

An Invitation to Cognitive Science
Daniel N. Osherson, Series Editor

Volume 1: Language

Edited by Lila R. Gleitman and Mark Liberman

Volume 2: Visual Cognition

Edited by Stephen M. Kosslyn and Daniel N. Osherson

Volume 3: Thinking

Edited by Edward E. Smith and Daniel N. Osherson

Volume 4: Conceptual Foundations

Edited by Saul Sternberg and Donald Scarborough

Visual Cognition

An Invitation to Cognitive Science
Second Edition

Volume 2

Edited by Stephen M. Kosslyn and
Daniel N. Osherson

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

Chapter 4

Visual Object Recognition

Irving Biederman

Try this experiment. Turn on your television with the sound off. Now change channels with your eyes closed. At each new channel, blink quickly. As the picture appears, you will typically experience little effort and delay (though there is some) in interpreting the image, even though it is one you did not expect and even though you have not previously seen its precise form. You will be able to identify not only the textures, colors, and contours of the scene but also the individual objects and the way in which the objects might be interacting to form a setting or scene or vignette. You will also know where the various entities are in the scene, so that you would be able to point or walk to any one of them if you were in the scene. Experimental observations confirm these subjective impressions (Intraub 1981; Biederman, Mezzanotte, and Rabinowitz 1982). People can usually interpret the meaning of a novel scene from a 100-millisecond (msec) exposure to it. However, they cannot attend to every detail; they attend to some aspects of the scene—objects, creatures, expressions, or actions—and not others. In this chapter, we focus primarily on our ability to recognize an object in a single glance on the basis of its shape.

Before we review the research and theory on object recognition, we will consider just what kinds of things a theory of object recognition might account for.

4.1 The Problem of Object Recognition

Object recognition is the activation in memory of a representation of a stimulus class—a chair, a giraffe, or a mushroom—from an image projected by an object to the retina. We would have very little to talk about in this chapter if every time we viewed an instance of a particular class it projected exactly the same image to the retina, as occurs, for example,

The writing of this chapter was supported by grants from the U.S. Air Force Office of Scientific Research (90-0274) and the McDonnell-Pew Foundation Program in Cognitive Neuroscience (T89-01245-029).

when the digits on a bank check are presented for reading by an optical scanner.

4.1.1 Pattern Variability

But there is a fundamental difference between reading digits on a check and recognizing objects in the real world: An object's orientation in depth can vary greatly, so that any one three-dimensional object can project an infinity of possible images onto the two-dimensional retina. We might see the object not only from a novel orientation but also when it is partially occluded by other surfaces—for example, behind foliage or draperies. Or the image of the object might fall on a different part of the retina or be of a different size. An object may be a novel instance of its class that does not exactly correspond to our previous experience, as, for example, when we view a new model of a chair or car. It is precisely this variation—and the apparent success of our visual system and brain at achieving recognition in the face of it—that makes the problem of pattern recognition so interesting.

4.1.2 Level of Classification

When we defined object recognition in the previous section as "the activation in memory of a stimulus class . . ." we did not specify just what constitutes a class. If we look at an elephant, we can classify it at many levels of abstraction: as an entity, as a living thing, as an animal, as an elephant, as an Asian elephant, as Jumbo. You probably feel that elephant is the most natural class. But why?

Linguists have developed the concept of *basic level* to refer to the initial classification given to individual visual entities, for example a chair, a bird, or a mushroom. When shown a picture of a sparrow, most people answering quickly call it a bird not a sparrow or an animal. The basic level (bird) is a level of abstraction of visual concepts that maximizes between-category distinctiveness and within-category informativeness (Rosch et al. 1976). It can be distinguished from *subordinate* (sparrow) and *superordinate* (animal) levels of classification. Most of our knowledge of the visual world can be accessed through the basic level. Specifying the subordinate-level class—for example, that something is an African versus an Indian elephant or is a particular style of sofa—provides only a slight increase in informativeness at an enormous loss of distinctiveness. That is, the difference between an African and an Asian elephant is much smaller (and less significant) than the difference between an elephant and a sofa. (Face recognition, a special form of subordinate-level recognition, is discussed in Chapter 3.) The superordinate level, which classifies something as, for example, an animal or an article of furniture, sacrifices informativeness with only a

slight gain in distinctiveness. The difference between (the classes) animals and furniture may be slightly greater than the differences, say, between an elephant and a rabbit or a sofa and a lamp, but that slight gain in distinctiveness comes at an enormous loss in the additional information we obtain from knowing that something is a lamp and not just an article of furniture, or an elephant and not just an animal. Basic-level terms are the first to enter a child's vocabulary, are used to a much greater extent than any other terms to describe objects, and are the highest level of abstraction whose objects share a characteristic shape (Rosch et al. 1976).

There are exceptions to the finding that people classify images more rapidly at the basic than at subordinate levels. Although a picture of a sparrow is classified as bird rather than a sparrow, a picture of a penguin is classified more quickly as a penguin than as a bird (Jolicoeur, Gluck, and Kosslyn 1984). The same holds true for ostrich, duck, and a number of other atypical instances of basic-level categories. Jolicoeur and his colleagues coined the term *entry level* to accommodate cases in which exemplars are initially classified at what would be, technically, a subordinate-level class. To a large extent, these exceptions have a different shape than the typical instances of the basic-level category. Some bird books display silhouettes of the entry-level subfamilies—ducks, songbirds (the prototypical basic-level class), hawks, and so on—to key the sections containing the subordinate-level information. In this chapter, we focus on the classification of an image into entry-level classes but also consider how subordinate-level classification might be accomplished.

4.1.3 How Many Entry-Level Objects Have to Be Modeled?

There are approximately three thousand entry-level terms for familiar concrete objects that can be identified on the basis of their shape rather than on surface properties of color or texture or on their position in a scene. These criteria, therefore, eliminate terms such as *fur* or *sand*. I arrived at this estimate by calculating the average number of entries per page meeting the criteria on a random sample of dictionary pages and multiplying it by the number of pages in the dictionary (Biederman 1987). This procedure yielded an estimate of approximately 1,600 terms, a result roughly consistent with linguists' estimates of the number of entry-level terms and naming words in the vocabulary of a six-year-old child. (A child of this age has a vocabulary of about ten thousand words, 10 percent of which are concrete nouns.) I then doubted this value to allow for idiosyncratic classes and objects not captured by the dictionary sample for a rough estimate of three thousand entry-level terms (or classes).

There may be an average of ten perceptual models for each of the three thousand entry-level, shape-based classes because (a) most objects require a few models for different orientations (such as the front and back of a

house), and (b) some entry-level terms (such as lamp, house, or chair) have several readily distinguishable object models (Biederman 1988). Six-year-old children reveal full adult competence in naming the objects in their visual world; indeed, they often achieve naming competence by the age of three. As the six-year-old has been awake for about thirty thousand hours, my estimate indicates that the child learns a new object model at a rate of one per waking hour.

4.2 Representing the Image

The initial sensing of visual information is performed by the photosensitive cells (rods and cones) of the retina, which are activated by individual photons reflected by an object. Each receptor responds to photons from only a tiny portion (a few minutes of arc) of the visual field. The exact same pattern of receptor activation is never duplicated from one occasion of looking at an object to the next. Indeed, as noted in section 4.1.1, recognition can be quite tolerant of the considerable variability in an object's image caused by differences in viewpoint or occlusion. The object does not even have to be identical to one seen previously for us to achieve relatively effortless classification of its image. How is the activity of individual photoreceptors employed by the brain to create a representation of an object that allows it to be recognized under such highly varied conditions?

4.2.1 Representation of Shape Information in V1

Ganglion-cell neurons arising in the retina synapse with neurons in the lateral geniculate body that in turn send fibers to V1, the first visual cortical area to receive information employed for shape perception. Although the cells in the retina and lateral geniculate body have a center-surround organization—in that they respond best to a spot of light (or darkness) at an extremely small area (a few minutes of arc in the fovea) of the visual field—*simple cells* in V1, the first cortical visual area, respond to variation in luminance at a particular orientation (e.g., to a bar at a vertical orientation but not at a horizontal or oblique orientation). The tuning to orientation could arise from a mapping in which a V1 cell receives inputs from a collinear array of geniculate cells.

Simple cells respond to a restricted region (e.g., 0.5 to 2 degrees) of the visual field, for example a vertical dark "bar" with light-colored flanks 1 degree in length and 0.3 degrees in width that is centered 2 degrees left of fixation. (Cells that are tuned to larger-scale variations in luminance; e.g., a bar 1.5 degrees in length, would have larger receptive fields.) *End-stopped cells* in V1 respond maximally to an oriented stimulus (such as a bar) only

if the stimulus terminates within the receptive field of the cell. End-stopped cells would presumably be maximally activated by contours that end at corners (or vertices).

Activation of simple and end-stopped cells are generally believed to provide the initial cortical activity of shape representation. Indeed, it would be possible to distinguish different shapes according to the differential activity of such cells. However, the identical shape presented at another position, size, or orientation also activates different cells; so we need some basis for representing shape that is not dependent on the particular V1 cells activated.

4.2.2 Invariant-Image Description

By a number of theoretical accounts, two related problems have to be addressed in order to form a representation allowing for invariant recognition. One problem is *grouping* or *binding*. When viewing an object like the one shown in Figure 4.1, we subjectively group contours *a* and *b* as part of one component, the brick, and *c* and *d* as part of another component, the cone, even though *a* and *c* are closer together and more similar in orientation than *a* and *b*. What principles allow such grouping?

A second problem is that of invariant description. It is particularly useful to have a representation that is the same whatever the viewpoint. We could then be fairly confident of what the object in Figure 4.1 looks like when it is rotated 30 degrees in depth. What information do we need to

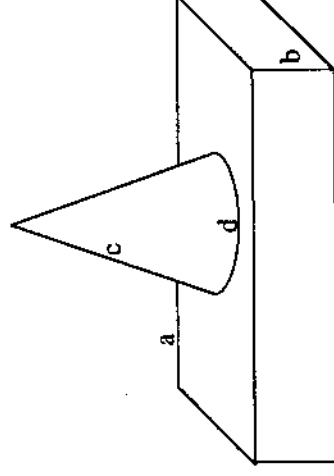


Figure 4.1

A vertical cone on a horizontal brick. This chapter concerns how we identify this image even though we probably have never seen it before. Why do we group segments *a* and *b* as part of one entity and *c* and *d* as part of another, despite the greater proximity of *a* and *d* (or *a* and *c*) and their greater similarity in orientation? (Adapted by permission of the publisher and authors from J. E. Hummel and I. Biederman, *Dynamic binding in a neural network for shape recognition*, 1987, *Psychological Review* 99, Figure 10, 489. Copyright 1987 by the American Psychological Association.)

do this? Objects in the real world have color, texture, and surface markings, but these sources of information are absent in the figure. There is some evidence that our capacity to recognize an object from different viewpoints is dependent on discontinuous edges of two types. Edges marking *orientation discontinuities* are formed by a sharp change in the orientation of abutting surfaces, such as occurs with the adjacent sides of a brick (segments *b* or *d* in Figure 4.1). Edges marking *depth discontinuities* are typically formed when one's line of sight grazes (that is, is tangent to) a curved surface so that there is a sharp jump in depth from surface to the background, as occurs with segment *c* in Figure 4.1. Sometimes the two types of edges coincide, as they do in segment *a*. A line drawing representing only these kinds of discontinuities (which can arise from differences in luminance, texture, color, etc.) can convey much of the three-dimensional shape of an object, as Figure 4.1 readily demonstrates. But how is it that a line drawing can convey the shape of an object? Or the fact that Figure 4.1 is a cone on top of a brick?

Subsections 4.2.3 and 4.2.4 describe the image information that might be employed to solve problems of viewpoint invariance and part structure. In section 4.3 we review theories about how neural computations might exploit this information.

4.2.3 Viewpoint-Invariant Properties

Viewpoint-invariant properties play a significant role in deriving a three-dimensional world from a two-dimensional image. Figure 4.2 illustrates several properties of image edges that are extremely unlikely to be a consequence of the particular alignment of eye and object. If the observer changes viewpoint or the edge or edges change orientation, assuming that the same region of the object is still in view, the image will still reflect that property. For example, a straight edge in the image is perceived as being a projection of a straight edge in the three-dimensional world. The visual system ignores the possibility that a (highly unlikely) accidental alignment of eye and a curved edge is projecting the image. Hence such properties have been termed *nonaccidental* (Lowe 1984). On those rare occasions when an accidental alignment of eye and edge does occur—for example, when a curved edge projects an image that is straight—a slight alteration of viewpoint or object orientation readily reveals that fact.

