PART III

Visual Object Processing

CHAPTER 7

Visual Object Processing

PETER ULRIC TSE AND STEPHEN E. PALMER

Perceptual systems have evolved to provide animals with accurate information about their environments and their bodies in real time, allowing them to plan and execute biologically relevant tasks. Without perception, an organism would quickly die from starvation, predation, or injury. In all sensory systems, some type of energy in the environment—whether from light, heat, or vibrations in the air—is transduced into neuronal action potentials that are sent to cortex and other brain areas for processing. Somewhat surprisingly, the information an organism perceives about its environment does not come just from the passive registration of sensory input, but from an active construction based on that input.

In this chapter, we focus on visual perception—the transformation of light energy to receptor activity to highly structured perceptions of meaningful objects, relations, and events—because it is the best understood of the perceptual systems and the most important for our species. The chapter is divided into four related subtopics: First we present an overview of what is known about the neural basis of vision. Second, we discuss how people organize visual input into perceptual objects. Third, we consider how shape representations are constructed. And fourth, we analyze how people identify or recognize those objects as instances of known, meaningful categories such as people, houses, trees, and cars.

Visual perception begins when photons enter the eye, but it culminates in what may be the greatest remaining puzzle confronting neuroscience, namely, the conscious experience of objects and events. Neuroscientists take it for granted that experience is realized in neural activity, but how this occurs remains a mystery. It is as if we have two sides of a bridge: We know a great deal about how neurons behave, and we know a lot about the structure of psychological experience. What we lack is a deep understanding of how physical and mental events are related. They even seem to have mutually exclusive properties. For example, neuronal activity is a physical, publicly observable phenomenon, whereas the mental events that arise from that activity are purely private.

Nonetheless, we do know quite a lot. We know that the information processing that culminates in visual experience begins as a pattern of local activations among the approximately one hundred million photoreceptors in each eye. After initial processing in the retina, image information is transduced into the action potentials of ganglion cells. This permits the transmission of image information to other areas of the brain via the optic nerve: a bundle of about a million ganglion cell axons per eye. Through some as yet poorly understood neural processing, this information is rapidly transformed into conscious experiences of three-dimensional objects that have particular sizes, shapes, motions, surface colors, and locations within the three-dimensional environment.

None of this information about the properties of environmental objects is explicitly present in the retina, because retinal information is inherently ambiguous. For example, an individual photoreceptor that responds to red light will respond equally to a red surface illuminated with white light and a white surface illuminated with red light. By itself, this cell cannot specify the surface color, much less the shape of an environmental object. In an important sense, object information has to be 'inferred' from the information available in the retinal image. Here, we do not mean the same thing that most people generally mean by '*inference*', which is based on what is perceived

and follows it in time. For example, when you go outside and see that, say, a bench is wet, you may then cognitively infer that it must have rained. This kind of inference follows perception in time and is necessarily based on the prior perception of environmental objects. However, the great German scientist and mathematician Hermann von Helmholtz spoke of a different kind of inference that temporally and logically must precede cognitive inference. He called it 'unconscious inference' (Helmholtz, 1867). By this he meant that our initial perception that the bench is wet is not cognitively inferred—like the inference that it must have rained—but perceptually inferred without conscious deliberation. Rather, we just see it as wet. Our conscious experience is rich with the conclusions of such unconscious inferences. Not only do we see the bench as wet, we see it as having a particular surface color, an intrinsic size and shape, and perhaps even as being made of a particular material. None of these perceptions was the result of a conscious, cognitive decision. We cannot help but see things as they are presented to us by unconscious processes that convert patterns of local activations at the retina into rich, integrated visual experiences of objects and events in an external physical world.

The unconscious inferences built into the transformations that convert retinal information into perceptions of environmental objects must be both accurate and rapid, so that we can see the world with reasonable accuracy in real time. The problem is that the image is ambiguous in multiple ways, and this, in turn, means that perceptual processes are highly underconstrained by the input. To overcome this ambiguity, the transformations that convert retinal information into perceptions of the environment must be based on many implicit assumptions that constrain the mapping between them. Usually, these assumptions are valid enough to recover events as they really are, but sometimes the visual system comes to the wrong conclusion, and we see something that we know could not be the case. One such example is the illusion of seeing the flashing lights on top of an ambulance move laterally back and forth. Cognitively, we may know that there is no lightbulb actually sliding back and forth, but because the outputs of the informational transformations that produced this experience are not influenced by cognitive knowledge, we cannot help but see the lights as moving. Far from being amusing anomalies, visual illusions offer some of the deepest insights into the nature of the processing that creates visual experience. By examining such "'mistakes'" we can often determine what information processing steps occur in the construction of our experience.

Nonetheless, given that most people obviously do manage to identify visually perceived objects as members of known, functional classes under a wide range of viewing conditions, how might this result be achieved? There are many possible models of object recognition, but within a modern, computational framework, all of them require four basic components: (1) The relevant characteristics of the to-be-categorized object must be perceived and represented within the visual system in an *object representation*. (2) Each known category must be represented in memory in a *category representation* that is accessible to the visual system. (3) There must be *comparison processes* through which the object representation is matched against possible category representations. (4) There must be a *decision process* that uses the results of the comparison process to determine the category to which a given object belongs. We will consider each of these components later and then describe contrasting theories about how object recognition might be performed.

In this chapter, we consider objects as a type of thing to be recognized. However, this is actually an impoverished view of objects, because they are more than just physical things; they have meanings that go beyond their physical functions and uses. Faces, in particular, are rich in social meaning and allow us to infer the thoughts, intentions and desires of other minds. The fact that neurons can be tuned to head orientation (Desimone, Albright, Gross, & Bruce, 1984; Perrett et al., 1985), gaze direction (Perrett et al., 1985), and facial expressions (Hasselmo, Rolls, & Baylis, 1989) as confirmed by brain imaging (Hadj-Bouziane, Bell, Knusten, Ungerleider, & Tootell, 2008; Hoffman, Gothard, Schmid, & Logothetis, 2007), suggests that the face processing system is inextricably linked, not only with the recognition of faces as objects, but also with extracting social meaning and deciphering other people's beliefs, desires, intentions, and other mental states. To keep the story simple, however, we will ignore the emotional, social, and semantic aspects of object processing here. Moreover, we will not consider the roles of attention in object processing nor the manner in which object learning influences object representations (but see Op de Beeck & Baker, 2009, for a review).

BRAIN AREAS INVOLVED IN OBJECT PROCESSING

This section present a short overview of the neuronal processing that begins with registration of image information in the retina and culminates with object recognition in the temporal lobes. The reader who would like a more detailed description of these processes should consult a standard neuroscience textbook (e.g., Bear, Connors, &

Paradiso, 2007; Gazzaniga, 1995; Kandel, Schwartz, & Jessell, 2000) or a comprehensive review such as that of Ungerleider and Bell (2011) or Orban (2011).

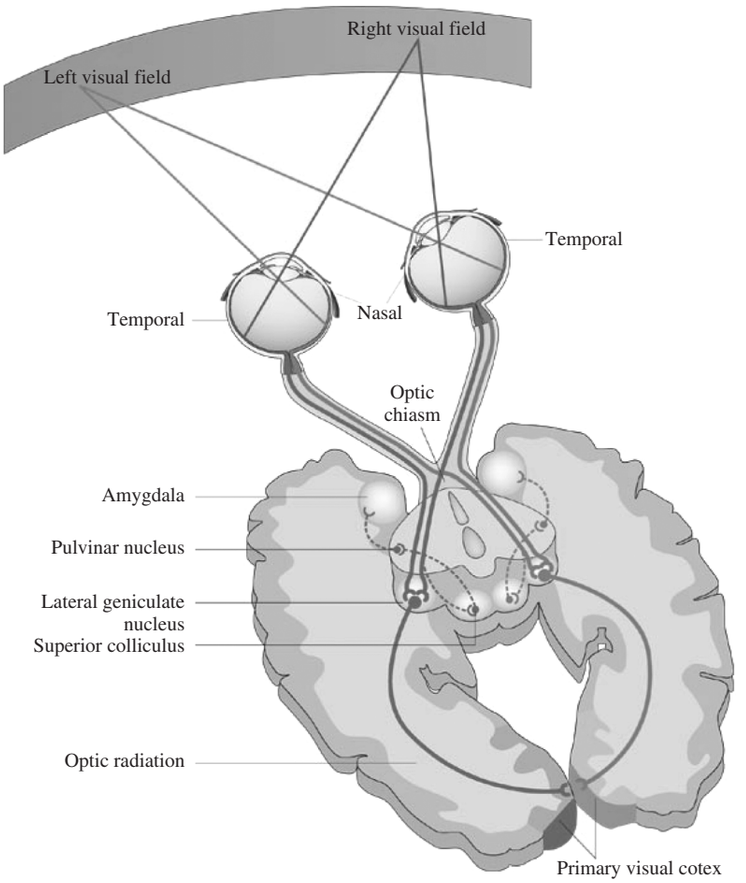
The Retina and Retinotopic Areas

The processing of information about color, space, and motion begins in the retina, which is technically part of the brain. After light has been transduced into electrochemical signals by cones and rods and subjected to additional processing by various classes of other retinal cells (amacrine, bipolar, and horizontal cells), light information reaches neurons in the retina called ganglion cells, which then convert this information into discrete neuronal firings or action potentials that are transmitted more centrally for further processing (Bear et al., 2007). Although there are many classes of ganglion cells, the most fundamental division is between magnocellular (large-celled) and parvocellular (small-celled) cells. Magnocellular cells are highly sensitive to luminance contrast and are particularly driven by fast changes in luminance but are insensitive to color information. They tend to have large receptive fields, making them sensitive to large regions of the visual field, and relatively insensitive to fine detail. Parvocellular ganglion cells have the opposite properties: They have relatively small receptive fields and so are sensitive to fine spatial detail. They are not very sensitive to sudden changes, but are sensitive to color. This retinal division of labor culminates in two very different streams of visual processing, one for determining where objects are and where they are going, and the other for determining what objects are and what they mean.

Certain ganglion cells have center-surround, opponent receptive fields that confer upon these cells a sensitivity to edges (abrupt discontinuities in luminance or color), while rendering them relatively insensitive to uniform regions. In effect, these center-surround cells filter out uniformity to some degree because regions of uniformity convey less useful information than boundaries. The information that the retina transmits to the cortical and subcortical structures for further processing is, thus, an informationally compressed' version of the image that emphasizes border information at multiple spatial scales. Efficient data compression requires preserving those aspects of an image that are most informative, whereas less informative aspects are discarded or left implicit. The fact that retinal processing emphasizes contour information suggests that contours play a crucial role in generating later representations of two-dimensional image shape and three-dimensional object shape. Data compression is necessary to efficiently encode and transmit the information available in 6–7 million cones and 120–130 million rods along an optic nerve comprised of only about 1.1 million ganglion cell axons for each eye.

These retinal signals reach the cortex by several parallel pathways (see Figure 7.1) originating within the thalamus—including the dorsolateral division of the lateral geniculate nucleus (LGNd), the pulvinar nuclear complex, the intralaminar thalamic nuclei and the thalamic reticular nucleus—as well as the basal ganglia. The LGNd segregates the inputs of magnocellular and parvocellular processing streams coming out of the retina into different laminae. Until the mid 1960s, it was believed that projections from the LGNd provided the only pathway through which visual information could gain access to the cortex and that the LGNd functioned primarily as a relay station for information passing from the retina en route to the cortex. Because the anatomical evidence indicated that in the primate, the cortical projections of the LGNd were confined to the striate cortex (also known as primary visual cortex, area 17, and V1), this cortex, with its fine-grained retinotopic representation of the visual world, was viewed as the first crucial stage in the cortical processing of visual information. Cortex is said to be retinotopically organized when neighbouring locations within a visual quadrant are mapped onto neighbouring portions of cortex (Horton & Hoyt, 1991; van Essen, 2004; van Essen, Drury, Joshi, & Miller, 1998). Each visual neuron has a receptive field that responds to stimuli falling within a well-defined region of this retinotopic space. Figure 7.2 depicts a schematic illustration of human visual area 1 (V1) and its corresponding retinotopic map of visual space. Neighboring neurons in the cortical sheet exhibit receptive fields that are overlapping in visual space.

The extrastriate cortex is now known to receive a subcortical afferent input independent of the geniculostriate system. This input exists in all mammalian species examined thus far and includes the retinal projection to the superior colliculus, an ascending projection from the superficial laminae of the colliculus to the pulvinar complex of the thalamus, and from there to the extrastriate cortex. In some mammals, including the cat, extrastriate cortex also receives a sizable afferent input from the geniculate itself. The parallel geniculocortical and tectopulvinar (also known as colliculopulvinar) cortical projections are not strictly independent. For example, the superior colliculus projects to the LGNd as well as to the pulvinar, and there exist extensive feed-forward and feedback corticocortical pathways between striate and extrastriate cortical areas (Felleman & Van Essen, 1991). In addition, all of these cortical areas contribute corticofugal projections to the LGNd, the pulvinar and the superior



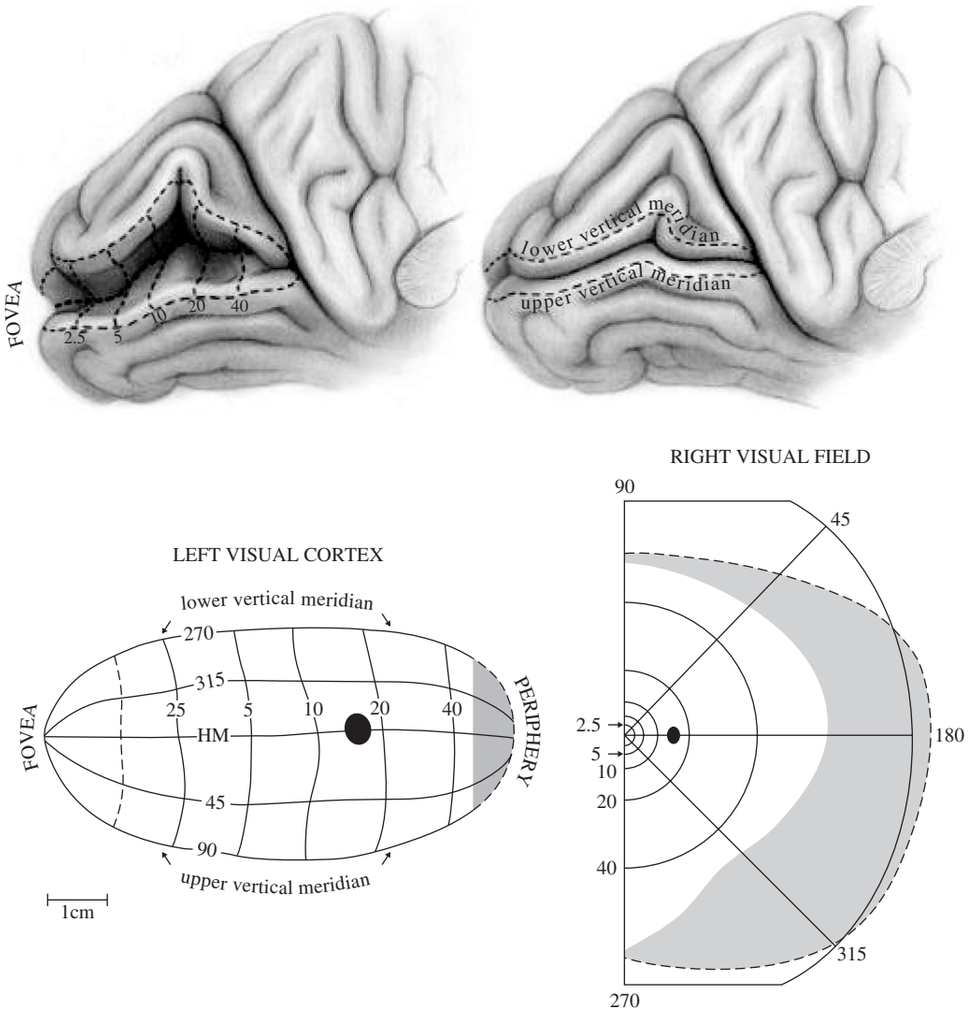
The geniculostriate pathway (solid lines) projects information from the retina to the lateral geniculate nucleus of the thalamus and then on to primary visual cortex. The tectopulvinar pathway (dashed line) is thought to project information from the retina to the superior colliculus to the pulvinar and other posterior thalamic nuclei, and on to the amygdala and cortex (Hannula, Simons, & Cohen, 2005).

