Chapter 26
Human Object Recognition: Appearance vs. Shape

Irving Biederman

26.1 Cortical Pathways for Visual Processing

High-resolution visual information is conveyed from the retina to the cortex via the lateral geniculate nucleus of the thalamus. The first cortical stage, V1, performs, essentially, a multiscale, multiorientation, Gabor filtering of the local contrast in the image through cells with small, local receptive fields (~0.5–2°). Activation is then fed forward through two major pathways. The ventral pathway, which we will focus on in this chapter, mediates recognition—how we know what we are looking at. This pathway extends from V1, through a series of stages, V2, V3, V4, and, in the macaque, the inferior temporal (IT) region (Fig. 26.1). The human homologue to IT—the final visual stage in the ventral pathway—in the macaque appears to be a region termed the lateral occipital complex (LOC), which will be discussed in detail below. A dorsal pathway mediates vision for purposes of motor interaction (or "how to") and extends from V1 (and V3 and V4) to the parietal cortex, which has extensive connections to the premotor cortex in the frontal lobe. This pathway specifies where objects are and their characteristics for, say, reaching and grasping. Another visual function is that of motion perception, with a region termed MT critical for its specification. MT is often considered to be a dorsal function but a case can be made that it be considered independently of its role in motor interaction.

DOI 10.1007/978-1-4471-5195-1_26, © Springer-Verlag London 2013
26.2 The Ventral Pathway

26.2.1 The Lateral Occipital Complex (LOC): An Area Critical for Perceiving Shape

LOC is operationally defined as the difference in the fMRI BOLD response to those regions of the cortex that show a heightened BOLD response to intact images of objects compared to their scrambled counterparts. In a typical study to localize LOC, subjects view 12 s blocks of images of 24 intact objects, with each image shown for 500 msec. After a 10 s blank period during which nothing is shown (to allow the BOLD signal to return to its resting state), the same objects can be shown, except that they are scrambled so that they resemble texture. (For another participant, the textured block would precede the intact block.) This procedure, when repeated for several cycles, yields a greater BOLD response in the lateral occipital cortex and the posterior portion of the fusiform gyrus (Fig. 26.2), which collectively comprise LOC [22]. LOC is functionally equivalent to the anterior inferior temporal area (IT) in the macaque in that it is considered the final stage in the ventral cortical pathway mediating shape recognition. Evidence for this equivalence derives, in part, from the similarity in the coding of object classes in IT and LOC. Kriegeskorte et al. [20] showed that the tuning of cells in IT of the macaque revealed the same similarity structure as fMRI-defined voxels in LOC. For example, if a cell responded strongly to a face it tended to respond strongly to other faces but not household appliances. A voxel could be found in human LOC that showed the same similarity relations.

That LOC is not just responsive to images of objects but is critical for their recognition is documented by studies of patient DF, who suffered bilateral lesions to LOC with sparing of other cortical areas, as a consequence of carbon monoxide poisoning at the age of 34 [23]. DF does not report seeing objects yet she shows normal motor interaction with them in that her hand normally conforms to an object prior to its grasping. Specifically, when picking up an object, the span of her prehension is accurately tuned to the position and width of the object that she is about to grasp, and her grasping points are optimal in that they pass through the center of mass of the object, in a manner that is virtually indistinguishable from normal control subjects [11]. She is normal or near normal in perceiving the color and texture of an object and whether it is moving or stationary.
DF cannot copy line drawings of simple common objects such as an apple or a house yet she can draw them reasonably well from memory, as we would if we had to draw the object in the dark. Her visual memory is excellent and she can process previously learned shape information in novel tasks, such as judging which capital letters of the alphabet have vertical lines on their left side and which have curved segments or whether particular objects are taller than they are wide. In all cases she is unable to visually recognize the letters or objects that she is judging.

It is possible to get the complementary deficit from a lesion in the parietal cortex that affects the dorsal pathway. Such individuals have no difficulty in recognizing objects but they are unable to achieve the fluid and efficient motor interactions of normal subjects. In picking up an object, they grope for it with open hand, the way we would when attempting to pick up an object in the dark.