Figure 4.2 illustrates several nonaccidental properties. In the two-dimensional image, if an edge is straight (collinear) or curved, it is perceived as a straight or curved edge, respectively. If two or more two-dimensional image edges terminate at a common point, or are approximately parallel or symmetrical, then the edges projecting those images are similarly interpreted. For reasons that will be apparent when we consider some theories of object recognition, Figure 4.2 presents these viewpoint-

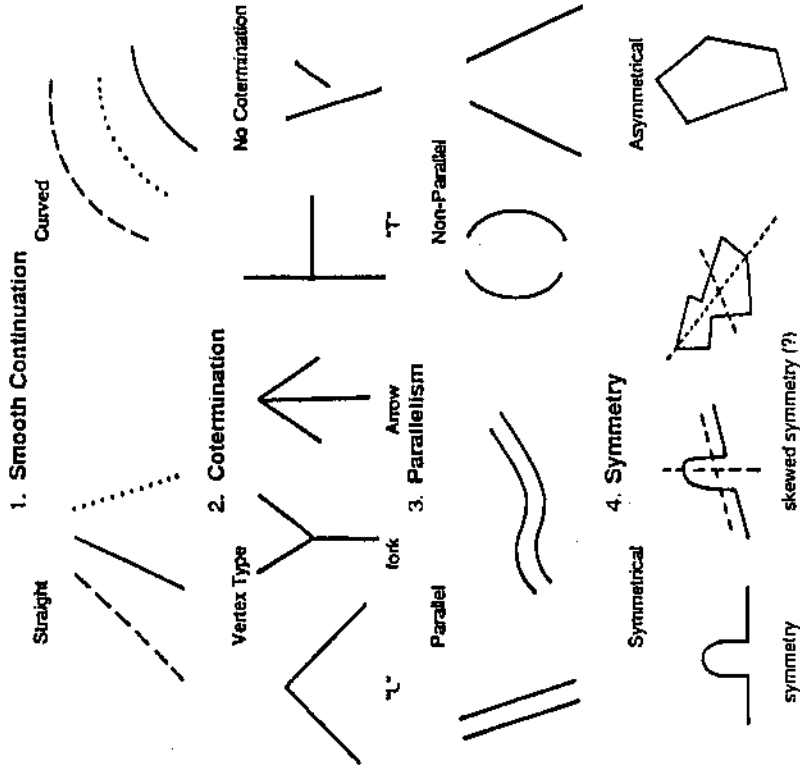
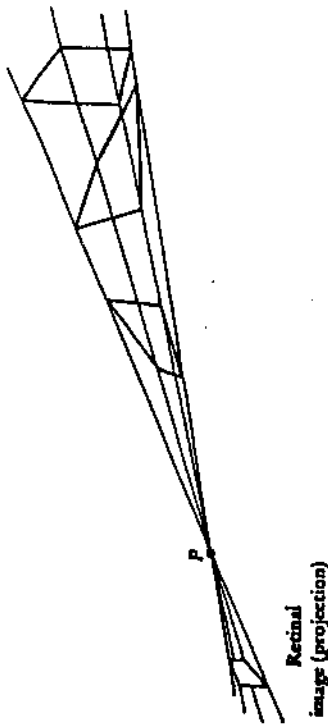


Figure 4.2

Contrasts in some viewpoint-invariant relations. In the case of parallelism, biases toward parallel and symmetrical percepts when images are not exactly parallel or symmetrical are evidenced. (Adapted by permission from D. Lower, Perceptual organization and visual recognition, Unpublished doctoral dissertation, Stanford University, 1984, Figure 4.2, 77.)

invariant properties as dichotomous *contrasts* (or differences). Any one edge can be characterized as straight or curved. We can describe the relation of two or more edges as coterminating or noncoterminating or parallel or nonparallel. The number of coterminating edges and whether they contain an obtuse angle also does not vary with viewpoint and can serve as a viewpoint-invariant classification of vertex type—L, Y (or fork), or arrow (or their curved counterparts) in Figure 4.2. In a strict sense, parallelism and symmetry varies with viewpoint and orientation, as occurs, for example, with perspective convergence. But there is a clear bias toward interpreting approximately parallel edges as parallel, especially when the surfaces are perceived as varying in depth (Ittelson 1952; King, Meyer, Tangney, and Biederman 1976). Within a tolerance range defined by the



Ambiguous Image (Inverse Optics Problem)

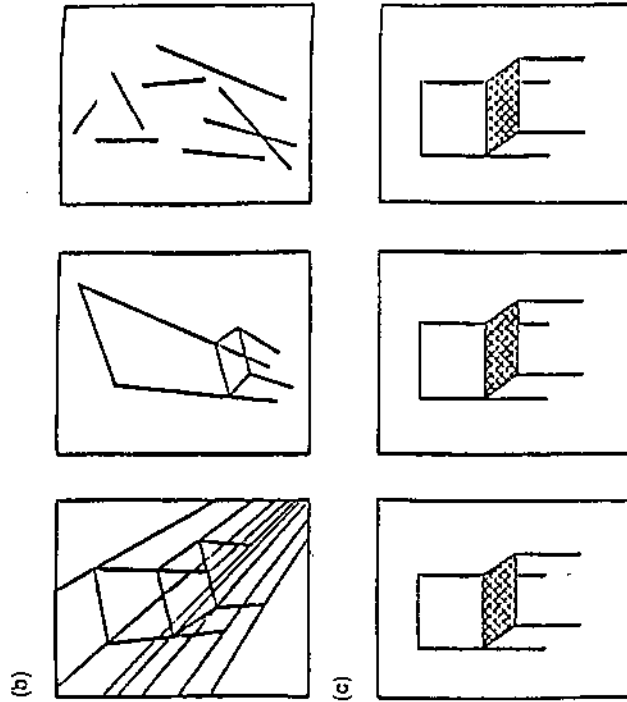


Figure 4.3
The Ames peephole perception demonstrations. (a) Illustration of the inverse-optics problem: A single image can be produced by an infinity of possible real-world objects. (b) Three stimulus arrangements constructed by the Ames group. The upper left panel shows the perspective lines from the peephole at the lower right. (c) The percepts from the stimuli in the upper panels. The three stimulus arrangements produce identical percepts. (Adapted by permission of publisher and author from R. N. Haber and M. Hershenson, *The psychology of visual perception*, 1981, Figure 12.5, 284. Copyright 1981 by Holt, Rinehart, & Winston.)

cues for surface slant, pairs of image edges that could be parallel or symmetrical, given uncertainty as to the actual orientation of the edges to the eye, are interpreted as parallel or symmetrical (King et al. 1976), as suggested by Figure 4.3.

The psychological potency of these viewpoint-invariant properties was demonstrated when Ames and his associates constructed a set of peephole perception demonstrations in which subjects viewed three arrangements of wires through a peephole, as shown in Figure 4.3b (Ittleson 1952). Although all three stimulus arrangements shown projected the identical image of a chair, as shown in Figure 4.3c, in only one of them (the left-hand one) did the wires actually form a chair. In the middle arrangement the segments all had the same cotermination points as the chair, except that the surfaces were no longer parallel. In the right-hand arrangement the segments did not even coterminate, yet the perception of this stimulus was indistinguishable from the other two. (Peephole viewing eliminates cues for stereoscopic vision, motion parallax, and image variation that would have resolved the accidents of viewpoint.) These results provide strong evidence that the viewpoint-invariant properties shown in Figure 4.1 and the biases toward parallelism and symmetry are immediate and compelling and could thus serve as a basis for characterizing image edges for purposes of recognition.

4.2.4 Decomposing Complex Objects into Parts

Complex visual entities almost always invite a decomposition of their elements into simple parts. We readily distinguish the legs, tail, and trunk of an elephant or the shade from the base of a lamp. People's spontaneous descriptions of basic-level classes almost always include a specification of distinctive parts (Tversky and Hemenway 1984). The manner of the decomposition into parts does not depend on familiarity with the object in that different observers agree on the part decompositions of nonsense shapes (Biederman 1987; Connell 1985; Kimia, Tannenbaum, and Zucker 1992). Nor does the part decomposition depend on surface color or texture as the part structure is readily perceived in line drawings.

In general, whenever there is a pair of matched cusps (discontinuities at minima of negative curvature), people will express a strong intuition that the object should be segmented at that region (Connell 1985). This tendency of the visual system to segment complex objects at regions of matched concavities is not an arbitrary bias. Hoffman and Richards (1985) note a result from projective geometry—the *transversality principle*—that whenever two shapes are combined, their join is almost always marked by matched cusps, as illustrated in Figure 4.4a. (The cusp projects an L-vertex that will be largely viewpoint invariant.) Segmenting at such regions provides a basis for appreciating the part structure of objects, as shown for the

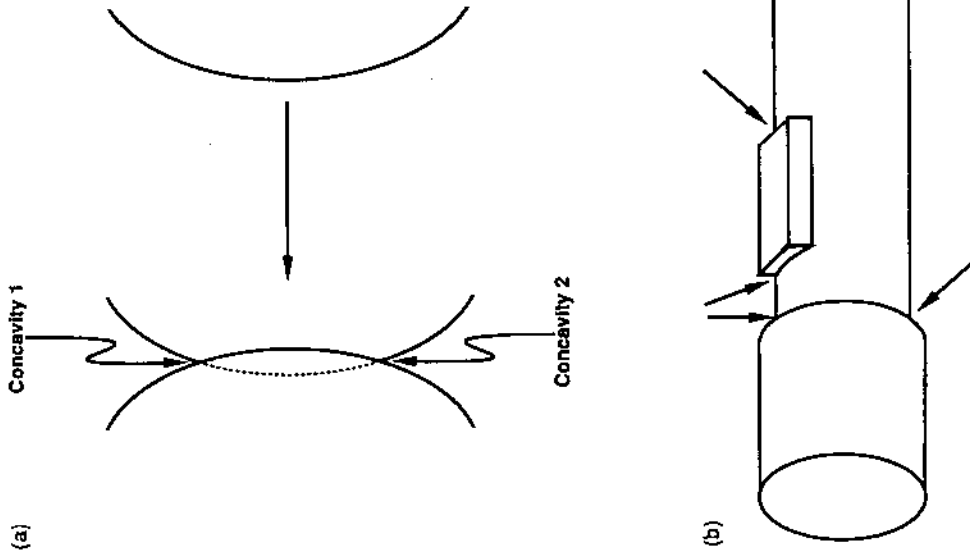


Figure 4.4
An illustration of the transversality regularity and how it can be applied to the segmentation of an object's parts.

flashlight in Figure 4.4b. Siddiqi, Tresness, and Kimia (1994) provide evidence that a narrowing of a shape without minima of negative curvature, which they call a *neck*, provides another basis for part decomposition. Indeed, an animal's neck provides a natural parsing region for separating the shoulders from the head. Matched cusps (or, more weakly, minima at negative curvature) and necks may provide much of the basis for the Gestalt principle of a good figure. If a shape is segmented at paired cusps or necks, the resulting parts will be convex or only singly concave. Such parts appear simple.

4.3 Theories of Object Recognition

Two major problems must be addressed by any complete theory of object recognition. The first is how to represent that information in the image so that it can activate a representation in memory under varied conditions. The second problem is how that stimulus representation is matched against—or indexes or activates—a representation of an object in memory.

With respect to the issue of representation of information in the image, different theories can be ordered along a continuum, according to the degree to which the image information is elaborated prior to matching (Dickinson, Pentland, and Rosenfeld 1992). Dickinson and colleagues refer to this continuum as *primitive complexity*. At one end of the continuum are simple points. Theories that posit the matching of such points exploit few of the principles of invariance or part decomposition described in section 4.2. Next on the continuum are schemes in which points are grouped into contours to provide a more complex primitive; even more complex primitives are groups of contours or surfaces; the most complex primitives of all are simple volumes. Dickinson et al. note a trade-off among various models between the ease of determining the indexing primitive and the ease of indexing an object model: the simpler the indexing primitive, the easier it is to determine that primitive but the more difficult it is to index an object from it. Thus the luminance of a small patch of points is easy to determine, but it is difficult to index an object from that patch. Once we know the convex volumes (parts and their relations) that might comprise an object, it is relatively easy to determine which object has those parts, but it can be very difficult for current vision systems to determine the convex volumes present in the image.

Theorists who have opted for simple primitives tend to focus on developing models that more readily allow activation of object representations from those primitives. Those who assume more complex primitives focus on schemes for more efficient and accurate extraction of the primitives from the image.

In this section we consider theories from three points along this continuum: (a) models that attempt recognition based on the outputs of simple cells (activated by small patches of pixels), either directly (Lades, Vorbrüggen, Buhmann, Lange von der Malsburg, Würtz, and Konen (1993) or with an intervening layer (Poggio and Edelman 1990); (b) a model by David Lowe (1987) of object recognition based on nonaccidental configurations of edges; and (c) a model by Irving Biederman and associates (Biederman 1987; Hummel and Biederman 1992) that assumes simple volumetric primitives roughly corresponding to an object's parts. The theories differ not only in the complexity of their matching primitives but in other characteristics as well. We also comment on these other characteristics when describing the theories in our overview in section 4.5.

One of the major advances in cognitive science over the past decade has been the development of theoretical formalisms, *neural networks*, that allow the expression of symbolic activity in terms of a pattern of activation over an aggregate of connected neuron-like elements. Several of the models considered in this section are of this type.

4.3.1 Matching of Simple Cell Outputs

The Lades et al. Face-Recognition System

Christoph von der Malsburg and his associates (Lades et al. 1993) initially developed their model as a face-recognition system, and it has enjoyed considerable success at that task. The model can be represented as a two-layer network, as illustrated in Figure 4.5. The input (or image) layer consists of an array of columns of individual units (or kernels), each roughly corresponding to a V1 simple cell. As described in section 4.2.1, a particular cell is tuned to variation in luminance at a particular orientation at a particular scale (i.e., spatial frequency) at a particular position of the visual field. The tuning of a simple cell can be approximated mathematically by a Gaussian-damped, sinusoidal filter termed a *Gabor filter*. A column of these filters, each tuned to different orientations and scales but with maximum responsiveness centered on the same region of the visual field (e.g., all those cells whose receptive fields are centered at 2 degrees left of fixation), is termed a *Gabor jet*. It roughly corresponds to the simple cells of a V1 hypercolumn.