Figure 7.1

colliculus. Although the tectopulvinar projection appears to play a role in the rapid object processing exhibited by the amygdala (Hannula et al., 2005), it does not appear to play an important role in cortical object recognition. In particular, lesioning the pulvinar and superior colliculus have no measurable effect on visual object discrimination (Chow, 1951; Ungerleider & Pribram, 1977). We will thus only consider the geniculostriate processing stream.

Cortical Processing

The study of visual perception has focused largely on the visual cortex and the multiple areas in which visual information is processed. The so-called "early" visual cortical areas, which receive a direct projection from LGNd, contain neurons selectively sensitive to changes in certain properties of the stimuli. For example, their levels of activity depend on features of the retinal image such as contour orientation, contour scale (or spatial frequency), binocular disparity, and the direction and velocity of movement (Hubel & Wiesel, 1962). Thus, area 17 in tree-shrews, lemurs, and primates and areas 17–18 in cats are considered "early" or "primary" visual areas, which further filter aspects of the compressed messages transmitted by the retina and extract a set of functional features or primitives. Early cortical processing thus appears to consist of a neural description of various image primitives and their locations within the scene.



Schematic illustration of the retinotopic organization of the primary visual cortex in the human brain. The upper panels depict the medial view of the left hemisphere with approximate isoeccentric radii marked on the surface of the calcarine sulcus. Note that the central 2.5 degrees of visual angle around the foveal representation of the right visual field (lower panel) activate a much larger region of visual cortex than an equivalently sized portion of the peripheral visual field. This anisotropic mapping in known as cortical magnification and reflects the relative importance given to the central visual field in cortical processing. The black spot in the lower panel represents the location and extent of the blind spot (lower right panel) and its monocular representation on the cortical surface (adapted from Horton and Hoyt 1991, with permission).

Q2 **Figure 7.2**

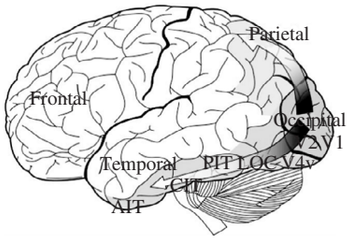
This description is still a long way from explicit identification of the three-dimensional structure of the visible world. Much additional computation is required. Trying to understand how such a complex system operates is a formidable task. As a result, simplifying concepts have emerged as guides. A keystone in thinking about the neural mechanisms of visual perception is the concept of hierarchical processing of the details of the visual image. A widely held view is that this processing occurs in stages, the first of which performs an analysis or filtering of the retinal image by extracting different elementary features (primitives), or classes of image "energy." It is generally believed that different primitive features may be processed by relatively independent modules that specialize in extracting and interpreting particular classes of visual information (e.g., Bear et al., 2007).

It is commonly held that later or "higher" stages of visual processing combine aggregates of primitive features

into progressively more complex representations. Two generally dichotomous characterizations of these elements should be mentioned. One is that contours are primitives that are used to define surfaces and object boundaries. Interpretations of receptive fields in terms of trigger "features," such as oriented bars or edges, provide a potential neural implementation of this conceptualization (Hubel & Wiesel, 1962). An alternative view is that receptive fields operate as localized, spatial frequency filters (De Valois & De Valois, 1990). The basic idea here is that cells with different receptive field dimensions are tuned to different spatial frequencies, such that cells tuned to spatial variations in light intensity that span small visual angles are said to be tuned to high spatial frequency information, whereas cells tuned to spatial variations that span large visual angles are said to be tuned to low spatial frequency information. Ganglion cells with similar tuning characteristics are distributed throughout the retina, and can be thought of as bandpass channels that accomplish a decomposition of the image using something analogous to piecewise (or local) Fourier analysis (e.g., De Valois & De Valois, 1990) or, in more recent conceptualizations, a wavelet decomposition (Wilson & Wilkinson, 1997). Convolution of the image with a variety of bandpass channels operating in parallel thus produces multiple bandpass filtered representations of the image. Low-pass filtering can convey information about overall image structure, and may contribute to certain gestalt-like operations such as grouping, closure, and good-continuation. High-pass filtered information emphasizes details within the image, particularly contours. Grouping procedures that compare, for example, contour orientations across the image, might take the high-passed output as input. Subsequent processes then operate on these multiple representations. It is not known which of these two general formulations (edge detection or spatial frequency analysis) is more accurate. Both types of detectors appear to exist and may constitute the extremes of a spectrum of processing types. It is important to recognize, however, that neither point of view suggests specific solutions to the most difficult conceptual problems raised by human pattern and form perception.

Higher Level Visual Processing

A basic tenet of modern neuroscience is that processing runs in multiple parallel streams (Bear et al., 2007; Ungerleider & Bell, 2011). The most basic of these is the distinction between the dorsal and ventral streams (Ettlinger, 1990; Goodale & Milner, 1992; Mishkin & Ungerleider, 1982; see Figure 7.3). The dorsal or "'where'" stream



A depiction of the left hemisphere's dorsal or "where" processing stream (upper arrow) running from visual input areas of the occipital lobe (depicted in blue) to the posterior parietal lobe (depicted in green), and the ventral or "what" processing stream (lower arrow) running from the occipital lobe along the ventral temporal lobe (depicted in lilac). V1=visual area 1, primary visual cortex, striate cortex. V2 and V4v are extrastriate retinotopic areas. LOC = lateral occipital complex. IT = inferotemporal cortex. AIT, CIT and PIT represent anterior, central and posterior regions of IT, respectively.

Figure 7.3

Q3
Q4

extends from the occipital lobe (V1) to the parietal lobe and is involved in the processing of where things are and where they are going in space for the purpose of computing bodily actions such as grasping. In contrast, the parallel ventral or 'what' stream extends from the occipital lobe (also from V1) along the lower portion of the temporal lobes, and is involved fundamentally in form and object representations, and is tightly linked with processing in the medial temporal lobe involved in memory storage and retrieval. The ventral stream appears to underlie our visual experience of the world. The dorsal stream takes primary input from the magnocellular processing stream, and the ventral system takes primary input from the parvocellular processing stream, although both processing streams receive both types of input. Because the dorsal stream is primarily involved in spatial and motion processing for action, it will not be our focus. This is unfortunate because, in fact, dorsal and ventral processing cannot be segregated (Farivar, 2009). For example, attention relies fundamentally on a spatial salience map realized in neuronal processing in posterior parietal cortex, and attention fundamentally influences object processing in the ventral stream. Because of space limitations, however, we will focus on shape and object processing in the ventral stream in isolation of other inputs.

As ventral stream information passes from V1 to V2 and V4v, each of which has a complete retinotopic map of the contralateral visual hemifield, receptive fields increase in size. Beyond V4v retinotopy begins to break down, and activity begins to show evidence of viewpoint and size invariance, as is the case in the occipitotemporal junction

area just anterior to V4v called the lateral occipital complex (Grill-Spector et al., 1999). This area is thought to be where localistic processing of visual features transitions into more global or configural processing of shape (e.g., Grill-Spector et al., 1998; Grill-Spector, Kourtzi, & Kanwisher, 2001; Lerner, Hendler, Ben-Bashat, Harel, & Malach, 2001). As information proceeds further along the ventral stream to posterior, central, and anterior inferotemporal cortex, retinotopic location is no longer represented, receptive fields increase even more in size, and tuning properties become ever more complex. In more anterior portions of the ventral temporal lobe, neurons can have receptive fields that span both visual hemifields (Gross, Rocha-Miranda, & Bender, 1972) and that are tuned to complex objects (e.g., Bruce, Desimone, & Gross, 1981; Desimone et al., 1984; Kobatake & Tanaka, 1994; Perrett, Rolls, & Caan,1982) in a way that is invariant to changes of position or size (Desimone et al., 1984) within the receptive field.

There does not appear to be a monolithic, multipurpose general recognition system. Instead, multiple representations of an object appear to be formed, each specialized for the purpose and coordinate transformation required for some perceptual or behavioral task. For example, face processing appears to take place in part in a region on the underside of the temporal lobes called the fusiform gyrus (Kanwisher, McDermott & Chun, 1997; Puce, Allison, Gore, & McCarthy, 1995, Puce, Allison, Asgari, Gore, & McCarthy 1996; Sergent, Ohta, & McDonald, 1992; for reviews see Tsao & Livingstone, 2008, and Rolls, 2007) as well as in the superior temporal sulcus (Haxby et al., 1999), anterior occipital lobe (Ishai, Ungerleider, Martin, Schouten, & Haxby, 1999), and anterior temporal lobe (Kriegeskorte, Formisano, Sorger, & Goebel, 2007; Rajimehr, Young, & Tootell, 2009; Tsao, Moeller, & Freiwald, 2008) of humans, and in corresponding cortical areas in monkeys (Bell, Hadj-Bouziane, Frihauf, Tootell, & Ungerleider, 2009; Pinsk, DeSimone, Moore, Gross, & Kastner, 2005; Pinsk et al., 2009; Tsao, Freiwald, Knutsen, Mandeville, & Tootell, 2003; Tsao, Freiwald, Tootell, & Livingstone, 2006). It has been shown that face-tuned inferotemporal neurons respond biphasically (Sugase, Yamane, Ueno, & Kawano, 1999): Their initial response, approximately 100–150 ms after face image onset, carries information about global face information useful for discriminating faces as members of a face category versus other categories, but immediately subsequent responses carry information about fine spatial detail that would permit recognition of particular facial expressions.

Other functional "modules" specialized for particular categories of input have been found in recent years. A "parahippocampal place area" (Epstein & Kanwisher, 1998; Epstein, DeYoe, Press, & Kanwisher, 2001) and a region of the retrosplenial cortex (Epstein, Parker, & Feiler, 2008; Park & Chun, 2009; Walther, Caddigan, Fei-Fei, & Beck, 2009) seem to be specialized for processing spatial layout of a scene. And an area near the occipotemporal junction and the motion-processing area "MT" was found to respond strongly to human body-parts (Downing, Jiang, Shuman, & Kanwisher, 2001). This extrastriate body-part area may, in fact, be specialized for the processing of limb and goal-directed movements (Astafiev, Stanley, Shulman, & Corbetta, 2004).

At some point in the ventral information processing stream, object representations must be compared with representations stored in memory. A decision must then be made concerning the best match between seen and stored representations. This recognition and decision process may happen at several stages in parallel, but certainly involves anterior inferotemporal neurons. Eifuku, De Souza, Tamura, Nishijo, & Ono (2004) trained monkeys to match an image of a face to one of many possibilities taken from different viewpoints. They found that the neural latency of anterior inferotemporal face-tuned neurons was correlated with the monkey's behavioral latency, suggesting that these neurons play a role in the decision process. Indeed, microstimulation of such face-tuned cells can lead a monkey to report that it has seen a face given ambiguous input more often than is the case in the absence of microstimulation (Afraz, Kiani, & Esteky, 2006).

Visual Agnosias

A fascinating neuropsychological phenomenon of object recognition is *visual agnosia* (see Farah, 2004, and Humphreys & Riddoch, 2006, for reviews), a perceptual deficit due to brain damage in which patients are unable to correctly categorize common objects with which they were previously familiar. ("*Agnosia*" is a term derived from Greek, which means "not knowing.") Damage within the temporal lobes and/or occipital lobes can produce profound deficits in object recognition of many, but not necessarily all, categories. In some cases, the patients cannot name an object when presented visually, but can name the object when they touch it. In other cases they cannot copy drawings of objects, but can name the same object they cannot draw. This suggests that there are specific areas of cortex that are specialized for high-level shape processing, object categorization, and matching to memory.

There are many different forms of visual agnosia, and the relations among them are not well understood. Some appear to be primarily due to damage to the later stages of precategorization sensory processing (termed "*apperceptive agnosia*" by Lissauer, 1890/1988). Such patients appear unable to recognize objects because they do not see them normally. In particular, apperceptive agnosics probably fail to recognize objects because they fail to process shape in a normal way. They cannot distinguish a triangle from a rectangle, and cannot even copy a shape. They typically have damage to shape-processing areas such as the lateral occipital complex located near the occipitotemporal junction. It appears that some local shape processing remains intact, whereas global analyses of shape and configural analyses of relationships among local shapes have been lost. Other patients have fully intact perceptual abilities, yet still cannot identify the objects they see, a condition Lissauer called "*associative agnosia*." Teuber (1968) described their condition as involving "a normal percept stripped of its meaning" due to an inability to categorize it correctly. Associative agnosics appear to have relatively intact shape processing. For example, they can copy images, but they lack the ability to match such representations with memories and, therefore, cannot recognize objects. Some associative agnostics can describe scenes and even describe certain aspects of an object, such as its use, but will then mislabel the object. Associative agnosia makes apparent that the processes involves in seeing can be dissociated from those involved in matching to memory, recognizing, or determining meaning on the basis of what is seen.

The case of a patient known as GL is a good example of associative agnosia (Ellis & Young, 1988). GL suffered a blow to his head when he was 80 years of age, after which he complained that he could not see as well as before the accident. The problem was not that he was blind or even impaired in basic visual function, for he could see the physical properties of objects quite well. Indeed, he could even copy pictures of objects that he could not identify. Even so, he mistook pictures for boxes, his jacket for a pair of trousers, and generally could not categorize even the simplest everyday objects correctly.

Patients with visual agnosia suffer from a variety of different symptoms. Some have deficits specific to particular classes of objects or properties. One classic example is *prosopagnosia*: the inability to recognize faces (Landis, Regard, Bliestle, & Kleihues, 1988; Rossion et al., 2003). Prosopagnosia typically arises from bilateral damage to the fusiform face area or adjacent occipitotemporal areas. Prosopagnosics can describe in detail the facial features of someone they are looking at, yet be completely unable to recognize the person, even if it is their spouse, their child, or their own face in a mirror. This is probably because they have a preserved ability to process the local features of a face, but have lost the ability to process configural relationships among local features. Indeed, neuronal responses in face selective neurons in areas that correspond to the fusiform face areas in monkeys are highly responsive to the global face pattern as opposed to local parts of a face (Desimone et al., 1984; Hasselmo et al., 1989; Perrett et al., 1982). Such patients will typically react to a relative as a complete stranger—until the person speaks, at which time the patient can recognize his or her voice. The fact that pure prosopagnosics cannot recognize faces but can recognize other classes of objects implies that face recognition is a functional module of object recognition (Damasio, Damasio, & Van Hoesen, 1982).

Other agnosic patients have been studied who have problems with object categories such as living things. Patient JBR, for example, was able to identify 90% of the pictures depicting inanimate objects, but only 6% of those depicting plants and animals (Warrington & Shallice, 1984). Even more selective deficits have been reported, including those confined to body parts, objects found indoors, and fruits and vegetables, although some of these deficits may be linguistic in nature rather than perceptual (Farah, 2004).

One problem for many visual agnosics that has been studied experimentally is their particular inability to categorize objects presented in "unusual" perspective views. Warrington and Taylor (1973, 1978) found that many agnosics who are able to categorize pictures of common objects taken from a usual or "canonical" perspective are unable to do so for unusual or "noncanonical" views, as shown in Figure 7.4. This phenomenon in agnosics bears a striking resemblance to perspective effects found in normal individuals (Palmer, Rosch, & Chase, 1981), except

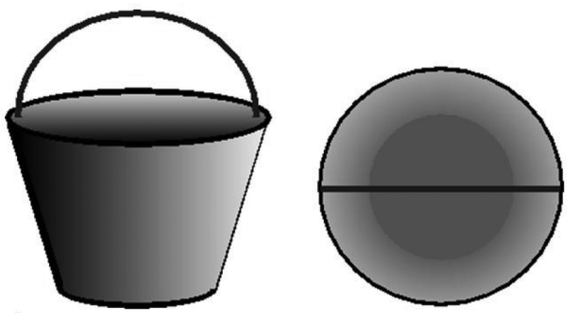


Figure 7.4 A canonical view of a bucket is shown on the left, and a noncanonical view from above is shown on the right

that instead of simply taking longer to arrive at the correct answer, these patients are unable to perform the task at all, even in unrestricted viewing conditions.

