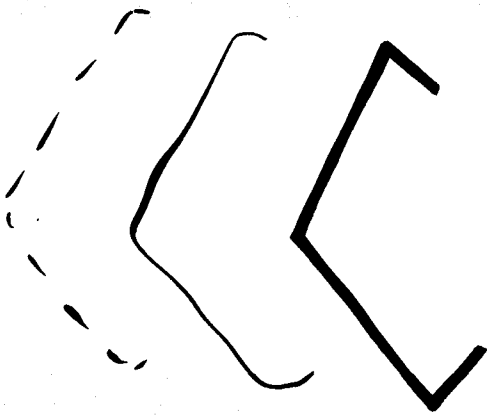
Chapter 2

Higher-Level Vision

Irving Biederman



Try this demonstration. Turn on your TV 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 should you be transported into the scene itself. Experimental observations confirm these subjective impressions (Intraub 1981; Biederman, Mezzanotte, and Rabinowitz 1982). In a 100-millisecond exposure of a novel scene, people can usually interpret its meaning. However, they cannot attend to every detail; they attend to some aspects of the scene—objects, creatures, ex-

The writing of this chapter was supported by grants 86-0106 and 88-0231 from the Air Force Office of Scientific Research.

pressions, or actions—and not others. In this chapter we will primarily focus on our ability to recognize a pattern in a single glance and our limitations in attending to simultaneous entities in our visual field.

2.1 The Problem of Pattern Recognition

2.1.1 The Nature 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 an instance of a particular class was viewed it projected the same image to the retina, as occurs, for example, with the digits on a bank check when they are presented for reading by an optical scanner.

But there is a fundamental difference between reading digits on a check and recognizing objects in the real world. The orientation in depth of an object can vary so that any one three-dimensional object can project an infinity of possible images onto a two-dimensional retina, as shown in figure 2.2a. Not only might the object be viewed from a novel orientation, it might be partially occluded behind another surface, or it might be broken into little pieces, as when viewed behind light foliage or drapes, or it might be a novel instance of its class, as for example when we view a new model of a chair. But it is precisely this variation—and the apparent success of our visual system and brain in achieving recognition in the face of it—that makes the problem of pattern recognition so interesting.

Two major problems must be addressed by any complete theory of object recognition. The first is how to represent the information in the image so that it could activate a representation in memory under varied conditions. Here we will focus on the representation of three-dimensional objects, because the problems of stimulus representation have been most extensively studied for such inputs. The second problem is how that stimulus representation is matched against—or activates—a representation in memory. Here we will concentrate on the perception of words.

2.1.2 Representing the Image

Over half a century ago the Gestalt psychologists noted that there is strong agreement among observers concerning the organization of a given pattern. Their observations led to the development of several principles of perception, such as the principle of good continuation, which holds that points that are aligned in a straight line or a smooth curve are interpreted as belonging together, and the law of Prägnanz or good figure, which

holds that patterns are seen in such a way that the resulting structure is as simple as possible. The Gestalt demonstrations have become standard fare in most introductory books in psychology and perception.

For decades the Gestalt principles of perceptual organization stood as a curious phenomenon in most treatises on perception, with no explicit link to pattern recognition. Recently there has been considerable success in interpreting these organizational phenomena as special cases of *constraints* imposed by the visual system that (1) allow a solution to the problem of interpreting a three-dimensional world from the two-dimensional image, even when that image is perturbed by noise, and (2) reveal the part structure of an image. These constraints may offer a basis for the construction of a theory of object recognition.

Viewpoint-invariant image properties play a significant role in the task of interpreting a three-dimensional world from a two-dimensional image. Figure 2.1 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 was 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 out of the plane will readily reveal that fact.

Figure 2.1 illustrates several nonaccidental properties. In the two-dimensional image, if an edge is straight (collinear) or curved, then it is perceived as a straight or curved edge, respectively. These two constraints imply, of course, the Gestalt principle of good (or smooth) continuation. 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 a theory of object recognition, figure 2.1 presents these viewpoint-invariant properties as dichotomous *contrasts*. Given an edge, it can be characterized as straight or curved. For two or more edges, the relation can be described as coterminating or noncoterminating, parallel or nonparallel, symmetrical or asymmetrical. 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, T, or arrow (or their curved counterparts). In a strict sense, parallelism and symmetry will vary with viewpoint and orientation, as

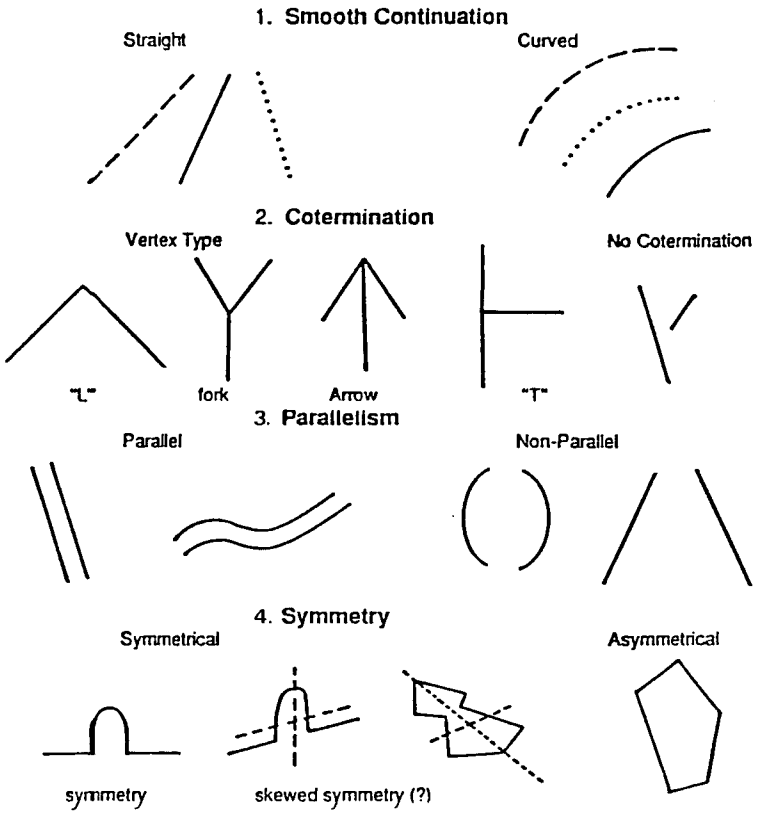


Figure 2.1
Contrasts in four viewpoint-invariant relations. In the case of Parallelism and Symmetry, biases toward parallel and symmetrical percepts when images are not exactly parallel or symmetrical are evidenced. (Adapted by permission of the author from D. Lowe, Perceptual organization and visual recognition, 1984, p. 77, fig. 5.2. Doctoral dissertation, Stanford University.)

occurs, for example, with perspective convergence. But there is a clear bias toward interpreting approximately parallel or symmetrical edges as parallel or symmetrical (Ittelson 1952; King et al. 1976). Within a tolerance range defined by the cues for surface slant, pairs of image edges that *could be* parallel or symmetrical, given uncertainty about the actual orientation of the edges to the eye, are interpreted as parallel or symmetrical (King et al. 1976).

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 wire edges through a peephole, as shown in figure 2.2 (Ittelson 1952). Although all three stimulus arrangements shown in figure 2.2b projected the identical image of a chair, as shown in figure 2.2c, in only one of them (the left-hand one) did the edges actually form a chair. In the middle arrangement the segments all had the same cotermination points as the segments of 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. These results provide strong evidence that the viewpoint-invariant properties shown in figure 2.1 and the biases toward parallelism and symmetry are immediate and compelling and thus could serve as a basis for characterizing image edges for the purposes of recognition.

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 and the base of a lamp. This decomposition does not depend on familiarity with an object or on differences in surface color or texture, as shown by the fact that it is readily performed on a line drawing. Even nonsense shapes elicit strong agreement among observers concerning their part decomposition.

In general, whenever there are a pair of matched cusps, observers 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) have noted a result from projective geometry—the *transversality* principle—that whenever two shapes are combined, their join is almost always marked by matched concavities, as illustrated in figure 2.3a. Segmenting at such regions provides a basis for appreciating the part structure of objects, as shown for the flashlight in figure 2.3b. The transversality principle also provides much of the basis for the Gestalt principle of good figure. If a shape is segmented at paired cusps, the resulting parts are convex or only singly concave. Such parts appear simple.

