probably with respect to a coordinate system centered on the neck axis. Each of these parts would in turn get its own model, also consisting of a coordinate axis centered on a part, plus a characterization of the parts subordinate to it. An example of a 3-D model for a human shape is shown in Fig. 3.

Employing a hierarchy of corrdinate systems solves the stability problems alluded to above, because even though the position and orientation of the hand relative to the torso can change wildly and unsystematically as a person bends the arm, the position of the hand relative to the arm does not change (except possibly by rotating within the range of angles permitted by bending of the wrist). Therefore the description of the shape of the arm remains constant only when the arrangement of its parts is specified in terms of angles and positions relative to the arm axis, not relative to the object as a whole (see Fig. 2). For this to work, of course, positions, lengths, and angles must be specified in terms of ranges (see Fig. 3d) rather than by precise values, so as to accommodate the changes resulting from movement or individual variation among exemplars of a shape. Note also that the hierarchical arrangement of 3-D models compensates for individual variation in a second way: a horse with a swollen or broken knee, for example, will match the 3-D model defining the positions of a horse's head, torso, limbs, and tail relative to the torso axis, even if the subordinate limb model itself does not match the input limb.

Organization and accessing of shape information in memory

Marr and Nishihara point out that using the 3-D model format, it is possible to define a set of values at each level of the hierarchy of coordinate systems that correspond to a central tendency among the members of well-defined classes of shapes organized around a single 'plan'. For example, at the top level of the hierarchy defining limbs with respect to the torso, one can define one set of values that most quadruped shapes cluster around, and a different set of values that most bird shapes cluster around. At the next level down one can define values for subclasses of shapes such as songbirds *versus* long-legged waders.

This modular organization of shape descriptions, factoring apart the arrangement of parts of a given size from the internal structure of those parts. and factoring apart shape of an individual type from the shape of the class of objects it belongs to, allows input descriptions to be matched against memory in a number of ways. Coarse information about a shape specified in a top-level coordinate system can be matched against models for general classes (e.g., quaduped) first, constraining the class of shapes that are checked the next level down, and so on. Thus when recognizing the shape of a person, there is no need to match it against shape descriptions of particular types of



		Origin location			Part orientation		
Shape	Part	p	r	θ	ι	ϕ	s
Human	head	DE	AB	NN	NN	NN	AB
	arm	DE	CC	EE	SE	EE	BC
	arm	DE	CC	WW	SE	WW	BC
	torso	CC	AB	NN	NN	NN	BC
	leg	CC	CC	EE	SS	NN	CC
	leg	CC	CC	WW	SS	NN	CC
Arm	upper arm	AA	AA	NN	NN	NN	CC
	lower arm	CC	AA	AA	NE	NN	CC
Lower Arm	forearm	AA	AA	NN	NN	NN	DD
	hand	DD	AA	NN	NN	NN	BB
Hand	palm	AA	AA	NN	NN	NN	CC
	thumb	AA	BB	NN	NE	NN	BC
	finger	CC	BB	NN	NN	NN	CC
	finger	CC	AB	NN	NN	NN	CC
	finger	CC	AB	SS	NN	NN	CC
	finger	CC	BB	SS	NN	NN	CC

guppies, parakeets, or beetles once it has been concluded that the gross shape is that of a primate. (Another advantage of using this scheme is that if a shape is successfully matched at a higher level but not at any of the lower levels, it can still be classified as failing into a general class or pattern, such as being a bird, even if one has never encountered that type of bird before). An alternative way of searching shape memory is to allow the successful recognition of a shape in a high-level model to trigger the matching of its subordinate part-models against as-yet unrecognized parts in the input, or to allow the successful recognition of individual parts to trigger the matching of their superordinate models against the as-yet unrecognized whole object in the input containing that part. (For empirical studies on the order in which shape representations are matched against inputs, see Jolicoeur *et al.* 1984a; Rosch *et al.* 1976; Smith *et al.* 1978. These studies suggest that the first index into shape memory may be at a 'basic object' level, rather than the most abstract level, at least for prototypical exemplars of a shape.)

Representing shapes of parts

Once the decomposition of a shape into its component axes is accomplished, the shapes of the components that are centered on each axis must be specified as well. Marr and Nishihara conjecture that shapes of parts may be described in terms of *generalized cones* (Binford, 1971). Just as a cone can be defined as the surface traced out when a circle is moved along a straight line perpendicular to the circle while its diameter steadily shrinks, a generalized cone can be defined as the surface traced out when *any* planar closed shape is moved along *any* smooth line with its size smoothly changing in *any* way. Thus to specify a particular generalized cone, one must specify

Figure 3. *Marr and Nishishara's 3-D model description for a human shape.* A *shows how the whole shape is decomposed into a hierarchy of models, each enclosed by a rectangle.* B *shows the information contained in the model description: the subordinate models contained in each superordinate, and the location and orientation of the defining axis of each subordinate with respect to a coordinate system centered on a part of the superordinate. The meanings of the symbols used in the model are illustrated in* C *and* D*: the endpoint of a subordinate axis is defined by three parameters in a cylindrical coordinate system centered on a superordinate part (left panel of* C*); the orientation and length of the subordinate axis are defined by three parameters in a spherical coordinate system centered on the endpoint and aligned with the superordinate part (right panel of* C*). Angles and lengths are specified by ranges rather than by exact values (*D*). From Marr and Nishihara (1978).*

the shape of the axis (i.e., how it bends, if at all), the two-dimensional shape of the generalized cone's cross-section, and the gradient defining how its area changes as a function of position along the axis. (Marr and Nishihara point out that shapes formed by biological growth tend to be well-modeled by generalized cones, making them good candidates for internal representations of the shapes of such parts.) In addition, surface primitives such as rectangular, circular, or bloblike markings can also be specified in terms of their positions with respect to the axis model.

Deriving 3-D descriptions from the 2½-D sketch

Unfortunately, this is an aspect of the Marr and Nishihara model that has not been developed in much detail. Marr and Nishihara did outline a limited process for deriving 3-D descriptions from the two-dimensional silhouette of the object. The process first carves the silhouette into parts at extrema of curvature, using a scheme related to the one proposed by Hoffman and Richards (1984). Each part is given an axis coinciding with its direction of elongation, and lines are created joining endpoints to neighboring axes. The angles between axes and lines are measured and recorded, the resulting description is matched against top-level models in memory, and the best-matched model is chosen. At that point, constraints on how a part is situated and oriented with respect to the superordinate axis in that model can be used to identify the viewer-relative orientation of the part axis in the 2½-D sketch. That would be necessary if the orientation of that part cannot be determined by an examination of the sketch itself, such as when its axis is pointing toward the viewer and hence is foreshortened. Once the angle of an axis is specified more precisely, it can be used in selecting subordinate 3-D models for subsequent matching.

The Marr and Nishihara model is the most influential contemporary model of three-dimensional shape recognition, and it is not afflicted by many of the problems that afflict the textbook models of shape representation summarized earlier. Nonetheless, the model does have a number of problems, which largely define the central issues to be addressed in current research on shape recognition. In the next section, I summarize some of these problems briefly.

Current problems in shape recognition research

Choice of shape primitives to represent parts

The shape primitives posited by Marr and Nishihara—generalized cones centered on axes of elongation or symmetry—have two advantages: they can easily characterize certain important classes of objects, such as living things, and they can easily be derived from their silhouettes. But Hoffman and Richards (1984) point out that many classes of shapes cannot be easily described in this scheme, such as faces, shoes, clouds, and trees. Hoffman and Richards take a slightly different approach to the representation of parts in a shape description. They suggest that the problem of *describing* parts (i.e., assigning them to categories) be separated from the problem of *finding* parts (i.e., determining how to carve an object into parts). If parts are only found by looking for instances of certain part categories (e.g., generalized cones) then parts that do not belong to any of those categories would never be found. Hoffman and Richards argue that, on the contrary, there is a psychologically plausible scheme for finding part boundaries that is ignorant of the nature of the parts it defines. The parts delineated by these boundaries at each scale can be categorized in terms of a taxonomy of lobes and blobs based on the patterns of inflections and extrema of curvature of the lobe's surface. (Hoffman (1983) has worked out a taxonomy for primitive shape descriptors, called 'codons', for two-dimensional plane curves). They argue not only that the decomposition of objects into parts is more basic for the purposes of recognition than the description of each part, but that the derivation of part boundaries and the classification of parts into sequences of codon-like descriptors might present fewer problems than the derivation of axis-based descriptions, because the projective geometry of extrema and inflections of curvature allows certain reliable indicators of these extrema in the image to be used as a basis for identifying them (see Hoffman, 1983).

Another alphabet of shape primitives that has proven useful in computer vision consists of a set of canonical volumetric shapes such as spheres, parallelopipeds, pyramids, cones, and cylinders, with parameterized sizes and (possibly) aspect ratios, joined together in various ways to define the shape of an object (see e.g., Hollerbach, 1975; Badler and Bajcsy, 1978). It is unlikely that a single class of primitives will be sufficient to characterize all shapes, from clothes lying in a pile to faces to animals to furniture. That means that the derivation process must be capable of determining prior to describing and recognizing a shape which type of primitives are appropriate to it. There are several general schemes for doing this. A shape could be described in parallel in terms of all the admissible representational schemes, and descriptions in inappropriate schemes could be rejected because they are unstable over small changes in viewing position or movement, or because no single description within a scheme can be chosen over a large set of others within that scheme. Or there could be a process that uses several coarse properties of an object, such as its movement, surface texture and color, dimensionality, or sound to give it an initial classification into broad categories such as animal *versus* plant *versus* artifact each with its own scheme of primitives and their organization (e.g., see Richards (1979, 1982) on "playing 20 questions" with the perceptual input).

Assigning frames of reference to a shape

In a shape representation, size, location, and orientation cannot be specified in absolute terms but only with respect to some frame of reference. It is convenient to think of a frame of reference as a coordinate system centered on or aligned with the reference object, and transformations within or between reference frames as being effected by an analogue of matrix multiplication taking the source coordinates as input and deriving the destination coordinates as output. However, a reference frame need not literally be a coordinate system. For example, it could be an array of arbitrarily labelled cells, where each cell represents a fixed position relative to a reference object. In that case, transformations within or between such reference frames could be effected by fixed connections among corresponding source and destination cells (e.g., a network of connections linking each cell with its neighbor to the immediate right could effect translation when activated iteratively; see e.g., Trehub, 1977).

If a shape is represented for the purpose of recognition in terms of a coordinate system or frame of reference centered on the object itself, the shape recognition system must have a way of determining what the object-centered frame of reference is prior to recognizing the object. Marr and Nishihara conjecture that a coordinate system used in recognition may be aligned with an object's axes of elongation, bilateral symmetry, radial symmetry (for objects that are radially symmetrical in one plane and extended in an orthogonal direction), rotation (for jointed objects), and possibly linear movement. Each of these is suitable for aligning a coordinate system with an object because each is derivable prior to object recognition and each is fairly invariant for a type of object across changes in viewing position.

This still leaves many problems unsolved. For starters, these methods only fix the orientation of one axis of the cylindrical coordinate system. The direction of the cylindrical coordinate system for that axis (i.e., which end is zero), the orientation of the zero point of its radial scale, and the handedness of the radial scale (i.e., whether increasing angle corresponds to going clockwise or counterclockwise around the scale) are left unspecified, as is the direction of one of the scales used in the spherical coordinate system specified within the cylindrical one (assuming its axes are aligned with the axis of the cylindrical system and the line joining it to the cylindrical system) (see Fig. 3c). Furthermore, even the choice of the orientation of the principal axis will be difficult when an object is not elongated or symmetrical, or when the principal axis is occluded, foreshortened, or physically compressed. For example, if the top-level description of a cow shape describes the dispositions of its parts with respect to the cow's torso, then when the cow faces the viewer the torso is not visible, so there is no way for the visual system to describe, say, the orientations of the leg and head axes relative to its axis.

There is evidence that our assignment of certain aspects of frames of reference to an object is done independently of its intrinsic geometry. The positive–negative direction of an intrinsic axis, or the assignment of an axis to an object when there is no elongation or symmetry, may be done by computing a global up–down direction. Rock (1973, 1983) presents extensive evidence showing that objects' shapes are represented relative to an up–down direction. For example, a square is ordinarily 'described' internally as having a horizontal edge at the top and bottom; when the square is tilted 45°, it is described as having vertices at the top and bottom and hence is perceived as a different shape, namely, a diamond. The top of an object is not, however, necessarily the topmost part of the object's projection on the retina: Rock has shown that when subjects tilt their heads and view a pattern that, unknown to them, is tilted by the same amount (so that it projects the same retinal image), they often fail to recognize it. In general, the up–down direction seems to be assigned by various compromises among the gravitational upright, the retinal upright, and the prevailing directions of parallelism, pointing, and bilateral symmetry among the various features in the environment of the object (Attneave, 1968; Palmer and Bucher, 1981; Rock, 1973). In certain circumstances, the front–back direction relative to the viewer may also be used as a frame of reference relative to which the shape is described; Rock *et al.* (1981) found that subjects would fail to recognize a previously-learned asymmetrical wire form when it was rotated 90° about the vertical axis.