There are many other visual disorders due to brain damage that are related to visual agnosia. They exhibit a wide variety of complex symptoms, are caused by a broad range of underlying brain pathologies, and are generally not well understood. Still, the case histories of such patients and the phenomenological descriptions of their symptoms make for fascinating reading (e.g., Sacks, 1985). The interested reader is referred to Farah (2000, 2004) for a neuropsychological overview of these and related disorders.

PERCEPTUAL ORGANIZATION

The difficulty in recovering a correct representation of the world from the image can most easily be appreciated by considering the output of the retinal mosaic simply as a numerical array, where each number represents the neural response of a single receptor. This array is ambiguous in many ways. A fundamental ambiguity arises because there are no explicit instructions in the image regarding what numbers in this array "go together" in the sense of corresponding to the same objects, parts, or groups of objects in the environment. This problem is hard enough for a static image, but the problem of determining what goes with what—often called segmentation and/or grouping problem—also arises across images, as objects move around us, as we move around objects, and as we make eye movements. In the context of motion, this matching problem across images is generally called "the correspondence problem." Whether matching regions within an image, regions across images, or both, neuronal processing mechanisms must match some things together and not others, and do it in ways that correspond adaptively to the structure of environmental objects and the ways in which we interact with them.

A different kind of ambiguity arises even in interpreting a single value measured at a point on the retina. For example, retinal stimulation at a given point in the images might register redness for many different reasons. Most likely it occurs because a red surface is illuminated by white light, but it could also arise if a white surface is illuminated with red light or if a white surface is seen through red glass or red smoke. The single color registered at this one location then has to be apportioned to the properties of the surface, illuminating light source, and any intervening layers of transparent material in a way that is not given explicitly in the image. Apportionment into sources, like grouping into objects within and across images, occurs accurately because prior assumptions concerning the mapping between image structure and world structure are built into the processes that accomplish this mapping, and those assumptions are generally correct.

Perception is biased to recover the intrinsic properties of objects rather than the properties of their projected images, which are subject to highly variable conditions of observation. Objects have, for example, an intrinsic shape, size, reflectance, and material. To recover these properties, it is necessary to discount various extraneous (or extrinsic) observational factors that partly determine the image, but are often of limited importance to the perceiving organism. For example, the retinal shape, size, and orientation of the image of an object are influenced by the slant, distance, and tilt of the object relative to the viewer's head. Our conscious experience of that object's properties as being constant when we change our viewing position or head tilt must, therefore, follow the application of certain unconscious "constancy" operations that generate a representation of an object as having an intrinsic size regardless of its size in the image (size constancy), an intrinsic shape (shape constancy) regardless of its shape in the image, and an intrinsic surface reflectance (color or lightness constancy), regardless of the local wavelengths measured at the level of the image. In the absence of such constancy operations, a person would seem to shrink as they walked away from us, or change their skin pigments as they walked under a shadow.

It is widely believed that early visual cortical areas are involved in grouping local information across the image into aggregate wholes. Gestalt psychologists described the principles (or criteria) used by the visual system to group parts into wholes. Although Gestalt psychologists did not quite talk in these terms, grouping principles can be understood as "heuristics" that arise from the combination of evolution and perceptual learning to produce mechanisms that are sensitive to the statistical regularities of real-world images. For instance, grouping procedures capture the fact that image regions that covary in certain ways tend to arise from common surfaces, objects, regions, and collections of objects in the world. Perceptual grouping is essential to the process of image segmentation—the process of determining which contours and textures belong to the same object. Because grouping involves decisions about what belongs with what in an image, it is relevant to making inferences about the state of the world. As such, grouping procedures do not merely extract information from the image, they create or construct new information—they make information that is only implicit at the level of the

retinal image explicit in relevant ways—that brings the organism closer to its goal of perceiving the environment in ways that are appropriate for effective action.

Perceptual Grouping

Thus far, we have focused on information that is available in the image that can be extracted or detected by filters or detectors tuned to useful image primitives, such as local, oriented lines and edges or local, oriented spatial-frequency content. At some point, these primitives must be used to construct representations of objects and three-dimensional surface layout. Retinal center-surround processing deemphasizes uniform regions, but it is still confined to extraction of local image-based (rather than object-based) information. Further shape processing requires the construction of information about three dimensions rather than two dimensions. Any inference of a three-dimensional surface from two-dimensional image cues is tantamount to the construction of information not present in the image. This type of construction was explored extensively by Helmholtz and later by Gestalt psychologists, who emphasized that visual perception must be subserved by rapid grouping procedures that link information across the image in a global fashion on the basis of heuristics such as contour continuity. Although there is strong evidence that such grouping procedures are carried out in the course of visual form processing, little progress has been made in discerning how such global grouping procedures are realized by local neuronal computations. Rather than give a necessarily speculative and sketchy description of how three-dimensional form is computed at the level of neuronal "hardware," it is more productive to consider how the construction of visual form might take place at the more abstract level of the information-processing algorithms used to construct information about three-dimensional form.

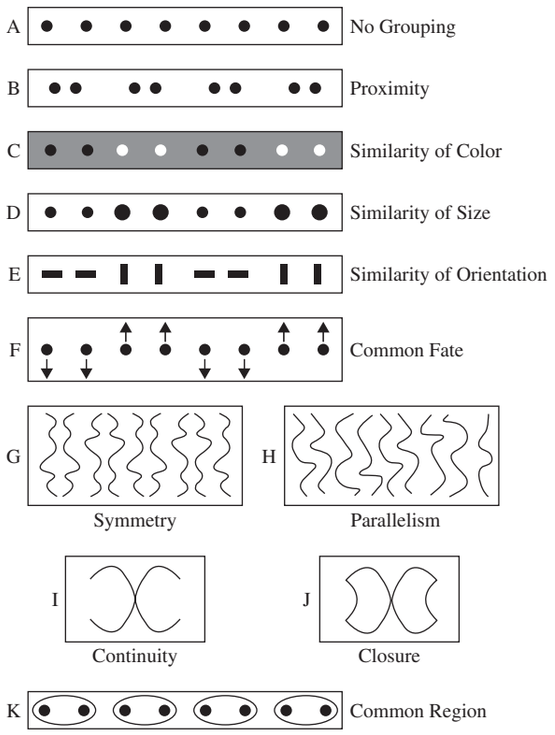
To solve this problem, multiple systems have evolved to recover three-dimensional shape from the various cues to form that are present in the image. Examples are the perception of shape from shading, the perception of shape from motion (often called "structure-from-motion") and the perception of shape from retinal disparity cues (that serve as the basis for random dot stereograms and "magic eye" books). Inferring shape using multiple strategies and cues has important advantages. Multiple circuits can compute shape solutions in parallel, reducing overall computational time. If one subsystem should reach a solution before others do, that shape solution can constrain the computations being carried out by the other subsystems. Parallel and concurrent processing also reduces the likelihood of attaining poor solutions because all shape subsystems must come to mutually consistent solutions, permitting a form of error checking and redundancy. Thus, under normal viewing conditions, the problem of recovering three-dimensional shape from the two-dimensional image can be solved using multiple mutually constraining shape-from-X cues (where X stands for the relevant channels of information, such as shading, motion, and texture). In some cases, however, just a single cue can be used to generate a distinct perception of three-dimensional form. For static images, contours probably offer the strongest constraints on three-dimensional form because other cues, such as shading or texture, appear to be dominated by contour cues (Tse, 2002).

Principles of Grouping

The visual phenomenon most closely associated historically with the concept of perceptual organization is *grouping:* the fact that observers perceive some elements of the visual field as "going together" more strongly than others. Indeed, perceptual grouping and perceptual organization are sometimes presented as though they were synonymous. They are not. Grouping is one particular kind of organizational phenomenon, albeit a very important one.

The Gestalt psychologist Max Wertheimer first posed the problem of perceptual grouping in his groundbreaking 1923 paper. He then attempted a solution at what would now be called the "computational level" by asking what stimulus factors influence perceived grouping of discrete elements. He first demonstrated that equally spaced dots do not group together into larger perceptual units, except as a uniform line (Figure 7.5a). He then noted that when he altered the spacing between adjacent dots so that some dots were closer than others, the closer ones grouped together strongly into pairs (Figure 7.5b). This factor of relative distance, which Wertheimer called *proximity*, was the first of his famous laws or (more accurately) *principles of grouping*.

Wertheimer went on to illustrate other grouping principles, several of which are portrayed in Figure 7.5. Parts c, d, and e demonstrate different versions of the general principle of *similarity*: All else being equal, the most similar elements (in color, size, and orientation for these examples) tend to be grouped together. Another powerful grouping factor is *common fate*: All else being equal, elements that move in the same way tend to be grouped together. Notice that both common fate and proximity can actually be considered special cases of similarity grouping in which the relevant properties are similarity of velocity



Classical principles of grouping: no grouping (a) versus grouping by proximity (b), similarity of color (c), similarity of size (d), similarity of orientation (e), common fate (f), symmetry (g), parallelism (h), continuity (i), closure (j), and common region (k).

Figure 7.5

and position, respectively. Further factors that influence perceptual grouping of more complex elements, such as lines and curves, include *symmetry* (Figure 7.5g), *parallelism* (Figure 7.5h), and *continuity* or *good continuation* (Figure 7.5i). Continuity is important in Figure 7.5i because observers perceive it as containing two continuous intersecting lines rather than as two angles whose vertices meet at a point. Figure 7.5j illustrates the further factor of *closure*: all else being equal, elements that form a closed figure tend to be grouped together. Note that this display shows that closure can overcome continuity because the very same elements that were organized as two intersecting lines in Figure 7.5i are organized as two angles meeting at a point in part Figure 7.5j.

Several other spatiotemporal grouping principles have been discovered beyond common fate. One is "extended common fate," an extension of common fate to grouping by common luminance changes (Sekuler & Bennett, 2001). When (unmoving) elements of a visual scene become brighter or darker together, observers have a powerful tendency to group those elements perceptually. It is as though the principle of common fate operates not only for the common motion of elements through three-dimensional physical space (i.e., common fate), but through luminance space as well (i.e., extended common fate). Another is grouping by synchrony: the tendency for elements that change simultaneously in some visible feature to be grouped together (Alais, Blake, & Lee, 1998; Lee & Blake, 1999). The changes do not have to be in the same direction, as in generalized common fate, however. A random field of black and white dots whose luminances flip in polarity over time against a gray background, for example, will segregate into two distinct regions if the dots in one area begin to change synchronously rather than randomly. Although there is some controversy about whether synchrony grouping might be an artifact (Farid & Adelson, 2001), it is particularly interesting because it has been suggested that synchrony of neural firing might be the implementation of all forms of perceptual grouping (e.g., Gray & Singer, 1989; Milner, 1974; Singer & Gray, 1995; von der Malsburg, 1987).

One might think that grouping principles are mere textbook curiosities only distantly related to anything that occurs in normal perception. On the contrary, they pervade virtually all perceptual experiences because they determine the objects and parts we perceive in the environment. Dramatic examples of perceptual organization going wrong can be observed in natural camouflage. Even perfect static camouflage is undone by the principle of common fate. The common motion of an animal whose contours and texture blends with the stationary background causes them to be strongly grouped together, providing an observer with enough information to perceive it as a distinct object against its unmoving background. Successful camouflage also reveals the ecological rationale for the principles of grouping. Camouflage results when the same grouping processes that would normally make an organism stand out from its environment as a separate object cause it to be grouped with its surroundings instead. This results primarily from similarity grouping of various forms, when the color, texture, size, and shape of the organism are similar enough to those of the objects in its environment to be misgrouped.

Integrating Multiple Principles of Grouping

The demonstrations of continuity and closure in Figure 7.5i and 7.5j illustrate that a grouping principle, as formulated by Wertheimer (1923), predicts the outcome of grouping with certainty only when no other grouping factor opposes its influence. We saw, for example, that continuity governs grouping when the elements do not form a closed figure,

but continuity can be overcome by closure when they are closed (Figure 7.5i versus 7.5j). The difficulty posed by independent grouping rules is that they provide no scheme for *integrating* multiple factors into an overall outcome—that is, for predicting the strength of their combined effects. For example, if proximity influences grouping toward one outcome and color similarity toward another, which grouping will be perceived depends heavily on the particular degrees of proximity and color similarity (e.g., Hochberg & Silverstein, 1956).

To determine how different grouping principles influence perceived grouping, Kubovy and Wagemans (1995) measured the relative strength of different groupings in dot lattices (Figure 7.6a) by determining the probability with which subjects reported seeing them organized into lines in each of the four orientations indicated in Figure 7.6b. After seeing a given lattice for 300 ms, subjects indicated which of the four organizations they perceived so that, over many trials, the probability of perceiving each grouping could be estimated. Consistent with the Gestalt principle of proximity, their results showed that the most likely organization is the one in which the dots are closest together, with other organizations being less likely as the spacing between the dots increased. Further experiments using lattices in which the dots differed in color similarity as well as proximity showed that the rule by which multiple grouping factors combine is multiplicative. This finding begins to specify general laws by which multiple factors can be integrated into a combined result.

Is Grouping an Early or Late Process?

If perceptual organization is to be understood as the result of computations, the question of where grouping occurs in the stream of visual processing is important. Is it an early process that works at the level of two-dimensional image structure or does it work later (Palmer, Brooks, & Nelson, 2003), after depth information has been extracted and perceptual constancy has been achieved? Wertheimer (1923/1950) discussed grouping as though it occurred at a low level, presumably corresponding to what is now called image-based processing (see Palmer, 1999). The view generally held since Wertheimer's seminal paper has been that organization must occur early to provide virtually all higher-level perceptual processes with discrete units as input (e.g., Marr, 1982; Neisser, 1967).

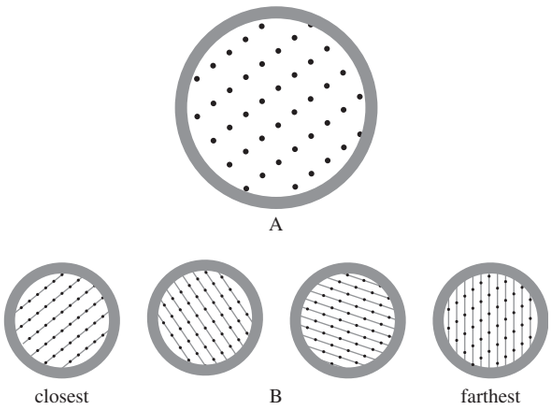


Figure 7.6 Dot lattice stimuli (a) and possible groupings (b) studied by Kubovy and Wagemans (1995). [From Palmer, 1999]

Rock and Brosgole (1964) reported evidence against the "early-only" view of grouping, however. They examined whether the relevant distances for grouping by proximity are defined in the two-dimensional image plane or in perceived three-dimensional space. They showed observers a two-dimensional rectangular array of luminous beads in a dark room either in the frontal plane (perpendicular to the line of sight) or slanted in depth, so that the horizontal dimension was foreshortened to a degree that depended on the angle of slant. The beads were actually closer together vertically, so that, when they were viewed in the frontal plane, observers saw them grouped into vertical columns rather than horizontal rows.

The crucial question was what would happen when the same lattice of beads was presented to the observer slanted in depth so that the beads were closer together horizontally when measured in the retinal image, even though they are still closer together vertically when measured in the three-dimensional environment. When observers viewed this slanted display with just one eye, so that binocular depth information was not available, they reported that the beads were organized into rows. However, when they perceived the slant of the lattice in depth by viewing the same display binocularly, their reports reversed; they now reported seeing the slanted array of beads organized into vertical columns. This finding thus supports the hypothesis that final grouping occurs after stereoscopic depth perception.

Rock, Nijhawan, Palmer, and Tudor (1992) addressed a similar issue in lightness perception. Their results showed that grouping followed the predictions of a late (post-constancy) grouping hypothesis: Similarity grouping in the presence of shadows and translucent overlays was governed by the perceived lightnesses of the elements rather than by their retinal luminances. Further findings using analogous methods have shown that perceptual grouping is also strongly affected by amodal completion (Palmer, Neff, & Beck, 1996) and by illusory contours

(Palmer and Nelson, 2000), both of which are believed to depend on depth perception in situations of occlusion (see Rock, 1983). Such results show that grouping cannot be attributed entirely to early, preconstancy visual processing, but they are also compatible with the possibility that grouping is a temporally extended process that includes components at both early and later levels of processing. A provisional grouping might be determined at an early, preconstancy stage of image processing, but it might be overridden if later, object-based information (from depth, lighting conditions, occlusion, and so forth) required it.