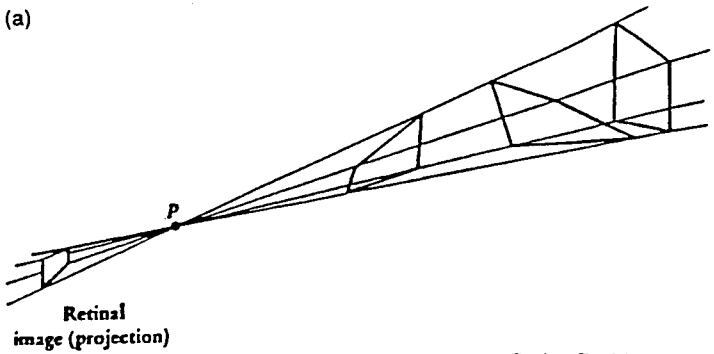


Figure 2.2
The Ames peephole perception demonstrations. (a) Illustration of the inverse optics problem: A single image can be produced by an infinity of possible real-world objects. (b) Three stimulus arrangements constructed by the Ames group. The left-hand panel shows the perspective lines from the peephole at the lower right. (c) The percepts from the stimuli in (b). (Adapted by permission of the publisher and author from R. N. Haber and M. Hershenson, *The psychology of visual perception*, 1981, p. 284, fig. 12.5. © 1981 by Holt, Rinehart and Winston.)

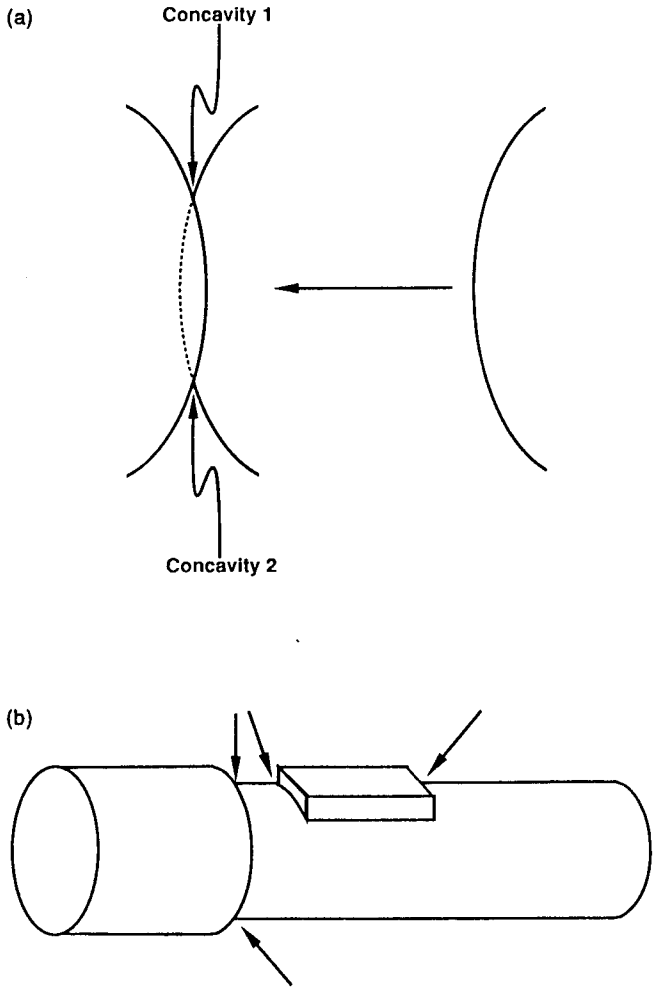


Figure 2.3
An illustration of the transversality principle and how it can be applied to the segmentation of an object's parts.

2.2 RBC: A Theory of Object Recognition

The theory of object recognition known as *recognition-by-components* (RBC) (Biederman 1987, 1988) explains how the edges that have been extracted from an image (as described in chapter 1) could activate an entry-level representation of that object in memory.

Entry level is a term coined by Jolicoeur, Gluck, and Kosslyn (1984) to refer to the initial classification of individual visual entities—for example, a chair, a giraffe, or a mushroom—that share a characteristic shape. Often the term that represents this classification (*chair, giraffe, mushroom*) will be the first that enters the child's vocabulary, and it will be used to a much greater extent than any other term to describe that entity. Entry-level classification is to be distinguished from *subordinate* classification, as for example when a particular subspecies of giraffe is specified. It is also to be distinguished from *superordinate* classification, in which shape descriptions are not specified; *mammal* and *furniture* are terms used at this level. If an entity is not typical for its class, such as penguins and ostriches for the class of birds, then entry-level classification is assumed to be at the individual level; that is, we would first classify the image as a penguin before we determined that it was a bird. Biederman (1988) has estimated that there are approximately 3,000 common entry-level terms in English for familiar concrete objects.

The central assumption of RBC is that a given view of an object is represented as an arrangement of simple primitive volumes, called *geons* (for *geometrical ions*). Five (of the 24) geons are shown in figure 2.4. The relations among the geons are also specified, so that the same geons in different relations will represent different objects (see the cup and pail in figure 2.4). The geons have the desirable properties that they can be distinguished from each other from almost any viewpoint and that they are highly resistant to visual noise. The objects shown in figure 2.4 also illustrate a derivation from the theory: An arrangement of three geons will generally be sufficient to classify any object. We will consider in greater detail the segmenting of the image into regions that will be matched to 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.

2.2.1 Geons from Viewpoint-Invariant Edge Descriptions

According to RBC, each segmented region of an image 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

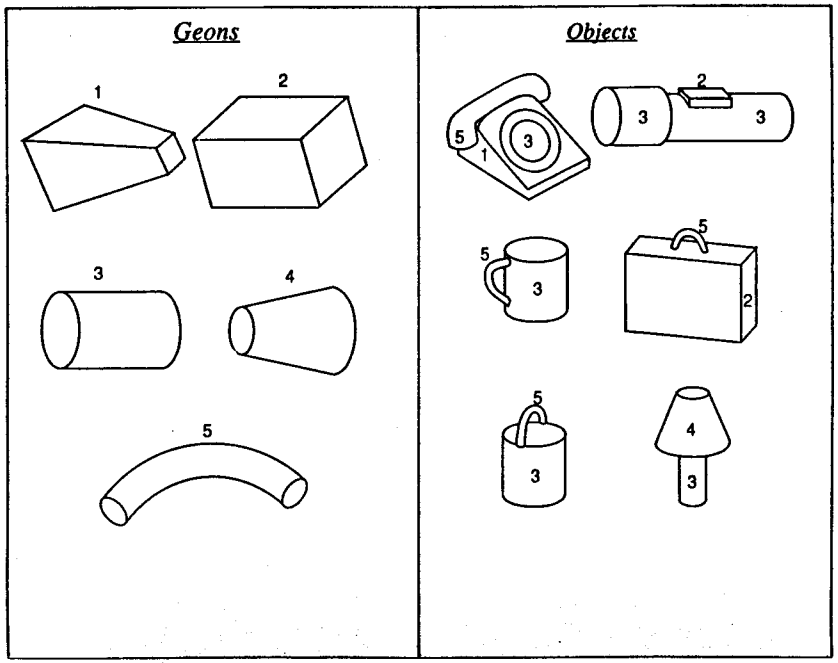


Figure 2.4
(Left) A given view of an object can be represented as an arrangement of simple primitive volumes, or geons, of which five are shown here. (Right) Only two or three geons are required to uniquely specify an object. The relations among the geons matter, as illustrated with the pail and cup.

cone is the volume swept out by a cross section moving along an axis. Marr (1977) showed that the contours generated by any smooth surface could be modeled by a generalized cone with a convex cross section.

The set of geons is so defined that they can be differentiated on the basis of dichotomous or trichotomous contrasts of viewpoint-invariant properties to produce 24 types of geons. The contrasts of the particular set of nonaccidental properties shown in figure 2.4 were emphasized because they may constitute a basis for generating this set of perceptually plausible components. Figure 2.5 illustrates the generation of a subset of the 24 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 *symmetry* (symmetrical versus asymmetrical). One attribute specifies the axis: straight versus curved.