What about the handedness of the angular scale in a cylindrical coordinate system (e.g., the θ parameter in Fig. 3)? One might propose that the visual system employs a single arbitrary direction of handedness for a radial scale that is uniquely determined by the positive–negative direction of the long axis orthogonal to the scale. For example, we could use something analogous to the 'right hand rule' taught to physics students in connection with the orientation of a magnetic field around a wire (align the extended thumb of your right hand with the direction of the flow of current, and look which way your fingers curl). There is evidence, however, that the visual system does *not* use any such rule. Shepard and Hurwitz (1984, in this issue; see also Hinton and Parsons, 1981; Metzler and Shepard, 1975) point out that we do not in general determine how parts are situated or oriented with respect to the left–right direction on the basis of the intrinsic geometry of the object (e.g., when we are viewing left and right hands). Rather, we assign the object a left–right direction in terms of our own egocentric left and right sides. When an object's top and bottom do not correspond to an egocentric or gravitational top–bottom direction, we mentally rotate it into such an orientation, and when two unfamiliar objects might differ in handedness, we rotate one into the orientation of the other (taking greater amounts of time for greater angles of rotation. Mental rotation is discussed further later in this paper). Presumably this failure to assign objects intrinsic left and right directions is an evolutionary consequence of the fact that aside from human artifacts and body parts, virtually no class of ecologically significant shapes need be distinguished from their enantiomorphs (Corballis and Beale, 1976; Gardner, 1967).

To the extent that a shape is described with respect to a reference frame that depends on how the object is oriented with respect to the viewer or the environment, shape recognition will fail when the object moves with respect to the viewer or environment. In cases where we do succeed at recognizing objects across its different dispositions and where object-centered frames cannot be assigned, there are several possible reasons for such success. One is that multiple shape descriptions, corresponding to views of the object with different major axes occluded, are stored under a single label and corresponding parts in the different descriptions are linked. Another is that the representation of the object is rotated into a canonical orientation or until the description of the object relative to the frame matches a memorized shape description; alternatively, the reference frame or canonical orientation could be rotated into the orientation of the object. Interestingly, there is evidence from Cooper and Shepard (1973) and Shepard and Hurwitz (1984) that the latter option (rotating an empty reference frame) is difficult or impossible for humans to do: advance information about the orientation of an upcoming visual stimulus does not spare the perceiver from having to rotate the stimulus mentally when it does appear in order to judge its handedness.[2] A third possibility stems from Hoffman and Richards's (1984) suggestion that part segmentation may be independent of orientation, and that only the representations of spatial relations among parts are orientation-sensitive. If so, recognition of an isolated part can be used as an index to find the objects in memory that contain that part. Finally, in some cases recognition might fail outright with changes in orientation but the consequences might be innocu-

[2] Hinton and Parsons (1981) have shown that when the various stimuli to be judged all conform to a single shape schema (e.g., alphanumeric characters with a vertical spine and strokes attached to the right side of the spine, such as 'R', 'L', and 'F'), advance information about orientation saves the subject from having to rotate the stimulus. However, it is possible that in their experiment subjects rotated a concrete image of a vertical spine plus a few strokes, rather than an empty reference frame.

ous. Because of the pervasiveness of gravity, many shapes will rarely be seen in any position but the upright (e.g., faces, trees), and many of the differences in precise shape among objects lacking axes of symmetry, movement, rotation, or elongation are not ecologically significant enough for us to distinguish among them in memory (e.g., differences among bits of gravel or crumpled newspaper). Naturally, to the extent that any of the suggestions made in this paragraph are true, the importance of Marr and Nishihara's argument for canonical object-centered descriptions lessens.[3]

Frames of reference for the visual field

We not only represent the shapes of objects internally; we also represent the locations and orientations of objects and surfaces in the visual field. The frames of reference that we use to represent this information will determine the ease with which we can make various spatial judgments. The relevant issues are the *alignment* of the frames of reference, and the *form* of the frames of reference.

Early visual representations are in a viewer-centered and approximately spherical frame of reference; that is, our eyes give us information about the world in terms of the azimuth and elevation of the line of sight at which the features are found relative to the retina, and their distance from the viewing position (this is the coordinate system used for the 2½-D sketch). Naturally, this is a clumsy representation to use in perceiving invariant spatial relations, since the information will change with eye movements. The system can compensate for eye movements by superimposing a head-centered coordinate system on top of the retinal system and moving the origin of that coordinate system in conjunction with eye movement commands. Thus every cell in the 2½-D sketch would be represented by the fixed 'address' defined with respect to the retina, and also by its coordinates with respect to the head, which would be dynamically adjusted during eye movements so that fixed locations in the world retain a constant coordinate address within the head-centered system. A third coordinate system, defined over the same information, could represent position with respect to the straight ahead direction of the body

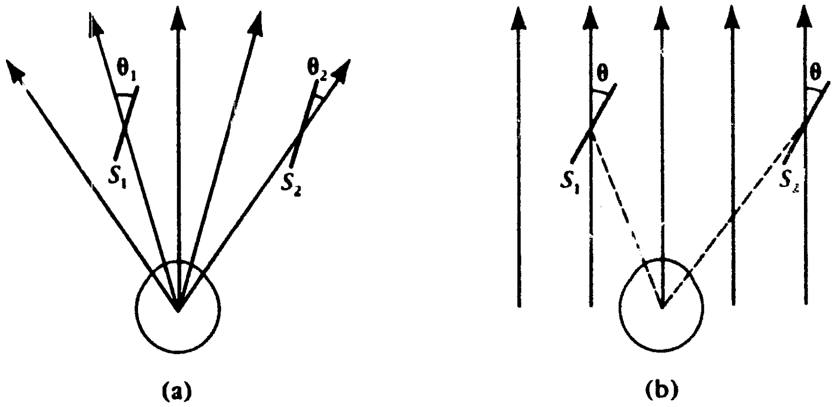
[3] Specifying the origin of the object-centered coordinate system presents a slightly different set of issues than specifying the orientation of its axes. An origin for an object-centered frame can be determined by finding its visual center of mass or by assigning it to one end of a principal axis. It is noteworthy that there are no obvious cases where we fail to recognize an object when it is displaced, where we see a shape as ambiguous by virue of assigning different 'centers' or 'reference locations' to it (analogous to the diamond/tilted square ambiguity), or where we have to mentally translate an object in order to recognize it or match it against a comparison object. This indicates either that the procedure that assigns an origin to an object on the basis of its intrinsic geometry always yields a unique solution for an object, or that, as Hinton (1979a) suggests, we do not compute an origin at all in shape descriptions, only a set of significant directions.

and it could be updated during head movements to represent the invariant position of surfaces across those movements. Other coordinate systems could be defined over these visible surface representations as well, such as coordinate systems aligned with the gravitational upright and horizontal ground (see Shepard and Hurwitz, 1984), with fixed salient landmarks in the world, or with the prevailing directions of large surfaces (e.g., the walls in a tilted room). These coordinate systems for objects' positions with respect to one's body or with respect to the environment could be similar to those used to represent the parts of an object with respect to the object as a whole. Presumably they are also like coordinate systems for objects' shapes in being organized hierarchically, so that a paper clip might be represented by its position with respect to the desk tray it is in, whose position is specified with respect to the desk, whose position is specified with respect to the room. Beyond the visual world, the origin and orientation of large frames of reference such as that for a room could be specified in a hierarchy of more schematic frames of reference for entities that cannot be seen in their entirety, such as those for floor plans, buildings, streets, neighborhoods and so on (see e.g., Kuipers, 1978; Lynch, 1960; McDermott, 1980).

The possible influence of various frames of reference on shape perception can be illustrated by an unpublished experiment by Laurence Parsons and Geoff Hinton. They presented subjects with two Shepard–Metzler cube figures, one situated 45° to the left of the subject, another at 45° to the right. The task was to turn one of the objects (physically) to whatever orientation best allowed the subject to judge whether the two were identical or whether one was a mirror-reversed version of the other (subjects were allowed to move their heads around the neck axis). If objects were represented in coordinate systems centered upon the objects themselves, subjects would not have to turn the object at all (we known from the Shepard and Metzler studies that this is quite unlikely to be true for these stimuli). If objects are represented in a coordinate system aligned with the retina, subjects should turn one object until the corresponding parts of the two objects are perpendicular to the other, so that they will have the same orientations with respect to their respective lines of sight. And if shapes are represented in a coordinate system aligned with salient environmental directions (e.g., the walls), one object would be turned until its parts are parallel to those of the other, so that they will have the same orientations with respect to the room. Parsons and Hinton found that subjects aligned one object so that it was nearly parallel with another, with a partial compromise toward keeping the object's retinal projections similar (possibly so that corresponding cube faces on the two objects would be simultaneously visible). This suggests that part orientations are represented primarily with respect to environmentally-influenced frames.

The choice of a reference object, surface, or body part is closely tied to the format of the coordinate system aligned with the frame of reference, since rotatable objects (such as the eyes) and fixed landmarks easily support coordinate systems containing polar scales, whereas reference frames with orthogonal directions (e.g., gravity and the ground, the walls of a room) easily support Cartesian-like coordinate systems. The type of coordinate system employed has effects on the ease of making certain spatial judgments. As mentioned, the 2½-D sketch represents information in a roughly spherical coordinate system, with the result that the easiest information to extract concerning the position of an edge or feature is its distance and direction with respect to the vantage point. As Marr (1982) points out, this representation conceals many of the geometric properties of surfaces that it would be desirable to have easy access to; something closer to a Cartesian coordinate system centered on the viewer would be much handier for such purposes. For example, if two surfaces in different parts of the visual field are parallel, their orientations as measured in a spherical coordinate system will be different, but their orientations as measured in a coordinate system with a parallel component (e.g., Cartesian) will be the same (see Fig. 4). If a surface is flat, the represented orientations of all the patches composing its surface will be identical in Cartesian, but not in spherical coordinates. Presumably, size constancy could also be a consequence of such a coordinate system, if a given range of coordinates in the left–right or up–down directions always stood for

Figure 4. *Effects of rectangular ve·sus polar coordinate systems on making spatial judgments. Whether two surfaces are parallel can be assessed by comparing their angles with respect to the straight ahead direction in a rectangular coordinate system (b), but not by comparing their angles with respect to the lines of sight in a polar system (a). From Marr (1982).*



a constant real world distance regardless of the depth of the represented surface.

One potentially relevant bit of evidence comes from a phenomenon studied by Corcoran (1977), Natsoulas (1966), and Kubovy *et al.* (1984, Reference note 1). When an asymmetric letter such as 'd' is traced with a finger on the back of a person's head, the person will correctly report what the letter is. But when the same letter is traced on the person's forehead, the mirror image of that letter is reported instead (in this case, 'b'). This would follow if space (and not just visible space) is represented in a parallel coordinate system aligned with a straight ahead direction, such as that shown in Fig. 4b. The handedness of a letter would be determined by whether its spine was situated to the left or right of the rest of its parts, such that 'left' and 'right' would be determined by a direction orthogonal to the straight ahead direction, regardless of where on the head the letter is drawn. The phenomenon would not be expected in an alternative account, where space is represented using spherical coordinates centered on a point at or behind the eyes (e.g., Fig. 4a), because then the letter would be reported as if 'seen' from the inside of a transparent skull, with letters traced on the back of the head reported as mirror-reversed, contrary to fact.

In many experiments allowing subjects to choose between environmental, Cartesian-like reference frames and egocentric, spherical reference frames, subjects appear to opt for a compromise (e.g., the Parsons and Hinton and Kubovy *et al.* studies; see also Attneave, 1972; Gilinsky, 1955; Uhlarik *et al.* 1980). It is also possible that we have access to both systems, giving rise to ambiguities when a single object is alternatively represented in the two systems, for example, when railroad tracks are seen either as parallel or as converging (Boring, 1952; Gibson, 1950; Pinker, 1980a), or when the corner formed by two edges of the ceiling of a room can be seen both as a right angle and as an obtuse angle.

Deriving shape descriptions

One salient problem with the Marr and Nishihara model of shape recognition in its current version is that there is no general procedure for deriving an object-centered 3-D shape description from the 2½-D sketch. The algorithm proposed by Marr and Nishihara using the two-dimensional silhouette of a shape to find its intrinsic axes has trouble deriving shape descriptions when axes are foreshortened or occluded by other parts of the object (as Marr and Nishihara pointed out). In addition, the procedures it uses for joining up part boundaries to delineate parts, to find axes of parts once they are delineated, and to pair axes with one another in adjunct relations rely on some limited heuristics that have not been demonstrated to work other than for objects composed of generalized cones—but the perceiver cannot in general know prior to recognition whether he or she is viewing such an object. Furthermore, there is no explicit procedure for grouping together the parts that belong together in a single hierarchical level in the 3-D model description. Marr and Nishihara suggest that all parts lying within a 'coarse spatial context' surrounding an axis can be placed within the scope of the model specific to that axis, but numerous problems could arise when unrelated parts are spatially contiguous, such as when a hand is resting on a knee. Some of these problems perhaps could be resolved using an essentially similar scheme when information that is richer than an object's silhouette is used. For example, the depth, orientation, and discontinuity information in the 2½-D sketch could assist in the perception of foreshortened axes (though not when the blunt end of a tapering object faces the viewer squarely), and information about which spatial frequency bandpass channels an edge came from could help in the segregation of parts into hierarchical levels in a shape description.

A general problem in deriving shape representations from the input is that, as mentioned, the choice of the appropriate reference frame and shape primitives depends on what type of shape it is, and shapes are recognized via their description in terms of primitives relative to a reference frame. In the remainder of this section I describe three types of solutions to this chicken-and-egg problem.