Figure-Ground Organization

For every bounding contour in a segmented image there is a region on both sides. Because most visible surfaces are opaque, the region on one side usually corresponds to a closer, occluding surface, and the region on the other side to a farther, occluded surface that extends behind the closer one. Boundary assignment is the process of determining to which region the contour "belongs," thus determining the shape of the closer surface, but not that of the farther surface. Because of its effects on perceived shape, boundary assignment has profound implications for object perception and recognition.

To demonstrate the remarkable difference that alternative boundary assignments can make, consider Figure 7.7. Region segmentation processes will partition the square into two uniformly connected regions, one white and the other black. However, to which side does the central boundary belong? If you perceive the edge as belonging to the white region, you will see a white object with "rounded fingers" protruding in front of a black background. If you perceive the edge as belonging to the black region, you will see a black object with "pointed claws" in front of a white background. This particular display is highly ambiguous, so that sometimes you see the white fingers and other times the black claws. This boundary-assignment aspect of perceptual organization is known in the classical perception literature as *figure-ground organization* (Rubin, 1921). The "thing-like" region is referred to as the *figure* and the "background-like" region as the *ground*.

Figure 7.7 Ambiguous edge assignment and figure-ground organization. [From Rock, 1983]

Principles of Figure-Ground Organization

Figure 7.7 is highly ambiguous in its figure-ground organization because it is about equally easy to see the black and white regions as figure, but this is not always, or even usually, the case. The visual system has distinct preferences for perceiving certain kinds of regions as figural, and these are usually sufficient to determine figure-ground organization. Studies have determined that the following factors are relevant, all of which bias the region toward being seen as figural: surroundedness, smaller size, horizontal and/or vertical orientation, higher contrast, greater symmetry, greater convexity (Kanizsa & Gerbino, 1976), parallel contours, meaningfulness (Peterson & Gibson, 1991), extremal edges (Ghose & Palmer, 2010; Palmer & Ghose, 2008), lower region (Vecera, Vogel, & Woodman, 2002), wider base (Hulleman & Humphreys, 2004), edge-region grouping (Palmer & Brooks, 2008), and voluntary attention (Driver & Baylis, 1996; Vecera, Flevaris, & Filapek, 2004). Recently, a number of dynamic factors in figure-ground organization have also been identified, such as advancing (versus retreating contour motion) (Barenholtz & Tarr, 2009) and contour motion in deforming shapes (Barenholtz, 2010; Barenholtz & Feldman, 2006). Analogous to the gestalt principles of perceptual grouping, these principles of figure-ground organization are rules in which a given factor has the stated effect, if all other factors are "equal" (i.e., eliminated or otherwise neutralized). As such, they have the same weaknesses as the principles of grouping, including the inability to predict the outcome when several conflicting factors are at work in the same display.

In terms of information-processing structure, Palmer and Rock (1994a, 1994b) proposed a process model of perceptual organization in which figure-ground organization occupies a middle position, occurring after region segmentation but before grouping and parsing (see Figure 7.8). They argued that figure-ground processing logically must occur after region-segmentation processing because segmented regions are required as input by any algorithm that discriminates figure from ground. The reason is that most of the principles of figure-ground organization—e.g., surroundedness, size, symmetry, and convexity—are properties that are only defined for two-dimensional regions, and thus require two-dimensional regions as input. More speculatively, Palmer and Rock (1994a, 1994b)

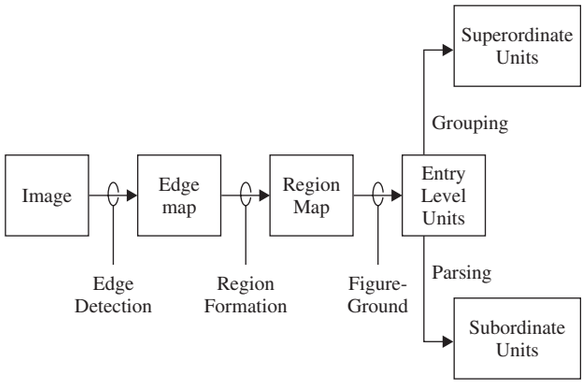


Figure 7.8 A computational theory of visual organization. See text [From Palmer and Rock, 1994a]

also claimed that figure-ground organization must logically precede grouping and parsing. The reason is that the latter processes, which apparently depend on certain shape-based properties of the regions in question—e.g., concavity/convexity, similarity of orientation, shape, size, and motion—require prior boundary assignment. Grouping and parsing thus depend on shape properties that are logically well defined for regions only after boundaries have been assigned, either to one side or perhaps initially to both sides (Peterson & Gibson, 1991).

THEORIES OF OBJECT RECOGNITION

After the image has been organized into midlevel shape representation, the perceptual objects thus defined are very often identified as instances of known, meaningful types, such as people, houses, dogs, and cars, or as meaningful particular individuals of those types, such as me, my house, my dog, and my car. This process, which is generically referred to as "object recognition" is also called "object identification" or "object categorization," with the latter term usually reserved for recognizing objects at the level of categories rather than individuals. Once a representation has been specified for the to-be-identified objects and the set of known categories, a process has to be devised for comparing the object representation with each category representation. This could be done serially across categories, but it makes much more sense for it to be performed in parallel. Parallel matching could be implemented, for example, in a neural network that works by spreading activation, where the input automatically activates all possible categorical representations to different degrees, depending on the strength of the match (e.g., Hummel & Biederman, 1992).

In all cases, the presumed goal of object (as opposed to person or other mind) recognition is the perception of *function*, enabling the observer to know, simply by looking, what objects in the environment are useful for what purposes. The general idea behind perceiving function via object recognition is to match the perceived properties of a seen object with some internal representation of the perceivable properties of the appropriate, known category of object. Once the object has been thus identified, its function can then be determined by retrieving associations between the object category and its known uses. This matching process will not make *novel* uses of the object available—additional problem solving processes are required for that purpose—but only ones that have been previously understood and stored in memory with that category.

Because the schemes for comparing representations are rather specific to the type of representation, in the following discussion we will simply assume that a parallel comparison process can be defined that has an output for each category that is effectively a bounded, continuous variable representing how well the target object's representation matches the category representation. The final process is then to make a decision about the category to which the target object belongs. Several different rules have been devised to perform this decision, including the *threshold*, *best fit*, and *best-fit-over-threshold* rules.

The threshold approach is to set a criterial value for each category that determines whether a target object counts as one of its members. The currently processed object is then assigned to whatever category, if any, exceeds its threshold matching value. This scheme can be implemented in a neural network in which each neural unit that represents a category has its own internal threshold, such that it begins to fire only after that threshold is exceeded. The major drawback of a simple threshold approach is that it may allow the same object to be categorized in many different ways (e.g., as a fox, a dog, and a wolf), because more than one category may exceed its threshold at the same time.

The best-fit approach is to identify the target object as a member of whatever category has the highest match among a set of mutually exclusive categories. This can be implemented in a "winner-take-all" neural network in which each category unit inhibits every other category unit among some mutually exclusive set. Its main problem lies in the impossibility of deciding that a novel target object is not a member of any known category. This is an issue because there is, by definition, always *some* category that has the highest similarity to the target object.

The virtues of both decision rules can be combined—with the drawbacks of neither—using a hybrid decision strategy: the best-fit-over-threshold rule. This approach is to set a threshold below which objects will be perceived as novel, but above which the category with the highest matching value is chosen. Such a decision rule can be implemented in a neural network by having internal thresholds for each category unit as well as a winner-take-all network of mutual inhibition among all category units. This combination allows for the possibility of identifying objects as novel without resulting in ambiguity when more than one category exceeds the threshold. It would not be appropriate for deciding among different hierarchically related categories (e.g., collie, dog, and animal), however, because they are not mutually exclusive.

Before pursuing the topic of object recognition in depth, it is worth mentioning that there are alternative approaches to understanding the perception of function. The primary competing view is J. Gibson's (1979) theory of *affordances*, in which opportunities for action are claimed to be perceived directly from visible structure in the dynamic optic array. Gibson claimed, for example, that people can literally see whether an object affords being grasped, or sat upon, or walked upon, or used for cutting without first identifying it as, say, a baseball, a chair, a floor, or a knife. This is possible, however, only if the relation between an object's form and its affordance (the function or use it offers an organism) is transparent enough that the relevant properties are actually *visible*. If this is not the case, then category-mediated object recognition appears to be the only route for perception of function. The two views are not necessarily mutually exclusive. For example, it is possible that category-mediated object recognition proceeds in the ventral "what" processing stream of the temporal lobes, which, also, incidentally, seems to subserve visual awareness, whereas the processing of affordances proceeds in the dorsal "where" processing stream (Norman, 2002).

Prototypes and Basic Level Categories

The first fact that must be considered about identifying or recognizing objects is that it is an act of classification or categorization (see Goldstone et al., this volume). Although most people typically think of objects as belonging to just one category—something is either a dog or a house or a tree or a book—all objects are actually members of many categories. Lassie, for example, is a collie, a dog, a mammal, an animal, a living thing, a pet, a TV star, and so on. The categories of human perception and cognition are quite complex and interesting psychological structures. Many (but not all) categories of objects are structured into hierarchies (e.g., Lassie/collie/dog/mammal /animal/living thing), with many categories nested within each category at any given level (e.g., collies, beagles, Chihuahuas, and Saint Bernards are all members of the dog category).

One of the most important modern discoveries about human categorization is the fact that categories do not seem to be defined by sets of *necessary and sufficient conditions*. The key idea of "norm-based encoding" posits that objects are represented in terms of psychological distance from one or more prototype objects or experiences that are the "best" (i.e., prototypical) instances of a category (Rosch, 1973, 1975a,b). The prototypical dog, for example, would be the "doggiest" possible dog: probably a standard "mutt" of some sort rather than a pure-bred dog, about average in size, having standard brownish sort of coloring and the usual doggy sort of shape. When Rosch asked people to rate various members of a category, such as particular breeds of dogs, in terms of how "good" or "typical" they were as examples of dogs, she found that they systematically rated beagles quite high and Saint Bernards and Chihuahuas quite low. Similarly, robins and sparrows are good examples of birds, whereas penguins and ostriches are poor examples. These *typicality* (or goodness-of-example) ratings turn out to be good predictors of how quickly subjects can respond "true" or "false" to verbal statements, such as, "A robin is a bird," versus, "A penguin is a bird" (Rosch, 1975b). Thus, the time required to identify an object as a member of a category depends on how "typical" it is as an example of that category.

Rosch made another major discovery about the structure of people's natural categories, this one concerning differences among levels within the categorical hierarchy. For example, at which level does visual identification first occur: at some low, specific level (e.g., collie), at some high, general level (e.g., animal), or at some intermediate level (e.g., dog)? The answer is that people generally recognize objects first at an intermediate level in the categorical hierarchy. Rosch called categories at this level of abstraction "basic level categories" (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Later research, however, has shown the matter to be somewhat more complex.

Jolicoeur, Gluck, and Kosslyn (1984) studied this issue by having subjects name a wide variety of pictures with the first verbal label that came to mind. They found that objects that were *typical* instances of categories, such as robins and sparrows, were indeed identified as members of

a basic level category, such as birds. *Atypical* ones, such as penguins and ostriches, tended to be classified at the next lower, subordinate level. This pattern of naming was not universal for all atypical category members, however. It occurs mainly for members of basic level categories that are relatively diverse. Consider some basic level categories from the superordinate categories of fruit (e.g., apples, bananas, and grapes) versus animals (dogs, birds, and monkeys). Most people would agree that the shape variation within the categories of apples, for instance, is more constrained than that within the categories of dogs. Indeed, most people would be hard-pressed to distinguish between two different kinds of apples, bananas, or grapes from shape alone, but consider how different dachshunds are from greyhounds, penguins are from ostriches, and goldfish are from sharks. Not surprisingly, then, the atypical exemplars from diverse basic-level categories are the ones that tend to be named according to their subordinate category. Because the categories into which objects are initially classified is sometimes different from the basic level, Jolicoeur et al. (1984) called them "entry level categories."

It is worth noting that, as in the case of basic level categories, the entry-level category of an object can vary over different observers and perhaps over different contexts as well. To an ornithologist or even to an avid bird watcher, for instance, BIRD may be the entry-level category for very few, if any, species of bird. Through a lifetime of experience at discriminating different kinds of birds, their perceptual systems may become so finely tuned to the distinctive characteristics of different kinds of birds that they first perceive robins as robins and sparrows as sparrows rather than just as birds (Tanaka & Taylor, 1991).

Many computational models conceive of object representation in terms of the responses of a population of neurons (e.g., Agam et al., 2010). Information, on this account, is carried in the pattern of neural responses across a population. One possibility is that patterns of neural activity reflect an encoding of an object along the multiple dimensions defining some object space. A prototype-based encoding could be thought of as an object representation that is encoded as a vectorial response within a neuronal ensemble. Presumably, the more similar the population response patterns are to two images, the more similar the object representations they encode will be. Inclusion in a category, thus, depends on the degree to which an object is similar to the nearest prototype in the relevant representational space, which can be thought of as the origin of the vector. Indeed, the number of dimensions of a representational space (e.g., a "face space") may be given by the number of criteria that are placed on inputs to determine whether something counts as a face. On this account, each individual criterion for some aspect of faceness defines a dimension of face space. Since criteria may look nothing like our concepts of what counts as a face, it is an empirical question what the dimensions of face space are. For example, there may exist neurons that are tuned to the ratio of the distance between the nose and lips versus the distance between the eyes. If so, that relationship would constitute a dimension of face space. In this sense, criterial encoding entails that at least some representational spaces have a quasivectorial organization, in which multidimensional vectors are specified by the degree to which various criteria are met, triggering the firing of neurons that impose those criteria on inputs (cf. Hsieh & Tse, 2009; Leopold, Bondar, & Giese, 2006; Leopold, O'Toole, Vetter, & Blanz, 2001; Loffler, Yourganov, Wilkinson, & Wilson, 2005; Rhodes, Michie, Hughes, & Byatt, 2009; Rhodes, Watson, Jeffrey, & Clifford, 2010; Rhodes et al., 2011). Norm-based and vector-based models offer a good, but rough, first approximation of certain types of mental representation.

Norm-based encoding expresses several fundamental traits of human cognition. Similarity and categorization fall out of such an architecture, as, for example, countless fonts and sizes of the letter A will all meet the criteria that define "A-ness" in terms of the configuration of lower level features such as T-junctions, L-junctions, lines, and terminators. Two things will be similar to the extent that they meet the same criteria, and they will lie in different categories when there is not a spectrum of criterial satisfaction but a clear boundary. Thus, criteria underlie both the formation of categories and a distance metric among categories of things that meet criteria. This permits the emergence of representational spaces based on a multidimensional similarity/difference metric. Norm-based encoding also affords the possibility of generalization, such that all things that obey certain criteria are tokens of a general class, even if they are very different at the level of defining sensory features. Prototyping also emerges as that which best meets particular criteria. Criteria can also be assessed in parallel at multiple levels within a criterial hierarchy, permitting constraint satisfaction across the hierarchy and a path to recognition exploiting many parallel avenues of matching input to memory (e.g. simultaneously on the basis of various featural, configural, and contextual cues). In particular, partial inputs can nonetheless lead to criterial satisfaction, thus explaining the brain's ability to complete degraded, partial, or occluded patterns. Norm-based encoding can even account

for the fact that objects or meanings "prime" or potentiate other objects or meanings, as *bank* primes *money*. For example, imagine a neuron that fires if and only if the word *soldier* is spoken or read. Depending on how this neuron is connected to other neurons in a semantic associative network of neurons, the tendency of neurons to fire that were tuned to *tank, gun, platoon, peace,* and so forth could become more or less likely.