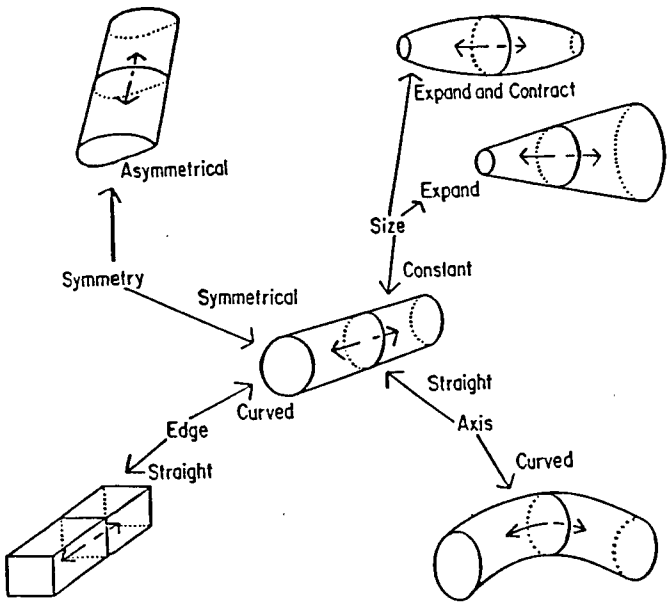


Figure 2.5
An illustration of how variations in three attributes of a cross section (curved versus straight edges; constant versus expanded versus expanded and contracted size; symmetrical versus asymmetrical shape) and one attribute of the shape of the axis (straight versus curved) can generate a set 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. 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 the publisher and author from I. Biederman, Recognition-by-Components: A theory of human image understanding, 1987, *Psychological Review* 94, p. 122, fig. 6. © 1987 by the American Psychological Association.)

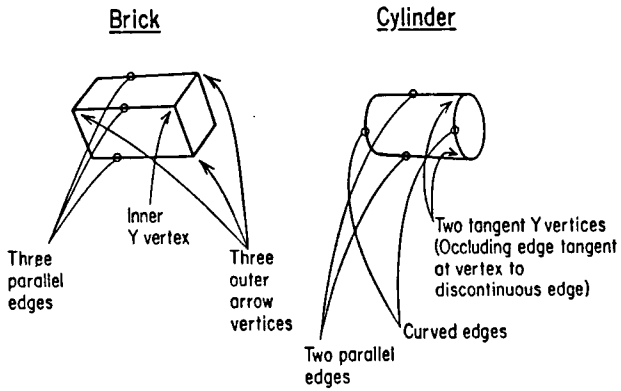


Figure 2.6
Some nonaccidental differences between a brick and a cylinder. (Reprinted by permission of the publisher and author from I. Biederman, Recognition-by-Components: A theory of human image understanding, 1987, *Psychological Review* 94, p. 121, fig. 5. © 1987 by the American Psychological Association.)

When the contrasts in the generating functions are translated into image features, it is apparent that the geons have a larger set of distinctive nonaccidental image features than the four that might be expected from a direct mapping of the contrasts in the generating function. Figure 2.6 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. By contrast, the vertices of the silhouette of the cylinder alternate between a pair of Ls and a pair of tangent Ys. The internal Y vertex is not present in the cylinder (or 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. The geons are modal types. It is possible that a given region of the image might weakly activate two or more geons if some of the distinguishing image features (vertices and edges) were missing or ambiguous.

Being derived from contrasts in viewpoint-invariant properties, the geons themselves will be invariant under changes in viewpoint. Because the geons are simple (namely, 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. Since geon activation requires only categorical classification of edge characteristics, processing can be completed quickly and

accurately. A representation that would require 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 process for object recognition.

2.2.2 Geon Relations and Three-Geon Sufficiency

According to RBC, the capacity to represent the tens of thousands of object images that people can rapidly classify derives from the employment of several viewpoint-invariant relations among geons (for example, TOP-OF, SIDE-CONNECTED, LARGER-THAN). These relations are defined for joined pairs of geons such that the same geons represent different objects if they are in different relations to each other, as with the cup and the pail in figure 2.4. How the relations among the parts of an object are to be described is still an open issue. The current version of RBC specifies 108 possible combinations of six types of relations. Also specified is a categorization of the relative aspect ratio of the geon (axis larger than, smaller than, or equal to the cross section).

With 24 possible geons, the variations in relations and aspect ratio can produce 186,624 possible two-geon objects ($24^2 \times 108 \times 3$). A third geon with its possible relations to another geon yields over 1.4 billion possible three-geon objects.

Although there are only about 3,000 common entry-level object names in English, people are probably familiar with approximately ten times that number of object models in that (1) many objects require a few models for different orientations and (2) some entry-level terms (such as *lamp* and *chair*) have several readily distinguishable object models (Biederman 1988). An estimate of the number of familiar object models would thus be on the order of 30,000. If these familiar models were 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 was 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.

2.2.3 Stage Model

Figure 2.7 presents an overall architecture for RBC. An initial edge extraction stage, responsive to differences in surface characteristics, such as sharp

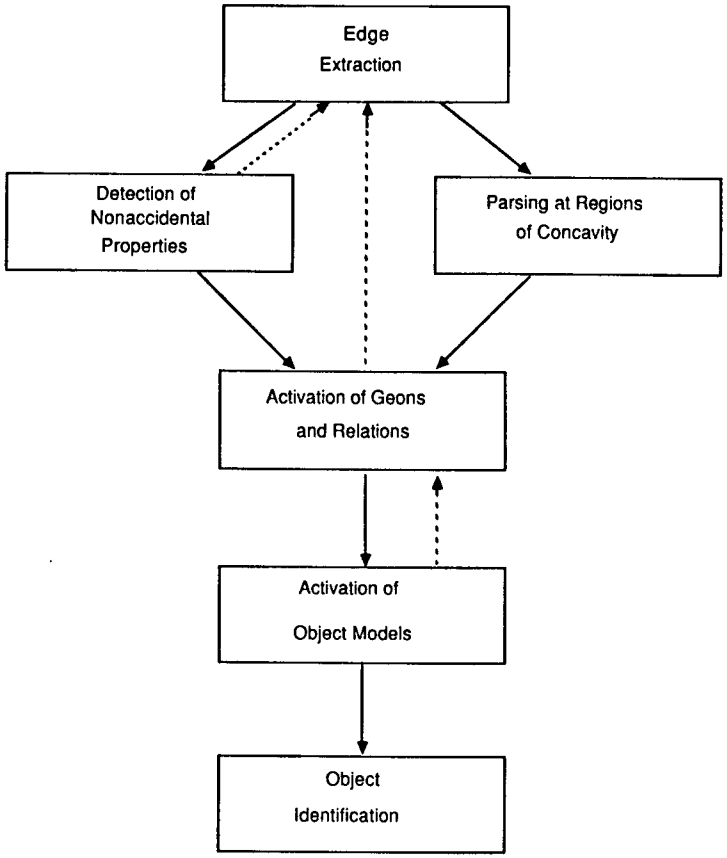
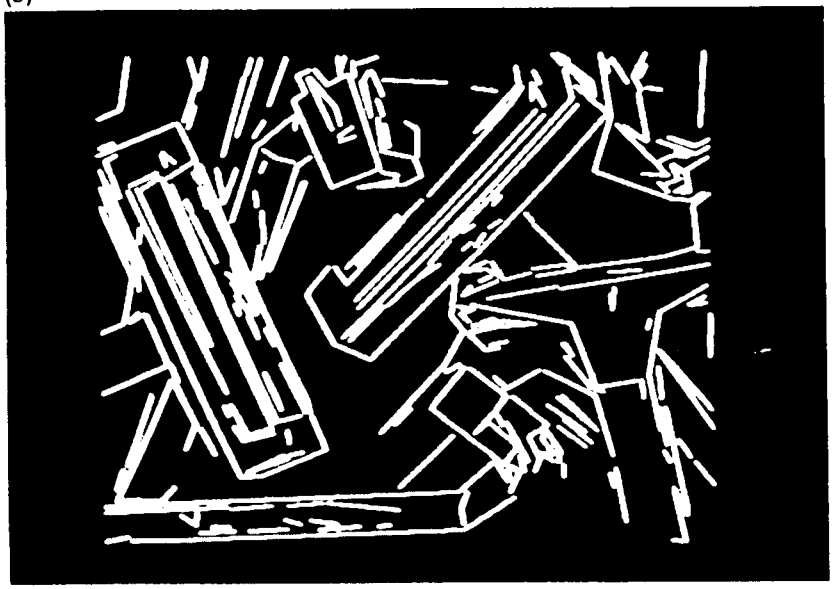


Figure 2.7
RBC's processing stages for object recognition. Possible top-down routes are shown with dashed lines. (Reprinted by permission of the publisher and author from I. Biederman, Recognition-by-Components: A theory of human image understanding, 1987, *Psychological Review* 94, p. 118, fig. 2. © 1987 by the American Psychological Association.)