Top–down processing

One response to the inherent difficulties of assigning descriptions to objects on the basis of their retinal images is to propose that some form of ancillary information based on a person's knowledge about regularities in the world is used to choose the most likely description or at least to narrow down the options (e.g., Gregory, 1970; Lindsay and Norman, 1977; Minsky, 1975; Neisser, 1967). For example, a cat-owner could recognize her cat upon seeing only a curved, long, grey stripe extending out from underneath her couch, based on her knowledge that she has a long-tailed grey cat that enjoys lying there. In support of top–down, or, more precisely, knowledge-guided perceptual analysis, Neisser (1967), Lindsay and Norman (1977), and Gregory (1970) have presented many interesting demonstrations of possible retinal ambiguities that may be resolved by knowledge of physical or object-specific regularities, and Biederman (1972), Weisstein and Harris (1974) and Palmer (1975b) and others have shown that the recognition of an object, a part of an object, or a feature can depend on the identity that we attribute to the context object or scene as a whole.

Despite the popularity of the concept of top–down processing within cognitive science and artificial intelligence during much of the 1960s and 1970s, there are three reasons to question the extent to which general knowledge plays a role in describing and recognizing shapes. First, many of the supposed demonstrations of top–down processing leave it completely unclear what kind of knowledge is brought to bear on recognition (e.g., regularities about the geometry of physical objects in general, about particular objects, or about particular scenes or social situations), and how that knowledge is brought to bear (e.g., altering the order in which memory representations are matched against the input, searching for particular features or parts in expected places, lowering the goodness-of-fit threshold for expected objects, generating and fitting templates, filling in expected parts). Fodor (1983) points out that these different versions of the top-down hypothesis paint very different pictures of how the mind is organized in general: if only a restricted type of knowledge can influence perception in a top–down manner, and then only in restricted ways, the mind may be constructed out of independent modules with restricted channels of communication among them. But if all knowledge can influence perception, the mind could consist of an undifferentiated knowledge base and a set of universal inference procedures which can be combined indiscriminately in the performance of any task. Exactly which kind of top–down processing is actually supported by the data can make a big difference in one's conception of how the mind works; Fodor argues that so far most putative demonstrations of top–down phenomena are not designed to distinguish among possible kinds of top-down processing and so are uninformative on this important issue.

A second problem with extensive top–down processing is that there is a great deal of information about the world that is contained in the light array, even if that information cannot be characterized in simple familiar schemes such as templates or features (see Gibson, 1966, 1979; Marr, 1982). Given the enormous selection advantage that would be conferred on an organism that could respond to what was really in the world as opposed to what it expected to be in the world whenever these two descriptions were in conflict, we should seriously consider the possibility that human pattern recognition has the most sophisticated bottom-up pattern analyses that the light array and the properties of our receptors allow. And as Ullman (1984, this issue) points out, we do appear to be extremely accurate perceivers even when we have no basis for expecting one object or scene to occur rather than another, such as when watching a slide show composed of arbitrary objects and scenes.

Two-stage analysis of objects

Ullman (1984) suggests that our visual systems may execute a universal set of 'routines' composed of simple processes operating on the 2½-D sketch, such as tracing along a boundary, filling in a region, marking a part, and sequentially processing different locations. Once universal routines are executed, their outputs could characterize some basic properties of the prominent entities in the scene such as their rough shape and spatial relationships. This characterization could then trigger the execution of routines specific to the recognition of particular objects or classes of objects. Because routines can be composed of arbitrary sequences of very simple but powerful processes, it might be possible to compile both a core of generally useful routines, plus a large set of highly specific routines suitable for the recognition of very different classes of objects, rather than a canonical description scheme that would have to serve for every object type. (In Ullman's theory visual routines would be used not only for the recognition of objects but also for geometric reasoning about the surrounding visual environment such as determining whether one object is inside another or counting objects.) Richards (1979, 1982) makes a related proposal concerning descriptions for recognition, specifically, that one might first identify various broad classes of objects such as animal, vegetable, or mineral by looking for easily sensed hallmarks of these classes such as patterns of movement, color, surface texture, dimensionality, even coarse spectral properties of their sounds. Likely reference frames and shape primitives could then be hypothesized based on this first-stage categorization.

Massively parallel models

There is an alternative approach that envisions a very different type of solution from that suggested by Richards, and that advocates very different types of mechanisms from those described in this issue by Ullman. Attneave (1982), Hinton (1981) and Hrechanyk and Ballard (1982) have outlined related proposals for a model of shape recognition using massively parallel networks of simple interconnected units, rather than sequences of operations performed by a single powerful processor (see Ballard *et al.* 1983; Feldman and Ballard, 1982; Hinton and Anderson, 1981), for introductions to this general class of computational architectures).

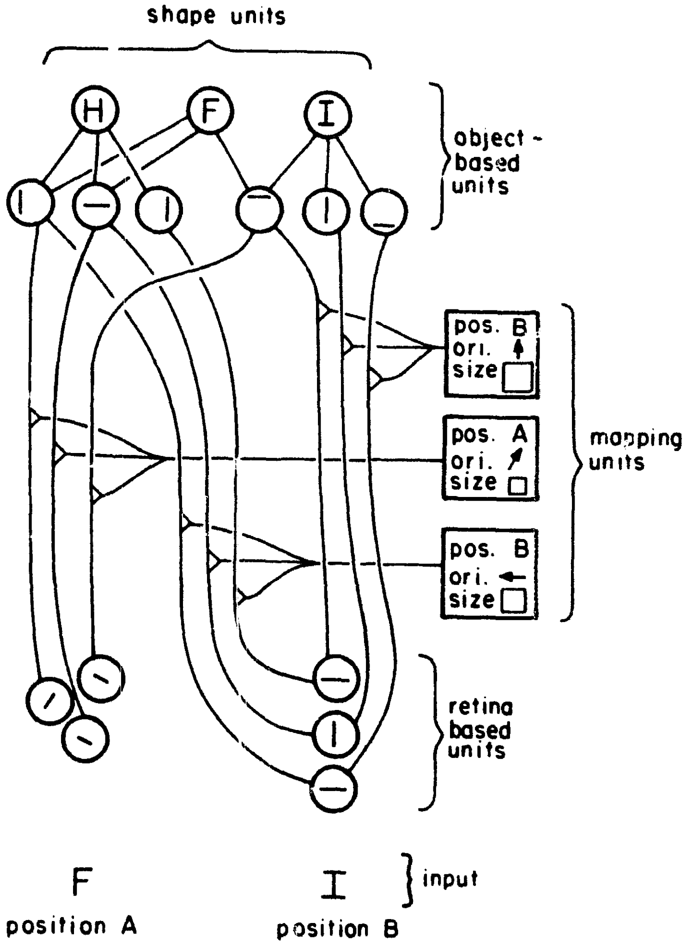
A favorite analogy for this type of computation (e.g., Attneave, 1982) is the problem of determining the shape of a film formed when an irregularly shaped loop of wire is dipped in soapy water (the shape can be characterized by quantizing the film surface into patches and specifying the height of each patch). The answer to the problem is constrained by the 'least action' principle ensuring that the height of any patch of the film must be close to the heights of all its neighboring patches. But how can this information be used if one does not know beforehand the heights of all the neighbors of any patch? One can solve the problem iteratively, by assigning every patch an arbitrary initial height except for those patches touching the wire loop, which are assigned the same heights as the piece of wire they are attached to. Then the heights of each of the other patches is replaced by the average height of its neighbors. This is repeated through several iterations; eventually the array of heights converges on a single set of values corresponding to the shape of the film, thanks to constraints on height spreading inward from the wire. The solution is attained without knowing the height of any single interior patch *a priori*, and without any central processor.

Similarly, it may be possible to solve some perceptual problems using networks of simple units whose excitatory and inhibitory interconnections lead the entire network to converge to states corresponding to certain geometric constraints that must be satisfied when recognition succeeds. Marr and Poggio (1976) proposed such a 'cooperative' model for stereo vision that simultaneously finds the relative distance from the viewer of each feature in pair of stereoscopic images *and* which feature in one image corresponds with a given feature in the other. It does so by exploiting the constraints that each feature must have a single disparity and that neighboring features mostly have similar disparities.

In the case of three-dimensional shape recognition, Attneave, Hinton, and Hrechanyk and Ballard point out that there are constraints on how shape elements and reference frames may be paired that might be exploitable in parallel networks to arrive at both simultaneously. First, every part of an object must be described with respect to the same object-centered reference frame (or at least, every part of an object in a circumscribed region at a given scale of decomposition; see the discussion of the Marr and Nishihara model). For example, if one part is described as the front left limb of a animal standing broadside to the viewer and facing to the left, another part of the same object cannot simultaneously be described as the rear left limb of that animal facing to the right. Second, a description of parts relative to an object-centered frame is to be favored if that description corresponds to an existing object description in memory. For example, a horizontal part will be described as the downward-pointing leg of a chair lying on its back rather than as the forward-facing leg of an unfamiliar upright object.

These constraints, it is argued, can be used to converge on a unique correct object-centered description in a network of the following sort. There is a *retina-based unit* for every possible part at every retinal size, location, and orientation. There is also an *object-based unit* for every orientation, location, and size of a part with respect to an object axis. Of course, these units cannot be tied to *individual* retina-based units, but each object-based unit can be connected to the entire set of retina-based units that are geometrically consistent with it. Every shape description in memory consists of a *shape unit* that is connected to its constituent object-based units. Finally, all the pairs of

Figure 5. *A portion of a massively parallel network model for shape recognition. Triangular symbols indicate special multiplicative connections: the product of activation levels of a retina-based and a mapping unit is transmitted to an object-based unit, and the product of the activation levels in those retina-based and object-based units is transmitted to the mapping unit. From Hinton (1981).*



object- and retina-based units that correspond to a single orientation of the object axis relative to the viewer are themselves tied together by a *mapping unit*, such that the system contains one such unit for each possible spatial relation between object and viewer. An example of such a network, taken from Hinton (1981), is shown in Fig. 5.

The system's behavior is characterized as follows. The visual input activates retina-based units. Retina-based units activate all the object-based units they are connected to (this will include all object-based units that are geometrically compatible with the retinal features, including units that are inappropriate for the current object). Object-based units activate their corresponding shape units (again, both appropriate and inappropriate ones). Joint activity in particular retina- and object-based units activate the mapping units linking the two, that is, the mapping units that represent vantage points (relative to an object axis) for which those object-based features project as those retina-based features. Similarly, joint activity in retina-based and mapping units activate the corresponding object-based units. Shape units activate their corresponding object-based units; and (presumably) shape units inhibit other shape units and mapping units inhibit other mapping units. Hinton (1981) and Hrechanyk and Ballard (1982) argue that such networks should enter into a positive feedback loop converging on a single active shape unit, representing the recognized shape, and a single active mapping unit, representing the orientation and position of its axis with respect to the viewer, when a familiar object is viewed.

In general, massively parallel models are effective at avoiding the search problems that accompany serial computational architectures. In effect, the models are intended to assess the goodness-of-fit between all the transformations of an input pattern and all the stored shape descriptions in parallel, finding the pair with the highest fit at the same time. Since these models are in their infancy, it is too early to evaluate the claims associated with them. Among other things, it will be necessary to determine: (a) whether the model can be interfaced to preprocessing systems that segregate an object from its background and isolate sets of parts belonging to a single object-centered frame at a single scale; (b) whether the initial activation of object-based units by retina-based units is selective enough to head the network down a path leading toward convergence to unique, correct solutions; (c) whether the number of units and interconnections among units needed to represent the necessary combinations of shapes, parts, and their dispositions is neurologically feasible; and (d) whether these networks can overcome their current difficulty at representing and recognizing *relations* among parts in complex objects and scenes, in addition to the parts themselves.

Visual imagery

Visual imagery has always been a central topic in the study of cognition. Not only is it important to understand our ability to reason about objects and scenes that are remembered rather than seen, but the study of imagery is tied to the question of the number and format of mental representations, and of the interface between perception and cognition. Imagery may also be a particularly fruitful topic for study among the higher thought processes because of its intimate connection with shape recognition, benefitting from the progress made in that area. Finally, the subject of imagery is tied to scientific and literary creativity, mathematical insight, and the relation between cognition and emotion (see the papers in Sheikh, 1983); though the scientific study of imagery is mostly concerned with more pedestrian spatial abilities such as memory for literal appearance, spatial transformations, and matching images against visual stimuli, it has been argued that an understanding of these abilities may give us the toehold we need to approach the less tractable topics in the future (Shepard, 1978; Shepard and Cooper, 1982).

Imagery is in some ways a more difficult topic to study than recognition, and progress in the area is slower and consensus rarer. Unlike recognition, the direct inputs and outputs of the imagery system are not known beforehand; they must be discovered at the same time as the operation of the system itself. This has two implications. First, it is difficult to characterize antecedently the 'function' that the imagery system computes (cf. Marr, 1982). Second, there is the practical problem that the imagery system cannot be engaged automatically in experimental settings by presenting a person with particular physical inputs; imagery must be engaged through more indirect pathways involving a person's conceptual system (e.g., presenting him or her with certain types of spatial problems, giving instructions to use imagery). Thus it can be difficult to determine when the imagery is used or even whether there is a distinct imagery system.

Philosophical objections to imagery

During much of this century the coherence of the concept of an image itself has been called into question, and there have been claims that there is no such thing as an image—that talk of imagery is just a figure of speech (e.g., Ryle, 1949). However, most of the arguments within this debate really concerned the use of representations and processes in explanations of intelligence and mental life. Now that there is something close to a working consensus among cognitive scientists that intelligence can be characterized as computations over data structures or representations (see Block, 1980; Fodor, 1975; Haugeland, 1981; Pylyshyn, 1980), many of the criticisms of the concept of imagery are now moot, or at least can be absorbed into the debate over the representational and computational theories of mind in general (see Block (1981, 1983) for further arguments). In particular, I think it would be wise to avoid worrying about the following three non-issues:

(1) *The homunculus problem.* How can the mind contain 'images' unless there was some little man in the head to look at the images? This is simply not a problem under the computational theory of mind: images may be construed as data structures, and there is no more of a conceptual problem with positing mechanistic operations that can access those data structures than there is in positing mechanistic operations that access other mental representations such as linguistic or logical structures, or positing operations that access data in a computer. In particular, the study of shape recognition has led people to posit many types of operations that take as input array-like data structures created by sensory receptors, such as the 2½-D sketch, and it would be a short step to claim that the same processes could access such data structures generated from memory rather than from the eyes (whether or not this is *true* is, of course, a different matter).