In Figure 4.5 the Gabor jets are illustrated as a stack of disks centered at a single position in the visual field. The jets are arranged in a lattice, with each node of the lattice designating the center of the receptive field for a jet. In the implementation described here, each jet consists of filters at five scales and eight orientations (therefore, at 45-degree intervals) so that at each node forty filters comprise each jet. There are 5×9 nodes (jets) in the lattice. Other parameters could have been employed but these are

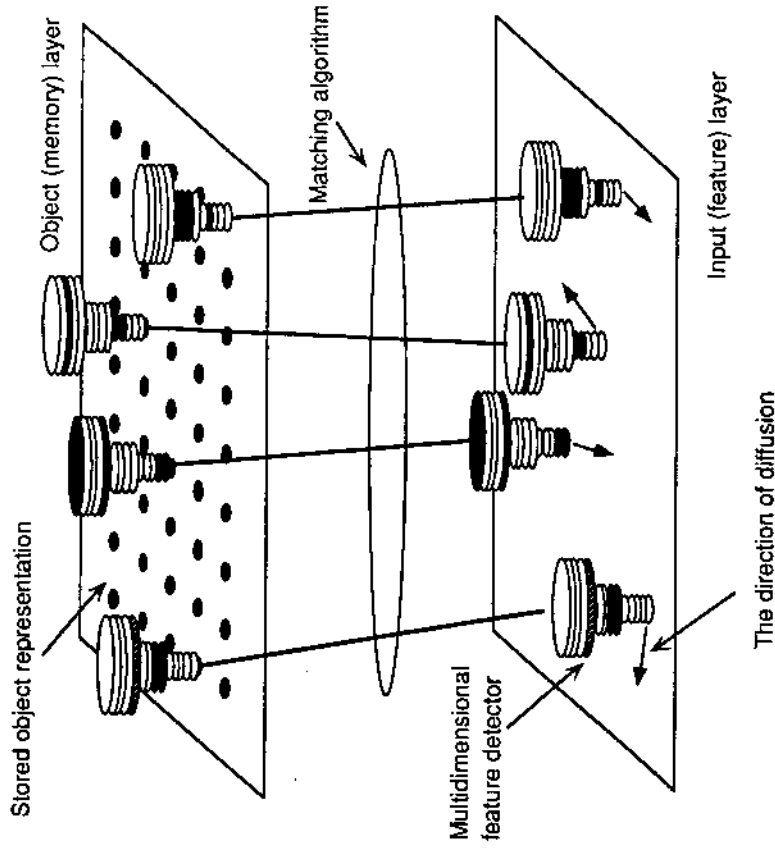


Figure 4.5

The architecture of the Lades et al. (1989) recognition system. Shown here are four of the Gabor jets, each composed of a set of filters (each represented by a disk in the stack) from a regular 5×9 matrix of jets. The filters differ in scale (spatial frequency) and orientation tuning. Activation values of the original image are stored in the object layer. The figure depicts the diffusion of the jets in the input layer, as indicated by the arrows, when that layer is activated by a new image and the matching algorithm attempts to find the best match against a previously stored image. (Reprinted with permission of the authors from Fiser, Biederman, and Cooper [1994]. Copyright by József Fiser and Irving Biederman.)

sufficient for reasonably accurate face recognition, which was the original goal of the system. The receptive fields of the largest filters are considerably larger than those indicated in Figure 4.5 in that they are affected by luminance variation approximately two nodes away from the center of their receptive fields.

A particular image results in activation of the different filters to various extents. These values are stored along with the relative positions of the adjacent jets. Figure 4.6a shows an image, a face, with the lattice superimposed over it. A new image is matched against the original by the

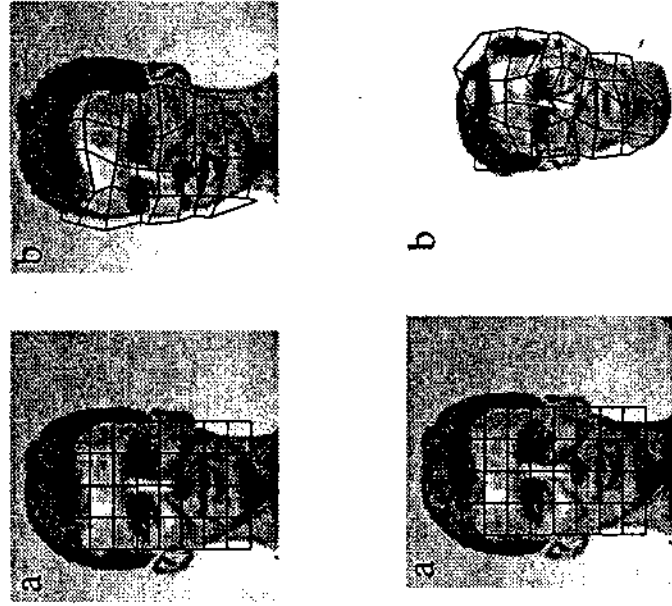


Figure 4.6

An example of the deformation of the Gabor-jet lattice when matching faces in the Lades et al. (1993) recognition system. (a) The target face stored in a gallery of images of the faces of 56 individuals, showing the original positioning and regularity of a lattice. (b) The upper face is an image of the same individual from a different orientation and slightly changed expression. The lower face is of a different individual at approximately the same orientation as the original. The superimposed meshes when the *b* images are matched against the *a* image show the degree of deformation for each of the matches. In this example, the system correctly matched the upper (b) face with the (a) face. (Reprinted with permission of the authors from Fisher, Biederman, & Cooper [1994]. Copyright by József Fiser and Irving Biederman.)

individual jets that, when activated by the new image, diffuse (gradually change their positions) to determine their own best fit, as illustrated by the arrows on the jets in the input layer in Figure 4.5. With faces, the same individual can be in a different orientation and expression (as shown in the upper panel of Figure 4.6b) or be the image could be of a different individual (as shown in the lower panel of Figure 4.6b). Although details of this matching algorithm are beyond the scope of this chapter, we note that similarity of a pair of images is (a) a positive function of the similarity of the activation values of the Gabor filters for corresponding jets (i.e., the jet in the third row, fourth column), and (b) a negative function of the degree to which a given jet has to be displaced, relative to its immediate

neighbors, to find its best match in a new image. To the extent that the jets move independently, the resultant positions will depart from the original, regular positions, as suggested by the different directions of movement of the jets in Figure 4.5 and as shown in the deformed mesh in the upper and lower panels of Figure 4.6b (see also Figure 4.16). Typically, the greater the deformation of the lattice, the lower the similarity of the match. A test image is compared against a number of stored images. The most similar image is taken to be the recognition choice. There is no reduction in similarity if the face or object appears at a position in the visual field other than where it first appeared and or if it is of a different size. In that case the lattice just has to be repositioned or expanded or contracted with little or no distortion of the original positions of the jets. Similarly, variations in the overall illumination levels are factored out, although differences in the direction of illumination for two images of the same person reduce similarity.

In a test of the Lades et al. system, researchers prepared fifty-six pairs of images of the faces of fifty-six individuals. One image of each person was sorted in a gallery of faces and recognition was attempted with the other member of the pair (which could differ in expression and orientation). The average ranking of the correct face was 1.4 (chance would have been 27.5) (Fiser, Biederman, and Cooper 1994). In section 4.5 we will evaluate this system as an object recognizer.

Because activation values are dependent on the specific view or aspect of the object, the Lades et al. and the Poggio and Edelman (1990) models are said to be *view (or aspect) based*. As different views of an object are encountered, the system builds up different patterns of activation of the hidden units that represent the different poses. What happens if the object is seen from a slightly different view? To the extent that the new image is similar—in terms of the pattern of filter activation values—to a previously learned view, the model might exhibit graded generalization to the new view if the unit in the output layer corresponding to that object is more activated than units representing other objects. It is important to note that in the test of face recognition the rotation in depth was limited to approximately 30 degrees (as illustrated in Figure 4.6). With rotations beyond that value, the accuracy of face recognition declined significantly. The Poggio and Edelman (1990) model was designed, in part, to increase the capacity of a filter-matching model to handle greater variations of rotation in depth.

The Poggio and Edelman Radial Basis Function Model

The Lades et al. model attempts to match filter outputs directly to an object representation layer. Poggio and Edelman (1990) assume a first stage that is similar to that of the Lades et al. model (in that it does a

simple filtering of the image); in addition, they assume a single hidden layer between that input stage and an output stage. Units in the hidden layer self-organize to take weighted activation values of the L1 filters to distinguish among a set of stimuli learned by the network. In the hidden layer of the network proposed by Poggio and Edelman (1990) these units are termed *radial basis functions* (RBFs); as they are designed to allow optimal classification of an image, a minimal number of these units allow classification of a large number of possible images. This model, then, provides a basis for determining when a new representation might be needed. In one exercise (Poggio and Edelman 1990), only two RBF units were sufficient to recognize ten to forty views of a bent paper clip over a 90-degree range of orientation. The object layer in the Lades et al. model is a representation of a particular view of an object, whereas the RBFs that emerge from experience with a series of views of an object need not (and typically do not) correspond to any particular view. The RBF thus constitutes a prototype for a modest range of views or deformations of an object.

The RBFs belonging to a single object are linked so that together they form a set of prototypes for an object. A significant challenge for the Poggio and Edelman model, however, is determining which object is projecting a new image so that an existing RBF can be modified, a new one created, or different RBFs linked. Currently, the model must be informed of this by some other system (or the programmer). In section 4.5 we consider several empirical tests of whether human object recognition can be predicted from filter outputs in the manner assumed by this class of models.

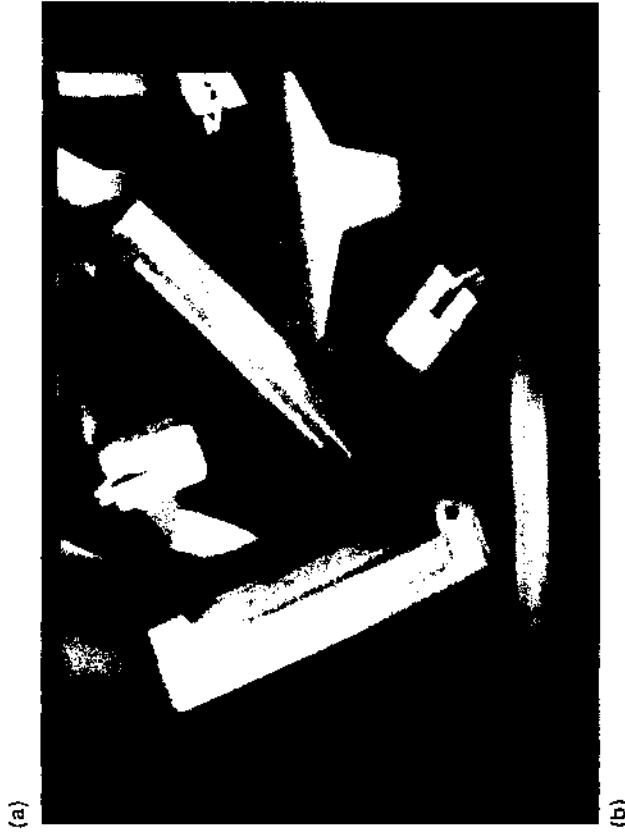
4.3.2 Model-Based Matching of Edges

The two models described in the previous section are pure "bottom-up" systems in that they assume a one-way flow of information from the initial image filtering to the representation of an object (or face). With an extremely large set of possible objects in a gallery or with variations in the shape and orientation of test objects to be matched against stored images, the speed and accuracy of correct recognition can decline greatly. A number of theorists have proposed schemes that reduce the degree of matching required by considering only those objects in memory that share certain features, which are initially extracted from the image, and only those poses of the object that are consistent with those features. Lowe (1987) offers a detailed proposal for how such a system might work. Whereas Lowe's proposal is limited to images with straight edges, Ullman's (1989) model, which has somewhat similar characteristics, has the potential for recognizing a broader class of objects, including those with curved surfaces. With both systems, a fully three-dimensional model

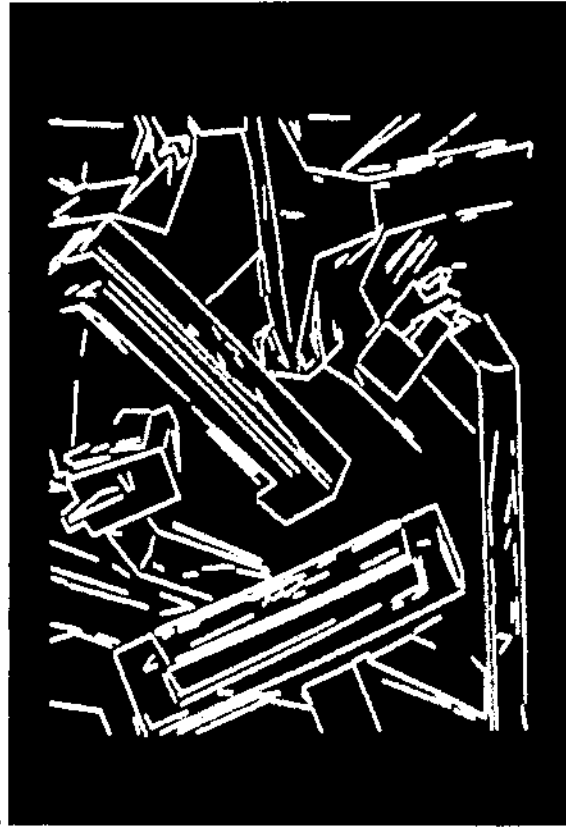
of an object is stored rather than a large set of representations each based on a different view. Ullman's model employs an initial extraction of features to determine the precise orientation and scale of the object model to be matched against the image.