(a)

(b)



(c)

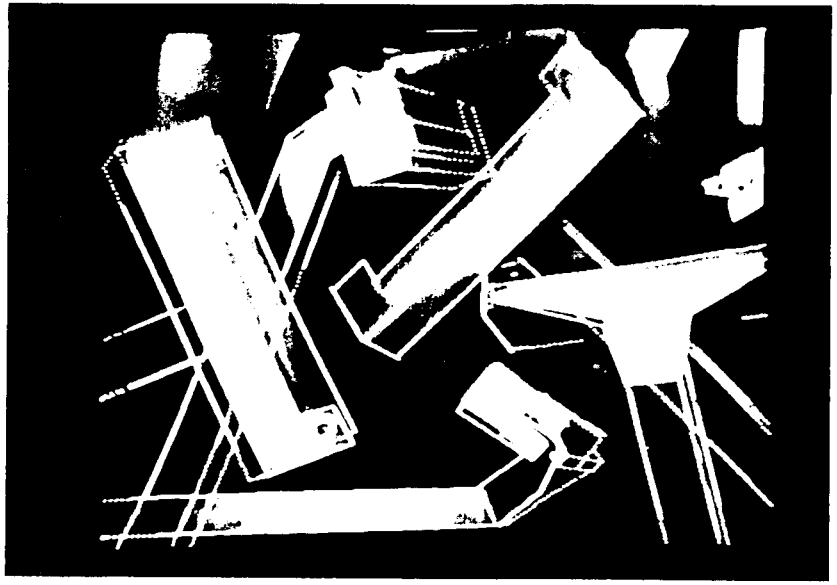


Figure 2.8
Lowe's viewpoint consistency model can find objects at arbitrary orientations and occlusions. (a) The original image of a bin of disposable razors. (b) The straight line segments that SCERPO derived from the image. (c) Final set of successful matches between sets of image segments and five 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, pp. 66 and 70, figs. 4, 5, and 8. © 1987 by Kluwer Academic Publishers.)

changes in luminance or texture, extracts the edges in the image. The image is then segmented at matched concavities and its edges characterized in terms of their viewpoint-invariant properties. (RBC also specifies additional, albeit weaker, principles for parsing (Biederman 1987). For example, a change in a viewpoint-invariant property, such as the change in parallelism from the base cylinder to the nose cone of a rocket, can also provide a (weaker) basis for parsing.) The geons and their relations are then activated, and this representation in turn activates a like representation in memory. RBC assumes that the activation of the geons and relations occurs in parallel, with no loss in capacity when matching objects with a large number of geons. Partial activation is possible, with the degree of activation assumed to be proportional to the overlap in the geon descriptions of a representation of the image and the representation in memory. Thus, an object missing some of its parts would produce weaker activation of its representation. An image from which it was difficult to determine the

geons—for example, because of low contrast—would suffer a delay in the activation of its geons. However, once the geons are activated, the activation of the object models in memory should proceed as with a sharp image.

2.2.4 Top-Down Effects and Model-Based Matching

As shown in figure 2.7, RBC is a one-way, bottom-up model proceeding from image to activation of the representation of the object. Edge extraction is assumed to be accomplished by a module that can proceed independently of the later stages, save for likely effects of the viewpoint-invariant property of smooth curvature.

Does object recognition always proceed as a largely one-way street? Probably not. When edge extraction is difficult, it is likely that top-down effects will be revealed. Such effects could be of two types: (1) they could stem from the viewpoint-invariant properties of cotermination, parallelism, and symmetry or from the geons themselves, and (2) they could stem from object models. The latter route is termed *model-based matching*. Two detailed proposals for such matching have been advanced by Lowe (1987) and by Huttenlocher and Ullman (1987). In section 2.3 we will examine how top-down effects may play a role in word perception.

Lowe's *Spatial Correspondence, Evidential Reasoning, and Perceptual Organization (SCERPO) model* is primarily directed toward determining the orientation and location of objects, even when they are partially occluded by other objects, under conditions where exact object models are available. The model takes as input an image such as the one shown in figure 2.8a, a number of disposable razors in arbitrary orientations. The model detects edges by finding sharp changes in image intensity values as reflected in the zero crossings of a $\nabla^2 G$ convolution across a number of scales, as discussed in chapter 1. The results of this edge detection stage are shown in figure 2.8b. The edges are then grouped according to viewpoint-invariant properties of collinearity, parallelism, and cotermination. A few of these image features are then tentatively matched against a component of the object model in which the orientation of the object is determined that would maximize the fit of those image features. From this initial hypothesis, the location of additional image features (edges) is proposed and their presence in the image evaluated. Figure 2.8c shows the successful final matches for five orientations of the razors. These matches provide segments not detected initially by the zero crossings (figure 2.8b) and discard edges that were initially detected but are not part of the object model, such as 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 where the initial extraction of image edges may be uncertain, as when visibility is poor or the orientation of an object is unfamiliar.

Huttenlocher and Ullman's *alignment model* first reorients all the object models that might be possible matches for the image and tests for the fit of the image against the aligned models in memory. The alignment capitalizes on a result that is similar to the structure from motion constraint proposed in theories of low-level vision: namely, that three noncoplanar 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 place where edges coterminate. However, any salient points or even general features would be sufficient for alignment. Although it would appear unlikely that people rotate (align) all possible candidate models in memory prior to matching, the alignment model offers a possible account of those cases where recognition depends on reorientation of a mental model.

2.2.5 Empirical Studies of Object Recognition

A number of experiments have been performed exploring human object recognition in general and various aspects of RBC in particular. In most of these experiments the subject names briefly presented object pictures (where "briefly" is, say, 100 milliseconds). The flash of the picture is followed by a *mask*, an array of meaningless straight and curved line segments, to reduce persistence of the image. Naming reaction times and errors are the primary dependent variables.

1. *Partial objects.* When only two or three geons of a complex object (such as an airplane or elephant) are visible, recognition can be fast and accurate (though, predictably, not as fast as when the complete image is available). This supports the principle of three-geon sufficiency. You can try this for yourself by covering up parts of pictures of common objects. See whether the object remains recognizable to a friend (who did not see the original) if three geons remain in view.

2. *Effect of object complexity.* Complex objects, defined as those such as an airplane or elephant that require six or more geons to appear complete, do not require more time for their recognition than simple objects such as a flashlight or cup (Biederman 1987). This lack of a disadvantage for complex objects is consistent with a model positing parallel activation of the geons rather than a serial trace of the contours of the object. Often a single-geon model is appropriate for several entry-level objects. Other information such as surface color or texture, small details, or context is then required to classify these objects (Biederman and Ju 1988). For example, distinguishing among a peach, a nectarine, and a plum requires that surface color and texture be specified. RBC would predict that identifying such objects would require more time than identifying objects with distinctive geon models.

3. *When does an object become unrecognizable?* Images can be rendered unrecognizable if the contour is deleted so that the geons cannot be

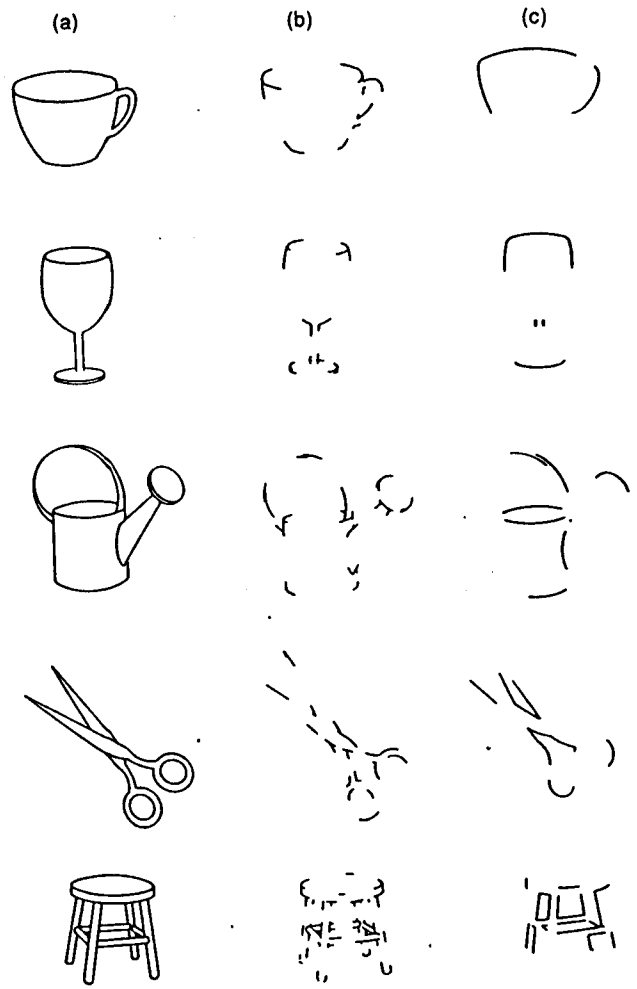


Figure 2.9
Example of five stimulus objects in the experiment on the perception of degraded objects. Column (a) shows the original intact versions. Column (b) shows the recoverable versions. The contours have been deleted in regions where they can be replaced through collinearity or smooth curvature. 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 (for example, from Ys to Ls). (Modified by permission of the publisher and author from I. Biederman, Recognition-by-Components: A theory of human image understanding, 1987, *Psychological Review* 94, p. 135, fig. 16. © 1987 by the American Psychological Association.)