(2) *The epiphenomenon problem.* Maybe images exist, but they are epiphenomenal to the actual computations that constitute thinking about visual objects. That is, they play no causal role, and are just like the lights that flash on and off on the front panel of a computer. The problem here is an ambiguity in the word 'image'. It could be taken to refer either to the subjective experience that people have when they report thinking in images, or to a mental representation that a theory might posit to help explain visual thinking. If 'image' is used in the former sense, then it is noncontroversially epiphenomenal if one subscribes to the computational theory of mind in general. In no computational theory of a mental process does subjective experience *per se* play a causal role; only representations and processes do, and the subjective experience, if it is considered at all, is assumed to be a correlate of the processing. If, on the other hand, 'image' is meant to refer to a representation posited in a theory of spatial memory and reasoning, then no one could hold that it is epiphenomenal: any theory positing a representation that never played a causal role would be rejected as unlikely to be true using ordinary parsimony criteria.

(3) *The subjectivity issue.* It is dangerous to take people's introspective reports about the contents or properties of their images as evidence in characterizing imagery, because such reports can be unreliable and subject to bias, because the referents of people's descriptions of imagery are unclear, and because interesting cognitive processes are likely to take place beneath the level of conscious accessibility. Again, a non-issue: all parties agree that much of image processing, whatever it is, is inaccessible to consciousness, that experimental data and computational and neurological feasibility are the proper source of constraints on imagery theories, and that introspective reports are psychological phenomena to be accounted for, not accurate descriptions of underlying mechanisms.

Imagery theories

I believe we have reached the point where it is particular *theories* of imagery, not analyses of the concept of imagery, that are in question, and that the 'imagery debate' is now a scientific controversy rather than a philosophical conundrum. This controversy largely concerns two questions. First, does the architecture of the mind contain any structures and processes that are specific to imagery, or does imagery simply consist of the application of general cognitive processes to data structures whose content happens to be about the visual world? Second, if it does make sense to talk of imagery as a dedicated module, does it consist of the processing of pixels in an array with properties similar to the 2½-D sketch, or does it consist of the processing of structural descriptions?

The most forceful proponent of the view that imagery is not a distinct cognitive module is Zenon Pylyshyn (1981). Pylyshyn argues that imagery simply consists of the use of general thought processes to simulate physical or perceptual events, based on tacit knowledge of how physical events unfold. For example, mental rotation effects (e.g., Shepard and Hurwitz (1984); also discussed below) occur because subjects are thinking in real time about the course of a physical rotation. They know implicitly that physical objects cannot change orientation instantaneously and that the time taken by a rotation is equal to the angle of rotation divided by the rotation rate. They perform the relevant computation, then wait the corresponding amount of time before indicating their response. Pylyshyn argues that virtually all known demonstrations of imagery phenomena, experimental or informal, can be interpreted as the use of tacit knowledge to simulate events rather than as the operation of a dedicated processor. In particular, he argues, the representation of space or of movement in images does not tell us anything about the format of imagery representation, least of all that there is anything 'spatial' or 'moving' in the image itself; the demonstrations just tell us about the content of information that can be represented in imagery.

Though Pylyshyn has not proposed an explicit model of the general purpose cognitive mechanisms that subserve imagery, the type of mechanism that would be congenial to his view would be a structural description. As mentioned, structural descriptions use a symbolic format similar to that used in the semantic networks proposed as representations of general knowledge.

A second class of theories has been proposed by Allan Paivio (1971) Roger Shepard (1981) and Stephen Kosslyn (1980, 1983) (see also Kosslyn *et al.* (1984, this issue) and Farah (1984, this issue)). They have proposed that imagery uses representations and processes that are ordinarily dedicated to visual perception, rather than abstract conceptual structures subserving thought in general. Furthermore, it is proposed that at least one of these representations used in perception and imagery has a spatial or array-like structure. By an array-like structure, I mean the following: images are patterns of activation in a structure consisting of units (or cells) that represent, by being on or off (or filled or unfilled) the presence or absence of a part or patch of the surface of an object at a particular disposition in space (orientation or location). The medium is structured so that each cell is adjacent to a fixed set of other cells, in such a way that the metric axioms are satisfied.[4] This makes it meaningful to speak of properties like 'position', 'proximity', and 'direction' of cells or pairs of cells; these properties are defined by the adjacencies among cells within the medium and need not correspond to physical position, distance, or direction within the neural instantiation of the array (though presumably they are related to proximity measured in number of neural connections). The processes that occur within that medium are sensitive to the location of a cell within the medium; that is, there are primitive operations that access a particular cell by absolute or relative location, that move the contents of one cell to a neighboring one within the medium, and so on. Furthermore, location within the medium is systematically related to the disposition in space of the represented object or feature, so that adjacent cells represent dispositions of the represented object or feature that differ by some minimal amount. An array of values in a computer memory functioning as a graphics bit map is an example of the type of array-like medium characterized in this paragraph, but such media can also be quite different from bit maps.[5]

Shepard (1981) has proposed an elegant though only partially worked-out set of conjectures, according to which a shape is represented by a two-dimensional manifold curved in three-dimensional space to form a closed surface, such as a sphere. Each position within the manifold corresponds to one orien-

[4] The metric axioms are (a) the distance between a point and itself is less than the distance between a point and any other point; (b) the distance between point *a* and point *b* is the same as the distance between point *b* and point *a*; (c) the distance between point *a* and point *b* plus the distance between point *b* and point *c* must be greater than or equal to the distance between point *a* and point *c*.

[5] Pylyshyn (1980, 1981) rightly emphasizes that it is crucial to distinguish among the *representation* of geometric properties like distance, position, and direction, the corresponding physical properties in the world themselves, and these properties defined over the surface of the brain. Pylyshyn accuses certain theorists of confusing these notions, but in my opinion Pylyshyn's argument loses its force by failing to acknowledge another sense of notions like distance, namely that defined by the intercell adjacencies in an array representation and respected by the processes that operate within such an array. According to the theories outlined in the text, position and distance in the array represent position and distance in the world, and possibly (depending on details of the neural instantiation of these mechanisms) rough position and distance within certain regions of the brain. Thus rather than confusing distance in the world, the internal representation of distance in the world, distance among cells in the internal structure representing the world, and distance in the brain, these theorists are making assertions about how these different senses of distance are related.

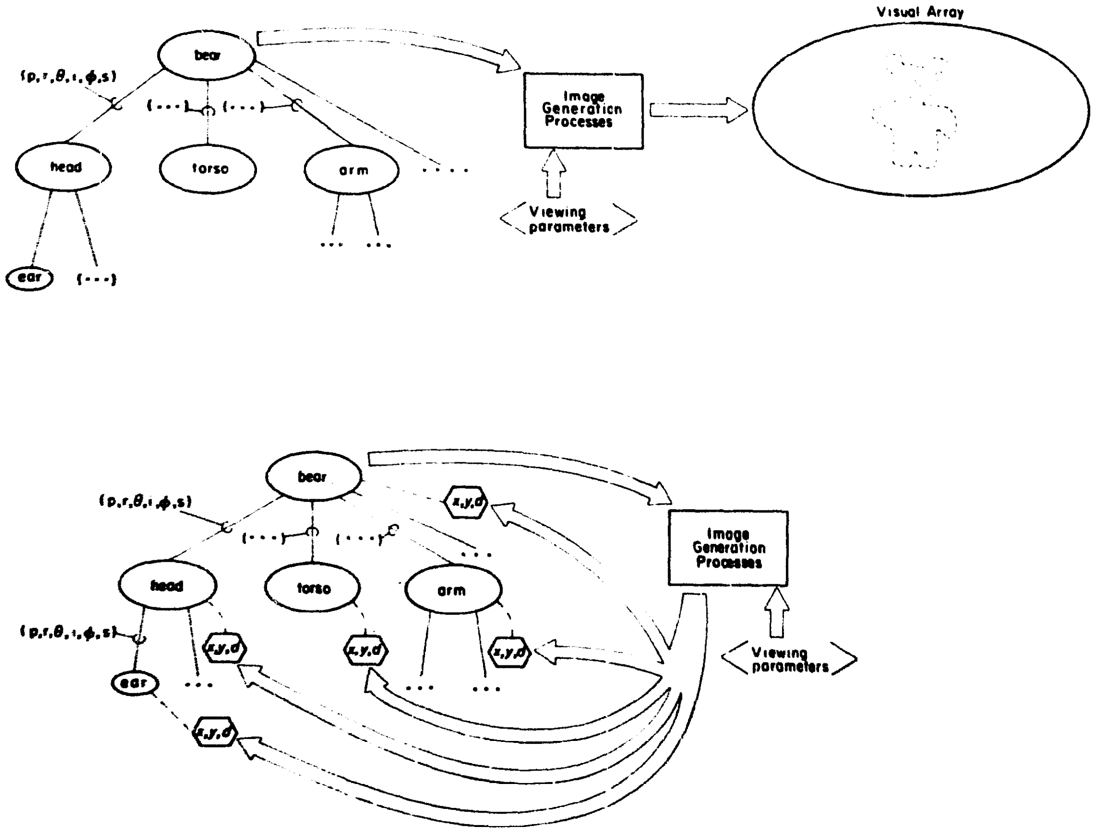
tation of the shape, with nearby positions corresponding to nearby orientations. The shape of the manifold captures the various symmetries in the represented object, so that the point representing the letter 'H' rotated 175° will be closer to the point representing its upright orientation than either of them is to the point representing its 90° orientation. When points are activated in the medium (i.e., when one perceives or imagines a shape at a particular orientation), activation spreads through the manifold, and when fronts of activation originating from different sources meet (i.e., when two views of a shape are seen or imagined), a straight-line path between the sources is activated, corresponding to the shortest angular trajectory that the shape would have to pass through to rotate one orientation into the other. Shepard uses this proto-model to make sense of a wide set of findings on mental rotation and apparent movement; see Shepard (1981) for details, and Pinker and Kosslyn (1983) for explication and commentary.

Kosslyn has proposed a different theory, instantiated in an explicit and detailed computational model (Kosslyn, 1980, 1983; Kosslyn and Shwartz, 1977; also briefly summarized in this issue by Kosslyn *et al.* (1984) and by Farah (1984)). Here the medium, which Kosslyn calls the "visual buffer", is two-dimensional and Euclidean, and the position of cells within the array corresponds to position within the visual field (i.e., a line of sight from the vantage point of the viewer; in this regard it is similar to a 2½-D sketch—see Pinker (1980b)). Cells, when activated, represent patches of the surface of a represented shape, so that the pattern of activation within the buffer is isomorphic to the shape of the visible surfaces of the object. As in Shepard's proposal, this medium can be filled from information arriving from the visual system (subsequent to preprocessing such as the detection of edges and surface properties); or from information in long-term memory—this is what generating a mental image consists of. In Kosslyn's theory the long term memory representations of objects' shapes and surface properties used in imagery are assumed to be shared with those used in recognition (see Farah (1984) for neuropsychological evidence that these representations are the same), and are assumed to have the format of a structural description augmented with whatever information is necessary to reconstruct the appearance of the surfaces of the object. The image generation processes (which are the focus of the article by Farah (1984)) can fill the buffer with patterns representing several objects from different long-term memory files, and can place the objects in different positions, orientations, and sizes with respect to one another. Once in the buffer, the pattern of activated cells can be rotated, scaled in size, or translated, and the resulting patterns can be examined by operations that detect shapes and spatial configurations (presumably, mechanisms similar to the visual routines proposed by Ullman

(1984) in this issue). The upper panel in Fig. 6 is an illustration of the general architecture of Kosslyn's model. (A slightly different hybrid model in this class has been proposed by Attneave (1974), who suggests that a two- or three-dimensional array can specify the locations of objects by containing labels or pointers to symbolic descriptions, rather than edge or surface primitives depicting the object's shape.)

Hinton (1979a, b) has proposed a model for imagery belonging to a third class. It shares the following assumptions with Kosslyn's model: there are processes dedicated to the manipulation of spatial information in imagery;

Figure 6. *Two ways of representing viewer-specific information in imagery. The top panel schematically illustrates an array model; the lower panel schematically illustrates a structural description model. The parameters within parentheses symbolize the disposition of a part in local coordinate systems centered on superordinate parts (see Fig. 2 and 3); parameters within hexagons in the lower panel symbolize the horizontal and vertical directions and depth with respect to a single vantage point.*



there is a special format for information represented in imagery, involving a global, viewer-centered reference frame; and there is an array-like scale within which the spatial disposition of the represented shape is specified. However, in Hinton's model there is no array whose cells correspond to local portions of the visual field and represent local patches of the shape. Rather, imagery consists of information appended to a structural description of the object's shape. In this long-term memory representation, units correspond to entire parts defined by a hierarchical decomposition of the shape of the object, and the spatial relations between adjacent parts are defined with respect to a frame of reference centered on one of the parts (thus it is similar to Marr and Nishihara's 3-D model description). Image generation consists of activating a certain set of part nodes, and appending to them a *second* representation of their spatial dispositions. This second set of coordinates is *not* specified with respect to a local, object-centered frame of reference, but with respect to a global reference frame centered on the viewer. A set of processes can then operate on these temporary viewer-centered coordinates appended to the activated nodes in order to ascertain spatial relations holding among parts of an object at different levels of its hierarchical decomposition or among parts of different objects. The lower panel in Fig. 6 is a sketch of the general properties of Hinton's model. An additional feature of Hinton's model is that the various quantitative parameters used in specifying spatial dispositions are encoded as pointers to one-dimensional arrays within which an activated cell represents a particular position, orientation, or size. (See Anderson (1983) who also presents a model that is centered around structural descriptions but which contains special processes for the manipulation of spatial information in images).