Lowe's SCERPO model is directed primarily toward determining the orientation and location of objects, even when they are partially occluded by other objects, under conditions in which exact three-dimensional object models are available. The SCERPO takes as input an image such as the one shown in Figure 4.7a, an image of a number of disposable razors in arbitrary orientations. The model detects edges by finding sharp changes in image intensity values across a number of scales (as discussed in Chapter 1). The results of this edge-detection stage are shown in Figure 4.7b. The edges are then grouped according to the viewpoint-invariant properties of collinearity, parallelism, and cotermination. A few of these image features are then tentatively matched against image features of the object model generated from a particular orientation of the object that would maximize the fit of those image features. From this initial hypothesis, the locations of additional image features (edges) are proposed and their presence in the image evaluated. Figure 4.7c shows the successful final matches for five orientations of the razors. These matches provide segments not detected initially by the edge finder (middle panel) and discard edges initially detected but not part of the object model (e.g., the glare edges on the handle of the razor extending horizontally in the lower part of the figure). SCERPO may provide a plausible scheme for characterizing human performance under conditions in which the initial extraction of image edges is uncertain, as in conditions of poor visibility or where the orientation of an object is unfamiliar.

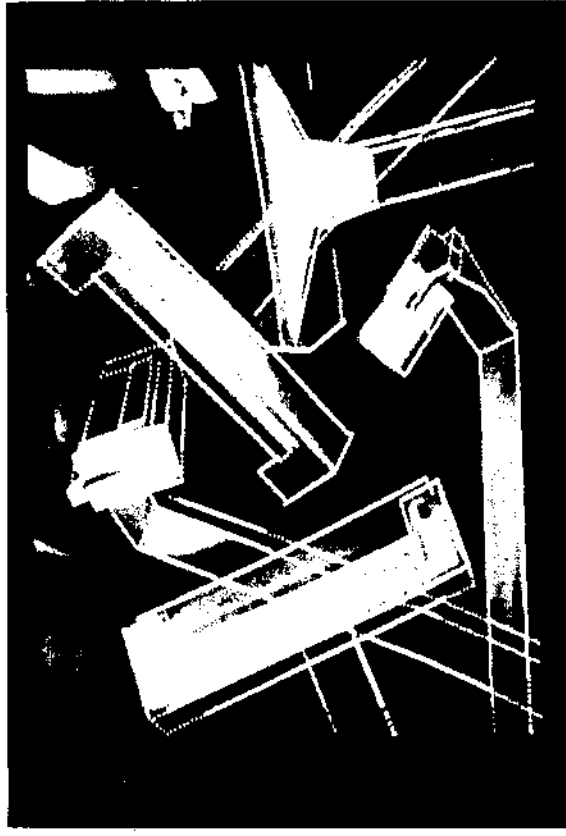
Ullman's (1989) Alignment model first reorients all the object models that might be possible matches for the image and tests them for the fit of the image against the aligned models in memory. The alignment capitalizes on the formal result that three non-coplanar points are generally sufficient to determine the orientation of any object. In practice, the three points are typically viewpoint invariant in that they are selected at a point where there is a cotermination of edges. However, any salient points, or even general features, would be sufficient for alignment. Although it appears unlikely that people rotate (align) all possible candidate models in memory prior to matching, the alignment model offers a possible account of those cases in which recognition depends on reorienting a mental model. Ullman and Basri (1990) present a general theory of how a three-dimensional object can be represented as a combination of two-dimensional images so that it is recognized under such transformations as rotation in depth and non-rigid transformations.



(a)



(b)



(c)

Figure 4.7 (cont.)

Matching Viewpoint-Invariant Parts

Biederman (1987; Hummel and Biederman 1992) proposed a theory of entry-level object recognition that assumes that a given view of an object is represented as an arrangement of simple, viewpoint-invariant, volumetric primitives called *geons*. Five (of the twenty-four) geons are shown in the left panel of Figure 4.8. The relationships among the geons are specified, so that the same geons in different relations will represent different objects, as with the cup and pail in the right panel of Figure 4.8. The geons have two particularly desirable properties: they can be distinguished from each other from almost any viewpoint, and their identification is highly resistant to visual noise. We will consider in greater detail the segmentation of the image into regions to be matched with geons, the description of the image edges in terms of viewpoint-invariant properties, and the geon arrangement that emerges from the parsing and edge processing.

Geons from Viewpoint-Invariant Edge Descriptions

According to RBC, each segmented region is approximated by a geon. Geons are members of a particular set of convex or singly concave volumes that can be modeled as *generalized cones*, a general formalism for representing volumetric shapes (Binford 1971; Brooks 1981). A generalized cone is the volume swept out by a cross section moving along an axis.

Figure 4.7

Lowe's viewpoint consistency model can find objects at arbitrary orientations and occlusions. (a) The original image of a bin of disposable razors. (b) The straight line segments that SCERPO derived from the image. (c) Final set of successful matches between sets of image segments and five particular viewpoints of the model (shown as bright dotted lines). (Reprinted by permission of the publisher and author from D. Lowe, The viewpoint consistency constraint, 1987, *International Journal of Computer Vision* 1, 66, 70, Figures 4, 5, and 8. Copyright 1987 by Kluwer Academic Publishers.)

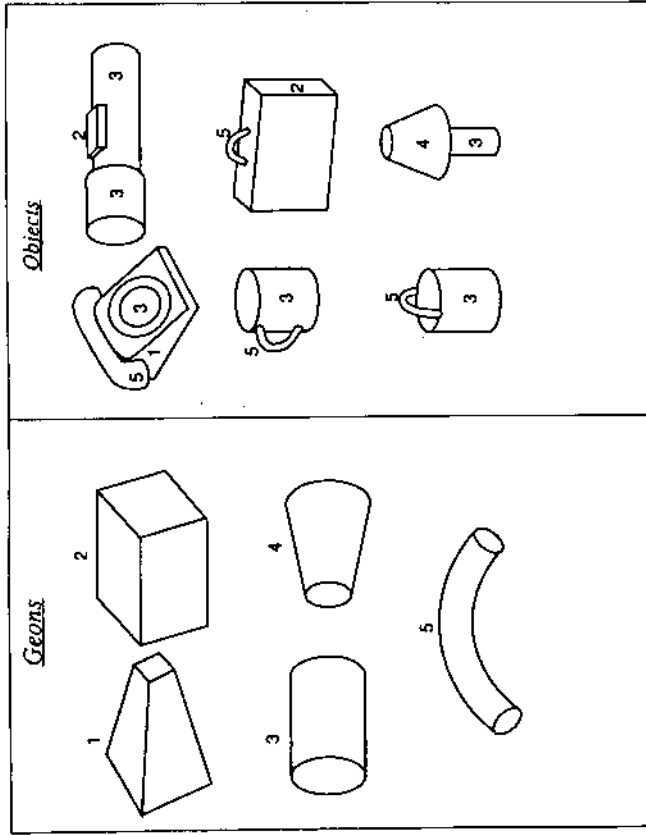


Figure 4.8 (Left) Five geons. (Right) Only two or three geons are required to uniquely specify an object. The relations among the geons matter, as illustrated by the pail and the cup.

The set of geons is defined so that they can be differentiated on the basis of dichotomous or trichotomous contrasts of viewpoint-invariant properties to produce twenty-four types of geons. The contrasts of the particular set of nonaccidental properties shown in Figure 4.2 were emphasized because they may constitute a basis for the generation of this set of perceptually plausible components. Figure 4.9 illustrates the generation of a subset of the twenty-four geons from contrasts in the nonaccidental relations of four attributes of generalized cones. Three of the attributes specify characteristics of the cross section: curvature (straight versus curved), size variation (constant [parallel sides], expanding [nonparallel sides], expanding and contracting [nonparallel sides with a point of maximum convexity]), and whether nonparallel-sided geons terminate in a point (in which case they have an L-vertex) or are truncated (in which case they will have arrow vertices). One attribute—straight versus curved—specifies the axis. These are modal types. It is possible that a given region of the image activates two or more geons, depending on the presence of particular image features.

When the contrasts in generating functions are translated into image features, it is apparent that the geons have a larger set of distinctive

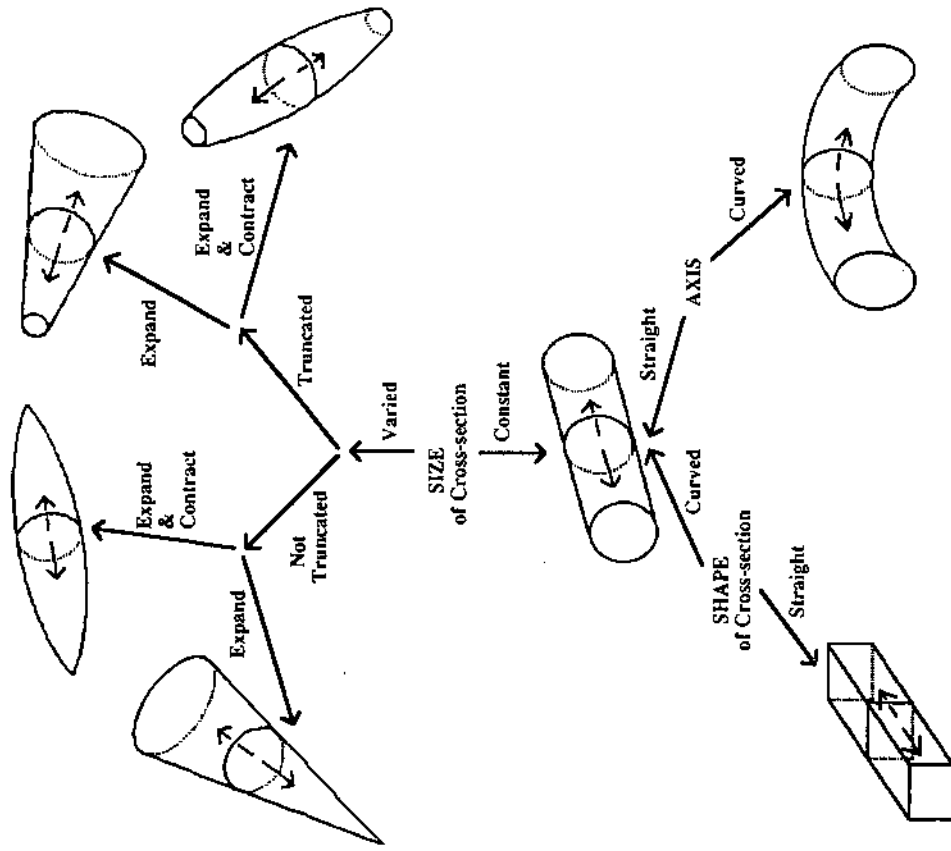


Figure 4.9

An illustration of how variations in three attributes of a cross section (curved versus straight edges; constant versus expanded and contracted size; symmetrical versus asymmetrical) and one attribute of the shape of the axis (straight versus curved) can generate a set of generalized cones differing in nonaccidental relations. Constant-sized cross sections have parallel sides; expanded or expanded and contracted cross sections have sides that are not parallel. When the sides are not parallel they could be truncated (as with the cone) or end at a point (L-vertex). Curved versus straight cross sections and axes are detectable through collinearity or curvature. Shown here are the neighbors of a cylinder. The full family of geons has 24 members. (Adapted by permission of publisher and author from I. Biederman, *Recognition-by-components: A theory of human image understanding*, 1987, *Psychological Review* 94, 122, Figure 6. Copyright 1987 by the American Psychological Association.)

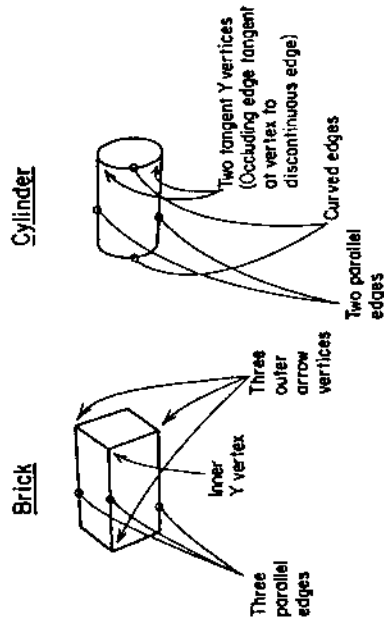


Figure 4.10

Some nonaccidental differences between a brick and a cylinder. (Reprinted by permission of the publisher and author from I. Biederman, Recognition-by-components: A theory of human image understanding, 1987. *Psychological Review* 94, 121, Figure 5. Copyright 1987 by the American Psychological Association.)

nonaccidental image features than the four that might be expected from a direct mapping of the contrasts in the generating function. Figure 4.10 shows some of the nonaccidental contrasts distinguishing a brick from a cylinder. The silhouette of a brick contains a series of six vertices, which alternate between Ls and arrows, and an internal Y-vertex. The vertices of the silhouette of the cylinder, by contrast, alternate between a pair of Ls and a pair of tangent Ys. The internal Y-vertex is not present in the cylinder (or in any geon with a curved cross section). These differences in image features would be available from a general viewpoint and, thus, could provide, along with other contrasting image features, a basis for discriminating a brick from a cylinder. Dickinson, Pentland, and Rosenfeld (1992) provide an extensive account of how nonaccidental configurations can be employed to determine a particular geon. Zerroug and Nevatia (1995) have derived a number of viewpoint-invariant properties of generalized cylinders that allow recovery of volumetric shape from a gray-level image. Interestingly, a number of these properties only hold if the volume can be described as a generalized cylinder with a cross section orthogonal to the axis.