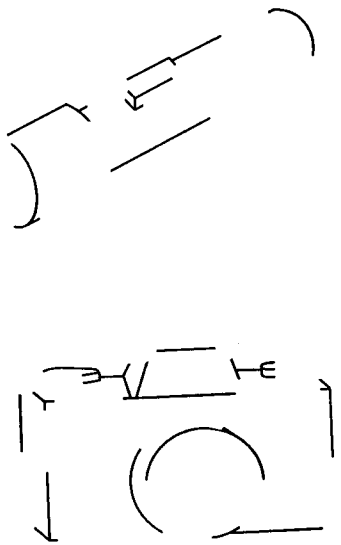


Figure 2.10
Contour-deleted images of two objects

recovered from the image. One technique is to delete the cusps to the point where the remaining contours would bridge the cusp through smooth continuation, as with the handle of the cup in figure 2.9a. Another technique is to alter vertices, as with the stool, and suggest inappropriate ones, as with the watering can. If the same amount of contour is deleted but in regions where the geons can still be activated, as shown in figure 2.9b, objects remain identifiable. Actually, even more contour can be removed from the images in figure 2.9b and they will still remain recognizable. You can test this for yourself by covering up parts of an object (say, the right or left half) and determining whether you or a friend who has not seen the original version can still identify the object.

4. *Features or geons?* According to RBC, an object is represented in terms of its geons, which are activated by image features such as vertices and edges. But if the geons are activated by image features—vertices and edges—why not just represent an object in terms of image features?

To see why this may not work, first identify the recoverable contour-deleted images shown in figure 2.10. Now look at figure 2.11. Do the images in this figure look the same as those you just viewed? Now compare them. You will note that these are actually complementary images, with each member of a pair having alternating vertices and edges. If we were to represent objects in terms of image features, we would need a different representation for each arrangement of occluding contour or for each

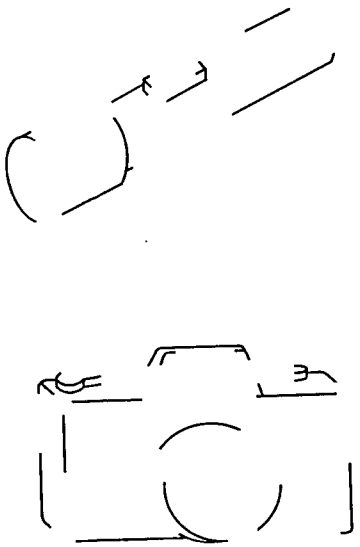


Figure 2.11
The complements to the contour-deleted images of the two objects in figure 2.10. These images contain almost all of the missing edges and vertices of the objects in 2.10, with almost no overlap between the two figures.

slightly altered orientation of the object. There is no doubt that people could code the individual image features, in that they could learn to distinguish the various versions of the complementary images. But the expectation from RBC would be that relying on such coding would slow their identification performance. That is, subjects might more readily identify the complementary version of the camera if they did not attempt to determine whether it contained the particular vertices and segments present in the original version.

5. *Rotation.* Rotation of the object in the plane slows recognition to a much greater extent than rotation in depth (Jolicoeur 1985; Bartram 1974). This result is contrary to what would be expected from the SCERPO and alignment models. According to RBC, rotation in the plane affects the TOP-OF relation, but rotation in depth leaves the geon descriptions themselves largely unaffected. At the heart of RBC is a representation that is invariant with changes of viewpoint in depth. As long as the same geon model can be activated by the image, no loss in recognition latency should be evident. However, mental rotation functions are evident in recognition performance when the TOP-OF relation is violated (Jolicoeur 1985).

6. *An extension to scene perception.* The mystery about the perception of scenes is that the exposure duration an observer requires to have an

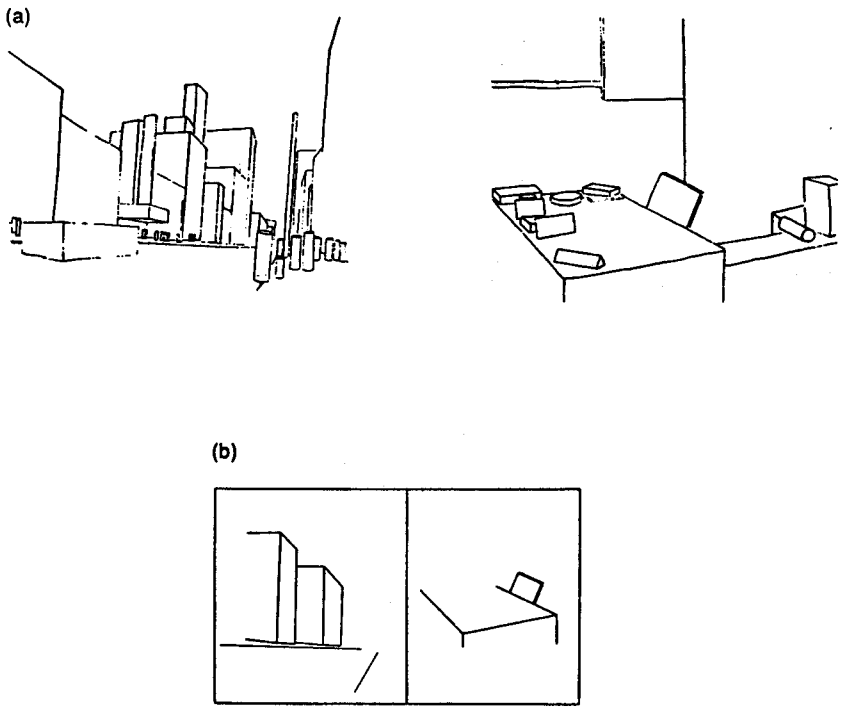


Figure 2.12
(a) Two of Mezzanotte's scenes: "City Street" and "Office." (b) Possible geon clusters for the scenes in (a).

accurate perception of an integrated real-world scene is not much longer than what is typically required to perceive individual objects. Recognizing a visual array as a scene requires not only identifying the various entities but also semantically specifying the interactions among the objects and providing an overall semantic specification of the arrangement.

However, the perception of a scene is not necessarily derived from an initial identification of the individual objects making up 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 conclude that we are looking at a kitchen.

Some demonstrations and experiments suggest that RBC may provide a basis for explaining rapid scene recognition. Mezzanotte (described in Biederman 1988) has shown that a readily interpretable scene can be constructed from arrangements of single geons that just preserve the overall aspect ratio of the objects, such as those shown in figure 2.12a. In this kind of scene none of the entities, when shown in isolation, could be

identified as anything other than a simple volumetric body, such as a brick. Most important, such settings could be recognized sufficiently quickly to interfere with the identification of intact objects that were inappropriate to the setting.

It is possible that quick understanding of a scene is mediated by the perception of *geon clusters*. A geon cluster is an arrangement of geons from different objects that preserves the relative size, 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, so an arrangement of two or more geons from different objects may produce a recognizable combination. The cluster acts very much like a large object. Figure 2.12b shows possible geon clusters for the scenes in figure 2.12a. If this account is correct, fast scene perception should only be possible in scenes where such familiar object clusters are present. This account awaits rigorous experimental test, but you may try to gauge it for yourself with the TV "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 where a familiar geon cluster is not present.