The advantage of having chains of successive, discrete criterial detectors is that different criteria can be set at different stages of neural processing, allowing for ever more complex criteria to be set. For example, photoreceptors initially detect light at points. Then center-surround receptive fields of ganglion and LGN cells detect small regions of light, particularly near borders. In V1, one finds cells with novel tuning properties: Simple cells detect oriented edges and bars at a location, complex cells detect moving edges and bars within a region, and hypercomplex cells detect edges and bars of certain lengths (Hubel & Wiesel, 1959, 1968). This, in turn, allows curvature detectors, closed region detectors, and perhaps even eye and mouth detectors in anterior occipital regions, and hand and face detectors (Perrett et al., 1982; Rolls, 1984; Yamane, Kaji, & Kawano, 1988) in the fusiform gyrus and other regions sensitive to configurations of features. This cascade may lead to neuronal populations sensitive to many other kinds of complex forms in AIT (Logothetis & Sheinberg, 1996; Gross, 2000). Initially it was thought that such a hierarchy would culminate in a stage of "grandmother cells" (Barlow, 1972; Gross, 2002). Very high level criteria decoders do indeed exist in the hippocampus and AIT (Connor, 2005; Gross, 2000; Kreiman, Fried, & Koch, 2002; Kreiman, Koch, & Fried, 2000, Quiroga, Reddy, Kreiman, Koch, & Fried, 2005; Quiroga, Kreiman, Koch, & Fried, 2008). However, because they remain criterial, a wide range of inputs will satisfy them, including things that are not your grandmother. For example, anything that satisfies the holistic criteria of what counts as a face (e.g. two dark eyelike blobs and a mouthlike blob within a round closed region) might trigger some face detectors (Freiwald, Tsao, & Livingstone, 2009; Jagadeesh, 2009). This might be why we sometimes find that we unwillingly see a face in a house's facade, or the front end of a car and why we can other times will ourselves to see a face in objects that are not faces.

The pandemonium model (Selfridge, 1959) is an instance of a hierarchical architecture of neuron-like units responding criterially to ever more complex features. It was proposed before Hubel and Wiesel discovered very similar detectors in area V1 of visual cortex. A more recent model that proposes higher level criterial detectors that pool over inputs from progressively more complex lower level detectors is described in the HMAX model of Riesenhuber and Poggio (1999, 2000, 2002), which is quite explicitly built on known neural structures. Whereas the pandemonium model involves linear pooling of responses from one stage in the hierarchy of feature detection to the next, HMAX involves nonlinear pooling, but the general strategy is similar. Another model suggests that object shape is coded criterially by neurons that respond to equivalence classes of basic shape attributes such as "convex in the lower left" (Yamane, Carlson, Bowman, Wang, & Connor, 2008). On this account, the population response to a complex input shape would be reflected in the responses of a "basis set" of neuronal responses to particular configural attributes.

Part-Based Structural Theories

Object recognition requires that some type of object representation is derived from the image and then matched to a corresponding type of object representation in memory according to a classification or matching rule or decision process. Structural description theories purport to describe both how objects are represented perceptually and in memory. Various versions were developed by computer scientists and computationally oriented psychologists, including Binford (1971), Biederman (1987), Marr (1982; Marr & Nishihara, 1978), and Palmer (1975b). Of the specific theories that have been advanced within this general framework, we will describe only one in detail: Biederman's (1987) *recognition by components theory*, sometimes called geon theory. It is not radically different from several others, but it is easier to describe and has been developed with more attention to the results of experimental evidence. We present it as representative of this class of models rather than as the "correct" or even the "best" one.

Recognition by components (RBC) theory is Biederman's (1987) attempt to formulate a single, psychologically motivated theory of how people recognize and classify objects. It is based on the idea that objects can be specified as spatial arrangements of a small set of volumetric primitives, which Biederman called *geons*. Object categorization then occurs by matching a geon-based structural description of the target object with corresponding geon-based structural descriptions of object categories. It was later implemented as a neural network (Hummel and Biederman, 1992), but we consider it here at the more abstract algorithmic level of Biederman's (1987) original formulation.

Geons. The first important assumption of RBC theory is that both the stored representations of categories and the representation of a currently attended object are volumetric structural descriptions. RBC representations are functional hierarchies whose nodes correspond to a discrete set of three-dimensional volumes (geons) and whose links to other nodes correspond to relations among these geons. Geons are generalized cylinders that have been partitioned into discrete classes by dividing their inherently continuous parameters (see later) into a few discrete ranges that are easy to distinguish from most vantage points. From the relatively small set of 108 distinct geons, a huge number of object representations can be constructed by putting two or more geons together, much as an enormous number of words can be constructed from a relatively small number of letters. A few representative geons are illustrated in Figure 7.9 along with some common objects constructed by putting several geons together to form recognizable objects.

Biederman defined the set of 108 geons by making discrete distinctions in the following variable dimensions of generalized cylinders: *cross-sectional curvature* (straight versus curved), *symmetry* (asymmetrical versus reflectional symmetry alone versus both reflectional and rotational symmetry), *axis curvature* (straight versus curved), *cross-sectional size variation* (constant versus expanding and contracting versus expanding only), and *aspect ratio* of the sweeping axis relative to the largest dimension of the cross-sectional area (approximately equal versus axis greater versus cross-section greater). The rationale for these particular distinctions is that, except for aspect ratio, they are qualitative, rather than merely quantitative, differences that result in qualitatively different retinal projections. The image features that characterize different geons are, therefore, relatively (but not completely) insensitive to changes in viewpoint.

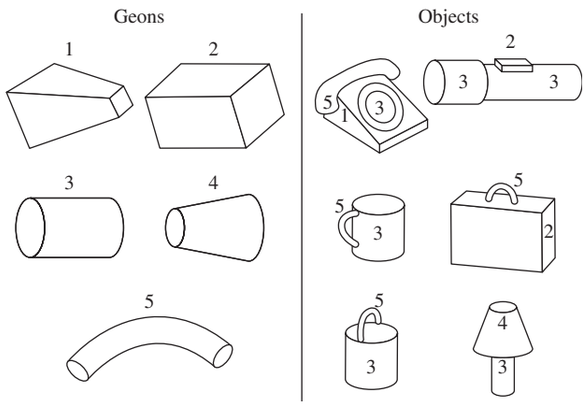


Figure 7.9 Examples of geons and their presence in objects (see text). [From Biederman, 1987]

Because complex objects are conceived in RBC theory as configurations of two or more geons in particular spatial arrangements, they are encoded as structural descriptions that specify both geons and their spatial relations. It is, therefore, possible to construct different object types by arranging the same geons in different spatial relations, such as the cup and pail (Figure 7.9). RBC uses 108 qualitatively different geon relations. Some of them concern how geons are attached (e.g., side-connected and top-connected) whereas others concern their relational properties, such as relative size (e.g., larger-than and smaller-than). With 108 geon relations and 108 geons, it is logically possible to construct more than a million different two-geon objects. Adding a third geon pushes the number of combinations into the billions. Clearly, geons are capable of generating a rich vocabulary of different complex shapes. Whether it is sufficient to capture the power and versatility of visual categorization is a question to which we will return later.

Once the shape of an object has been represented via its component geons and their spatial relations, the problem of object categorization within RBC theory reduces to the process of matching the structural description of an incoming object with the set of structural descriptions for known entry-level categories. RBC theory proposes that this process takes place in several stages, the details of which are complex and will not be described here. (The interested reader should consult Biederman's papers.) In the original formulation the overall flow of information was depicted in the flowchart of Figure 7.10. (a) An *edge-extraction* process initially produces a line drawing of the edges present in the visual scene. (b) The image-based properties needed to identify geons are extracted from the edge information by *detection of nonaccidental properties*. The crucial features are the nature of the edges (e.g., curved versus straight), the nature of the vertices (e.g., Y-vertices, K-vertices, L-vertices), parallelism (parallel versus nonparallel), and symmetry (symmetric versus asymmetric). The goal of this process is to provide the feature-based information required to identify the different kinds of geons (see stage d). (c) At the same time as these features are being extracted, the system attempts to *parse objects at regions of deep concavity*, as suggested by Hoffman and Richards (1984; Hoffman & Singh, 1997). The goal of this parsing process is to divide the object into component geons without having to match them explicitly on the basis of edge and vertex features. (d) The combined results of feature

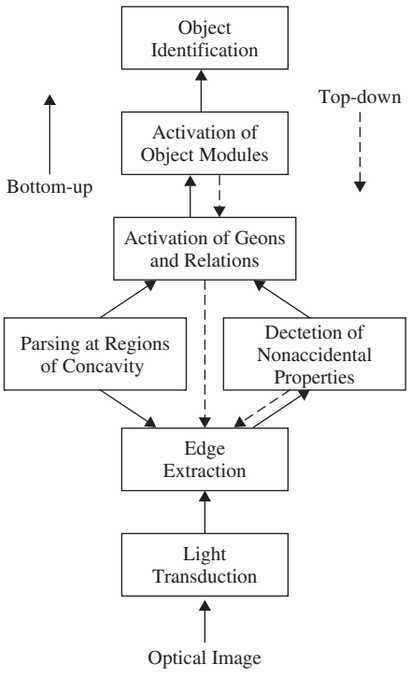


Figure 7.10 Processing stages in RBC theory (see text). [From Biederman, 1987]

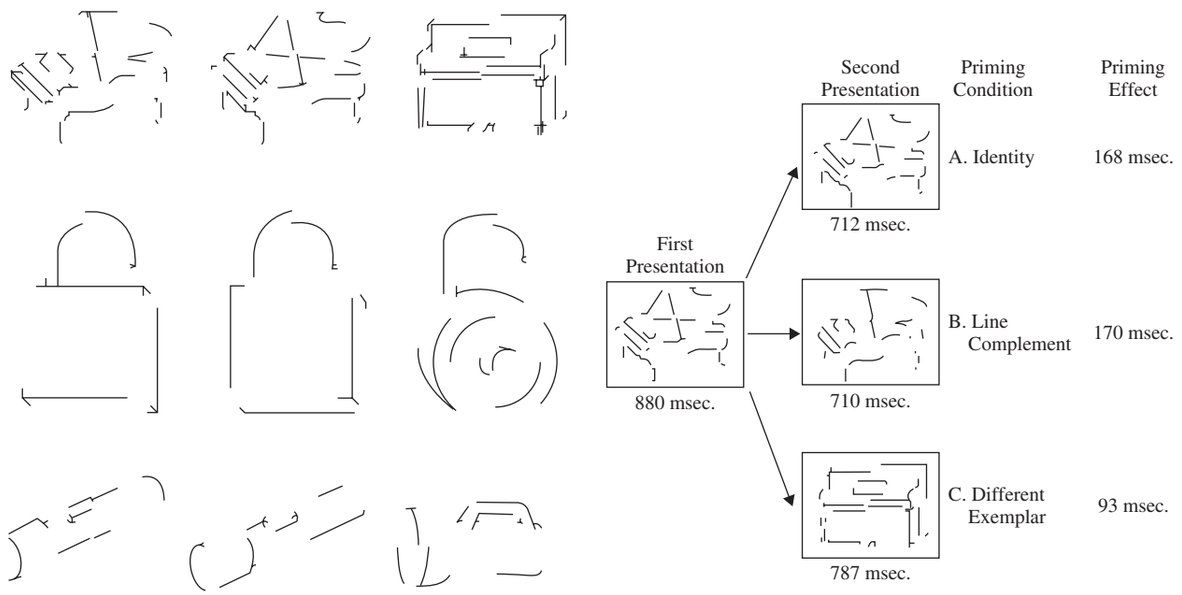
detection (b) and object parsing (c) are used to *activate the appropriate geons and spatial relations* among them. (e) Once the geon description of the input object is constructed, it automatically causes the *activation of similar geon descriptions stored in memory*. This matching process is accomplished by activation spreading through a network from geon-nodes and relation-nodes present in the representation of the target object to similar geon-nodes and relation-nodes in the category representations. This comparison is a fully parallel process, matching the geon description of the input object against all category representations at once and using all geons and relations at once. (f) Finally, *object recognition* or recognition occurs when the target object is classified as an instance of the entry-level category that is most strongly activated by the comparison process, provided it exceeds some threshold value.

Although the general flow of information within RBC theory is generally bottom-up, it also allows for top-down processing. If sensory information is weak—for example, noisy, brief, or otherwise degraded images—top-down effects are likely to occur. There are two points in RBC at which they are most likely to happen: feedback from geons to geon features and feedback from category representations to geons. Contextual effects could also occur through feedback from prior or concurrent object recognition to the nodes of related sets of objects, although this level of processing was not actually represented in Biederman's (1987) original model.

A central idea of part-based structural accounts of object recognition is that perceptual organization is centrally aimed at structuring the perceived world into part-whole hierarchies. It is undeniable that human bodies have heads, arms, legs, and a torso; tables have a flat top surface and legs; an airplane has a fuselage, two main wings, and several smaller tail fins. One important question is whether these parts play a significant mediating role in object recognition or recognition, as many theorists have claimed (e.g., Biederman, 1987; Marr, 1982; Neisser, 1967). The most revealing studies on this issue were performed by Biederman and Cooper (1991) using a priming paradigm. They showed that identification of degraded line drawings in the second (test) block of trials was facilitated when subjects had seen the *same parts* of the same objects in the initial (priming) block, but not when they had seen *different parts* of the same object in the priming block. This result implies that the process of identifying or recognizing objects is mediated by perceiving their parts and spatial interrelations because, otherwise, it is not clear why priming occurs only when the same parts were seen again.

The drawings Biederman and Cooper (1991) used were degraded by deleting half the contours in each stimulus. In the first experiment, subjects were shown a priming series of contour-deleted drawings and then a test series in which they saw either the identical drawing (Figure 7.11a), its line complement (Figure 7.11b), or a different object from the same category (Figure 7.11c). The results showed that the line-complement drawings produced just as much priming (170 ms) as the identical drawings (168 ms), and much more than the same-name drawings (93 ms). Biederman and Cooper (1991) argued that the stronger priming in the first two conditions was due to the fact that the same parts were perceived in both the identical and the line-complement drawings.

To be sure that this pattern was not due merely to the fact that the same object was depicted in the same pose in both of these conditions, they performed a second experiment in which half of the *parts* were deleted in the initial priming block (Figures 7.12a–c). Then, in the test block, they found that priming by the part-complement drawings was much less (108 ms) than priming by the identical drawings (190 ms). In fact, part-complement priming was no different from that in the same-name control condition (110 ms). Thus, the important feature for obtaining



Three of the object classes used in experiments on complementary feature priming. Left and middle columns: Examples of contour-deleted complementary images. From each component for each image, alternate vertices and edges have been removed so that each component had 50% of the contour of the original. When superimposed, the members of a complementary pair would make an intact figure with no overlap in contour. Assuming that the image in the left column was originally shown on the first (priming) block, the figure in the middle column would be an instance of complementary priming and the right figure would be a different exemplar (same name) control. [From Palmer, 1999]

Figure 7.11 A line-complement priming experiment (see text)

significantly more priming than for mere response repetition is that the same parts must be visible in the priming and test blocks. This result supports the inference that object recognition is mediated by part perception.

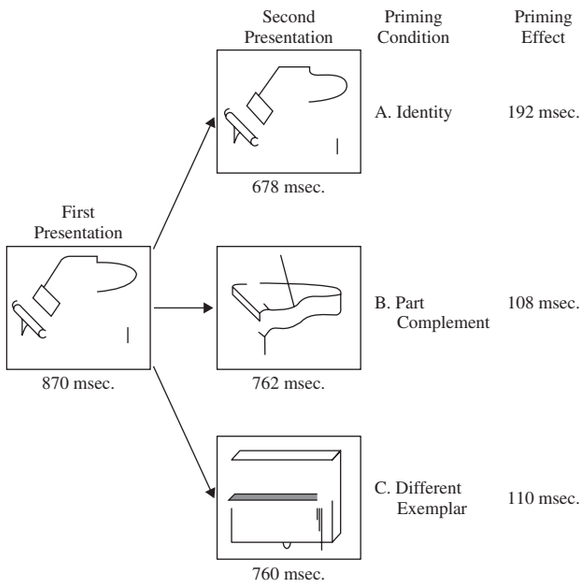


Figure 7.12 A part-complement priming experiment (see text). [From Palmer, 1999]

Viewpoint Effects

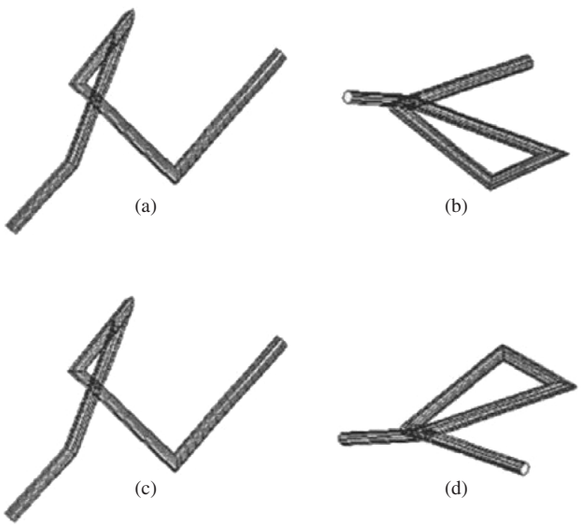
One of the seemingly obvious facts about identifying three-dimensional objects is that people can do it from almost any viewpoint. The living room chair, for example, seems to be easily perceived as such regardless of whether one is looking at it from the front, side, back, top, or any combination of these views. Thus, one of the important phenomena that must be explained by any theory of object classification is how this is possible.

