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 doubled 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](image)

**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 L. 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 \( h \) 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 luminence, 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-
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
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 (Littleton 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
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 Ladet 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 Józef 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.
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.)
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 segmenting 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.