2.3 Activating a Representation: McClelland and Rumelhart's Interactive Activation Model of Word Recognition as an Example of Parallel Distributed Processing

2.3.1 The Interactive Activation Model

The recognition stage of RBC had only been sketched when it was asserted that the representation of the image might "activate" or be "matched against" a like representation in memory. We now turn to one of the few models of image understanding that provides a detailed account of the time course of memorial activation in perceptual recognition: the *interactive activation model* (IAM) of word recognition of McClelland and Rumelhart (1981). This model also provides an account of how information in memory could affect, top-down, the course of perceptual recognition.

Although we have focused in this chapter on object recognition, McClelland and Rumelhart's work on word recognition may have broad applicability to all cases of image understanding. At the very least, humans have only one visual system and it is unlikely that special-purpose mechanisms for reading have developed in the evolutionarily insignificant 5,000 years since the invention of the alphabet. We will consider IAM in some detail because it presaged much current theorizing in cognitive science

under the rubrics of *connectionism* or *parallel distributed processing* (Rumelhart and McClelland 1986).

IAM was initially designed to model why the perception of a target letter in a brief, masked presentation of a word was more accurate than when the same letters in the word formed a nonword or even when the individual target letter was presented by itself. This result was found with strict controls for guessing a word (Reicher 1969; Wheeler 1970). For example, following the presentation of a word such as READ, a nonword such as AEDR, or the letter E, the subject might be asked (in the first two cases) whether E or O was in the second letter position or (in the third case) whether the single letter was E or O. Note that either response alternative would produce a common word if inserted in R_AD. Before reading further, you might wish to try your own hand (or mind) at explaining this effect. But you should be warned that it took over a decade of intense research on this problem before a single sufficient account of these results was proposed.

IAM posits three levels of representation arranged in a hierarchy: features, letters, and words. As illustrated in figure 2.13, each level consists of a number of *nodes* at various states of activation for the entities relevant to that level. Each node is connected to a large number of other nodes from

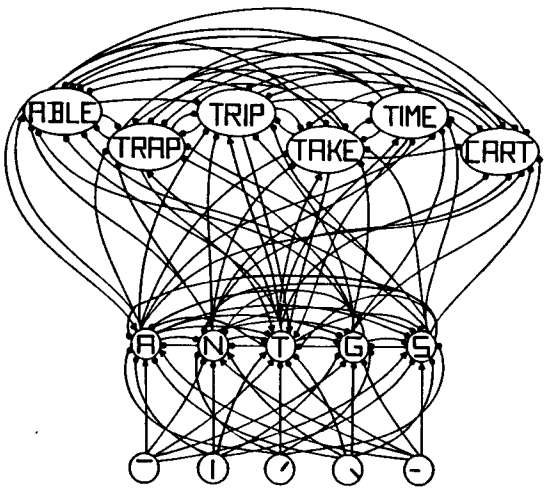


Figure 2.13
The three levels of the interactive activation model for the word superiority effect. (Reprinted by permission of the publisher and author from J. L. McClelland and D. E. Rumelhart, An interactive activation model of context effects in letter perception: Part 1. An account of basic findings, 1981, *Psychological Review* 88, p. 380, fig. 3. © 1981 by the American Psychological Association.)

which it can receive either excitatory inputs (designated by an arrow at the end of the connection in figure 2.13), which raise its activation level, or inhibitory inputs (designated by a small disk in figure 2.13), which lower its activation level. Each node, in turn, transmits its activation as excitatory or inhibitory inputs to other nodes.

The presentation of a letter (actually, the letter's features) causes excitation of the nodes consistent with that letter's features and inhibition of the nodes for those features that are inconsistent with that letter. The nodes whose activity has been increased transmit their excitation by increasing the activation of letter nodes that contain those features and inhibiting letter nodes that do not contain those features. Similarly, the activation of the letter nodes results in excitation of word nodes that contain those letters and inhibition of word nodes that do not contain those letters. At all levels there is strong intralevel inhibition. Each node at a given level inhibits the other nodes at that level. There is also top-down excitation. Activity at the word level can excite or inhibit activity at the letter level.

We can now trace the time course of activation of a given node as a word—WORK, for example—is presented. Assume that the subject successfully detected the first three letters, WOR_, but detected only some of the features of the fourth letter, as indicated in figure 2.14a. Initially there is an increase in the activation level of those letters consistent with the features actually detected (we are only considering the letter nodes corresponding to the fourth position in the word). These would be R and K. These nodes transmit their excitation to the word level. Although WORK can benefit from activation of the K node, there is no word WORR to receive activation from R. As the activation of WORK increases, it starts to excite, top-down, the K node and inhibit the R node. R starts to weaken and the activation level of K grows until it clearly exceeds the activation level of R.

We can now see how IAM handles the major phenomenon of the word superiority effect as well as the advantage of a letter within a word over the letter itself. A nonword would not have a node at the word level. Consequently, there would be less chance for a letter in that string to benefit from top-down activation. It is possible, however, that words sharing letters with the nonword might generate, through their partial activation, some top-down excitation. The individual letter, like a nonword, has less chance to benefit from top-down facilitation from the word level.

2.3.2 General Features of Parallel Distributed Models

Parallel distributed models—the class of models of which IAM was a precursor—have generated much interest in current theorizing in cognitive science, and it is worthwhile to consider some of their general characteristics:

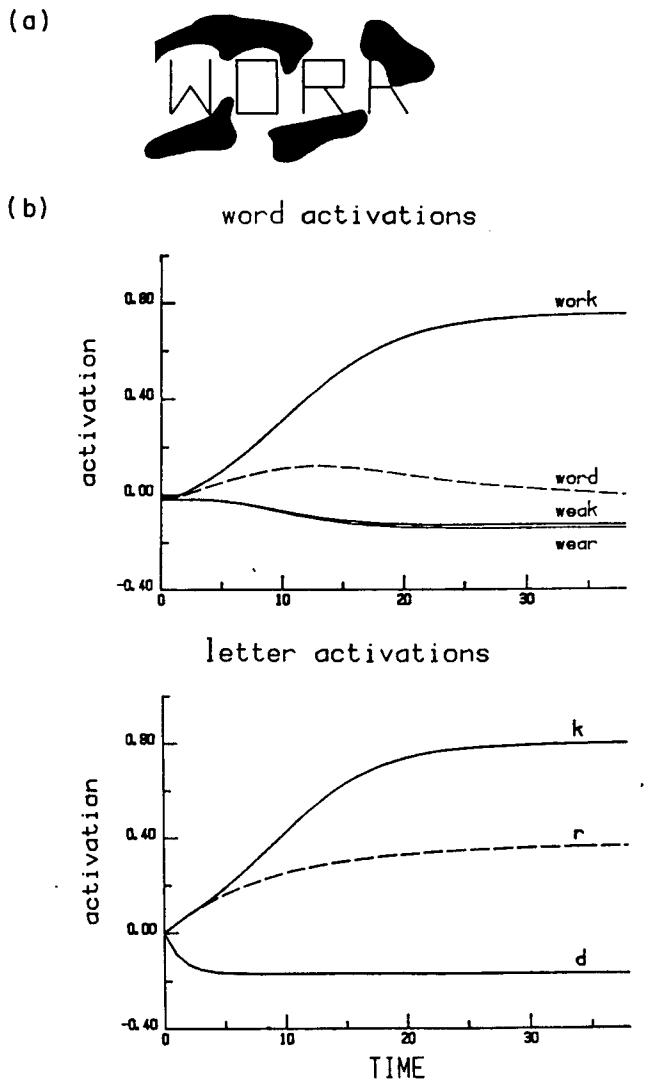


Figure 2.14
The interactive activation model's characterization of the activation of the letter K in the fourth position, given that WORK was presented and the features that were extracted were as shown in part (a). (Reprinted by permission of the publisher and author from J. L. McClelland and D. E. Rumelhart, An interactive activation model of context effects in letter perception: Part 1. An account of basic findings, 1981, *Psychological Review* 88, pp. 383 and 384, figs. 5 and 6. © 1981 by the American Psychological Association.)

1. As their name indicates and as illustrated in figure 2.13, these systems are highly parallel. Simultaneous activation is possible for a large number of nodes at a given level and at different levels.

2. The nodes are richly connected. Each node is connected to a large number of other nodes.

3. Partial activation can be transmitted. A node need not exceed a high threshold for it to affect the activation of nodes to which it is connected. It is through this characteristic that IAM was able to handle the advantage of a word over a letter. If the nodes at the word level never became activated until the letters were maximally activated (to the point where they could be identified), then the identification of letters within words could not be superior to the identification of individual letters.