However, given the fact that people *can* categorize objects from various perspective views, it is all too easy to jump to the conclusion that object categorization is *invariant* over perspective views. Closer study indicates that this is not true. Palmer, Rosch, & Chase (1981) systematically investigated and documented perspective effects in object recognition. They began by having subjects view many pictures of the same object and make subjective ratings of how much each one looked like the objects they depicted using a scale from 1 (very like) to 7 (very unlike). Other subjects were then asked to name the entry-level categories of these pictures, as quickly as possible, for five perspectives ranging from the best to the worst, based on the ratings made by other participants. Pictures rated as the best (i.e. the "canonical") perspective were

named fastest, and naming latencies gradually increased as the "goodness" of the views declined, with the "worst" (i.e. noncanonical) ones being named much more slowly than the "best" ones.

It seems possible that such perspective effects could be explained by familiarity: perhaps canonical views are simply the most frequently seen views. More recent studies have examined perspective effects using identification of novel objects to control for frequency effects. For example, Edelman and Bülthoff (1992) found canonical view effects in recognition time for novel bent-paperclip objects that were initially presented to subjects in a sequence of static views that produced apparent rotation of the object in depth (Figure 7.13). Because each single view was presented exactly once in this motion sequence, familiarity effects should be eliminated. Even so, recognition performance varied significantly over viewpoints, consistent with the perspective effects reported by Palmer et al. (1981).

Further studies have shown that familiarity *does* matter, however. When only a small subset of views were displayed in the initial training sequence, later recognition performance was best for the views seen during the training sequence and decreased with angular distance from these training views (Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992). These results suggest that subjects may be storing specific two-dimensional views of the objects and matching novel views to them via processes that deteriorate with increasing disparity between the novel and stored views.



Stimuli used in an experiment on object recognition from different viewpoints. Representative "best" and "worst" views for one of the test objects. (a) View with shortest response time (711 msec). (b) View with longest response time (1,405 msec). (c) View with lowest error rate (0%). (d) View with highest error rate (27%). [From Bülthoff and Edelman, 1992]

Figure 7.13

Q8

Further experiments demonstrated that when multiple views of the same objects were used in the training session, recognition performance improved, but the improvement depended on the relation of the test views to the training views (Bülthoff & Edelman, 1992). In particular, if the novel test views were related to the training views by rotation about the *same* axis through which the training views were related to each other, recognition was significantly better than for novel views that were rotations about an *orthogonal* axis. This suggests that people may be interpolating between and extrapolating beyond specific two-dimensional views in recognizing three-dimensional objects. This possibility will be important when view-based theories of object categorization are described below (e.g., Poggio & Edelman, 1990; Ullman, 1996; Ullman & Basri, 1991).

A different method of study, known as the "priming paradigm," has produced interesting, but somewhat contradictory, results about perspective views. The basic idea behind this experimental design is that categorizing a particular picture of an object will be faster and more accurate if the same picture is presented a second time, because the processes that accomplish it initially are in a state of heightened readiness for the second presentation (Bartram, 1974). The priming effect is defined as the difference between the naming latencies in the first block of trials and those in the second block of repeated pictures. What makes priming experiments informative about object categorization is that the "repetitions" in the second block of trials can differ from the initial presentation in different ways. For example, repetitions can be of the same object, but with changes in its position within the visual field (e.g., left versus right visual hemifield), its retinal size (large versus small), its mirror-image reflection (as presented initially or left-right reversed), or the perspective from which the object is viewed.

The results of such studies show that the magnitude of the object priming effect does not diminish when the second presentation shows the same object in a different position or reflection (Biederman & Cooper, 1991) or even at a different size (Biederman & Cooper, 1992). Showing the same object from a different perspective, however, has been found to reduce the amount of priming (Bartram, 1974). This perspective effect is thus consistent with the naming latency results reported by Palmer et al. (1981)

and the recognition results by Edelman and Bülthoff (1992) and Bülthoff and Edelman (1992). Later studies on priming with different perspective views of the same object by Biederman and Gerhardstein (1993), however, failed to show any significant decrease in priming effects due to depth rotations.

To explain this apparent contradiction, Biederman and Gerhardstein (1993) went on to show that priming effects did not diminish when the same *parts* were visible in the different perspective conditions. This same-part visibility condition is not necessarily met by the views used in the other studies, which often include examples in which different parts were visible from different perspectives. Visibility of the same versus different parts may thus explain why perspective effects have been found in some experiments but not in others. The results of these experiments on perspective effects, therefore, suggest care in distinguishing two different kinds of changes in perspective: those that do not change the set of visible parts and those that do.

Orientation Effects

Object recognition also varies with changes in an object's orientation when it is rotated about the observer's line of sight rather than in depth. These orientation effects cannot be explained by differences in the visibility of parts, however, because the same parts are visible in all cases. Jolicoeur (1985) has shown that subjects are faster at categorizing pictures of objects in their normal, upright orientation than when they are when misoriented in the picture plane. Naming latencies increase almost linearly with angular deviation from their upright orientation, as though subjects were mentally rotating the objects to upright before recognizing it and making their response.

Interestingly, orientation effects diminish considerably with extended practice. Tarr and Pinker (1989) studied this effect using novel objects so that the particular orientations at which subjects saw the objects could be precisely controlled. When subjects received extensive practice with the objects at *several* orientations, rather than just one, naming latencies were fast at *all* the learned orientations. Moreover, response times at novel orientations increased with distance from the nearest familiar orientation. Tarr and Pinker, therefore, suggested that people may actually store multiple representations of the same object at different orientations rather than a single representation that is orientation invariant. This possibility will become particularly important now that view-specific theories of categorization will be considered.

View-Specific Theories

There is a large body of empirical evidence that shows that object recognition is at least partially view dependent, in contrast to the predictions of structural theories such as geon theory. Researchers who emphasize this evidence argue that objects are represented and stored as a series of views. However, what comprises a view is not clear. At one extreme, a view might just be a two dimensional image. Such theories have difficulty accounting for the various constancies (i.e., indifferences to image transformations) expressed by the visual system. For example, an object defined by contours alone, motion alone, or texture alone will tend to look like it has the same shape across these cues. Moreover an object viewed from various distances and under various lighting conditions will generally appear to have the same shape, although particular images will be very different from each other. A more moderate stance is that a view is a collection of features. Such features, even if they do not explicitly represent three-dimensional shape or depth information, may implicitly capture three-dimensional information, because viewpoint invariant recognition could emerge if all views of an object are matched to the same node in a distributed neural network. If network models can be built that match correctly, it may become difficult to experimentally distinguish whether the visual system constructs explicit representations of three-dimensional shape or whether it only acts as if it did. At the other extreme, a view might include an explicit representation of three-dimensional shape. Tarr and Kriegman (2000), for example, suggest that a view is a span of viewpoints over which the qualitative shape description, in terms of occluding contour relationships, does not change. This converges to a certain extent with a revised version of geon theory (Biederman & Gerhardstein, 1995; Hummel & Biederman, 1992), according to which recognition will be view invariant only over a set of views for which a given collection of geons is visible. Three-dimensional structures (such as holes, protrusions, parts, corners, valleys, indentations, etc.) and the particular spatial relationships that hold among them (e.g., hole below pinnacle above bulge) which can be discerned from a given viewpoint are intrinsic to the object and can underlie a viewpoint-invariant representation of shape because these same structures will be visible from many other viewpoints. Even if a geon description per se is not utilized by the visual system for recognition, it is likely that some other structural description is employed. In general, the visual system attempts to recover the intrinsic properties of objects (e.g., surface reflectance, material substance,

three-dimensional shape) because these are more or less constant, whereas extrinsic properties (e.g., lighting, shading, shadows, distance, orientation) are constantly changing. Both intrinsic and extrinsic information can be derived from the image, and probably both types are stored and utilized for various tasks, including recognition.

In many ways, the starting point for view-specific theories of object recognition is the existence of the perspective effects described previously. The fact that recognition and categorization performance is not actually invariant over different views (e.g., Palmer et al., 1981) raises the possibility that objects might be identified by matching two-dimensional input views directly to some kind of view-specific category representation. It cannot be done with a single, specific view (such as one canonical perspective) because there is simply not enough information in any single view to identify other views. A more realistic possibility is that there might be *multiple* two-dimensional representations from several different viewpoints that can be employed in recognizing objects. These multiple views are likely to be those perspectives from which the object has been seen most often in past experience. Evidence supporting this possibility has come from a series of experiments that studied the identification of two-dimensional figures at different orientations in the frontal plane (Tarr & Pinker, 1989) and of three-dimensional figures at different perspectives (Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992).

Q10

There are several theories of object recognition that incorporate some degree of view specificity. One is Koenderink and Van Doorn's (1979) aspect- graph theory,

Q11

which is a well-defined elaboration of Minsky's (1975) frame theory of object perception. An aspect graph is a network of representations containing all topologically distinct two-dimensional views (or aspects) of the same object. Its major problem is that it cannot distinguish among different objects that have the same edge topology. All tetrahedrons are equivalent within aspect-graph theory, for example, despite the existence of large metric differences that are easily distinguished perceptually. This means that there is more information available to the visual system than is captured by edge topology, a conclusion that led to later theories in which projective geometry plays an important role in matching input views to object representations.

One approach was to match incoming two-dimensional images to internal three-dimensional models by an alignment process (e.g., Huttenlocher & Ullman, 1987; Lowe, 1985; Ullman, 1989). Another was to match incoming two-dimensional images directly against stored two-dimensional views, much as template theories advocate (e.g., Poggio & Edelman, 1990; Ullman, 1996; Ullman & Basri, 1991). The latter, exclusively two-dimensional, approach has the same problem that plagues template theories of recognition: An indefinitely large number of views would have to be stored. However, modern theorists have discovered computational methods for deriving many two-dimensional views from just a few stored ones, thus suggesting that template-like theories may be more viable than had originally been supposed.

Ullman and Basri (1991) demonstrated the viability of deriving novel two-dimensional views from a small set of other two-dimensional views, at least under certain restricted conditions, by proving that all possible views of an object can be reconstructed as a linear combination from just three suitably chosen orthographic projections of the same three-dimensional object. Figure 7.14 shows some rather striking examples based on this method. Two actual two-dimensional views of a human face (models M1 and M2) have been combined to produce other two-dimensional views of the same face. One is an intermediate view that has been interpolated *between* the two models (linear combination LC2), and the other two views have been extrapolated *beyond* them (linear combinations LC1 and LC3). Notice the close resemblance between the interpolated view (LC2) and the actual view from the corresponding viewpoint (novel view N).

This surprising result only holds under restricted conditions, however, some of which are ecologically unrealistic. Three key assumptions of Ullman and Basri's (1991) analysis are that (1) all points belonging to the object must be visible in each view, (2) the correct correspondence of all points between each pair of views must be known, and (3) the views must differ only by rigid transformations and/or uniform size scaling (dilations). The first assumption requires that none of the points on the object be occluded in any of the three views. This condition holds approximately true for wire objects, which are almost fully visible from any viewpoint, but it is violated by virtually all other three-dimensional objects due to occlusion. The linear combinations of the faces in Figure 7.14, for example, actually generate the image of a "mask" of the facial surface itself rather than of the whole head. The difference can be seen by looking carefully at the edges of the face where the head ends rather abruptly and unnaturally. The linear combination method would not be able to derive a profile view of the same head, because the back of the head is not present in either of the model views (M1 and M2) used to extrapolate other views.

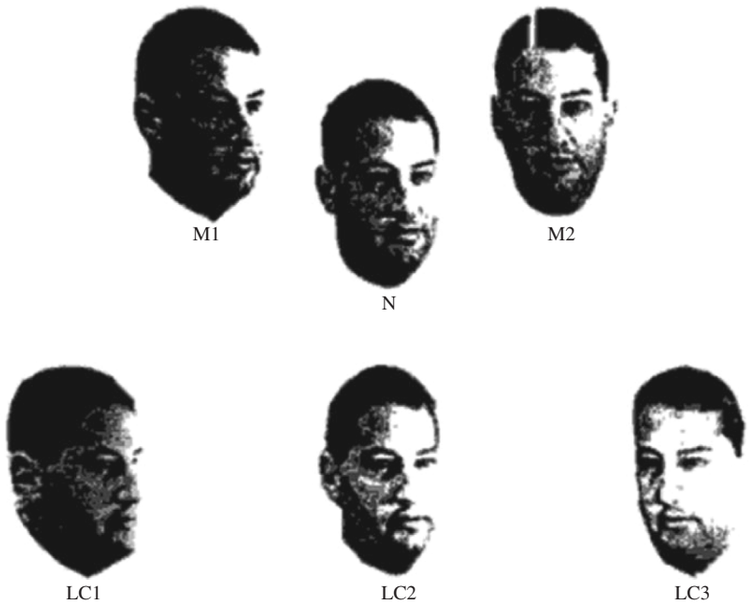


Figure 7.14 Novel views obtained by combination of gray-scale images (see text). [From Ullman, 1996]

The second assumption requires that the correspondence between points in stored two-dimensional views be known before the views can be combined. Although solving the correspondence problems is a nontrivial computation for complex objects, it can be derived "off-line" rather than during the process of recognizing an object. The third assumption means that the view combination process will fail to produce an accurate combination if the different two-dimensional views include plastic deformations of the object. If one view is of a person standing and the other of the same person sitting, for instance, their linear combination will not necessarily correspond to any possible view of the person. This restriction thus, can cause problems for bodies and faces of animate creatures as well as inanimate objects made of pliant materials (e.g., clothing) or having a jointed structure (e.g., scissors). Computational theorists are currently exploring ways of solving these problems (see Ullman, 1996, for a wide-ranging discussion of such issues), but they are important limitations of the linear combinations approach.

The results obtained by Ullman and Basri (1991) prove that two-dimensional views can be combined to produce new views under the stated conditions, but it does not specify how these views can be used to recognize the object from an input image. Further techniques are required to find a best-fitting match between the input view and the linear combinations of the model views as part of the object recognition process. One approach is to use a small number of features to find the best combination of the model views. Other methods are also possible, but are too technical to be described here. (The interested reader can consult Ullman's, 1996, book for details.)

Despite the elegance of some of the results that have been obtained by theorists working within the view-specific framework, such theories face serious problems as a general explanation of visual object recognition: (a) They do not account well for people's perceptions of three-dimensional structure in objects. Just from looking at an object, even from a single perspective, people generally know a good deal about its three-dimensional structure, including how to shape their hands to grasp it and what it would feel like if they were to explore it manually. It is not clear how this can occur if all they have access to is a structured set of two-dimensional views. (b) Most complex objects have a fairly clear perceived hierarchical structure in terms of parts and subparts. The view-specific representations considered earlier do not contain any explicit representation of such hierarchical structure because they consist of sets of unarticulated points or low-level features, such as edges and vertices. It is not clear, then, how such theories could explain Biederman and Cooper's (1991) priming experiments on the difference between line and part deletion conditions. Ullman (1996) has suggested that parts as well as whole objects may be represented separately in memory. This proposal serves as a reminder that part-based recognition schemes, like RBC, and view-based schemes are not mutually exclusive, but can be combined into various hybrid approaches (e.g.,

Hummel & Stankiewicz, 1996). (c) Finally, it is not clear how the theory could be extended to handle object recognition for entry-level categories. The situations to which view-specific theories have been successfully applied thus far are limited to identical objects that vary only in viewpoint, such as recognizing different views of the same face. The huge variation among different exemplars of chairs, dogs, and houses poses a serious problem for view specific theories.