Current issues in the study of visual imagery

Distinguishing among the three classes of models just described is not the only focus of current research on imagery (for example, Kosslyn *et al.* (1984) and Farah (1984) examine the decomposition of imagery into modules, and do not directly address the format of the short-term memory structure underlying images). The following is a summary of some important current topics in imagery research; I will highlight how data bear on the models described above but will also describe classes of research that are independent of that controversy. For more general literature reviews on imagery research, see Kosslyn (1980), Kosslyn and Shwartz (1981), Shepard and Cooper (1982), and Finke and Shepard (In press).

Cognitive penetration

These phenomena are relevant to the distinction between theories appealing to tacit knowledge alone *versus* theories appealing to dedicated imagery processes. Pylyshyn (1979, 1981) cites cases where people's knowledge and belief may influence the time and ease with which they use imagery in certain tasks. Pylyshyn argues that if the operation of a putative processing module is sensitive to the contents of a person's beliefs, then it cannot be a primitive component of the cognitive architecture inasmuch as the mode of operation of the primitative architecture is by definition sensitive only to the syntactic form of representations, not their content. Thus sensitivity to the contents of beliefs is evidence that the process in question is a manifestation of whatever mechanisms manipulate the representation underlying knowledge in general.

Although the form of Pylyshyn's argument is generally accepted (all things being equal), its application to particular sets of data, especially in the case of imagery, is controversial. The problem is that the penetrability criterion pertains to individual information processing *components*, but we can only gather direct evidence that beliefs are penetrating individual *tasks* involving many components (Fodor, 1983; Kosslyn *et al.* 1979). If a person's beliefs influence the rate of mental rotation, the time to generate an image, and so on, it could simply be that the executive has access to certain parameters that can be set prior to the execution of an operation, such as transformation rate, decision criteria, or the choice of shapes to imagine or transformations to execute (for example, the rotation operator might have a rate parameter that can be set externally, but all other aspects of its operation might be fixed). Which processing stage is the one influenced by a person's beliefs makes all the difference in the world, but identifying such stages is as difficult as making any other claim about representations and processes based on experimental data. (There is also controversy over the facts of which imagery tasks actually are penetrable by beliefs; see e.g., Finke, 1980; Kosslyn, 1981; Pinker, Choate and Finke, 1984a; Reed *et al.*, 1983).

Constraints on imagery

If imagery is nothing but the use of tacit knowledge to simulate physical events, then the only constraints on what we can do in our images should stem from what we know can or cannot occur in the world. However, there have been many reports, some introspective, some experimental, that people cannot form images of arbitrary specifications of spatial properties and relations. For example, we cannot imagine a shape whose orientation, location, subjective size, or direction with respect to the vantage point are simply indeterminate, undefined, or unspecified; each image must make commitments to particular values of these parameters. Similarly, we cannot imagine two objects that are next to one another without one being to the left of the other; nor can we imagine the front and back of an opaque intact object simultaneously, nor the visual space in front of and behind the head (see e.g., Fiske *et al.* 1979; Hinton, 1979b; Pinker and Finke, 1980; Poincaré, 1913; Johnson-Laird, 1983; Keenan and Moore, 1979).

Note that these possible constraints stand in contrast to long-term memory for visual information in general. As Pylyshyn (1973) has pointed out, we often remember that an object was in a room or near another object without being able to recall where in the room or on what side of the object it was; similarly, Nickerson and Adams (1979) have shown that people are quite inaccurate at remembering spatial relations among the parts of familiar objects such as a Lincoln penny. The constraints also contrast with the optionality of other properties in imagery, such as color, surface texture, local parts, and details of edges, which are often reported as being totally unspecified in images. This means that the constraints are not just constraints on which properties are defined in the world, because just as an object must have an orientation when viewed, it must have a certain color and texture.

When there are constraints on which geometric properties are optional in images and which are obligatory, when these constraints hold of imagery in particular and not of long term memory about visual information in general, and when they are not predictable from physical and geometric constraints on objects in the world, we have evidence that imagery is represented by special mechanisms. In particular, in a structural description, any geometric attribute that can be factored out of a shape description (e.g., orientation, size, relative location) can be lost and hence undefined, and abstract spatial relations (e.g., 'next to') can be specified easily. In contrast, in an array model it is impossible to form an image representation lacking a size, absolute or relative location, or orientation, because shapes are represented only by being placed somewhere in the array medium, thereby receiving some specification of location, size, and so on automatically. Thus the constraints, if they are robust, would speak against a totally factored structural description.

Mental transformations and couplings among between geometric properties

The most famous body of research on imagery in the past decade has been concerned with image transformations (Cooper and Shepard, 1973; Shepard and Metzler, 1971; Shepard and Hurvitz, 1984); see also the section on "Assigning reference frames" above). Shepard and his collaborators have shown that when people have to decide whether two 3-D objects have the same shape, the time they take is a linear function of the difference in their depicted orientations. When they have to judge the handedness of alphanumeric characters or random polygons (i.e., whether one is normal or mirror-reversed), time increases monotonically with degree of deviation in orientation from the upright. Shepard's interpretation of these findings is that subjects engage in a smooth, continuous process of mental rotation, transforming the orientation of an imagined shape until it coincides in a template-like manner with a second, perceived shape or with a shape stored in a canonical upright orientation in memory. By itself, the increase in reaction time with orientation would not necessarily support the claim that a continuous rotation is imagined, but Shepard and Cooper have independent evidence that the rotation process computes intermediate representations in the angular trajectory (see Shepard and Cooper (1982) for a review). There have also been demonstrations of phenomena interpretable as mental translation or scanning (Finke and Pinker, 1982, 1983; Kosslyn *et al.*, 1978; Pinker *et al.*, 1984a), and size scaling (Bundesen and Larsen, 1975; Kosslyn, 1980).

In interpreting these data, it is important to separate two aspects of the phenomenon: why transformations are necessary at all, and why the transformations are gradual (e.g., why people take increasing time for greater orientation disparities, rather than simply taking a constant additional amount of time when there is any difference in orientation at all). I think that the *necessity* of performing image transformations tells us about which pairs of geometric attributes are obligatorily coupled in images, rather than being factored apart, leading to similar conclusions to those suggested in the previous section on imagery constraints. Consider the following structural description of a viewed object:

Object X:
Shape:
$\left[\text{(Object-centered 3-D model)}\right]$

Viewer-relative
location: (x, y, d)

Viewer-relative
orientation: (s, t)

Size: (z)

Now consider what would happen if one had to verify that two stimuli had the same shape, or whether one stimulus did or did not correspond in shape to a standard in memory. If the judgment could be made on the basis of structural descriptions such as this one, exploiting the explicit decoupling of geometric attributes in it, then depicted orientation should make no difference. All one has to do is examine the part of the structural description that specifies shape, and ignore the parts specifying orientation, location, and size. In fact, the factoring apart of orientation, location, and size in structural descriptions, allowing processes to ignore selectively the geometric attributes that are irrelevant to their tasks, is considered one of the chief selling points of this format. However, the facts of mental transformations indicate that this account cannot be completely correct: when judging shape, people are systematically affected by the irrelevant attributes of orientation and size. Similarly, when verifying whether an imagined object has a part, people are affected by the size of the object or of the part (Kosslyn, 1980); and when verifying whether one point lies in a certain direction with respect to another, they are affected by the distance between them (Finke and Pinker, 1983). There is also evidence that the imagined size of an object affects the rate of rotation (Shwartz, 1979; Pinker, unpublished data; though see also Bundesen *et al.*, 1981). Finally, when matching an imagined shape against a physically presented one, differences in orientation, size, location, and combinations of these differences all affect the speed of the match (Bundesen and Larsen, 1975; Cooper and Shepard, 1973; Farah, In press; Kosslyn, 1980; Shepard and Cooper, 1982).

The phenomena of mental image transformations, then, suggest that the completely factored structural description as shown above cannot be the one used in imagery. Exactly what about it is wrong is not completely clear. The phenomena are consistent with the spatial models of Shepard and Kosslyn in that in those models, shape and orientation (and, in Kosslyn's model, size and location) are *not* factored apart; they are 'in' one and the same set of activated cells. Thus the value of one attribute may affect the accessing of another when two representations differing in the first attribute are compared; the comparison process might be similar in some ways to template matching. Hinton and Parsons (1981) argue otherwise; they suggest that shape and orientation are factored apart in image representations except for the handedness of the object-centered reference frame, which is determined in the viewer-centered reference frame (see the earlier section on "Assigning reference frames"). Hence normalization of orientation is necessary whenever shapes must be discriminated from their mirror-reversed versions, the situation in most of the mental rotation experiments. Mental rotation also occurs, however, when the foils are not mirror-images (e.g., Cooper and Podgorny, 1976; Shwartz, 1981, Reference note 4); whether or not Hinton and Parson's account is correct in these cases will depend on whether the relevant shape representations depend on the handedness of their reference frame (e.g., whether random polygons are represented in terms of a list of their angles going clockwise from the topmost angle). Another, related possibility for the necessity of computing geometric transformations is that the shape description does not have a global specification of viewer-relative orientation, only specifications appended to each part. The description of the dispositions of the object's parts would change with orientation, requiring a normalization of orientation before shape can be extracted. In any case, the fact that mental transformations must occur when processing information in images indicates that images cannot consist of a viewpoint-invariant object-centered description plus a global specification of orientation.

Why the transformations must be executed incrementally is another issue. Logically there is no reason why the processes that update the represented orientation of a shape could not arrive at new orientations in one fell swoop. If orientation is simply a parameter appended to a shape description, one value could simply be replaced by another. Even in an array theory like Kosslyn's, the process that moves surface primitives from cell to cell (based on the coordinates of the first cell and the nature of the transformation) could calculate the coordinates of the target destination in one step rather than calculating a series of destinations separated by small increments.

Explanations for the gradualness of image transformations divide into three classes. In theories with an array component, neighboring cells are used to represent adjacent orientations and the orientation changes are accomplished by hardwired connections between cells within some bounded neighborhood (e.g., Hinton, 1979a; Shepard, 1981; Trehub, 1977). Since there are hardwired connections only between neighboring cells, larger transformations require the network of connections to be activated iteratively, transferring activation from initial to final state through intermediate states in bucket brigade fashion.[6]

The second class of account appeals not to constraints on the transformation mechanisms but on the executive processes that control it. For example, Kosslyn (1980) proposes that incremental transformations are selected because they minimize noise introduced into the depicted shape by the transformation operation, and because they allow a simple control strategy when the target orientation is not known in advance: the executive can monitor the successive representations and stop the transformation when the represented shape is familiar or matches some target (Marr (1982) makes a similar conjecture). In these accounts, the necessity of choosing gradual transformations

[6] The local nature of the wiring in these networks could either be an accidental consequence of principles of neural organization, or could have been selected during evolution to mirror the continuity of the motion of physical objects, as Shepard (1981) and Hayes-Roth (1979) have conjectured.

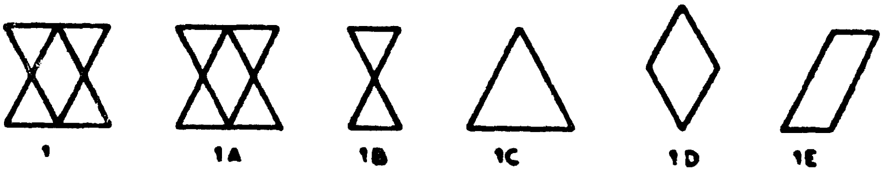
ultimately derives from the coupling of shape and orientation (size, location, etc.) in imagery, because only if they are coupled would the size of the transformation affect the process of recognizing a target shape.

The third class of account comes from Pylyshyn (1981) and Mitchell and Richman (1980), who also argue that it is executive processes that cause the transformation to be incremental, but attribute the source of this choice to a different factor, namely one's tacit knowledge that movement in the world is continuous and one's desire or tendency to simulate the time course of physical events. In particular, experiments in which subjects are told explicitly to transform a pattern (e.g., Kosslyn *et al.*, 1973); or in which subjects know that the experiment is about 'rotation', 'movement', and so on, are open to this explanation, it is argued, because subjects could literally construe their task as the mental simulation of motion. The account becomes less plausible in cases where subjects are left to their own devices and are asked to solve the task as quickly as possible, with no mention of imagery or of transformations (e.g., Cooper and Shepard, 1973; Finke and Pinker, 1982, 1983; Pinker *et al.*, 1984a). Instead, tacit knowledge, accounts would have to claim that subjects carry with them the habit of simulating physical events whenever they make spatial judgments. However, such an account would then need to explain what it is about the mind that would cause it to adopt such unnecessary habits (Kosslyn, 1981) and why mental transformations are not carried out when various sorts of advance information are provided to the subject (e.g., Cooper and Shepard, 1973; Finke and Pinker, 1983; Pinker *et al.*, 1984a).