4. There is no explicit appeal to rules or regularities. This point is important and deserves elaboration. In the present context the relevant rules would concern spelling regularities. For example, if we were presented with the word ____I?G, with uncertainty about whether the letter in the position of the question mark was N, R, or M, wouldn't it be reasonable to appeal to our knowledge that many words in English have an *-ing* ending and use that information as a rule to guess that the uncertain letter was N? Not according to McClelland and Rumelhart. They would argue that what appears to be rulelike behavior in perception is merely statistical behavior: the combined effect of our having been exposed to many instances of a given class, in this case the large number of *-ing* words in English.

5. Although a detailed study of connectionism is beyond the scope of this chapter, it should be noted that most current connectionist models do not assign an individual node to represent an individual entity, such as a feature, letter, or word, as in McClelland and Rumelhart's model. Instead, all the nodes at a given level might be used to code all the entities at that level. It is the *pattern* of activation over all the nodes that allows the system to discriminate various inputs. This is accomplished by adjusting the weights by which the nodes activate or inhibit other nodes.

2.4 Visual Attention

Despite our impressive capacities for recognizing an object or scene at a single glance, striking limitations in our abilities for identifying objects at different spatial locations are often evident. These limitations are often encountered, since it is rare that we are faced with only a single entity in our visual field. We often find it necessary to select for attention one from the multitude of objects that may be present in the visual field. The selection is necessary for at least three reasons.

First, on a peripheral level, only the *fovea*—which corresponds to the central 2 degrees of our retina—is capable of resolving fine detail. Even within the 2-degree area there is a marked decline in the capacity to resolve detail from an image at increasing eccentricities from fixation. One aspect of attention, then, consists of moving our eyes in a series of jumps, called *saccades*, from one part of the visual field to another. The locations fixated as we move our eyes around the scene (or page, when we are reading) are decidedly not random. We move our eyes to regions of interest. The maximum rate at which we can make saccades is about three or four per second. Most of the time is spent during the fixation itself; the actual saccade requires only 10 milliseconds.

Second, once our eye is fixated onto a region, we may benefit by shifting attention to another region within the first few hundredths of a second, even without moving our eyes, as demonstrated in an experiment by Sperling (1960). At brief durations after a display of three rows of four random-appearing letters was flashed for a few milliseconds, Sperling sounded a high, middle, or low tone as a cue for the subject to report either the top, middle, or bottom row. Subjects could use the tone to improve their accuracy on the cued row as long as the tone was not delayed beyond 200 milliseconds, at which time the trace (or icon) of the display would have disappeared. During the 200 milliseconds following the flash of the display, the benefit of the tone could not have been produced by eye movements, because the display was no longer present and would only have moved with the eyes if there was some residual retinal activity. Thus, the benefit from cueing a row could only have come from a redirection of attention to the cued row in the icon.

Reeves and Sperling (1986) have shown similar phenomena in a paradigm in which subjects monitored a stream of letters, presented at a high rate (4.6 to 13.4 letters per second) just to the left of a fixation point. When a given target letter occurred, they were to report the next digit (or digits) presented in a stream just to the right of fixation. Reeves and Sperling found that the switch took time, so that often a later letter was reported. As determined by the digit actually reported, the switch often was made by 200 milliseconds.

Third, Treisman and her associates (Treisman and Gelade 1980) have discovered that for a wide range of stimulus attributes, attention must be shifted serially from position to position when we attempt to *conjoin* independent attributes of a single stimulus (for example, color and shape).

In one of Treisman's experiments the subject's task was to hit a key if a predesignated target was present in the display (which it was on half the trials) and another key if it was absent. Two conditions will illustrate the type of display that Treisman used. In the *conjunction* search condition

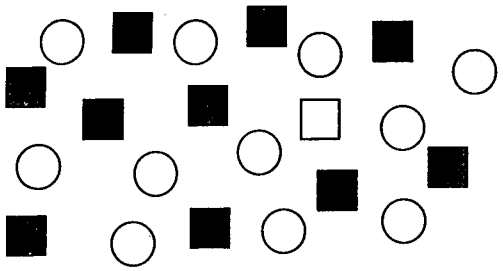


Figure 2.15
An example of a conjunctive search display of the type used by Treisman. The target is a white square (white *and* square, hence a conjunction) and the distractors are white circles and dark squares. Mean reaction times for detecting the target in such displays have been found to increase linearly as a function of the number of distractors.

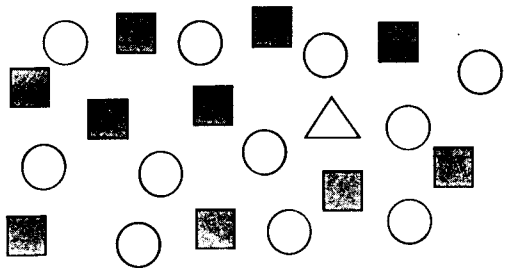


Figure 2.16
An example of a feature search display of the type used by Treisman. The target is a triangle. Mean reaction times for detecting the target in such displays have been found to be independent of the number of distractors.

the target is a white square. See whether you can find it in figure 2.15. You probably had to look carefully to find it among the white circle and dark square distractors. In the *feature* search condition the target is a triangle. Can you find it in figure 2.16? You probably felt that the triangle "popped out." Over trials, Treisman varied the number of items in the display between 1 and 38. In the feature search condition there was no increase in search times as a function of the number of distractors, indicating that all display positions could be examined simultaneously, with no loss in capacity for examining any single position. In the conjunctive condition reaction times increased linearly, suggesting a serial search of the various positions.

Treisman's explanation was that the colors (white versus dark) for the various items activate their locations on a separate "map" for each color. The shapes similarly activate their own maps for the various shape attributes. Treisman argues that we can detect activity on a given map—for

example, that there is activity on a map of diagonal edges signaling the presence of a triangle, independent of the number of locations that might possibly contain a triangle. To detect a conjunction, however, we must attend to a particular position on two or more maps. To determine whether a particular square is dark, we must attend to that location on both the map for squares and the map for color. Treisman likens this act to the movement of a narrow flashlight beam that can move through space, illuminating only a single position, on *all* the feature maps, at any one time. The search for a target defined by a conjunction must therefore be performed serially.

Treisman has documented these limitations for a variety of stimulus attributes (for instance, color and shape), and they impose a significant bottleneck on the human's capacity for attending to stimuli distributed across the visual field. It should be noted that not all combinations of stimulus attributes reveal conjunctive limitations (Nakayama and Silverman 1986). For example, given a display with red circles moving up and green circles moving down, a red circle moving down can be detected independent of the number of distractors in the display.

Is the notion of feature maps a farfetched idea physiologically? Not at all. Approximately a dozen regions of the primate visual cortex can be characterized as preserving the spatial relations among stimuli projected onto the retina. These maps have been discovered by presenting stimuli at a given part of the retina and recording the activity from single cortical neurons in a given region of the cortex. The evidence for a map is derived from the finding that adjacent regions of the retina produce activity in adjacent cells in the cortical region. Consistent with Treisman's findings, these maps appear to be specialized for particular attributes—say, a line at a given orientation—although it is common to find cells that appear tuned to several attributes.

Moran and Desimone (1985) have discovered a possible physiological substrate for Sperling's phenomenon of reallocation of spatial attention over a spatial map in the absence of eye movements. Recording from single neurons in the visual cortex (V4), these investigators found a neuron that was sensitive to a particular stimulus, for example, a vertical bar at a particular location on the retina. The monkeys in this experiment were trained to maintain their fixation on a given point (or else they were not rewarded) and to respond by releasing a switch if a second stimulus presented either within or outside of the attended region matched the one that they were attending. When a vertical bar was presented within the attended region, the cell with that receptive field fired. However, when the monkey was not attending to that region, the vertical bar did not produce a response in the neuron, even when it was presented at the identical position on the retina.

2.5 Spatial Processing and Other Visual Activities

In addition to recognition and attention there are other visual activities that might be considered in a chapter entitled "Higher-Level Vision." Other chapters will review imagery, visual development, and oculomotor control.

But much of visual processing entails the employment of space without recognition, as when we navigate in the environment, calculating a course that avoids obstacles and minimizes distances, or when we reach for an object. There is strong evidence that the areas of the brain involved in recognition are anatomically separate from the area involved in spatial processing. Mishkin and his associates (see, for example, Mishkin and Appenzeller 1987) have shown that damage to the monkey's inferior temporal (IT) cortex results in an inability to detect differences among objects but has little effect on its ability to use the knowledge of where an object was located. Damage to the posterior parietal region of the brain has the opposite effect.