We have already discussed neurophysiological evidence for a structural encoding of object features and configurations that would appear to be viewpoint invariant as long as object-defining features are visible. There is, however, additional neurophysiological evidence that supports a view-based or image-based encoding of objects. In particular, inferotemporal neurons respond in a view-dependent manner. Following Bülthoff and Edelman (1992), Logothetis, Pauls, Bülthoff, and Poggio (1994) found that monkeys recognized objects better when presented from a familiar viewpoint than an unfamiliar one. Corresponding IT neurons were viewpoint dependent (Logothetis, Pauls, & Poggio, 1995). They suggested that a view-invariant representation might derive from the responses of multiple view-dependent neurons.

One possible resolution would be that both part-based and view-based processes may be used, but for different kinds of tasks (e.g., Farah, 1992; Tarr & Bülthoff, 1995). View-specific representations seem well suited to recognizing the very same object from different perspective views because, in that situation, there is no variation in the structure of the object; all the differences among images can be explained by the variation in viewpoint. Recognizing specific objects is difficult for structural description theories because their representations are seldom specific enough to discriminate between different exemplars. In contrast, structural description theories, such as RBC, seem well suited to entry-level categorization because they have more abstract representations that are better able to encompass shape variations among different exemplars of the same category. This is exactly where view-specific theories have difficulty.

Contextual Effects

All the phenomena of object recognition considered thus far concern the nature of the target object itself: how typical it is of its category; the perspective from which it is viewed; its size, position, orientation, and visible parts. However, identification can also be influenced by *contextual factors*: the spatial array of objects that surround the target object. One well-known contextual effect can be demonstrated by the phrase **THE CAT**, which everyone initially perceives as THE CAT. This seems entirely unproblematic—until one realizes that the central letters of both words are actually identical and ambiguous, midway between an H and an A. It is, therefore, possible that the letter strings could be perceived as TAE CHT, TAE CAT, or THE CHT, but this almost never happens.

There have been several well-controlled experiments documenting that appropriate context facilitates identification, whereas inappropriate context hinders it. In one such study, Palmer (1975a) presented subjects with line drawings of common objects to be identified, which were preceded by either a brief presentation of a contextual scene (Figure 7.15) or a blank screen. The relation between the contextual scene and the target object was manipulated.

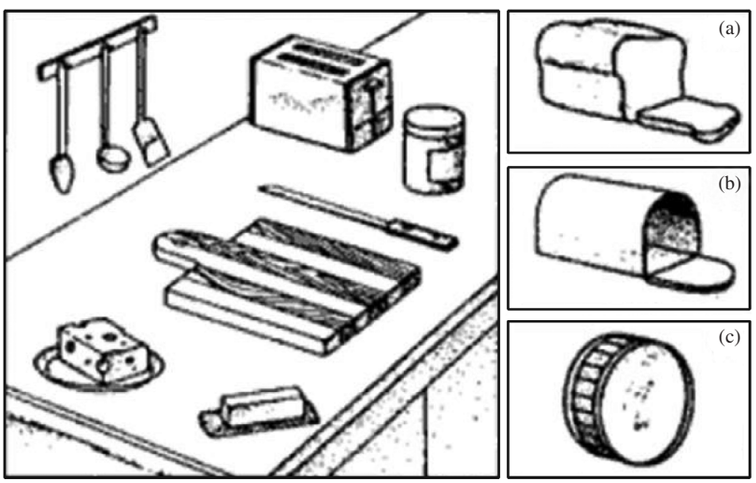


Figure 7.15 Stimuli from an experiment on contextual effects on object recognition (see text). [From Palmer, 1975a]

In the case of the kitchen counter scene, for example, the subsequently presented object could be either appropriate to the scene (a loaf of bread), inappropriate (a bass drum), or misleading in the sense that the target object was visually similar to the appropriate object (a mailbox). For the no-context control condition, the objects were presented following a blank field instead of a contextual scene. By presenting the objects and scenes in different combinations, all objects were equally represented in all four contextual conditions. The results of this experiment showed that appropriate contexts facilitated correct categorization relative to the no-context control condition, and that inappropriate contexts inhibited it. Performance was worst of all in the misleading context condition, in which subjects were likely to name the visually similar object appropriate to the scene. These differences demonstrate that recognition accuracy can be substantially affected by the nature of the surrounding objects in a simple identification task.

Biederman (1972; Biederman, Glass, & Stacy, 1973) used a different method to study context effects. He had subjects search for the presence of a given target object in a scene and measured how much time it took them to find the named target. In the first study, he manipulated context by presenting either a normal photograph or a randomly rearranged version. Subjects took substantially longer to find the target object in the rearranged pictures than in the normal ones.

These contextual effects indicate that relations among objects in a scene are complex and important factors for normal visual identification. Obviously, people can identify objects correctly even in bizarre contexts. A fire hydrant on top of a mailbox may take longer to identify—and cause a major double-take once it is—but people manage to recognize it even so. Rather, context appears to change the efficiency of identification. In each case, the target object in a "normal" context is processed quickly and with few errors, whereas one in an "abnormal" context takes longer to process and is more likely to produce errors. Since "normal" situations are, by definition, encountered more frequently than abnormal ones, such contextual effects are generally beneficial to the organism in its usual environment.

CONCLUSION

It is possible that both view-based and part-based schemes can be combined to achieve the best of both worlds. They are not mutually exclusive, and could even be implemented in parallel (e.g., Hummel & Stankiewicz, 1996). This approach suggests that when the current view matches one stored in view-based form in memory, recognition will be fast and accurate; when it does not, categorization must rely on the slower, more complex process of matching against structural descriptions. Which, if any, of these possible resolutions of the current conflict will turn out to be most productive is not yet clear. The hope is that the controversy will generate interesting predictions that can be tested experimentally, for that is how science progresses.

In this chapter, we have traversed the visual system from the level of initial extraction or detection of image primitives, to the stage where those primitives can be used as the input to complex algorithms that compute surface shape and layout. A great deal remains to be discovered about each level of analysis. We still do not understand how information is processed by neurons in a deep sense, and we certainly do not grasp how complex computations, such as those that presumably underlie gestalt grouping procedures or the generation of surfaces and volumes, are realized in the information processing of neuronal circuits. At a more abstract level of analysis, we do not understand the nature of the computations that generate veridical representations of shape within a fraction of a second, permitting matches to memory (recognition) and motoric behavior in response to the visual environment. Although much is already known, much more research needs to be done before we can say that we have even a basic understanding of how form is processed and represented in the nervous system.

Key processing is almost certainly happening at the level of neuronal circuits. However, we really do not have a good method for observing whole circuits. This would require measuring the firings of hundreds if not thousands of neurons at the same time that are known to be providing input to one another. Single unit work is the present gold standard, and it is certainly useful, but its application is akin to trying to understand architecture by observing a brick. Multi-unit techniques are available, but they do not provide information about whether the neurons that they are measuring are in direct communication with one another as part of a common circuit. Such methods increase data yield, and do allow an analysis of correlations between activity in pairs of neurons, but this is a far cry from a causal analysis of neuronal activity that a true circuit level of analysis would allow. Another important limitation in neuroscience is that although we can measure tuning properties and receptive fields via single cell recording quite well, we cannot measure *operations* very easily. Indeed, specifying the operations that are carried

out on input, rather than the detection of characteristics of input, is crucial in specifying neural information processing, and it is here that important advances are required. A technological breakthrough is needed that will allow the simultaneous measurement of hundreds of neurons forming a common circuit, ideally in freely moving animals and even humans. Until that point is reached, neuroscience will probably lack a way to answer the deepest questions about the information processing operations carried out by neural circuits. Unfortunately, the amount that has yet to be explained is humbling: We have not yet deciphered the neural code, we do not yet understand the neural basis of consciousness, and we do not yet have a clear understanding of how—or even whether—mental events could cause neuronal events.

The foregoing discussion concerning perceptual organization, shape formation, and object recognition barely scratches the surface of what needs to be known to understand the central mystery of vision: How are the responses of millions of independent retinal receptors processed to a level at which information is made explicit about the identities and spatial relations among meaningful objects in the environment? It is indisputable that people achieve such knowledge, that it is evolutionarily important for our survival as individuals and as a species, and that scientists do not yet understand how it arises. Despite the enormous amount that has been learned about low-level processing of visual information, the higher-level problems of organization and identification remain largely unsolved. It will take a concerted effort on the part of the entire vision science community—including psychophysicists, cognitive psychologists, physiologists, neuropsychologists, and computer scientists—to reach explanatory solutions. Only then will we begin to understand how the extraordinary feat of perception and recognition is accomplished by the visual nervous system.

REFERENCES

Afraz, S. R., Kiani, R., & Esteky, H. (2006). Microstimulation of inferotemporal cortex influences face categorization. *Nature*, *442*, 692–695.

Agam, Y., Liu, H., Pappanastassiou, A., Buia, C., Golby, A. J., Madsen, J. R., & Kreiman, G. (2010). Robust selectivity to two-object images in human visual cortex. *Current Biology*, *20*, 872–879.

Alais, D., Blake, R., & Lee, S.-H. (1998). Visual features that vary together over time group together over space. *Nature Neuroscience*, *1*, 160–164.

Astafiev, S. V., Stanley, C. M., Shulman, G. L., & Corbetta, M. (2004). Extrastriate body area in human occipital cortex responds to the performance of motor actions. *Nature Neuroscience*, *7*(5), 542–548.

Barenholtz, E. (2010). Convexities move because they contain matter. *Journal of Vision*, *10*, 1–12.

Barenholtz, E., & Feldman, J. (2006). Determination of visual figure and ground in dynamically deforming shapes. *Cognition*, *101*, 530–544.

Barenholtz, E., & Tarr, M. J. (2009). Figure-ground assignment to a translating contour: A preference for advancing vs. retreating motion. *Journal of Vision*, *9*, 1–9.

Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, *1*, 371–394.

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

Bear, M. F., Connors, B., & Paradiso, M. (2007). *Neuroscience: Exploring the brain*. Hagerstown, MD: Lippincott, Williams, & Wilkins.

Bell, A. H., Hadj-Bouziane, F., Frihauf, J. B., Tootell, R. B., & Ungerleider, L. G. (2009). Object representations in the temporal cortex of monkeys and humans as revealed by functional magnetic resonance imaging. *Journal of Neurophysiology*, *101*, 688–700.

Biederman, I. (1972). Perceiving real-world scenes. *Science*, *177*(4043), 77–80.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*, 115–147.

Biederman, I., & Cooper, E. E. (1991). Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology*, *23*, 393–419.

Biederman, I., & Cooper, E. E. (1992). Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 121–133.

Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, *19*, 1162–1182.

Biederman, I., & Gerhardstein, P. C. (1995). Viewpoint-dependent mechanisms in visual object recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 1506–1521.

Biederman, I., Glass, A. L., & Stacy, E. W. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology*, *97*, 22–27.

Binford, T. O. (1971). *Visual perception by computer*. Paper presented at the IEEE Conference on Systems and Control, Miami, FL.

Bruce, C., Desimone, R., & Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology*, *46*, 369–384.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional interpolation theory of object recognition. *Proceedings of the National Academy of Science, USA*, *89*, 60–64.

Chow, K. L. (1951). Effect of partial extirpations of the posterior association cortex on visually mediated behavior. *Comparative Psychological Monographs*, *20*, 187–217.

Connor, C. (2005). Friends and grandmothers. *Nature*, *435*(7045), 1036–1037.

Damasio, A. R., Damasio, H., & Van Hoesen, G. W. (1982). Prosopagnosia: Anatomic basis and behavioral mechanisms. *Neurology*, *32*(4), 331–341.

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, *4*, 2051–2062.

De Valois, R., & De Valois, K. (1990). *Spatial Vision. Oxford Psychology Series: No. 14*. New York, NY: Oxford University Press.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*(5539), 2470–2473.

Driver, J., & Baylis, G. C. (1996). Edge-assignment and figure-ground segmentation in short-term visual matching. *Cognitive Psychology*, *31*, 248–306.

Edelman, S., & Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research*, *32*, 2385–2400.

Eifuku, S., De Souza, W. C., Tamura, R., Nishijo, H., & Ono, T. (2004). Neuronal correlates of face identification in the monkey anterior temporal cortical areas. *Journal of Neurophysiology*, *91*, 358–371.

Ellis, A. W., & Young, A. W. (1988). *Human cognitive neuropsychology*. Hillsdale, NJ: Erlbaum.

Epstein, R., DeYoe, E. A., Press, D., & Kanwisher, N. (2001). Neuropsychological evidence for a topographical learning mechanism in parahippocampal cortex. *Cognitive Psychology*, *18*, 481–508.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*, 598–601.

Epstein, R. A., Parker, W. E., & Feiler, A. M. (2008). Two kinds of FMRI repetition suppression? Evidence for dissociable neural mechanisms. *Journal of Neurophysiology*, *99*, 2877–2886.

Ettlinger, G. (1990). "Object vision" and "spatial vision": The neuropsychological evidence for the distinction. *Cortex*, *26*, 319–341.

Farah, M. J. (1992). Is an object and object an object? Cognitive and neuropsychological investigations of domain specificity in visual object recognition. *Current Directions in Psychological Science*, *1*, 164–169.

Farah, M. J. (2000). *The cognitive neuroscience of vision*. Malden, MA: Blackwell.

Farah, M. J. (2004). *Visual agnosia* (2nd ed.). Cambridge, MA: MIT Press.

Farid, H., & Adelson, E. H. (2001). Synchrony does not promote grouping in temporally structured displays. *Nature Neuroscience*, *4*, 875–876.

Farivar, R. (2009). Dorsal-ventral integration in object recognition. *Brain Research Reviews*, *61*, 144–153.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*, 1–47.

Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. *Nature Neuroscience*, *12*, 1187–1196.

Gazzaniga, M. S. (Ed.) (1995). *The cognitive neurosciences*. Cambridge, MA: MIT Press.

Ghose, T., & Palmer, S. E. (2010). Extremal edges versus other principles of figure-ground organization. *Journal of Vision*, *10*, 1–17.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neuroscience*, *15*, 20–25.

Gray, C. M., & Singer, W. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proceedings of the National Academy of Sciences, USA*, *86*, 1698–1702.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*, 1409–1422.

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron*, *24*, 187–203.

Grill-Spector, K., Kushnir, T., Hendler, T., Edelman, S., Itzchak, Y., & Malach, R. (1998). A sequence of object-processing stages revealed by fMRI in the human occipital lobe. *Human Brain Mapping*, *6*, 316–328.

Gross, C. G. (2000). Coding for visual categories in the human brain. *Nature Neuroscience*, *3*, 855–856.

Gross, C. G. (2002). Genealogy of the "grandmother cell". *Neuroscientist*, *8*(5), 512–518.

Gross, C. G., Rocha-Miranda, C. E., & Bender, D. B. (1972). Visual properties of neurons in inferotemporal cortex of the Macaque. *Journal of Neurophysiology*, *35*, 96–111.

Hadj-Bouziane, F., Bell, A. H., Knusten, T. A., Ungerleider, L. G., & Tootell, R. B. (2008). Perception of emotional expressions is independent of face selectivity in monkey inferior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 5591–5596.

Hannula, D. E., Simons, D. J., & Cohen, N. J. (2005). Imaging implicit perception: Promise and pitfalls. *Nature Reviews Neuroscience*, *6*, 247–255.

Hasselmo, M. E., Rolls, E. T., & Baylis, G. C. (1989). The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural Brain Research*, *32*, 203–218.

Haxby, J. V., Ungerleider, L. G., Clark, V. P., Schouten, J. L., Hoffman, E. A., & Martin, A. (1999). The effect of face inversion on activity in human neural systems for face and object perception. *Neuron*, *22*, 189–199.

