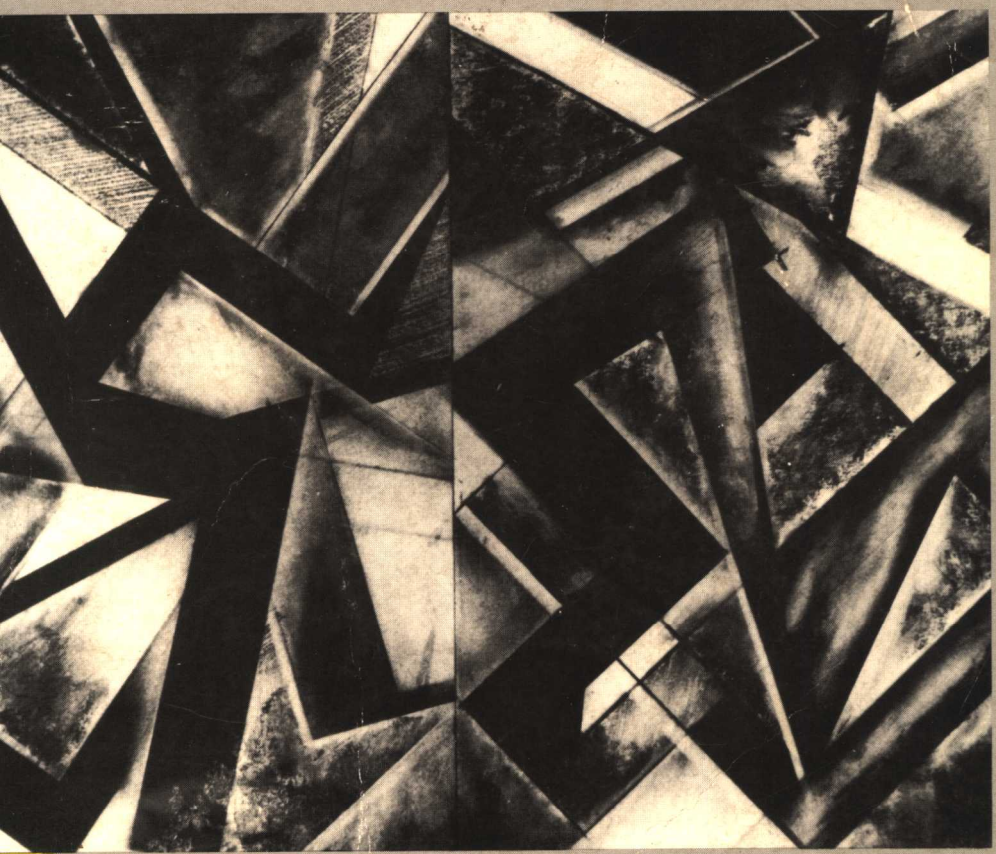
VISUAL COGNITION

edited by Steven Pinker



A COGNITION Special Issue

Visual Cognition

edited by
Steven Pinker

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

Second printing, 1986
First MIT Press edition, 1985

Reprinted from *Cognition: International Journal of Cognitive Psychology*, volume 18 (ISSN: 0010-0277). The MIT Press has exclusive license to sell this English-language book edition throughout the world.

Printed and bound in the United States of America

Library of Congress Cataloging in Publication Data

Main entry under title:

Visual cognition.

(Computational models of cognition and perception)
"A Bradford book."
"Reprinted from Cognition: international journal of cognitive psychology, volume 18"—T.p. verso.
Bibliography: p.
Includes index.
1. Visual perception. 2. Cognition. I. Pinker, Steven, 1954– . II. Series.
BF241.V564 1985 153.7 85-24155
ISBN 0-262-16103-6

Preface

This collection of original research papers on visual cognition first appeared as a special issue of *Cognition: International Journal of Cognitive Science*. The study of visual cognition has seen enormous progress in the past decade, bringing important advances in our understanding of shape perception, visual imagery, and mental maps. Many of these discoveries are the result of converging investigations in different areas, such as cognitive and perceptual psychology, artificial intelligence, and neuropsychology. This volume is intended to highlight a sample of work at the cutting edge of this research area for the benefit of students and researchers in a variety of disciplines. The tutorial introduction that begins the volume is designed to help the nonspecialist reader bridge the gap between the contemporary research reported here and earlier textbook introductions or literature reviews.

Many people deserve thanks for their roles in putting together this volume: Jacques Mehler, Editor of *Cognition*; Susana Franck, Editorial Associate of *Cognition*; Henry Stanton and Elizabeth Stanton, Editors of Bradford Books; Kathleen Murphy, Administrative Secretary, Department of Psychology, MIT; Loren Ann Frost, who compiled the index; and the ad hoc *Cognition* referees who reviewed manuscripts for the special issue. I am also grateful to Nancy Etcoff, Stephen Kosslyn, and Laurence Parsons for their advice and encouragement.

Preparation of this volume was supported by NSF grants BNS 82-16546 and BNS 82-19450, by NIH grant 1RO1 HD 18381, and by the MIT Center for Cognitive Science under a grant from the A. P. Sloan Foundation.

Contents

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.

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, individual 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-dimesional 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