Goodness and cohesiveness of parts

The definition of the primitive units represented in structural description and array theories is one of the key features that differentiate them. In structural descriptions, the primitives correspond to cohesive parts of objects and their dispositions with respect to reference frames centered on superordinate parts. In an array representation, the primitives correspond to local patches of surface or to edges, located with respect to a single reference frame centered on the viewer (see Fig. 6). Hence, in structural descriptions, it should be easy to make judgments about spatial relations among cohesive parts specified within the same reference frame, whereas in arrays, the parts do not have a special status and the difficulty of a judgment should be determined by factors such as distance and size rather than part cohesiveness or membership in a family of parts in the same reference frame. Thus it has been taken as evidence against array representations that it is extremely difficult to see parts in complex objects that are low in *Gestalt* 'goodness' or that do not correspond to one of the units in one's internal description or

Figure 7. *Effects of descriptions on the visualization of parts. Different parts are more or less easy to detect in an imagined pattern depending on whether the whole is conceptualized as two overlapping triangles, two overlapping parallelograms, two adjacent hourglasses, or a diamond contained in a large hourglass. From Reed (1974).*



conceptualization of the objects (Hinton, 1979a; Palmer, 1977; Reed, 1974; see also Fig. 7).

Unfortunately, these demonstrations become less than decisive in settling the array–*versus*–description issue because most explicit imagery theories have multiple mechanisms. For example, in Kosslyn's array theory, objects' shapes are stored in long-term memory as structural descriptions involving the objects' parts, and image generation consists of the successive activation in the array of patterns for those parts (Kosslyn *et al.*, 1983). Since imagined parts begin to fade as soon as they are generated, (Kosslyn, 1980), at any given time an image of a complex shape will contain only subsets of parts that were originally generated in close temporal proximity. These will usually be parts that have a distinct identity and that belong to the same reference frame in memory. Hence relations that cut across parts or reference frames will be difficult to make, just as in the structural description account. Conversely, in a structural description theory that allows a global set of viewer-centered coordinates to be appended to each part description (e.g., as shown in the lower panel of Fig. 6), relations among parts in different reference frames should *not* be difficult to perceive. In both theories, limitations on the number of parts kept active at one time, and the use of structural descriptions at one level of shape representation, are the source of the explanation for these phenomena. Discriminating among the theories will be easier when each makes more precise commitments as to how the limitations in capacity are measured, and what determines the choice of reference frames.

Imagery and perception

There is a large body of research exploring similarities between imagining a pattern and perceiving one (see Finke, 1980;Finke and Shepard, In press; Shepard and Podgorny, 1978; Shepard and Cooper, 1982). These include chronometric patterns in discriminating magnitudes of perceived and imagined objects; perceptual aftereffects following the formation of images, and parallels in the scaling of magnitudes or similarities among perceived and remembered stimuli. Finke (1980) and Finke and Shepaɪd (In press) discuss the relevance of this large body of findings to theories of imagery. Finke distinguishes among similarities between imagery and perception according to the locus of the relevant perceptual effect in the nervous system. For example, one can conceive of perceptual effects that are due to properties of the sensory receptors and lower-level feature analyzers and grouping processes (i.e., those processes leading to what Marr calls the Primal Sketch), those that are due to properties of the higher level analysis of objects' shapes, sizes, surface features, orientations, and so on; and those that are due to general knowledge and cognitive skills, such as a person's tacit knowledge of the workings of his or her perceptual system. Finke argues that among the phenomena that are similar in imagery and perception, there are some that can be attributed to the second of these three levels, that is, phenomena that reflect the operation of middle-level perceptual processes that are independent of a person's knowledge and beliefs. If so, it would argue that imagery is more than the application of one's general purpose knowledge about physical or perceptual processes.

In search of the imagery medium

The research most directly addressed to distinguishing between the array and structural description theories has attempted to discover evidence for the putative medium underlying images. According to the array theory there is a fixed medium underlying all visual images, regardless of their content; according to the alternative, the representation underlying the image of an object is simply the activated representation of that object's shape in memory (see Farah, 1984 for discussion). If an imagery medium exists, its parts should correspond to fixed points in the visual field, it should display the same eccentricity, isotropy, and contrast constraints as one would find in the corresponding perceptual medium, regardless of which objects are being represented, and it should have an identifiable locus or loci in the nervous system. There are inherent methodological difficulties in determining whether such a medium exists because people's tacit knowledge of their own perceptual experience could make it easy for them simply to remember how something originally appeared to them when forming an image, rather than forming a pure image and allowing the inherent structure of imagery representations to affect its "appearance" (Pylyshyn, 1981). Nonetheless, there have been several interesting lines of investigation, and attempts to overcome the methodological problems.

For example, Finke and Kosslyn (1980), Finke and Kurtzman (1981), and Pennington and Kosslyn (1981, Reference note 2; see the paper by Kosslyn *et al.* in this issue) have attempted to show that images have decreased resolution at peripheral eccentricities and oblique orientations, just like physical patterns, but that these effects were not known to the subjects, at least not consciously. Kosslyn (1983) has also shown that imagining a pattern at a particular location over a prolonged period makes it harder to imagine a new pattern at that location, as if a single neural substrate was being fatigued or habituated. Bisiach and Luzzati (1978) have shown that brain-injured patients suffering from attentional neglect of one visual hemifield also show signs of neglect in images of objects in that hemifield. For example, when imagining a piazza and describing the buildings facing them they fail to describe the buildings on one side—but fail to describe a different set of buildings, those that would be seen in the same part of the visual field, when imagining themselves facing in a different direction. And Farah (1984) argues that the process that converts long-term memory representations into the format corresponding to imagery can be selectively impaired and that this impairment is caused by damage only to certain parts of the brain.[7]

Computing and updating viewer-centered representations

The instrospective and experimental evidence on imagery suggests images represent the surfaces visible from a fixed vantage point, rather than just the intrinsic geometry of an object (see Hinton, 1979b; Pinker, 1980b; Pinker and Finke, 1980). The major models of imagery, both of the array and structural description variety, are designed to capture this fact. However, computing the dispositions and visible surfaces of parts from a description of their shape plus a specification of a vantage point, and updating that representation during image transformations, can be a nontrivial computational problem, especially as images are subject to transformations that alter the sets of surfaces that are visible (e.g., rotation in depth, translation or panning that brings new objects into view, zooming to reveal formerly blurred detail).

Furthermore, simply computing the rotation and translation transformations themselves can be problematic, especially in array theories. If the viewer-centered coordinates are specified in polar coordinates, then rotations in the frontal plane around the fovea are easy to compute, but not rotations in depth, about noncentral axes, or translations. If rectangular coordinates are used, sideways or up-and-down translations are easy to compute, but not diagonal translations or rotations. One possibility is that aside from the vie-

[7] Evidence on position-specific neglect following brain injury supports theories that have specific neural loci representing specific locations in the visual field. Hinton's (1979a) hybrid structural description model has this property, since location parameters are claimed to be given a 'place' representation in unidimensional neural arrays. There is, however, a prediction that differentiates Hinton's model from array theories: according to Hinton, entire objects or parts should be neglected when they are represented at particular locations, whereas according to Kosslyn, arbitrary chunks of a part (whatever material overlaps the missing region) can be inaccessible.

wer-centered coordinate system that defines the fixed 'address' of each array cell, rectangular and cylindrical coordinate systems can be dynamically centered on or linked to patterns for objects in the array, giving the cells a second, transient set of coordinates. Imagined transformations and the positioning of objects and parts with respect to one another could then be accomplished by processes that manipulate these transient, more appropriate coordinates (see the section on "Frames of reference for the visual field"; also Pinker, 1980c, Reference note 3; 1981; Trehub, 1977).

Thus the complexity of mechanisms needed in an imagery theory hinges crucially on which geometric transformations we actually can compute in generating and updating our images. It may turn out, for example, that the image generation processes (i.e., those discussed at length by Farah (1984) are not as powerful as one might think. I have found that subjects can accurately remember the projected positions of objects in scenes when they are told to visualize the way the scene looked from a vantage point that the subjects had actually experienced. However, when the subjects had to visualize the scene as it would appear from a different, hypothetical viewing distance, the subjects were unable to predict the projected positions of the objects accurately (unless a rich context framework was visible at the time of making the judgments). Instead, the subjects simply reconstructed the perspective they had actually witnessed and then uniformly scaled its size to approximate the novel perspective view (Pinker, 1983). In a different set of experiments, (Pinker *et al.*, 1984b), I found that subjects could not visualize in a single step the appearance of a three-dimensional object from an arbitrary viewing angle, even when they had advance information about the viewing angle. Instead they first visualized it in some canonical orientation, and then mentally rotated it into the target orientation.

These findings argue against an image generation process that takes as input an object-centered shape representation plus a specification of an arbitrary viewpoint, and computes as output the corresponding viewer-centered representation. Instead it suggests that the memory representation from which images are generated uses a viewer-centered format and that the generation process simply activates this representation intact (at which point image transformation processes could begin to operate). This could have interesting implications for the representation of shape in general. My experimental evidence suggests that long-term image representations are primarily viewer-centered. Both parsimony considerations and neurological evidence summarized by Farah (1984) suggest that the long term representations of shape used in recognition are the same as those used in imagery. And Marr and Nishihara argue that shape representations used in recognition are primarily object-centered. One of these three claims has to give.

What is imagery good for?

The question of whether we have cognitive representations and processes dedicated to imagery is closely related to the question of what advantages such representations bring to reasoning and thinking. If general knowledge systems could do everything that a putative imagery system could do and could do it as well or better, one would have to question why the imagery system is there. No one questions the need for memories and reasoning processes that deal with visual information or hypothetical scenes or physical interactions; the question is whether any of the special properties that have been attributed to the imagery system—such as sharing one of the representational media used in perception, representing information in a single viewer-centered coordinate system, conflating shape, orientation, and position, or executing transformations continuously—are computationally desirable.

There have been several conjectures about the possible usefulness of these and other properties of the imagery system. None of these conjectures is strongly grounded as yet in computational complexity theory or in experimental investigations, but all have a certain intuitive appeal and all are amenable to such investigations. Here is a sample of prominent suggestions:

(1) *Global coordinate system.* In the section on shape recognition, I reviewed compelling arguments for representing objects' shapes in distributed coordinate systems rather than a global one (see e.g., Fig. 2). Efficient though this scheme is for recognition, it can be problematical when one must reason about spatial relations among non-adjacent parts. Consider how one could answer a question such as "is the midpoint of a horse's tail higher or lower than its lips?" The position of the midpoint of the tail is specified with respect to the tail as a whole; the position, angle, and size of the tail are specified with respect to the torso; the position, angle, and size of the lips are specified with respect to the face, whose position, size, and angle are specified with respect to the head, whose position, size, and angle are specified with respect to the torso. One cannot simply look up any pair of coordinates to answer the question, because the coordinates of the midpoint of the tail and those of the lips have completely different meanings. Comparing the positions of the two objects could be done by transforming the geometric parameters of one of them into the coordinate system of the other, but that would require several successive coordinate transforms. Not only might such a series of transformations be time-consuming and error-prone, but noise in the representation of one of the parameters or noise introduced by the transformation processes could accumulate from transformation to transformation and the final comparison could be severely inaccurate. For example, errors of a few degrees in the representation of the angle of a giraffe's neck could lead to

large errors in the judgment of how far ahead of its feet its snout extends.

If the position, orientation, and size of each of a set of parts could be specified in terms of a single coordinate system, relations between parts within the same 3-D model would be a bit more difficult to compute, but relations among parts separated by many intervening models would be easier and more accurate. Hinton (1979b) has suggested that visual imagery consists of the use of a global viewer-centered reference frame to represent the dispositions of arbitrary collections of parts for the purpose of judging spatial relations among them (see Finke and Pinker (1983) for some relevant experimental evidence). Shepard and Hurvitz (1984) also point out that mental rotation can serve to bring one object into alignment with the reference frame of another, or into alignment with a viewer-centered reference frame, to facilitate the computation of reference-frame-specific predicates such as 'right' and 'left' (see also the section above on "Assigning reference frames").

This advantage is not tied to a single theory of imagery; it could be obtained whether the global coordinates are listed as tags on nodes in structural descriptions, of whether they are the addresses of the cells in an array medium occupied by the represented surfaces of the object (see Fig. 6). There are, however, diferences in the extent to which the advantages could be exploited in the two models; in an array model, the boundaries of various parts or details of their surfaces can be compared in the same global reference frame whenever an object or part is activated, whereas in the structural description account, only relations among entire parts are possible. Thus it has been argued that array representations are especially efficient when properties that cut across part boundaries, such as the empty space defined by a pile of objects on a table, must be computed (Funt, 1976; Hayes-Roth, 1979; see also Waltz, 1979).

(2) *Incidental propcrties.* When we learn about a new object, we record a number of facts about its structure, but not every potentially useful fact. For example, we do not record explicitly the shape of the ear of a sheep, or whether any of the parts of a car are triangular, or how many windows there are in one's house. Such information is often implicit in the information we do record, but it is simply not listed as an explicit proposition in memory. If there are routines analogous to the ones proposed in this issue by Ullman (1984) that can recognize these properties from visual input, then perhaps they can also recognize them from information stored in memory that is similar in form to the visual input. That is, we might store a relatively uncommitted, literal record of the appearance of objects from which we can compute properties that we could not anticipate the need for knowing when we initially saw the object. Kosslyn (1980) reports an extensive series of experiments and intuitive demonstrations showing that imagery is used when people are re-

quired to answer questions about parts of objects that are 'poorly associated' with the objects (that is, parts that we are not likely to have thought of in connection with that object before) and not deducible from the properties of the superordinate class to which the object belongs. Thus "does a bee have a dark head" requires the use of imagery (as assessed by effects of size and scanning, as well as by introspective reports), but "does a bee have a stinger" or "does a bee have wheels" do not.