Deriving the geons from contrasts in viewpoint-invariant properties renders the geons themselves largely invariant under changes in viewpoint. (Current theoretical work is exploring a redefinition of the geons in terms of those volumes that are maximally viewpoint invariant. Most likely the resultant volumes will largely correspond to the present set.) Because the geons are simple (*viz.*, convex or only singly concave), lack

sharp concavities, and have redundant image properties, they can be readily restored in the presence of visual noise. Therefore objects that are represented as an arrangement of geons will possess the same invariance to viewpoint and noise. Geon activation requires only categorical classification of edge characteristics for processing to be completed quickly and accurately. A representation that requires fine metric specification, such as the degree of curvature or length of a segment, cannot be performed with sufficient speed and accuracy by humans to be the controlling processing for object recognition.

Geon Relations and Geon Attributes

Much of the capacity to represent the tens of thousands of object images that people can rapidly classify from a small alphabet of geons derives from several viewpoint-invariant relations between pairs of geons and some coarse metric attributes of individual geons. Examples of relations that have been hypothesized are vertical position (above, below, beside), join type (end-to-end, end-to-middle centered, end-to-middle off-centered), relative size (larger, smaller, equal to), and relative orientation (parallel, orthogonal, oblique). These relations are defined for joined pairs of geons so that the same subset of geons represent different objects if they are in different relations to each other—like the cup and pail in Figure 4.8. Also specified are two coarsely coded metric aspects of the geons: (a) the relative-aspect ratio of the geon (five levels of the length of the axis compared to the diameter of the cross section) and (b) the orientation of the geon (e.g., vertical, horizontal, or oblique). There are eighty-one combinations of pairwise relations and fifteen attributes. A representation that specifies parts (geons), attributes, and relations independently and explicitly is termed a *structural description*.

Three-Geon Sufficiency

Object space and three-geon sufficiency. With twenty-four possible geons, eighty-one combinations of relations, and fifteen attributes, the variations in relations and aspect ratio can produce 10,497,600 possible two-geon objects ($24^2 \times 15^2 \times 81$). A third geon, with its possible attributes and its relations to one other geon, yields over 306 billion possible three-geon objects. This is two orders of magnitude greater than the number of seconds in a hundred-year lifetime.

If the 30,000 familiar object models estimated in section 4.1.3 are distributed homogeneously throughout the space of possible object models, then the extraordinary disparity between the number of possible two- or three-geon objects and the number of objects in an individual's object vocabulary—even if the estimate of 30,000 is short by an order of

magnitude—means that an arrangement of two or three geons would almost always be sufficient to specify any object.

The theory thus implies a *principle of geon recovery*: if an arrangement of two or three geons can be recovered from the image, objects can be quickly recognized even when they are occluded, rotated in depth, novel, extensively degraded, or lacking customary detail, color, and texture. Experimental results support this expectation of geon theory (Biederman 1987): When only two or three geons of a complex object (such as an airplane or elephant) are visible, recognition can be fast and accurate (although, predictably, not as fast as with a complete image). You can try this for yourself by covering up parts of pictures of common objects. See if the object remains recognizable to a friend (who did not see the original) when only two or three parts are in view. The simple line drawings of objects shown in Figure 4.8 illustrate this expectation of three-geon sufficiency.

Just as three geons are usually sufficient for classification, objects composed of a single geon are often appropriate for several entry-level objects. In these cases, other information—such as color, texture, small details, or context—are required for classification (Biederman and Ju 1988). For example, distinguishing among a peach, a nectarine, and a plum requires that surface color and texture be specified. The expectation from geon theory would be that the identification of single-part objects would require more time than objects with distinctive geon configurations, as well as, to a much greater extent, specification of color and texture. Biederman, Hilton, and Hummel (1991) confirmed these expectations.

A Neural-Net Implementation of Geon Theory

Hummel and Biederman (1992) proposed a neural-net implementation of geon theory.

Problems and goals of the implementation. As discussed in section 4.2.1, the representation of the image at the first cortical stage, V1, probably consists of activation of a large number of cells, each tuned to variation of luminance at a particular orientation in a small (0.5 to 2.0-degree) region of the visual field. How could a structural description specifying parts (geons) and relations be derived from this activity? Moreover, given the

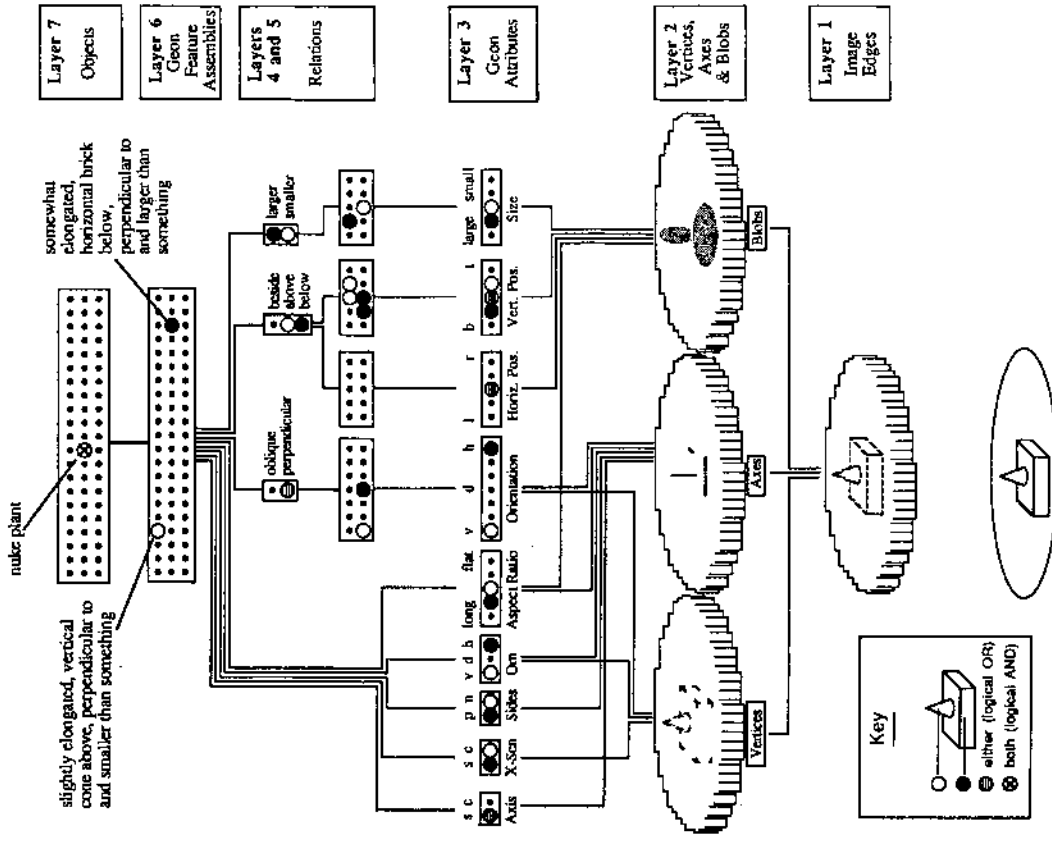


Figure 4.11

The architecture of the Hummel and Biederman (1992) neural-net implementation of geon theory, indicating the representation activated at each layer by the image in the key. In layers 3 and above, large circles indicate cells activated in response to the image and dots indicate inactive cells. Cells in layer 1 represent the edges (specifying discontinuities in surface orientation and depth) in an object's image. Layer 2 represents the vertices, axes, and blobs defined by conjunctions of edges in layer 1. Layer 3 represents the geons in an image in terms of their defining dimensions: Axis shape (Axis), straight (s) or curved (c); cross section shape (X-Str), straight (s) or curved (c); whether the Sides are parallel (p) or nonparallel (n); coarse orientation (Om), vertical (v), diagonal (d), or horizontal (h); aspect ratio, elongated (long) to flattened (flat); fine orientation (Orientation), vertical (v), two differ-

ent diagonals (d), and four different horizontals (h); horizontal position in the visual field (Horiz. Pos.), left (l) to right (r); vertical position in the visual field (Vert. Pos.), bottom (b) to top (t); and size, small (near 0 percent of the visual field) to large (near 100 percent of the visual field). Layers 4 and 5 represent the relative orientations, locations, and sizes of the geons in an image. Cells in layer 6 respond to specific conjunctions of cells activated in layers 3 and 5, and cells in layer 7 respond to complete objects, defined as conjunctions of cells in layer 6. (From J. E. Hummel, and I. Biederman, 1992, *Psychological Review*, 99, 486, Figure 7. Copyright 1992 by the American Psychological Association.)

evidence for position, size, reflection, and depth invariance (described in section 4.4.3), how could the same description be derived when completely different cells are activated because of a change in viewpoint? The Hummel and Biederman network offered a possible answer to this question.

The network, whose overall architecture is shown in Figure 4.11, takes as input a line drawing representing the orientation and depth discontinuities of an object and activates units in the seventh layer (L7) that represent a viewpoint-invariant structural description of the object specifying its geons, geon attributes and the relations between geons. This description is activated regardless of whether the model has previously been exposed to the object. The model is meant to be a working hypothesis and is admittedly incomplete.

The binding problem. In the network, V1 is roughly modeled by a lattice of units in L1 that code whether a given contour is straight or curved and whether it passes through or terminates (is end-stopped) in the receptive field of a particular end-stopped cell. The end-stopped activity of two or more contours at a common point in the visual field activates L2 units representing such vertices as forks, arrows, and Ls.

The binding problem—determining what goes with what—is a major problem that needs to be solved to achieve invariant recognition. Consider again Figure 4.1, which depicts a vertical cone on top of a horizontal brick. We readily segment this object into two parts, the brick and the cone. In section 4.2.4 we considered what information the visual system might employ to segment an object into its parts. Here we focus on how grouping itself might be represented. If we have, in an oversimplified case, four neurons activated by *a*, *b*, *c*, and *d*, how does the neural activity code *a* and *b* to one group and *c* and *d* to another? Hummel and Biederman's solution (1992) was to induce the units activated by the contours of one geon to fire (approximately) synchronously and the units activated by another geon to fire synchronously but out of phase with each other. The signal that induces the synchronization is passed through links of nearby units that have collinear or parallel receptive fields or, for end-stopped cells, units that coterminate or have complementary orientations. In Figure 4.1, the synchronization would run around the vertices of each of the geons (through end-stopped cells) but would not pass from segment *a* of the brick to segment *d* of the cone because *a* and *d* do not coterminate. Thus the activity does not pass from the stem to the top of a T-junction. Instead, end-stopped cells activated by *a* and its extension on the other side of the cone would be synchronized because they have complementary orientations. Details of the algorithms that induce the synchrony are discussed in Hummel and Biederman (1992).

Using temporal asynchrony for representing geon attributes and relations. The units activated in L1 activate units in L2 representing vertices, axes, and blobs, providing information about the approximate size of the part and its center of mass. The L2 units are enumerated throughout the visual field so that a given vertex detector—for example a Y-vertex at a given orientation—will be available for each hypercolumn. All the L2 units, in turn, activate a single set of geon attribute units in L3. For example, all Y-vertex units send activation to the unit representing a straight cross section. The temporal correlation in the firing of the units activated by one object part induced in L1 and L2 is maintained through the first six layers of the model. For the example shown in Figure 4.11, all the units marked by filled circles represent information about the part that is a brick. All these fire together and out of phase with the units marked by the open circles representing the cone. Only thirty-six cells in the L3 layer are required to specify information for each part such as the geon type (one of eight possibilities) its orientation, and aspect ratio. The binding is thus achieved without positing additional units for “anding” that would, for example, posit a small vertical cone detector for each position of the visual field. Because the binding is temporary, these same cells can be used to code other parts of the object as well as the parts of other objects, no matter where they are in the visual field.

L4 and L5 derive invariant relations of vertical position, size, and relative orientation. The temporal correlation is maintained so that the “above” cell fires in phase with the units representing the cone. The same pattern results if the object is presented at another region in the visual field or at another site. The outputs of L3 that represent the distributed values of a geon, its orientation and aspect ratio, and the outputs of L5 representing their interrelations, provide a vector that self-organizes a unit in L6 termed a *geon feature assembly*. Units in L7 are object cells that self-organize to an integration over successive outputs from L6. These operations produce a parts-based structural description that is subsequently used directly as a basis for viewpoint-invariant recognition. The model's recognition performance conforms well to the results from the shape-priming experiments described in section 4.4.3 in that it manifests invariance to translation, rotation, and orientation in depth.

Binding through temporal correlation may provide some insight into the underlying neural basis of the attentional bottlenecks, discussed in Chapter 2. As the number of objects or object parts increases, insufficient temporal resolution to keep them out of phase with each other may be available and accidental simultaneous firing of the units representing two or more parts could occur. At this point, attention would be required to inhibit some of the activity producing the accidental synchrony.

4.4 Empirical Tests of Geon Theory

In this section, we describe experimental tests of two assumptions that distinguish geon theory from other theories of object representation. These assumptions are (1) that objects are represented in terms of their simple parts, and (2) that the parts are characterized by differences in viewpoint-invariant properties. In most of these experiments, the subjects named briefly presented (e.g., 100-msec.) object pictures. The flash of the picture was followed by a mask, an array of meaningless straight and curved line segments, to reduce persistence of the image. Naming reaction times and errors were the primary dependent variables.

4.4.1 Is the Representation Part-Based?

According to RBC, an object is represented in terms of its geons, which are activated by such local image features as vertices and edges. But, if the geons are activated by image features, why not just represent an object in terms of those features? In the real world, objects are often partially occluded by other surfaces, as when we view a car behind some light foliage. The pattern of occlusion of small regions of the parts can vary dramatically when we change viewpoint or when the wind shakes the leaves, but the various parts of the car likely would remain identifiable. To represent objects in terms of local image features, we would need a different representation for each arrangement of occluding contour or for each slightly altered orientation of the object.