The general picture of modularity that emerges from the study of behavioral phenomena in vision is apparent when the brain areas underlying a behavior as fundamental as orientation are explored. Goodale and Milner (1982) have shown that anatomically separate regions of the rodent's brain control orientation toward a goal in central vision and orientation toward stimuli in peripheral vision. Moreover, the brain areas responsible for orientation toward a goal are not the ones that appear to be involved in the avoidance of obstacles.

One of the unsolved problems in vision is how relations among parts—for example, that a cylinder might be CENTERED-BELOW a brick—are derived from the earlier maps of the visual system that preserve retinal space so that we identify the same relation even though it might be on different parts of our retina or at different orientations in depth. It is generally accepted that cells in the visual cortex such as V1, which receive inputs from the retina via the lateral geniculate body, have smaller receptive fields (respond to stimuli over a more circumscribed area of the retina) than cells in later visual regions, such as V2 (which receives inputs from V1), V4, or IT. For V1, individual neurons are tuned to patterns of a given width and orientation (say, vertical) within a 0.5-degree region. Those slightly later, in V2, have receptive fields between 0.5 and 1 degree; those in V4 respond over an area of 1 to 4 degrees; and those in IT that control object recognition have receptive fields as large as 25 degrees. Whether these changes in receptive field size are associated with the human's ability to extract relations and identify objects anywhere on the retina and from arbitrary orientations in depth is yet to be determined.

Suggestions for Further Reading

An excellent treatment of many of the topics discussed in this chapter, and particularly of the interface between lower- and higher-level vision, can be found in chapter 12 of Stillings et al. 1987. Marr 1982 has become a classic in its statement of the issues of vision and its general computational approach to many of the problems of vision. Pinker 1985 provides a broad, critical overview of the more cognitive aspects of vision, including imagery and visual representation. A good general-purpose text on perception, such as Goldstein 1988 or Sekuler and Blake 1985, provides a useful background with which to consider the early constraints on higher-level vision.

Questions

2.1 Consider the features that might be used to recognize the capital letters of the English alphabet. To what extent might they be considered to reflect the workings of a system sensitive to viewpoint-invariant properties of edges?

2.2 Discuss how viewpoint-invariant properties might provide an account of the "good" and "bad" subparts of figure 3.1.

2.3 Discuss RBC as a model analogous to McClelland and Rumelhart's model for word recognition.

2.4 Treisman's experiments reveal striking limitations on the human's ability to attend to conjunctions of stimulus attributes. Discuss why these limitations are not apparent when (1) viewing complex objects that may enjoy a slight advantage in recognition speed compared to simple objects and (2) viewing a scene composed of many objects.

2.5 After reading chapter 4 on visual development, speculate on which aspects of RBC might be present at birth.

References

Bartram, D. (1974). The role of visual and semantic codes in object naming. *Cognitive Psychology* 6, 325–356.

Biederman, I. (1987). Recognition-by-Components: A theory of human image understanding. *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.* Norwood, NJ: Ablex.

Biederman, I., and G. Ju (1988). Surface vs. 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.

Binford, T. O. (1971). Visual perception by computer. Paper presented at IEEE Systems Science and Cybernetics Conference, Miami, December.

Brooks, R. A. (1981). Symbolic reasoning among 3-D models and 2-D images. *Artificial Intelligence* 17, 205–244.

Connell, J. H. (1985). Learning shape descriptions: Generating and generalizing models of visual objects. Master's thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA.

Goldstein, E. B. (1988). *Sensation and perception.* 3rd ed. Belmont, CA: Wadsworth.

Goodale, M. A., and A. D. Milner (1982). Fractioning orientation behavior in rodents. In D. Ingle, M. A. Goodale, and R. Mansfield, eds., *Analysis of visual behavior.* Cambridge, MA: MIT Press.

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*. Washington, DC: Computer Society Press of the IEEE.

Intraub, H. (1981). Identification and naming of briefly glimpsed visual scenes. In D. F. Fisher, R. A. Monty, and J. W. Senders, eds., *Eye movements: Cognition and visual perception*. Hillsdale, NJ: L. Erlbaum Associates.

Ittelson, W. H. (1952). *The Ames demonstrations in perception*. New York: Hafner.

Jolicoeur, P. (1985). The time to name disoriented natural objects. *Memory and Cognition* 13, 289–303.

Jolicoeur, P., M. A. Gluck, and S. M. Kosslyn (1984). Picture and names: Making the connection. *Cognitive Psychology* 16, 243–275.

King, M., G. E. Meyer, J. Tangney, and I. Biederman (1976). Shape constancy and a perceptual bias towards symmetry. *Perception and Psychophysics* 19, 129–136.

Lowe, D. G. (1984). Perceptual organization and visual recognition. Doctoral dissertation, Department of Computer Science, Stanford University, Stanford, CA.

Lowe, D. G. (1987). The viewpoint consistency constraint. *International Journal of Computer Vision* 1, 57–72.

McClelland, J. L., and D. E. Rumelhart (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review* 88, 375–407.

Marr, D. (1977). Analysis of occluding contour. *Proceedings of the Royal Society of London* B197, 441–475.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.

Mishkin, M., and T. Appenzeller (1987). The anatomy of memory. *Scientific American* 256, 80–89.

Moran, J., and R. Desimone (1985). Selective attention gates visual processing in the extrastriate cortex. *Science* 229, 782–784.

Nakayama, K., and G. H. Silverman (1986). Serial and parallel processing of visual feature conjunctions. *Nature* 320, 264–265.

Pinker, S. (1985). Visual cognition. An introduction. In S. Pinker, ed., *Visual cognition*. Cambridge, MA: MIT Press.

Reeves, A., and G. Sperling (1986). Attention gating in short-term visual memory. *Psychological Review* 93, 180–206.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of the stimulus material. *Journal of Experimental Psychology* 81, 275–280.

Rumelhart, D. E., J. L. McClelland, and the PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1: *Foundations*. Cambridge, MA: MIT Press.

Sekuler, R., and R. Blake (1985). *Perception*. New York: Knopf.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs* 74 (11, Whole No. 498).

Stillings, N. A., M. H. Feinstein, J. L. Garfield, E. L. Rissland, D. A. Rosenbaum, S. E. Weisler, and L. Baker-Ward (1987). Vision. In *Cognitive Science: An Introduction*. Cambridge, MA: MIT Press.

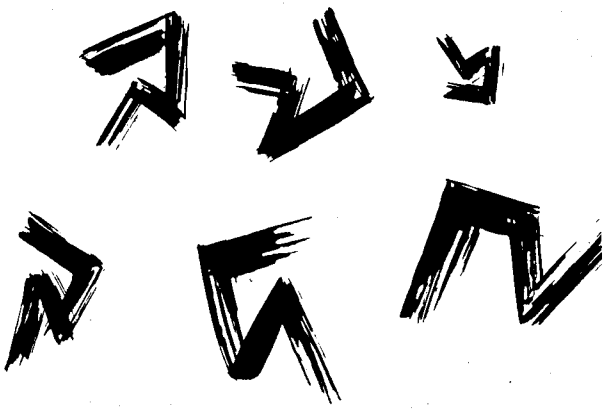
Treisman, A., and G. Gelade (1980). A feature integration theory of attention. *Cognitive Psychology* 12, 97–136.

Wheeler, D. D. (1970). Processes in word recognition. *Cognitive Psychology* 1, 59–85.

Chapter 3

Mental Imagery

Stephen Michael Kosslyn



What seems to happen when you try to decide which is higher off the ground, the tip of a racehorse's tail or its rear knees? Or when you think about how a new arrangement of furniture would look in your living room? Many people report that in performing these tasks, they "see" with their "mind's eye" the horse's tail and the furniture. Visual mental imagery is "seeing" in the absence of the appropriate immediate sensory input; imagery is a "perception" of remembered information, not new input. But if all one is doing is "mentally perceiving" what has already been perceived, what is the use of imagery? And in what way does it make sense to talk about "seeing," "hearing," and the like without actually perceiving? Indeed, the very idea of mental images is fraught with puzzles and possible paradoxes. What, exactly, is being "perceived"? Surely images cannot be actual pictures in the brain; there is no light in there, and who or what would look at the pictures, even if they were there? And, given that there are no hands in the brain, how do we "move things around" in images?

At first glance (and even at second glance!) these are knotty problems indeed. In this chapter we will explore the recent attempts to answer these and related questions. Because most research has focused on visual imagery,