(3) *Incremental representation in perception.* Ullman (1984) argues that certain types of visual recognition are solved by routines that can add information to the visual input, yielding "incremental representations" that subsequently are used by that routine or by other routines (e.g. 'coloring' regions, marking objects). Though Ullman suggests that visual routines are fast, unconscious, low-level recognition operations, it is also possible that some of these operations are identical to what are often called imagery operations (sequences of which, presumably, can become automatized with practice and hence could become fast and unconscious). Thus it is noteworthy that Ullman's operation of 'boundary tracing', which is similar to the operations of mental scanning and extrapolation proposed in connection with imagery, appears to occur at the same rate (in degrees of visual angle per second) as image scanning (Finke and Pinker, 1983; Jolicoeur *et al.*, 1984b). It is also possible that the generation and template-like matching of images against perceptual input (see Shepard and Cooper (1982) for a review) is a visual routine that can be used to recognize objects when one number of a small set of candidate objects is expected and when it is difficult to compute a 3-D model for the input. This could happen if the input pattern lacks stable axes or a stable decomposition into parts, when it must be discriminated from mirror-reversed versions, or when the input is severely degraded.

(4) *Reasoning in isomorphic domains.* If images are representations in a medium with certain fixed properties, and can be subjected to transformations such as rotation and scaling, then imagery could be used as an analogue computer, to solve problems whose entities and relations are isomorphic to objects and spatial relations. That is, certain abstract problems could best be solved by translating their entities into imagined objects, transforming them using available image transformations, detecting the resulting spatial relations and properties, and translating those relations and properties back to the problem domain.

For example, if every imagined object is constrained to have a single set of coordinates within a global coordinate system (e.g., the array proposed by Kosslyn), then it is impossible to represent the fact that one object is next to another without also committing oneself to which is to the left. (This is not true for abstract propositional representations, where the two-place predicate "next to" can be asserted of a pair of objects with no such commitment.) Furthermore, if there is a routine that can compute which of two objects is to the left of another on the basis of its global coordinates or the position of the cells it occupies in the array, then transitivity of left-to-right position falls out of the transitivity of the coordinates or of cell position within the array. The result of these properties is that problems such as three-term syllogisms (e.g., John is nobler than Bill, Sam is less noble than Bill, who is the noblest?) can be solved straightforwardly by imagining an object for each entity in the problem and placing them to the left and right of one another in an order corresponding to the dimension of comparison (Huttenlocher, 1968; Shaver *et al.*, 1975; see also Johnson-Laird, 1983).

Shepard (1978) and Shepard and Cooper (1982) also note that the use of imagery in mathematical and scientific problem solving may be effective because the medium used to represent images and the operations transforming them might embody physical and geometric constraints on terrestrial objects and space. When there are isomorphisms between physical objects in space and other domains (e.g., electromagnetic fields and elastic lines or surfaces), imagining a concrete analogue of an entity, transforming it, and then translating it back to the original domain could make explicit certain properties and equivalences in that domain that were only implicit beforehand.

A concluding remark

In this tutorial review, I have made no effort to conceal the disagreement and lack of resolution that surrounds many of the issues discussed. This should be taken not as a sign of disarray, but as a sign of the vigor of a newly revitalized branch of cognitive psychology. After a period of relative stagnation, researchers in visuospatial cognition are striving to synthesize a large number of new empirical findings, theoretical constructs, and external constraints. The past decade has seen Marr's important statements of the problems that visual recognition must solve and of the criteria of adequacy for theories of shape representation, his introduction into discussions in visual cognition of physical, optical, and geometric constraints that the visual system can exploit, and his concrete proposals on several classes of visual representations and algorithms. It has also witnessed a burgeoning of experimental data on imagery and recognition made possible by the chronometric methodology of Shepard and Kosslyn; tentative resolutions of the principal conceptual objections to theories of visual representations; the development of explicit computations and neural models of processes and structures that were previously characterized only in vague metaphors; and the application to visual imagery of concepts used in shape recognition such as distributed coordinate systems, object- and viewer-centered reference frames, and the 2½-D sketch. Most recently, we have seen an exposition of properties of alternative computational architectures, including the massively parallel systems that visual processing surely requires at some levels. Theories and data in visual cognition are being applied for the first time to neighboring disciplines that previously had been largely insulated from theoretical considerations, such as computer vision systems, individual difference psychology, and neuropsychology, and these disciplines are now in a position, in turn, to inform basic research on visual cognition. There are disagreements and confusion over specifics, to be sure, and syntheses between independent bodies of research that have yet to be made, but it seems clear what the problems to be solved are, what sorts of data and arguments are relevant, and what degrees of precision and explicitness it is reasonable to hope for in our theories.

References

Anderson, J.A. and Hinton, G.E. (1981) Models of information processing in the brain. In Hinton, G.E. and Anderson, J.A. (eds.), *Parallel Models of Associative Memory*. Hillsdale, NJ, Erlbaum.

Anderson, J.R. (1978) Arguments for representations for mental imagery. *Psychol. Rev.*, *85*, 249–277.

Anderson, J.R. (1983) *The Architecture of Cognition*. Cambridge, MA, Harvard University Press.

Attneave, F. (1968) Triangles as ambiguous figures. *Am. J. Psychol.*, *81*, 447–453.

Attneave, F. (1972) Representation of physical space. In A.W. Melton and E.J. Martin (eds.), *Coding Processes in Human Memory*. Washington, DC, V.H. Winston.

Attneave, F. (1974) How do you know? *Am. Psychol.*, *29*, 493–499.

Attneave, F. (1982) Pragnanz and soap bubble systems: A theoretical exploration. In J. Beck (ed.), *Organization and Representation in Perception*. Hillsdale, NJ, Erlbaum.

Badler, N. and Bajcsy, R. (1978) Three-dimensional representations for computer graphics and computer vision. *Comp. Graph. 12*, 153–160.

Ballard, D. and Brown, C. (1982) *Computer Vision*. Englewood Cliffs, NJ, Prentice Hall.

Ballard, D., Hinton, G.E. and Sejnowski, T.J. (1983) Parallel visual computation. *Nature*, *306*, 21–26.

Biederman, I. (1972) Perceiving real-world scenes. *Science*, *177*, 77–80.

Binford, T.O. (1971) Visual perception by computer. Presented to the IEEE conference on Systems and Control, December, Miami.

Bisiach, E. and Luzzatti, G.R. (1978) Unilateral neglect of representational space. *Cortex*, *14*, 129–133.

Block, N. (ed.) (1980) *Readings in Philosophy of Psychology*, *Vol. 1* Cambridge, MA, Harvard University Press.

Block, N. (ed.) (1981) *Imagery*. Cambridge, MA, MIT Press.

Block, N. (1983) Mental pictures and cognitive science. *Phil. Rev.*, *92*, 499–542.

Boring, E.G. (1952) Visual perception as invariance. *Psychol. Rev.*, *59*, 142–150.

Bundesen, C.C. and Larsen, A. (1975) Visual transformation of size. *J. exp. Psychol.: Hum. Percep. Perf.*, *1*, 214–220.

Bundesen, C.C., Larsen, A. and Farrell, J.E. (1981) Mental transformations of size and orientation. In A.D. Baddeley and J.B. Long (eds.), *Attention and Performance, Vol. 9*. Hillsdale, NJ, Erlbaum.

Campbell, F.W. and Robson, J.G. (1968) Application of Fourier analysis to the visbility of gratings. *J. Physiol.*, *197*, 551–566.

Cooper, L.A. and Shepard, R.N. (1973) Chronometric studies of the rotation of mental images. In W.G. Chase (ed.), *Visual Information Processing*. New York, Academic Press.

Cooper, L.A. and Podgorny, P. (1976) Mental transformations and visual comparison processes: Effects of complexity and similarity. *J. exp. Psychol.: Hum. Percep. Perf.*, *2*, 503–514.

Corballis, M.C. and Beale, I.L. (1976) *The Psychology of Left and Right*. Hillsdale, NJ, Erlbaum.

Corcoran, D.W.J. (1977) The phenomena of the disembodied eye or is it a matter of personal geography? *Perception*, *6*, 247–253.

Cornsweet, T.N. (1970) *Visual Perception*. New York, Academic Press.

Cutting, J.E. and Millard, R.T. (1984) Three gradients and the perception of flat and curved surfaces. *J. exp. Psychol.: Gen.*, *113*, 221–224.

Farah, M.J. (1984) The neurological basis of mental imagery: A componential analysis. *Cog.*, *18*, 245–272.

Farah, M.J. (In press) Psychophysical evidence for a shared representational medium for visual images and percepts. *J. exp. Psychol.: Gen.*

Feldman, J.A. and Ballard, D.H. (1982) Connectionist models and their properties. *Cog. Sci.*, *6*, 205–254.

Finke, R.A. (1980) Levels of equivalence in imagery and perception. *Psychol. Rev.*, *87*, 113–132.

Finke, R.A. and Kosslyn, S.M. (1980) Mental imagery acuity in the peripheral visual field. *J. exp. Psychol.: Hum. Percep. Perf.*, *6*, 244–264.

Finke, R.A. and Kurtzman, H.S. (1981) Area and contrast effects upon perceptual and imagery acuity. *J. exp. Psychol.: Hum. Percep. Perf.*, *7*, 825–832.

Finke, R.A. and Pinker, S. (1982) Spontaneous mental image scanning in mental extrapolation. *J. exp. Psychol.: Learn. Mem. Cog.*, *8*, 142–147.

Finke, R.A. and Pinker, S. (1983) Directional scanning of remembered visual patterns. *J. exp. Psychol.: Learn. Mem. Cog.*, *9*, 398–410.

Finke, R.A. and Shepard, R.N. (In press) Visual functions of mental imagery. In L. Kaufman and J. Thomas (eds.), *Handbook of Perception and Human Performance*. New York, Wiley.

Fiske, S.T., Taylor, S.E., Etcoff, N.L. and Laufer, J.K. (1979) Imaging, empathy, and causal attribution. *J. exp. soc. Psychol.*, *15*, 356–377.

Fodor, J.A. (1975) *The Language of Thought*, New York, Crowell.

Fodor, J.A. (1983) *Modularity of Mind*. Cambridge, MA, MIT Press/Bradford Books.

Funt, B.V. (1976) WHISPER: a computer implementation using analogues in reasoning. Ph.D. Thesis, University of British Columbia.

Gardner, M. (1967) *The Ambidextrous Universe*. London, Allen Lane/Penguin Books.

Gibson, J.J. (1950) *The Perception of the Visual World*. Boston, Houghton-Mifflin.

Gibson, J.J. (1966) *The Senses Considered as Perceptual Systems*. Boston, Houghton-Mifflin.

Gibson, J.J. (1979) *The Ecological Approach to Visual Perception*. Boston, Houghton-Mifflin.

Gilinsky, A. (1955) The effect of attitude on the perception of size. *Am. J. Psychol.*, *68*, 173–192.

Ginsburg, A.P. (1971) Psychological correlates of a model of the human visual system *Proceedings of the IEEE National Aerospace and Electronics Conference*, 283–290.

Ginsburg, A.P. (1973) Pattern recognition techniques suggested from psychological correlates of a model of the human visual system. *Proceedings of the IEEE National Aerospace and Electronics Conference*, 309–316.

Gregory, R.L. (1970) *The Intelligent Eye*. London, Weidenfeld and Nicholson.

Haugeland, J. (ed.) (1981) *Mind Design: Philosophy, Psychology, Artificial Intelligence*. Montgomery, VT, Bradford Books.

Hayes-Roth, F. (1979) Understanding mental imagery: Interpretive metaphors versus explanatory models. *Behav. Br. Sci.*, *2*, 553–554.

Hinton, G.E. (1979a) Some demonstrations of the effects of structural descriptions in mental imagery. *Cog. Sci.*, *3*, 231–250.

Hinton, G.E. (1979b) Imagery without arrays. *Behav. Br. Sci.*, *2*, 555–556.

Hinton, G.E. (1981) A parallel computation that assigns canonical object-based frames of reference. Proceedings of the International Joint Conference on Artificial Intelligence, Vancouver, Canada.

Hinton, G.E. and Anderson, J.A. (eds.) (1981) *Parallel Models of Associative Memory*. Hillsdale, NJ, Erlbaum.

Hinton, G.E. and Parsons, L.M. (1981) Frames of reference and mental imagery. In A. Baddeley and J. Long (eds.), *Attention and Performance IX*. Hillsdale, NJ, Erlbaum.

Hoffman, D.D. (1983) Representing shapes for visual recognition. Doctoral dissertation, MIT.

Hoffman, D.D. and Richards, M. (1984) Parts of recognition, *Cog.*, *18*, 65–96.

Hollerbach, J.M. (1975) Hierarchical shape description of objects by selection and modification of prototypes. MIT Artificial Intelligence Laboratory Technical Report #346.

Horn, B.K.P. (1975) Obtaining shape from shading information. In P.H. Winston (ed.), *The Psychology of Computer Vision*. New York, McGraw-Hill.

Hrechanyk, L.M. and Ballard, D.H. (1982) A connectionist model of form perception. Proceedings of the IEEE Special Workshop on Computer Vision, Rindge, NH.

Huttenlocher, J. (1968) Constructing spatial images: A strategy in reasoning. *Psychol. Rev.*, *75*, 550–560.

Jackendoff, R. (1983) *Semantics and Cognition*. Cambridge, MA, MIT Press.

James, W. (1890; reprinted 1980) *Principles of Psychology*. New York, Holt, Rinehart and Winston.