Subjective impression is consistent with a parts-based representation. To see this, identify the two contour-deleted images in the second column of Figure 4.12 while covering up the images in the third column. Now look at the images in the third column, covering up those in the second column. Do these images look the same as those you just viewed? Now compare them. You will note that they are actually different images, each member of a pair having different vertices and edges. Despite the differences at the vertex and contour level, however, you probably saw the members of each pair as identical.

To test this issue of features versus parts experimentally, Biederman and Cooper (1991a) created pairs of *complementary images* of object pictures by deleting every other edge and vertex from each geon to create the two images of each object shown in Figure 4.12. The two images, when superimposed, form the intact picture shown in the far left column with no overlap in contour. The complementary images were created in such a way that each part (or geon) of the object can be recovered (or fail to be recovered) from each of the images. Although the complementary images share no edges and vertices, they presumably activate the same compo-

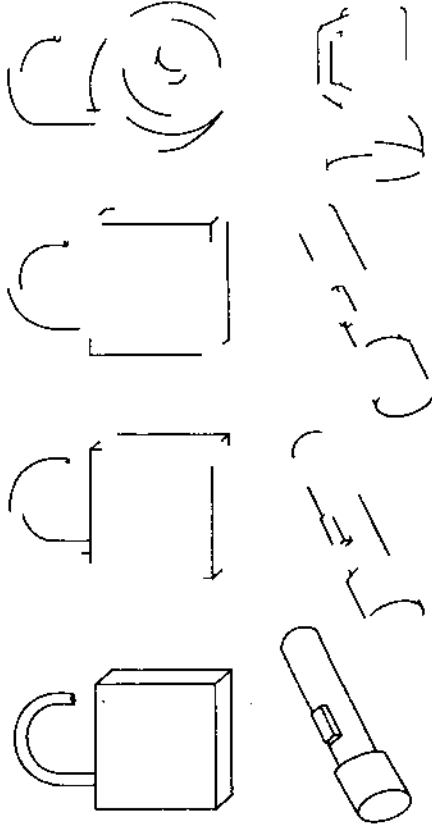


Figure 4.12

Complementary-feature images. From an original intact image (left column), two complementary-feature images (middle two columns) were formed by deleting every other vertex and edge from each part so that each image had 50 percent of its contour. When superimposed, the two complements comprise the original intact image with no overlap in contour. (Right column) Same-name, different-exemplar images from another complementary pair. Subjects never viewed the intact image. Assuming that an image in the second column was shown on the first (priming) block, the other member of the complementary pair (in the third column) would be an instance of complementary-feature priming and the right figure would be a different-exemplar control. (Adapted by permission of the author and publisher from I. Biederman and E. E. Cooper, Priming contour-deleted images: Evidence for intermediate representations in visual object recognition, 1991, *Cognitive Psychology* 23, 397, Figure 1. Copyright 1991 by Academic Press.)

nents. Because the amount of contour deleted from each image is substantial and includes vertices, it is unlikely that a local process of good continuation could not complete the contour of these images (see Biederman and Cooper 1991a, for a more complete discussion of this issue).

On a first block of trials, subjects viewed a number of brief (200-msec) presentations of one member from each complementary pair, naming it as quickly and as accurately as possible. On the second block, they saw either the identical image, its complement, or a same-name, different-exemplar image (also contour-deleted) from a category with the same name and basic-level concept but a different shape as shown in the far right column of Figure 4.12). The mean correct naming reaction times and error rates graphed in Figure 4.13 were markedly lower for the identical image than for the different exemplars, indicating that a portion of the priming was indeed visual and not just conceptual or lexical (i.e., the result of faster accessing of the name). The critical comparison, however, concerned the relative performance of the complementary condition. If priming was a

function of repetition of the specific vertices and edges in the image, then the complementary condition would have been equivalent to the different-exemplar condition, as neither shared any features with the original image. Remarkably, there was no difference in performance in naming complementary and identical images, indicating that none of the priming could be attributed to the specific vertices and lines actually present in the image.

What then caused the priming? Before we can attribute the priming to activation of the parts (in relations to other parts), which were common in the two conditions, we have to evaluate whether the priming could have been a consequence of activation of a semantic model of a subordinate category. Although the different-exemplar condition had the same basic-level category as the same-exemplar conditions, the identical and complementary conditions differed from the different-exemplar condition at the entry or subordinate level. If, for example, priming was due to activation of the concept of a grand piano or of a square lock rather than just piano or lock, then the advantage of the same-exemplar conditions could have been obtained without a contribution from activation of the parts. To test this possibility, we ran an experiment in which complementary images were created by deleting half the parts of the objects, as shown in Figure 4.14. With these stimuli, presumably, the same subordinate category would be activated from either member of a complementary pair, but through different parts. (This experiment required the use of objects that had at least four parts to look complete.) The design was otherwise identical to that of the previous study. As with the first experiment, performance with identical images was better than with different exemplars. Now, however, performance with the complements was equivalent to that with the different exemplars, indicating that none of the priming could be attributed to a subordinate semantic model. By eliminating subordinate-level concept priming as a factor in the first experiment, we obtained results suggesting that all the priming can be attributed to a representation of the parts of the object (and their interrelations) and none to the activation of the image features or subordinate-level concepts.

When does an object become unrecognizable? If object recognition is mediated by a representation of an object's parts, recognition should be particularly difficult if contour is deleted in locations that reduce the recoverability of the parts from the image. The images in the right-hand column of Figure 4.15 provide confirmation of this prediction. One method of reducing the recoverability of the parts is by deleting the cusps (the parsing regions discussed in section 4.2.4) to the point where the remaining contours bridge the cusp through smooth continuation. An example of how this method can interfere with the recovery of the parts is shown with the cup in Figure 4.15c. In this figure, the cusp between the top of the handle and the back of the lip of the bowl (shown in Figure 4.15a) has been

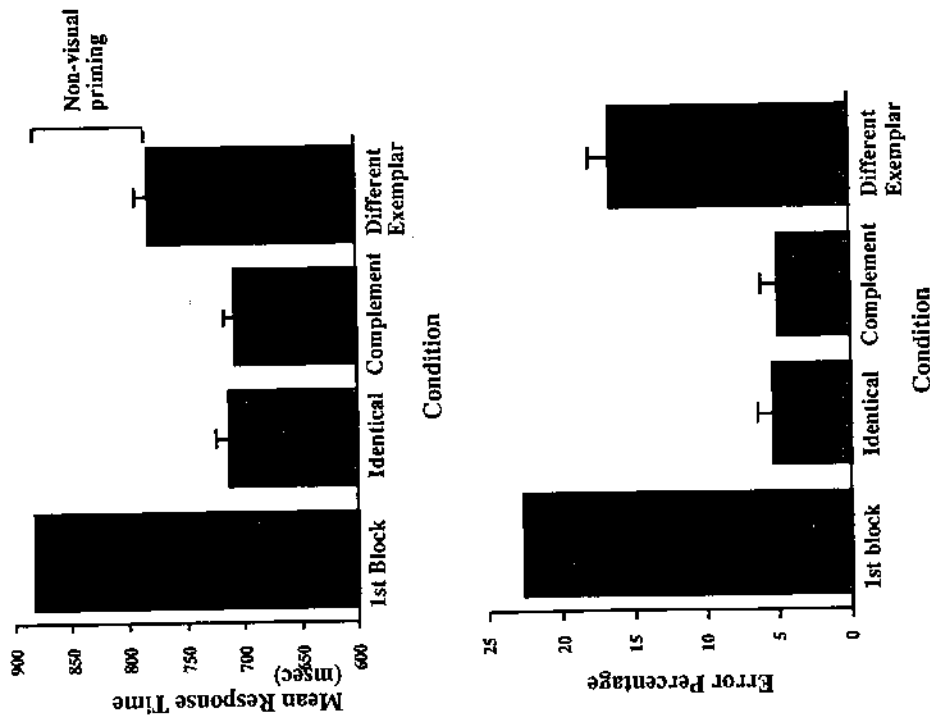


Figure 4.13 Mean correct reaction times and error rates of the complementary-feature priming experiment. The advantage in naming speed and accuracy of the identical image (when the second image of Figure 4.12 was also shown on the second block) compared to the different-exemplar condition provided a measure of visual priming. Any advantage of the identical over the complementary condition would be evidence of feature priming. There was none. Therefore, there was no contribution to the magnitude of priming from the representation of the lines and vertices. (Adapted by permission of the author and publisher from I. Biederman and E. E. Cooper, Priming contour-deleted images: Evidence for intermediate representations in visual object recognition, 1991, *Cognitive Psychology* 23, 399, Figure 2. Copyright 1991 by Academic Press.)

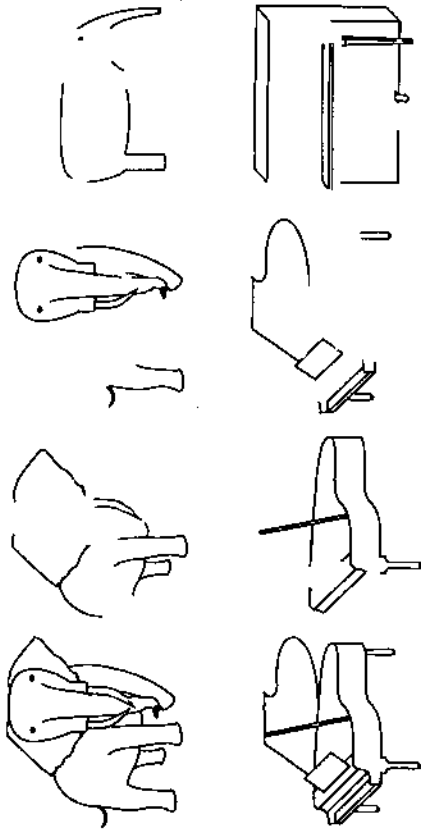


Figure 4.14

Complementary-part images. From an original intact image (left column), two complementary images (middle two columns), each composed of half the parts, were formed by deleting parts so that each image contained approximately 50 percent of the contour. If superimposed, the two complements would comprise the original intact image with no overlap in contour. (Right column) A same-name, different-exemplar image from another complementary pair. Subjects never viewed the intact image. Assuming that an image in the second column was shown on the first (priming) block, the other member of the complementary pair (in the third column) would be an instance of complementary-parts priming and the right figure would be a different-exemplar control. Unlike the results shown in Figure 4.13, the complementary and different-exemplar conditions were equivalent, both having reaction times and error rates that were considerably higher than the identical condition. This result indicates that none of the complementary-feature priming (Figures 4.12 and 4.13) could be attributed to the activation of a subordinate concept. (Adapted by permission of the author and publisher from I. Biederman and E. E. Cooper, Priming contour-deleted images: Evidence for intermediate representations in visual object recognition, 1991, *Cognitive Psychology* 23, 403. Figure 4. Copyright 1991 by Academic Press.)

deleted; the remaining contours would be continuous if joined through smooth continuation. Another technique is to delete a segment of a vertex so that a three-segment vertex, such as a fork or a tangent Y, or a T-junction, becomes a two-segment L-vertex. Examples of such operations are provided by the goblet in Figure 4.15C, in which the tangent Y-vertices at the junction of the sides and lip of the bowl have been converted to L-vertices by the deletion of the front of the lip of the goblet. Similarly, the T-vertices formed at the junction of the scissors blades or where the legs of the stool occlude the cross brace have been converted to L-vertices by deletion of one of the parts of the top of the T. Alternatively, the parsing regions can be deleted from several parts so that the remaining contours form an inappropriate vertex with more segments than

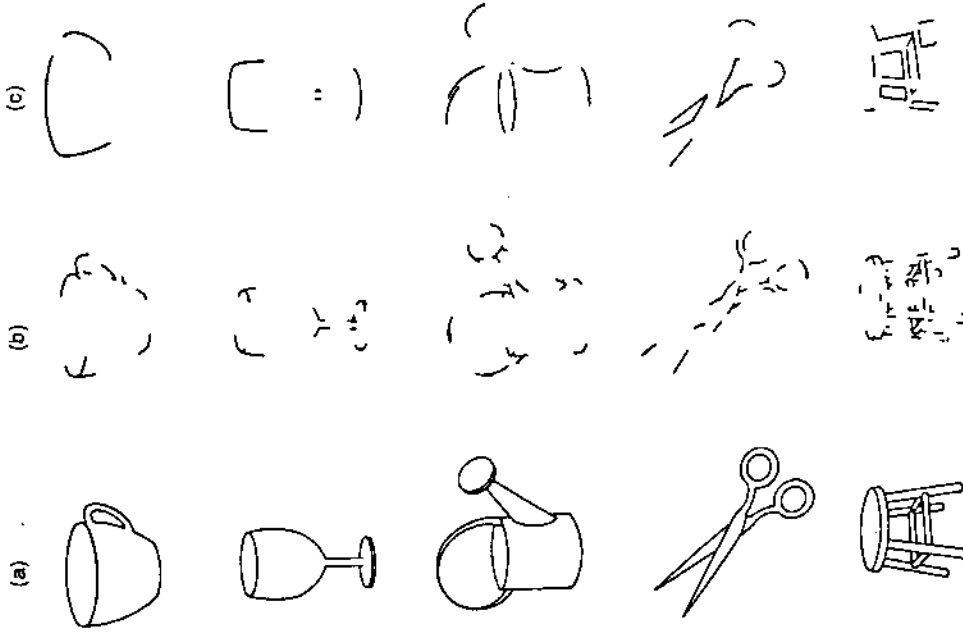


Figure 4.15

Example of five stimulus objects in the experiment on the perception of degraded objects. Column (a) shows the original intact versions. Column (b) shows the recoverable versions. The contours have been deleted in regions where they can be replaced through collinearity or smooth curvature. Column (c) shows the nonrecoverable versions. The contours have been deleted at regions of concavity so that collinearity or smooth curvature of the segments bridges the concavity. In addition, vertices have been altered (e.g., from Ys to Ls. (Modified by permission of the author and publisher from I. Biederman, Recognition-by-components: A theory of human image understanding, 1987, *Psychological Review* 94, 135. Figure 16. Copyright 1987 by the American Psychological Association.)

any of the original vertices, as with the watering can, where contours from the handle, opening, and base of the spout all appear to meet at a common vertex. If the same amount of contour is deleted, but in regions where the parts can still be activated, as in Figure 4.15b, objects remain identifiable.