Helmholtz, H. L. (1867/1910). *Handbuch der physiologischen Optik*. Leipzig, Germany: L. Voss. Reprinted In A. Gullstrand, J. von Kries, & W. Nagel (Eds), *Handbuch der physiologischen Optik* (3rd ed.). Hamburg and Leipzig, Germany: L. Voss.

Hochberg, J., & Silverstein, A. (1956). A quantitative index for stimulus similarity: Proximity versus differences in brightness. *American Journal of Psychology*, *69*, 480–482.

Hoffman, D. D., & Richards, W. A. (1984). Parts of recognition. Special Issue: Visual cognition. *Cognition*, *18*, 65–96.

Hoffman, D. D., & Singh, M. (1997). Salience of visual parts. *Cognition*, *63*, 29–78.

Hoffman, K. L., Gothard, K. M., Schmid, M. C., & Logothetis, N. K. (2007). Facial expression and gaze-selective responses in the monkey amygdala. *Current Biology*, *17*, 766–772.

Horton, J. C., & Hoyt, W. F. (1991). The representation of the visual field in human striate cortex. A revision of the classic Holmes map. *Archives of Ophthalmology*, *109*, 816–824.

Hsieh, P.-J., & Tse, P. U. (2009). Feature mixing rather than feature replacement during perceptual filling-in. *Vision Research*, *49*(4), 439–450.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, *148*, 574–591.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture of the cat's visual cortex. *Journal of Physiology (London)*, *160*, 106–154.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, *195*, 215–243.

Hulleman, J., & Humphreys, G. W. (2004). A new cue to figure-ground coding: Top-bottom polarity. *Vision Research*, *44*, 2779–2791.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*, 480–517.

Hummel, J. E., & Stankeiwicz, B. J. (1996). Categorical relations in shape perception. *Spatial Vision*, *10*, 201–236.

Humphreys, G. W., & Riddoch, M. J. (2006). Features, objects, action: The cognitive neuropsychology of visual object processing, 1984–2004. *Cognitive Neuropsychology*, *23*, 156–183.

Huttenlocher, D. P., & Ullman, S. (1987). *Object recognition using alignment* (MIT AI Memo 937). Cambridge, MA: MIT Press.

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences of the United States of America*, *96*, 9379–9384.

Jagadeesh, B. (2009). Recognizing grandmother. *Nature Neuroscience*, *12*(9), 1083–1085.

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

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

Kandel, E., Schwartz, J., & Jessell, T. (2000). *Principles of neural science* (4th ed.). New York, NY: McGraw-Hill Medical.

Kanizsa, G., & Gerbino, W. (1976). Convexity and symmetry in figure-ground organization. In M. Henle (Ed.), *Vision and artifact*. New York, NY: Springer.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*, 4302–4311.

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, *71*, 856–867.

Koenderink, J. J., & van Doorn, A. J. (1979). The internal representation of solid shape with respect to vision. *Biological Cybernetics*, *32*, 211–216.

Kreiman, G., Fried, I., & Koch, C. (2002). Single-neuron correlates of subjective vision in the human medial temporal lobe. *Proceedings of the National Academy of Science, USA*, *99*, 8378–8383.

Kreiman, G., Koch, C., & Fried, I. (2000). Category specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience*, *3*, 946–953.

Kriegeskorte, N., Formisano, E., Sorger, B., & Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 20600–20605.

Kubovy, M., & Wagemans, J. (1995). Grouping by proximity and multistability in dot lattices: A quantitative Gestalt theory. *Psychological Science*, *6*, 225–234.

Landis, T., Regard, M., Bliestle, A., & Kleihues, P. (1988). Prosopagnosia and agnosia for noncanonical views. An autopsied case. *Brain*, *111*, 1287–1297.

Lee, S.-H., & Blake, R. (1999). Visual form created solely from temporal structure. *Science*, *284*, 1165–1168.

Leopold, D. A., Bondar, I. V., & Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, *442* (7102), 572–575.

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, *4*, 89–94.

Lerner, Y., Hendler, T., Ben-Bashat, D., Harel, M., & Malach, R. (2001). A hierarchical axis of object processing stages in the human visual cortex. *Cerebral Cortex*, *11*, 287–297.

Lissauer, H. (1890/1988). A case of visual agnosia with a contribution to theory. *Cognitive Neuropsychology*, *5*, 157–192.

Loffler, G., Yourganov, G., Wilkinson, F., & Wilson, H. R. (2005). FMRI evidence for the neural representation of faces. *Nature Neuroscience*, *8*, 1386–1390.

Logothetis, N. K., Pauls, J., Bulthoff, H. H., & Poggio, T. (1994). View-dependent object recognition by monkeys. *Current Biology*, *4*, 401–414.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, *5*, 552–563.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, *19*, 577–621.

Lowe, D. G. (1985). *Perceptual organization and visual recognition*. Boston, MA: Kluwer Academic.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society London*, *200*, 269–294.

Milner, P. M. (1974). A model for visual shape recognition. *Psychological Review*, *81*, 521–535.

Mishkin, M., & Ungerleider. L. G. (1982). Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behavioral and Brain Research*, *6*, 57–77.

Neisser, U. (1967). *Cognitive psychology*. New York, NY: Appleton-Century-Crofts.

Norman, J. (2002). Two visual systems and two theories of perception: An attempt to reconcile the constructivist and ecological approaches. *Behavioral and Brain Sciences*, *25*, 73–144.

Op de Beeck, H. P., & Baker, C. I. (2009). The neural basis of visual object learning. *Trends in Cognitive Sciences*, *14*, 22–30.

Orban, G. A. (2011). The extraction of 3D shape in the visual system of human and nonhuman primates. *Annual Review of Neuroscience*, *34*, 361–388.

Palmer, S. E. (1975a). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, *3*, 519–526.

Palmer, S. E. (1975b). Visual perception and world knowledge: Notes on a model of sensory-cognitive interaction. In D. A. Norman & D. E. Rumelhart (Eds), *Explorations in cognition* (pp. 279–307). San Francisco, CA: W. H. Freeman.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. Cambridge, MA: Bradford Books/MIT Press.

Palmer, S. E., Brooks, J. L., & Nelson, R. (2003). When does grouping happen? *Acta Psychologica*, *114*, 311–330.

Palmer, S. E., & Ghose, T. (2008). Extremal edges: A powerful cue to depth and figure-ground organization. *Psychological Science*, *19*, 77–84.

Palmer, S. E., Neff, J., & Beck, D. (1996). Late influences on perceptual grouping: Amodal completion. *Psychonomic Bulletin & Review*, *3*, 75–80.

Palmer, S. E., & Nelson, R. (2000). Late influences on perceptual grouping: Illusory contours. *Perception and Psychophysics*, *62*, 1321–1331.

Palmer, S. E., & Rock, I. (1994a). On the nature and order of organizational processing: A reply to Peterson. *Psychonomic Bulletin & Review*, *1*, 515–519.

Palmer, S. E., & Rock, I. (1994b). Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin & Review*, *1*, 29–55.

Palmer, S. E., Rosch, E., & Chase, P. (1981). Cannonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds), *Attention and performance IX* (pp. 135–151). Hillsdale, NJ: Erlbaum.

Park, S., & Chun, M. M. (2009). Different roles of the parahippocampal place area (PPA) and retrosplenial cortex (RSC) in panoramic scene perception. *Neuroimage*, *47*, 1747–1756.

Perrett, D. I., Rolls, E. T., & Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Experimental Brain Research*, *47*, 329–342.

Perrett, D. I., Smith, P. A., Potter, D. D., Mistlin, A. J., Head, A. S., Milner, A. D., & Jeeves, M. A. (1985). Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *223* (1232), 293–317.

Peterson, M. A., & Gibson, B. S. (1991). The initial identification of figure-ground relationships: Contributions from shape recognition processes. *Bulletin of the Psychonomic Society*, *29*, 199–202.

Pinsk, M. A., Arcaro, M., Weiner, K. S., Kalkus, J. F., Inati, S. J., Gross, C. G., & Kastner, S. (2009). Neural representations of faces and body parts in macaque and human cortex: A comparative FMRI study. *Journal of Neurophysiology*, *101*, 2581–2600.

Pinsk, M. A., DeSimone, K., Moore, T., Gross, C. G., & Kastner, S. (2005). Representations of faces and body parts in macaque temporal cortex: A functional MRI study. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 6996–7001.

Poggio, T., & Edelman, S. (1990). A neural network that learns to recognize three-dimensional objects. *Nature*, *343*, 263–266.

Puce, A., Allison, T., Asgari, M., Gore, J. C., & McCarthy, G. (1996). Differential sensitivity of human visual cortex to faces, letter strings, and textures: A functional magnetic resonance imaging study. *Journal of Neuroscience*, *16*, 5205–5215.

Puce, A., Allison, T., Gore, J. C., & McCarthy, G. (1995). Face-sensitive regions in human extrastriate cortex studied by functional MRI. *Journal of Neurophysiology*, *74*, 1192–1199.

Quiroga, R. Q., Kreiman, G., Koch, C., & Fried, I. (2008). Sparse but not "grandmother-cell" coding in the medial temporal lobe. *Trends in Cognitive Science*, *12*, 87–91.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435* (7045), 1102–1107.

Rajimehr, R., Young, J. C., & Tootell, R. B. (2009). An anterior temporal face patch in human cortex, predicted by macaque maps. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 1995–2000.

Rhodes , G., Jaquet, E., Jeffery, L. Evangelista, E. Keane, J. & Calder, A. J. (2011). Sex-specific norms code face identity. *Journal of Vision*, *11* (1), 1–11.

Rhodes, G., Michie, P. T., Hughes, M. E., & Byatt, G. (2009). The fusiform face area and occipital face area show sensitivity to spatial relations in faces. *European Journal of Neuroscience*, *30* (4), 721–733.

Rhodes, G., Watson, T. L., Jeffery, L., & Clifford, C. W. G. (2010). Perceptual adaptation helps us identify faces. *Vision Research*, *50*, 963–968.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*, 1019–1025.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, *3* (Supplement), 1199–1204.

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, *12*, 162–168.

Rock, I. (1983). *The logic of perception*. Cambridge, MA: MIT Press.

Rock, I., & Brosgole, L. (1964). Grouping based on phenomenal proximity. *Journal of Experimental Psychology*, *67*, 531–538.

Rock, I., Nijhawan, R., Palmer, S., & Tudor, L. (1992). Grouping based on phenomenal similarity of achromatic color. *Perception*, *21*, 779–789.

Rolls, E. T. (1984). Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. *Human Neurobiology*, *3*, 209–222.

Rolls, E. T. (2007). The representation of information about faces in the temporal and frontal lobes. *Neuropsychologia*, *45*, 124–143.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*, 328–350.

Rosch, E. (1975a). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*, 192–233.

Rosch, E. (1975b). The nature of mental codes for color categories. *Journal of Experimental Psychology: Human Perception and Performance*, *104*, 303–322.

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

Rossion, B., Caldara, R., Seghier, M., Schuller, A. M., Lazeyras, F., & Mayer, E. (2003). A network of occipito-temporal face-sensitive areas besides the right middle fusiform gyrus is necessary for normal face processing. *Brain*, *126*, 2381–2395.

Rubin, E. (1921). Visuell Wahrgenommene Figuren [Visual perception of figures]. Copenhagen, Denmark: Glydenalske boghandel.

Sacks, O. W. (1985). *The man who mistook his wife for a hat and other clinical tales*. New York, NY: Summit Books.

Sekuler, A. B., & Bennett, P. J. (2001). Generalized common fate: Grouping by common luminance changes. *Psychological Science*, *12*, 437–444.

Selfridge, O. (1959). Pandemonium: A paradigm for learning. In Symposium on the mechanization of thought processes. London, UK: HM Stationary Office.

Sergent, J., Ohta, S., & MacDonald, B. (1992). Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain*, *115*, 15–36.

Singer, W., & Gray, C. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience*, *18*, 555–586.

Sugase, Y., Yamane, S., Ueno, S., & Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, *400*, 869–873.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, *23*, 457–482.

Tarr, M. J., & Bulthoff, H. H. (1995). Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 1494–1505.

Tarr, M. J., & Kriegman, D. J. (2001). What defines a view? *Vision Research*, *41*, 1981–2004.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, *21*, 233–282.

Teuber, H. L. (1968). Alteration of perception and memory in man. In L. Weiskrantz (Ed.), *Analysis of behavioral change*. New York, NY: Harper & Row.

Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B., & Tootell, R. B. (2003). Faces and objects in macaque cerebral cortex. *Nature Neuroscience*, *6*, 989–995.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, *311*, 670–674.

Tsao, D. Y., & Livingstone, M. S. (2008). Mechanisms of face perception. *Annual Review of Neuroscience*, *31*, 411–437.

Tsao, D. Y., Moeller, S., & Freiwald, W. A. (2008). Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 19514–19519.

Tse, P. U. (2002). A contour propagation approach to surface filling-in and volume formation. *Psychological Review*, *109*, 91–115.

Tse, P. U., Martinez-Conde, S., Schlegel, A. A., & Macknik, S. L. (2005). Visibility, visual awareness, and visual masking of simple unattended targets are confined to areas in the occipital cortex beyond human V1/V2. *Proceedings of the National Academy of Science, USA*, *102*, 17178–17183.

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, *32*, 193–254.

Ullman, S. (1996). *High level vision*. Cambridge, MA: MIT Press.

Ullman, S., & Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*, 992–1006.

Ungerleider, L. G., & Bell, A. H. (2011). Uncovering the visual "alphabet": Advances in our understanding of object perception. *Vision Research*, *51*, 782–799.

Ungerleider, L. G., & Pribram, K. H. (1977). Inferotemporal versus combined pulvinar-prestriate lesions in the rhesus monkey: Effects on color, object and pattern discrimination. *Neuropsychologia*, *15*, 481–498.

Van Essen, D. C. (2004). Surface-based approaches to spatial localization and registration in primate cerebral cortex. *Neuroimage*, *23* (Suppl 1), S97–107.

Van Essen, D. C., Drury, H. A., Joshi, S., & Miller, M. I. (1998). Functional and structural mapping of human cerebral cortex: solutions are in the surfaces. *Proceedings of the National Academy of Science, USA*, *95*, 788–795.

Vecera, S. P., Flevaris, A. V., & Filapek, J. C. (2004). Exogenous spatial attention influences figure–ground assignment. *Psychological Science*, *15*, 20–26.

Vecera, S. P., Vogel, E. K., & Woodman, G. F. (2002). Lower region: A new cue for figure-ground assignment. *Journal of Experimental Psychology: General*, *131*, 194–205.

von der Malsburg, C. (1987). Synaptic plasticity as a basis of brain organization. In J. P. Chaneaux & M. Konishi (Eds), *The neural and molecular bases of learning* (pp. 411–432). New York, NY: Wiley.

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *Journal of Neuroscience*, *29*, 10573–10581.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107*, 829–854.

Warrington, E. K., & Taylor, A. M. (1973). The contribution of the right parietal lobe to object recognition. *Cortex*, *9*, 152–164.

Warrington, E. K., & Taylor, A. M. (1978). Two categorical stages of object recognition. *Perception*, *7*, 695–705.

Wertheimer, M. (1923/1950). Untersuchungen zur Lehre von der gestalt. *Psychologische Forschung*, *4*, 301–350.

Wilson, H. R., & Wilkinson, F. (1997). Evolving concepts of spatial channels in vision: From independence to nonlinear interactions. *Perception*, *26*, 939–960.

Yamane, S., Kaji, S. & Kawano, K. (1988). What facial features activate face neurons in the inferotemporal cortex of the monkey? *Experimental Brain Research*, *73*, 209–214.

Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z., & Connor, C. E. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature Neuroscience*, *11*, 1352–1360.

Queries in Chapter 7

Q1. Caption for Figure 7.1 is missing. Please provide.

Q2. Caption for Figure 7.2 is missing. Please provide.

Q3. Caption for Figure 7.3 is missing. Please provide.

Q4. References to color in label for Figure 7.3; can these be changed?

Q5. Caption for Figure 7.5 is missing. Please provide.

Q6. Please add Reference entry for Palmer & Brooks, 2008.

Q7. Please provide reference for Brooks, 2008".

Q8. Caption for Figure 7.13 is missing. Please provide.

Q9. Please provide reference for Tarr and Kriegman 2000".

Q10. Please add Minsky, 1975 to References.

Q11. Please provide reference for Minsky's 1975".