Johnson-Laird, P.N. (1983) *Mental Models*. Cambridge, MA, Harvard University Press.

Jolicoeur, P., Gluck, M.A. and Kosslyn, S.M. (1984a) Pictures and names: Making the connection. *Cog. Psychol.*, *16*, 243–275.

Jolicoeur, P., Ullman, S. and Mackay, M.E. (1984b) Boundary tracing: A possible basic operation in the perception of spatial relations. Research Bulletin, Department of Psychology, University of Saskatchewan.

Julesz, B. (1971) Experiments in the visual perception of texture. *Scient. Am.*, *232*, 34–43.

Kabrisky, M. (1966) *A Proposed Model for Visual Information Processing in the Human Brain*. Urbana, IL, University of Illinois Press.

Keenan, J.M. and Moore, R.E. (1979) Memory for images of concealed objects: A re-examination of Neisser and Kerr. *J. exp. Psychol. Hum. Learn. Mem.*, *5*, 374–385.

Koffka, K. (1935) *Principles of Gestalt Psychology*. New York, Harcourt Brace Jovanovich.

Kohler, W. (1947) *Gestalt Psychology*. New York, Mentor/Liveright.

Kosslyn, S.M. (1980) *Image and Mind*. Cambridge, MA, Harvard University Press.

Kosslyn, S.M. (1981) The medium and the message in mental imagery: a theory. *Psychol. Rev.*, *88*, 46–66.

Kosslyn, S.M. (1983) *Ghosts in the Mind's Machine*. New York, Norton.

Kosslyn, S.M., Ball, T.M. and Reiser, B.J. (1978) Visual images preserve metric spatial information: evidence from studies of imagery scanning. *J. exp. Psychol.: Hum. Percep. Perf.*, *4*, 47–60.

Kosslyn, S.M., Brunn, J., Cave, K. and Wallach, R. (1984) Individual differences in mental imagery ability: A computational analysis. *Cog.*, *18*, 195–243.

Kosslyn, S.M., Pinker, S., Smith, G.E., and Shwartz, S.P. (1979) On the demystification of mental imagery. *Behav. Br. Sci.*, *2*, 535–548.

Kosslyn, S.M., Reiser, B.J., Farah, M.J. and Fliegel, S.L. (1983) Generating visual images: units and relations. *J. exp. Psychol. Gen.*, *112*, 278–303.

Kosslyn, S.M. and Shwartz, S.P. (1977) A simulation of visual imagery. *Cog. Sci.*, *1*, 265–295.

Kosslyn, S.M. and Shwartz, S.P. (1981) Empirical constraints on theories of mental imagery. In A.D. Baddeley and J.B. Long (eds.), *Attention and Performance, Vol. 9*. Hillsdale, NJ, Erlbaum.

Kuipers, B. (1978) Modeling spatial knowledge. *Cog. Sci.*, *2*, 129–141.

Lindsay, P.H. and Norman, D.A. (1977) *Human Information Processing: An Introduction to Psychology*, 2nd ed. New York, Academic Press.

Lynch, K. (1960) *The Image of the City*. Cambridge, MA, MIT Press.

Marr, D. (1982) *Vision*. San Francisco, Freeman.

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

Marr, D. and Poggio, T. (1977) Cooperative computation of stereo disparity. *Science*, *194*, 283–287.

McDermott, D. (1980) Spatial inferences with ground, metric formulas on simple objects. Research report #173, Yale University Department of Computer Science.

Metzler, J. and Shepard, R.N. (1975) Transformational studies of the internal representation of three-dimensional objects. In R. Solso (ed.), *Theories in Cognitive Psychology: The Loyola Symposium*. Hillsdale, NJ, Erlbaum.

Minsky, M. (1975) A framework for representing knowledge. In P.H. Winston (ed.), *The Psychology of Computer Vision*. New York, McGraw-Hill.

Minsky, M. and Papert, S. (1972) *Perceptrons*, 2nd ed. Cambridge, MA, MIT Press.

Mitchell, D.B. and Richman, C.L. (1980). Confirmed reservations: mental travel. *J. exp. Psychol. Hum. Percep. Perf.*, *6*, 58–66.

Natsoulas, T. (1966) Locus and orientation of the perceiver (ego) under variable, constant, and no perspective instructions. *J. Person. Soc. Psychol.*, *3*, 190–196.

Neisser, U. (1967) *Cognitive Psychology*. New York, Appleton-Century-Crofts.

Nickerson, R.S. and Adams, M.J. (1979) Long-term memory for a common object. *Cog. Psychol.*, *3*, 287–307.

Norman, D.A. and Rumelhart, D.E. (1975) *Explorations in Cognition*. San Francisco, Freeman.

Palmer, S.E. (1975a) Visual perception and world knowledge: Notes on a model of sensory–cognitive interaction. In D.A. Norman and D.E. Rumelhart (eds.), *Explorations in Cognition*. San Francisco, Freeman.

Palmer, S.E. (1975b) The effects of contextual scenes on the identification of objects. *Mem. Cog.*, *3*, 519–526.

Palmer, S.E. (1977) Hierarchical structure in perceptual representation. *Cog. Psychol.*, *9*, 441–474.

Palmer, S.E. and Bucher, N. (1981) Configural effects in perceived pointing of ambiguous triangles. *J. exp. Psychol. Hum. Percep. Perf.*, *7*, 88–114.

Paivio, A. (1971) *Imagery and Verbal Processes*. New York, Holt, Rinehart and Winston.

Persoon, E. and Fu, K. (1974) Shape description using Fourier descriptives. *Proceedings of the Second International Joint Congress on Pattern Recognition*, 126–130.

Pinker, S. (1980a) Mental imagery and the visual world. Occasional Paper #4. MIT Center for Cognitive Science.

Pinker, S. (1980b) Mental imagery and the third dimension. *J. exp. Psychol. Gen.*, *109*, 354–371.

Pinker, S. (1981) What spatial representation and syntax acquisition don't have in common. *Cog.*, *10*, 243–248.

Pinker, S. (1983) Perspective information in visualized scenes. Paper presented at the Annual Meeting of the Psychonomic Society, San Diego, November 17–19, 1983.

Pinker, S., Choate, P. and Finke, R.A. (1984a) Mental extrapolation in patterns constructed from memory. *Mem. Cog.*.

Pinker, S. and Finke, R.A. (1980) Emergent two-dimensional patterns in images rotated in depth. *J. exp. Psychol. Hum. Percep. Perf.*, *6*, 69–84.

Pinker, S. and Kosslyn, S.M. (1983) Theories of mental imagery. In A.A. Sheikh (ed.), *Imagery: Current Theory, Research and Application*. New York, Wiley, 1983.

Pinker, S., Stromswold, K., and Beck, L. (1984b) Visualizing objects at prespecified orientations. Paper presented at the annual meeting of the Psychonomic Society, San Antonio, November.

Poggio, T. (1984) Vision by man and machine. *Scient. Am.*, *250*, 106–116.

Poincaré, H. (1913) *The Foundations of Science*. Lancaster, PA, Science Press.

Pribram, K.H. (1971). *Languages of the Brain*. Englewood Cliffs, NJ, Prentice-Hall.

Pylyshyn, Z.W. (1973) What the mind's eye tells the mind's brain: a critique of mental imagery. *Psychol. Bull.*, *80*, 1–24.

Pylyshyn, Z.W. (1979) The rate of "mental rotation" of images: a test of a holistic analogue hypothesis. *Mem. Cog.*, *7*, 19–28.

Pylyshyn, Z.W. (1980) Computation and cognition: issues in the foundations of cognitive science. *Behav. Br. Sci.*, *3*, 382–389.

Pylyshyn, Z.W. (1981) The imagery debate: Analogue media versus tacit knowledge. *Psychol. Rev.*, *88*, 16–45.

Reed, S.K. (1974) Structural descriptions and the limitations of visual images. *Mem. Cog.*, *2*, 329–336.

Reed, S.K., Hock, H.S. and Lockhead, G. (1983) Tacit knowledge and the effect of pattern configuration on mental scanning. *Mem. Cog.*, *11*, 137–143.

Richards, W. (1979) Natural computation: Filling a perceptual void. Presented at the 10th Annual Pittsburgh Conference on Modeling and Simulation, April, University of Pittsburgh.

Richards, W. (1982) How to play twenty questions with nature and win. MIT Artificial Intelligence Laboratory Memo # 660.

Rock, I. (1973) *Orientation and Form*. New York, Academic Press.

Rock, I. (1983) *The Logic of Perception*. Cambridge, MA, MIT Press/Bradford Books.

Rock, I., di Vita, J. and Barbeito, R. (1981) The effect on form perception of change of orientation in the third dimension. *J. exp. Psychol. Hum. Percep. Perf.*, *7*, 719–732.

Rosch, E., Mervis, C.B., Gray, W., Johnson, D. and Boyes-Braem, P. (1976) Basic objects in natural categories. *Cog. Psychol.*, *8*, 382–439.

Ryle, G. (1949) *The Concept of Mind*. London, Hutchinson.

Selfridge, O.G. (1959) Pandemonium: A paradigm for learning. In *Symposium on the Mechanization of Thought Processes*. London: HMSO.

Selfridge, O.G. and Neisser, U. (1960) Pattern recognition by machine. *Scient. Am.*, *203*, 60–68.

Shaver, P., Pierson, L. and Lang, S. (1975) Converging evidence for the functional significance of imagery in problem-solving. *Cog.*, *3*, 359–376.

Sheikh, A.A. (ed.) (1983) *Imagery: Theory, Research, and Application*. New York, Wiley.

Shepard, R.N. (1978) The mental image. *Am. Psychol.*, *33*, 125–137.

Shepard, R.N. (1981) Psychophysical complementarity. In M. Kubovy and J. Pomerantz (eds.), *Perceptual Organization*. Hillsdale, NJ, Erlbaum.

Shepard, R.N. and Cooper, L.A. (1982) *Mental Images and their Transformations*. Cambridge, MA, MIT Press/Bradford Books.

Shepard, R.N. and Hurwitz, S. (1984) Upward direction, mental rotation, and discrimination of left and right turns in maps, *Cog.*, *18*, 161–193.

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

Shepard, R.N. and Podgorny, P. (1978) Cognitive processes that resemble perceptual processes. In W.K. Estes (ed.), *Handbook of Learning and Cognitive Processes, Vol. 5*. Hillsdale, NJ, Erlbaum.

Shwartz, S.P. (1979) Studies of mental image rotation: Implications for a computer simulation of visual imagery. Doctoral dissertation, The Johns Hopkins University.

Smith, E.E., Balzano, G.J. and Walker J. (1978) Nominal, perceptual, and semantic codes in picture categorization. In J.W. Cotton and R.L. Klatzky (eds.), *Semantic Factors in Cognition*. Hillsdale, NJ, Erlbaum.

Stevens, K.A. (1981). The information content of texture gradients. *Biol. Cybernet.*, *42*, 95–105.

Trehub, A. (1977) Neuronal models for cognitive processes: Networks for learning, perception and imagination. *J. theor. Biol.*, *65*, 141–169.

Waltz, D. (1979) On the function of mental imagery. *Behav. Br. Sci.*, *2*, 569–570.

Weisstein, N. (1980) The joy of Fourier analysis. In C.S. Harris (ed.), *Visual Coding and Adaptability*. Hillsdale, NJ, Erlbaum.

Weisstein, N. and Harris, C.S. (1974) Visual detection of line segments: An object superiority effect. *Science*, *186*, 752–755.

Winston, P.H. (1975) Learning structural descriptions from examples. In P.H. Winston (ed.), *The Psychology of Computer Vision*. New York, McGraw-Hill.

Uhlarik, J., Pringle, R., Jordan, K. and Misceo, G. (1980) Size scaling in two-dimensional pictorial arrays. *Percep. Psychophys.*, *27*, 60–70.

Ullman, S. (1979) *The Interpretation of Visual Motion*. Cambridge, MA, MIT Press.
Ullman, S. (1984) Visual routines. *Cog.*, *18*, 97–159.

Reference Notes

1. Kubovy, M., Turock, D., Best, T.L. and Marcus, J. (1984) The virtual vantage point for the identification of cutaneous patterns. Unpublished paper. Rutgers University.
2. Pennington, N. and Kosslyn, S.M. (1981) The oblique effect in mental imagery. Unpublished manuscript, Harvard University.
3. Pinker, S. (1980c) The mental representation of 3-D space. Unpublished manuscript. MIT.
4. Shwartz, S.P. (1981) The perception of disorientated complex objects. Unpublished manuscript. Yale University.

Resumé

Cet article est une revue didactique sur les questions essentielles de la cognition visuelle. Il est centré sur la reconnaissance des formes et sur la représentation des objets et des relations spatiales en perception et en imagerie. L'auteur donne d'abord un bref rapport sur l'état de la question puis fait une présentation plus approfondie des théories contemporaines, des données et des prospectives. Il discute différentes théories de la reconnaissance des formes telles que les descriptions structurales en termes de patrons, traits, Fourier, Marr–Nishihara, et les modèles parallèles. Il discute aussi les propositions du type cadres de reference, primitifs, traitements de haut en bas et architectures de calcul utilisées dans la reconnaissance spatiale. Suit une discussion sur l'imagerie mentale ou sont abordés les concepts utilisés dans les recherches sur l'imagerie, les théories de l'imagerie, les rapports entre imagerie et perception, les transformations d'image, les complexités de calcul dans le traitement des images, les questions neurologiques et le rôle fonctionnel possible de l'imagerie. On insiste sur les relations entre les théories de la reconnaissance et l'imagerie ainsi que sur la pertinence des articles de ce volume sur ces sujets.