The median accuracy of recognition of the nonrecoverable images, even after five seconds of viewing, was 0 percent (Biederman 1987). Given sufficient time (a few hundred msec), the recoverable images could be recognized perfectly. Actually, even when more contour is removed from the recoverable images than from the nonrecoverable images the former remain recognizable. You can test this by covering up parts of the objects in the middle column, say the right or left half, and determining whether you, or a person who has not seen the original versions, can still identify the objects. Recognition should be possible as long as enough contour remains to recover two or three parts of the object.

4.4.2 Viewpoint-Invariant versus Metric Properties

The complementary image and recoverable-nonrecoverable experiments provide evidence that the priming effects can be explained by a representation that specifies the parts of the object. But are these parts geons? A fundamental assumption of geon theory is that viewpoint-invariant differences, such as straight versus curved or parallel versus nonparallel, are given more weight than an equivalent amount of metric variation, such as aspect ratio. But how can metric and viewpoint-invariant contour variation be equated so that such an assumption can be investigated? Cooper and Biederman (1993) reasoned that the greater salience of viewpoint-invariant differences would be produced not in V1 but by the neural tissue more exclusively devoted to object recognition in extrastriate cortex, say in inferotemporal cortex (IT) (see Chapter 5). Consequently, they scaled the similarity of their stimuli according to the Lades et al. (1993) model of V1 simple-cell hypercolumns.

The positioning of the lattice over an original image is shown in the images in the left-hand column of Figure 4.16. Distorted lattices are shown in the middle and right columns of the figure, reflecting the reduced similarity of the images due to changes from the original in a viewpoint-invariant property (middle column) or a metric property (right column). In general, the two sources of error (filter similarity and lattice distortion) are correlated so that the more distorted the lattice, the less the similarity of the image to the original.

In the Cooper and Biederman (1993) experiment, subjects judged whether a pair of sequentially presented images of simple objects (containing only two or three parts), as illustrated in Figure 4.16, had the same name (and basic-level category). The images were each shown for 100

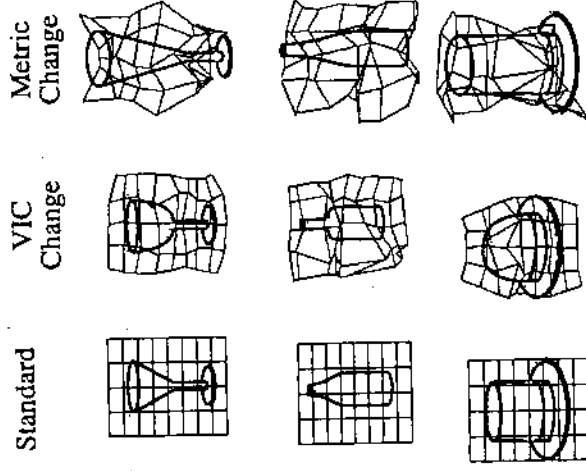


Figure 4.16

Examples of the Cooper and Biederman (1993) stimuli testing sensitivity in object recognition to viewpoint-invariant contrasts (VIC changes) compared to aspect-ratio differences (metric changes) of a single part. The superimposed grids over the VIC and metric-changed stimuli show the deformation of the Gabor jet lattices from the Lades et al. system when these images were matched against the standard. The grid over the standard shows the original positions of the Gabor jets. The greater the deformation, the lower the similarity between the standard and a changed image. Overall, the metric-changed stimuli were slightly less similar to the standard than the VIC-changed stimuli. (Adapted by permission from E. E. Cooper, and I. Biederman (1993). Geon differences during recognition are more salient than metric differences. Poster presented at the Meetings of the Psychonomic Society, Washington, D. C. Copyright 1993 by Eric E. Cooper.)

msec, with an intervening mask. When the images had the same name, which occurred on half the trials, one of parts differed in either a non-accidental property (as illustrated in the middle column of Figure 4.16) or in aspect ratio (right column). Dissimilarity, as specified by the Lades et al. model, between the original images and those that differed in a non-accidental property was slightly less than the dissimilarity between the original and metric changed images. If a difference in a viewpoint-invariant property results in greater dissimilarity for object classification than a difference in a metric property (when equated according to V1 similarity), then it should be more difficult (i.e., slower and less accurate) to judge whether two images belong to the same category when they differ in a viewpoint-invariant property, compared to two images that differ in aspect ratio. The results clearly supported this implication of geon theory.

4.4.3 Viewpoint Invariance and Its Neural Underpinnings

One of the most striking characteristics of human object recognition is its apparent invariance over changes in viewpoint. When we see an object at one position on the retina, at a given size and at a given orientation in depth, we can typically recognize it as the same object when, on a subsequent occasion, it is in another position, size, and orientation. Experimental results from name-priming studies confirm these impressions. The design of these studies are similar to that for the complementary-image experiment in that objects are presented in two blocks of trials, with the subject naming the objects as fast as possible. Biederman and Cooper (1991b; 1992) found that naming reaction times and error rates on the second block of trials were unaffected by a change in the position of the object—say from 2 degrees to the left of fixation to 2 degrees to the right of fixation—a 65 percent change in size, or a mirror reflection. As in the complementary-images experiment, slower naming times and higher error rates for images with the same name but a different shape documented the fact that the priming in these experiments was largely visual, rather than conceptual or lexical. Similarly, Biederman and Gerhardstein (1993) found that rotation in depth had no effect on name priming as long as the object could be readily described as an arrangement of distinctive geons and the original geons remained in view. The rotation, of course, altered the aspect ratio and degree of curvature of the objects.

M. Goodale (Chapter 5) provides an overview of the research that, over the past two decades, has established that there are at least two major cortical visual systems for the processing of shape. Both cortical pathways start in V1, but one extends dorsally to the posterior parietal region (PP), while the other extends ventrally, through V2, V4, and IT. Why would there be two systems for representing shape? It has been argued, by Goodale (Chapter 5) and Biederman and Cooper (1992), among others, that the dorsal pathway supports shape representations for motor interactions, where position, size, and orientation in depth must be precisely specified. To sit in a chair, we must specify where the chair is, its size, and its orientation in depth. In contrast, it would seem to be advantageous for a recognition system to be invariant over position, size, and orientation. Cells in both PP and IT have large receptive fields, but the shape of the receptive fields in PP—peaked, with just a partial overlap—may be most efficient for providing coarse coding of position, whereas the relatively flat, almost completely overlapping, receptive fields in IT may be designed to produce positional invariance (O'Reilly, Kosslyn, Marsolek, and Chabris 1990).

What are cells in IT tuned to? Tanaka and his associates (Tanaka 1993; Kobatake and Tanaka 1994) have recently shown that a number of cells in area TE, a part of IT in the macaque brain presumed to mediate object

recognition, respond to complex object features—such as two black horizontal bars superimposed over two adjacent corners of a square—but not to simple features—such as the horizontal bars or oriented lines—that form parts of complex features. (Cells earlier in the ventral pathway respond best to the simple features.) The set of complex features that Tanaka discovered appear to be largely viewpoint invariant in that it would be quite easy to distinguish most of them from almost any viewpoint. For example, other complex features are an upside-down T and a large circle with a small square attached to its bottom—both of which could be readily distinguished from the square with two horizontal bars. For the most part, these complex features appear to be the kinds of representations that would be created by the geon feature assembly layer (L6) in the Hummel and Biederman network.

4.4.4 Are the Experimental Results Predictable from Theories That Perform Matching Based on Filter Representations?

Fiser, Biederman, and Cooper (1994) assessed whether the Lades et al. (1989) recognition system could account for the empirical results presented in this section. It could not. Whereas the system's recognition performance was equivalent (and reasonably accurate) for recoverable and nonrecoverable stimuli, human subjects do not recognize nonrecoverable images. The system showed a marked decrement in recognition of complementary-feature images compared to identical images, but people show equivalent degrees of priming for both types of images. As discussed in section 4.4.2, the system also did not manifest the greater sensitivity that humans reveal to viewpoint-invariant differences relative to metric differences. It is likely that the Poggio and Edelman system would perform in an identical manner to the Lades et al. system. Because a rotation in depth always produces a change in initial filter values, the Lades et al. and Poggio and Edelman systems could not manifest the invariance documented by Biederman and Gerhardstein (1993) in their experiment with depth-rotated images.

4.5 An Extension of RBC to Scene Perception

The mystery about the perception of scenes is that the exposure duration required to have an accurate perception of an integrated real-world scene is not much longer than what is typically required to perceive individual objects. The recognition of a visual array as a scene requires not only identification of the various entities but also a semantic specification of the interactions among the objects as well as an overall semantic characterization of the arrangement. Perception of a scene is not, however, necessarily

derived from an initial identification of the individual objects comprising that scene (Biederman 1988). That is, in general we do not first identify a stove, refrigerator, and coffee cup, in specified physical relations and then come to the conclusion we are looking at a kitchen.

Some demonstrations and experiments suggest that geon theory may provide a basis for explaining rapid scene recognition. Robert Mezzanotte (described in Biederman 1988) showed that a readily interpretable scene could be constructed from arrangements of single geons that only preserved the approximate overall aspect ratio of the object. In these kinds of scenes, some examples of which are shown in the upper portion of Figure 4.17, none of the entities were identified as anything other than a simple volumetric body (e.g., a brick) when shown in isolation. Most important, Mezzanotte found that such settings can be recognized sufficiently quickly to interfere with the identification of intact objects inappropriate to the setting.

It is possible that quick understanding of a scene is mediated by the perception of *geon clusters*, an arrangement of geons from different objects

that preserve the relative size and aspect ratio and relations of the largest visible geon of each object. In such cases, the individual geon will be insufficient to allow identification of the object. However, just as an arrangement of two or three geons almost always allows identification of an object, an arrangement of two or more geons from different objects may produce a recognizable cluster. The cluster acts very much as a large object does. The lower section of Figure 4.17 shows possible geon clusters for the scenes in the upper section of that figure. If this hypothesis is true, fast scene perception should be possible only in scenes where such familiar object clusters are present. Although this account of scene recognition awaits rigorous experimental test, you may be able to gauge it for yourself with the television experiment described in the opening paragraph of this chapter. Are there some scenes that you cannot identify from a single glance? My own experience is that such scenes are those which do not contain a familiar geon cluster.

4.5 Overview of Theories of Object Recognition

Although we used the complexity of the primitive as an organizing theme to present the three proposals for object recognition in section 4.4, several other distinguishing characteristics of theories can help provide a general framework for understanding theorizing in this area. Some of these characteristics will be described in this section. We should also note that the three kinds of theories of object recognition we have presented in this chapter are not necessarily mutually exclusive. (Nor are they frozen entities; theorists are actively involved in advancing their development.) It is possible that they address different aspects of object recognition or object recognition under different conditions. Because there are many ways in which an image can be classified on the basis of its shape, there is probably no single route to classification. It might be more profitable to consider the conditions under which one or the other forms of processing might be involved. We do this while discussing the distinguishing characteristics of the theories.

4.5.1 Viewpoint Dependence and the Development of Invariant Representations

The representations of Lades et al. (1993) and Poggio and Edelman (1990) theories are viewpoint dependent in that recognition is achieved by template deformation or interpolation or extrapolation during matching. The template is created directly from a single image in the Lades et al. system and from a set of similar images in the Poggio and Edelman system. Both models are self-organizing systems in which connection weights are automatically determined by the stimuli presented.

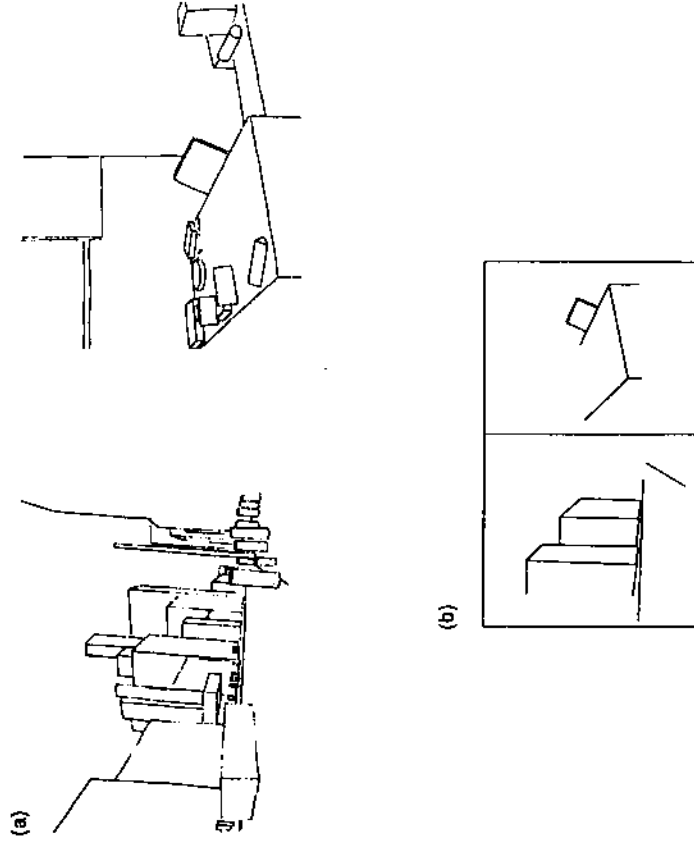


Figure 4.17
(a) Two of Mezzanotte's scenes, "City Street" and "Office." (b) Possible geon clusters for the scenes in (a).

The representation activated in L7 in the Hummel and Biederman (1992) geon model is viewpoint invariant. One cannot play the model backwards and determine the precise orientation of the image that resulted in the activation of layers 3 through 7. Whereas it is clear how the RBF units in the Poggio and Edelman model develop as a consequence of experience with a set of images, the development of the intermediate layers in the Hummel and Biederman model are largely unspecified. It is possible, of course, that the units allowing for structural description of an object are genetically determined, such an assumption, however, flies in the face of an enormous amount of recent evidence indicating that the development of neural connectivity is activity dependent. Genetics provides only a rough scaffolding within which certain kinds of very common stimulation, such as long edges, serve to determine the course of neuronal interactions.

The self-organizing systems of the kind proposed by Lades et al. and Poggio and Edelman could thus describe the development of a network that maps filter values onto hidden units that, in turn, allow the activation of a structural description of an object. Over the first months or year of life, these units would self-organize in response to recurrent activity driven by the statistical regularities of experience—for example, that co-terminating edges terminate at a common point in depth. Later units would respond to the part structure and viewpoint-invariant properties of an object, such as those expressed in the Hummel and Biederman system. This visual experience would necessarily be one of specific images, which is all anyone experiences, and the neural activity derived from those images would sculpt the connections to intermediate units. Because geons are largely viewpoint invariant, they will be better able to activate a unit tuned to them for a greater proportion of the viewing experience than a unit tuned to a shape that is highly irregular. Consequently, units tuned to geons might have a better chance of achieving a stable self-organization. Once developed, intermediate (geon) units might be employed for any new object in much the same way, perhaps, that units for phoneme (or syllable) recognition developed in infancy can be employed for representing novel words and names.

4.5.2. Bottom-Up versus Bottom-Up and Top-Down Systems

The Lades et al., Poggio and Edelman, and Biederman models assume that one-way, bottom-up processing proceeds from image to activation of the representation of the object. The Lowe and Ullman models assume model-based matching in which some initial processing constrains further testing. Does object recognition always proceed as a bottom-up, one-way street? Probably not. For example, there are times when we are stumped in our first pass at looking at an object. At such times, we might perform some mental operation (such as the mental rotation discussed by Kosslyn in

Chapter 7) to determine whether some possible object corresponds to the image. In general, however, the extraordinary speed of object recognition in the absence of any context argues against a central role for top-down information.

4.5.3. Metric versus Viewpoint Invariant Differences

Whereas geon theory assigns great weight to viewpoint-invariant properties, the filter-matching models do not. Lowe's SCERPO model employs viewpoint-invariant properties in its initial hypothesis selection but thereafter the matching is performed on representations that are metrically specified. Viewpoint-invariant differences appear to be of little use in distinguishing faces, whereas the fine metric detail captured by filter-matching models enjoy considerable success in this regard. Interestingly (as discussed by Farah in Chapter 3), object and face recognition may involve different cortical loci.

What about subordinate differences among objects, such as those that distinguish a Mazda 626 from a Honda Accord? Some investigators (e.g., Bulthoff and Edelman 1992), have argued that an application of RBF theory can provide an account of how human recognition performance degrades with the introduction of differences in orientation among highly similar objects, such as a set of paper clips that differ only in the angles between their segments. Although there may be occasions when we have to rely on fine metric detail, Biederman and Shiffrar (1988) argue that when faced with a problem of discriminating among highly similar objects, people most often search for some viewpoint-invariant difference, albeit at a small scale. In the case of cars, it may be the logo or nameplate. Imagine that you have to distinguish among a set of chairs from a dining room set that are all the identical model. How would you do it? Most likely you would search for a scratch or stain or irregularity that is viewpoint invariant in that you could employ it as long as a particular surface is in view.

4.5.4. Concluding Remarks and Future Directions

Shortly after I assumed a new position at the University of Minnesota in September 1987, I was interviewed by a reporter for the University newspaper. A student photographer listened quietly to the interview until I described a possible application of my research in object recognition to development of a robot vision system that could be employed for inventory control. She then blurted out, "I would love to have such a robot. It could pick up after me!" Despite intense research activity during the intervening seven years, we still do not have an object recognition system that can come anywhere close to matching the capabilities of the human.

But the photographer's remark underscores one reason why there is little doubt that this intense activity will continue. The potential payoffs of an artificial object-recognition system are enormous.

Closer to the hearts of most vision scientists, however, is the appreciation that visual recognition is simply too central and extraordinary an activity, as described in the first paragraph of this chapter, to remain a mystery. About 50 to 65 percent of the primate cortex is devoted to vision. A considerable fraction of this is in the extrastriate cortex, in regions presumed to be involved in higher-level vision. The absolute amount of cortex is, of course, much greater in humans than it is in the monkey, and the proportion of cortex devoted to higher level vision appears to be greater as well. By most accounts, demands on vision for information about the physical world, such as the detection of motion or visual representations for motor interactions, do not noticeably differ between human and monkey. What then is the function of the larger expanse of human extrastriate cortex? Nobel Prize winner Francis Crick has recently argued (1994) that vision can provide the basis for an attack on that most ineffable quality of mind: consciousness. Mysteries like these just won't go away.

The vision scientists involved in this mission are psychophysicists, computer scientists, neuroscientists, and cognitive neuroscientists. Discussions among investigators flow freely over issues of computational and neural-net modeling, perceptual data, and neural coding and structuring. Part of the joy in working on visual recognition is the great diversity of talent and knowledge that one is exposed to everyday. What are the likely directions of this activity? In theorizing, there is little doubt that symbolic psychological theories have given way to theories of subsymbolic operations, as expressed in neural-net modeling. This is a natural course of progression for an activity that is assumed to operate neurally. In my opinion this trend will continue, if not accelerate.

Suggestions for Further Reading

An excellent treatment of many of topics discussed in this chapter, and of presumed neural mechanisms, can be found in Kosslyn's recent book (1994). A somewhat more popular treatment of these issues is in Kosslyn and Koenig 1992. A delightful treatment of vision that, nonetheless, comes to full grips with the deep problems of neural representation and consciousness is in Crick 1994. A description of a fully self-organizing, neural-net system for recognition is presented in Waxman, Seibert, Bernardon, and Fay 1993.

Problems

4.1 Consider the features that might be used to distinguish between English capital letters say, between an A and an H, a C and a V, and a C and an H. Characterize their differences in terms of viewpoint-invariant properties. Which pair would you expect to be

least confusable? Why? Because print is displayed on flat surfaces there is little or no requirement for depth invariance. Why are viewpoint-invariant properties relevant?

4.2 After reading Chapter 3 on face recognition, discuss how the Lades et al. (1993) face-recognition system might provide a basis for creating an integrated (or holistic) representation of a face.

4.3 Which layer(s) of the Hummel and Biederman (1992) network (Figure 4.11) are likely to be the locus (or loci) of priming, as revealed in the complementary-feature and complementary-parts experiments described in section 4.4.1?

References

- Biederman, I. (1987). Recognition-by-components: A theory of human image interpretation. *Psychological Review* 94, 115–147.
- Biederman, I. (1988). Aspects and extensions of a theory of human image understanding. In Z. Pylyshyn, ed., *Computational processes in human vision: An interdisciplinary perspective*. New York: Ablex.
- Biederman, I., and E. E. Cooper (1991a). Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology* 23, 393–419.
- Biederman, I., and E. E. Cooper (1991b). Evidence for complete translational and reflectional invariance in visual object priming. *Perception* 20, 585–593.
- Biederman, I., and E. E. Cooper (1992). Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance* 18, 121–133.
- Biederman, I., and P. C. Gerhardstein (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance* 19, 1162–1182.
- Biederman, I., H. J. Hillon, and J. E. Hummel (1991). Pattern goodness and pattern recognition. In J. R. Pomerantz, and G. R. Lockhead, eds., *The perception of structure*, 73–95. Washington, D.C.: American Psychological Association.
- Biederman, I., and G. Ju (1987). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology* 20, 38–64.
- Biederman, I., R. J. Mezzanotte, and J. C. Rabinowitz (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology* 14, 143–177.
- Biederman, I., and M. M. Shiffrar (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13, 640–645.
- Brooks, R. A. (1981). Symbolic reasoning among 3-D models and 2-D images. *Artificial Intelligence* 17, 205–244.
- Binford, T. O. (1971). Visual perception by computer. *IEEE Systems Science and Cybernetics Conference*. Miami, December 1971.
- Bülthoff, H. H., and S. Edelman (1992). Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences* 89, 60–64.
- Connell, J. H. (1985). Learning shape descriptions: Generating and generalizing models of visual objects. Unpublished master's thesis, Cambridge: Massachusetts Institute of Technology.
- Cooper, E. E., and I. Biederman (1993). Geon differences during recognition are more salient than metric differences. Poster presented at the Meeting of the Psychonomics Society, Washington, D.C., November.

- Crick, F. (1994). *The Astonishing hypothesis: The scientific search for the soul*. New York: Scribner's.
- Dickinson, S. J., A. P. Pentland, and A. Rosenfeld (1992). From volumes to views: An approach to 3-D object recognition. *Computer Vision, Graphics, and Image Processing: Image Understanding* 55, 130-154.
- Edelman, S. (1993). Representation, similarity, and the torus of prototypes. Weizmann Institute (Rehovot, Israel) Technical Report CW93-10.
- Fiser, J., I. Biederman, and E. E. Cooper (1994). Are the direct outputs of Gabor filters sufficient for human object recognition or are they only the prior stage for intermediate representations? Poster presented at the Annual Meeting of the Association for Research in Vision and Ophthalmology, Sarasota, FL, May.
- Gibson, E. J. (1969). *Principles of perceptual learning and development*. New York: Appleton-Century-Crofts.
- Haber, R. N., and M. Hershenson (1981). *The psychology of visual perception*. New York: Holt, Rinehart, and Winston.
- Hoffman, D. D., and W. Richards (1985). Parts of recognition. *Cognition* 18, 65-96.
- Huttenlocher, D. P., and S. Ullman (1987). Object recognition using alignment. In *Proceedings of the First International Conference on Computer Vision*, IEEE Computer Society, 102-111. London, June.
- Hummel, J. E., and I. Biederman (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review* 99, 480-517.
- Intraub, H. (1981). Identification and naming of briefly glimpsed visual scenes. In D. F. Fisher, R. A. Monty, and J. W. Senders, eds. *Eyemovements: Cognition and Visual Perception*. Hillsdale, NJ: L. Erlbaum.
- Iitelson, W. H. (1952). *The Ames demonstrations in perception*. New York: Hafner.
- Jolicoeur, P., M. A. Gluck, and S. M. Kosslyn (1984). Picture and names: Making the connection. *Cognitive Psychology* 16, 243-275.
- Kimia, B. B., A. R. Tannenbaum, and S. W. Zucker (1995). Shapes, shocks, and deformations, I: The components of shape and the reaction-diffusion space. *International Journal of Computer Vision*.
- King, M., G. E. Meyer, J. Tangney, and I. Biederman (1976). Shape constancy and a perceptual bias toward symmetry. *Perception & Psychophysics* 19, 129-136.
- Kobatake, E., and K. Tanaka (1994). Neuronal selectivity to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology* 71, 856-867.
- Kosslyn, S. M. (1994). *Image and brain: The resolution of the imagery debate*. Cambridge, MA: MIT Press.
- Kosslyn, S. M., and O. Koenig (1992). *Wet mind: The new cognitive neuroscience*. New York: Free Press.
- Lades, M., J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers* 42, 300-311.
- Lowe, D. (1984). Perceptual organization and visual recognition. Unpublished doctoral dissertation, Stanford University.
- Lowe, D. G. (1987). The viewpoint consistency constraint. *International Journal of Computer Vision* 1, 57-72.
- Maer, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: W. H. Freeman.
- O'Reilly, R. C., S. M. Kosslyn, C. J. Marsolek, and C. F. Chabris (1990). Receptive field characteristics that allow parietal lobe neurons to encode spatial properties of visual input: A computational analysis. *Journal of Cognitive Neuroscience* 2, 141-155.
- Poggio, T., and S. Edelman (1990). A network that learns to recognize three-dimensional objects. *Nature* 343, 263-266.
- Rosch, E., C. B. Mervis, W. D. Gray, and P. Boyes-Braem (1976). Basic objects in natural categories. *Cognitive Psychology* 8, 382-439.
- Siddiqi, K., K. Tresness, and B. B. Kirma (1994). Parts of visual form: Ecological and psychophysical aspects. Laboratory for Engineering, TR LEMS-104, Brown University.
- Tanaka, K. (1993). Neuronal mechanism for object recognition (1993). *Science* 262, 685-688.
- Tversky, B., and K. Hemenway (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General* 113, 169-193.
- Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition* 32, 193-254.
- Ullman, S., and R. Basri (1990). Recognition by a linear combination of models. A. I. Memo No. 1152, Artificial Intelligence Laboratory, MIT.
- Waxman, A. M., M. Seibert, A. M. Bernardon, and D. A. Fay (1993). Neural systems for automatic target learning and recognition. *Lincoln Laboratory Journal* 6, 77-116.
- Zerroug, M., and R. Nevatia (1995). Volumetric descriptions from a single intensity image. *International Journal of Computer Vision*, in press.