26.2.2 Coding for Shape vs. Surface in LOC

Objects vary in shape, color, texture, and the luminance pattern over their surface depends on the direction of illumination. LOC represents the object as a line drawing that codes the orientation and depth discontinuities of an object—that is, its shape—to the exclusion of the surface features of the original image, such as color, texture, and direction of illumination. An fMRI experiment by Grill-Spector, Kourtzi, & Kanwisher [12] used an adaptation design in which subjects viewed a number of presentations of a sequence of two images. If the images were identical (Cup → Cup) then the BOLD response was smaller than if the images were different (Cup → Violin). The repetition of the identical images is said to produce adaptation and
the different images are said to produce a release from adaptation, indicating that the underlying neural representations of the two images differ. (Exactly why repetition of the identical image reduces the BOLD response, that is, adaptation, is still a subject of debate with non-exclusive interpretations including fatigue, narrowing of tuning, and competitive coding.) But what happens if, say, the first image is a photograph of a cup and the second is a line drawing of the same cup? Remarkably, the adaptation is unaffected, that is, there is no release of adaptation. Despite the drastic alteration of the image by the removal of its surface properties, the absence of an effect on adaptation indicates that the coding of LOC is independent of surface properties.

When we look at the tuning of single units in macaque IT, we can witness the same equivalence in the coding of line drawings and photographs. Given the preferences, i.e., spike rate, of a neuron over a set of colored photos of objects, that preference ordering is maintained over line drawings of the objects (e.g., [16, 18]). That is, if the cells fired at a higher rate to a color photograph of a chair over a lamp, that ordering was maintained for line drawings of those objects.

The representation of objects in terms of their orientation and depth discontinuities renders them invariant in terms of recognition performance to the direction of illumination [24]. This invariance to direction of illumination is also witnessed in the tuning of macaque cells in IT [29].

Although there is no question that humans can readily determine an object’s orientation and depth discontinuities, that is, they can produce an excellent line drawing by depicting those discontinuities, this capacity still remains a great challenge to the computer vision community. The design of a system for determining those discontinuities and distinguishing them from shadows, texture, surface markings, reflection highlights, etc. is still an unsolved problem. In fact, the major motivation for appearance-based approaches in the computer vision community may arise out of the inability to model how a line drawing can be extracted from an image of an object. This challenge is not restricted to those designing computer vision systems. Vision scientists do not know how line drawings of 3D shape are achieved by the visual system.

26.2.3 Different Subregions for the Processing Different Stimulus Dimensions

If LOC does not code for surface properties, where are they coded? Within the ventral pathway, different regions appear to be maximally activated when we attend to an object’s shape, color, or texture. Consider the unfamiliar shaped-objects that vary in shape, texture, and color in Fig. 26.3 [8]. Cant and Goodale had subjects view sequences of these objects and in separate blocks of trials, the subjects had to press a key if the object on the current trial matched the shape of the object in the prior trial. In such blocks, the color and texture could be ignored. In other blocks subjects judged if the prior object matched in texture and in still other blocks of trials.
Fig. 26.3 Stimuli from the Cant & Goodale [8] experiment that varied in shape, color, texture, and orientation. From *Cerebral Cortex*, v. 17 Fig. 26.2

In color (which was varied within textures, although not shown in Fig. 26.3), each time ignoring the other dimensions of stimulus variation. As noted previously, attention to shape maximally activated LOC. Attention to texture (or material properties) activated the collateral sulcus. Attention to color overlapped the collateral sulcus and LOC. One doesn’t have to give an explicit task to induce attention to a stimulus dimension. Merely passively viewing objects, such as those shown in Fig. 26.3, that vary in only one attribute, say texture, while shape and color are held constant, will be enough to differentially activate the collateral sulcus, the cortical area tuned to texture.

We should not be surprised to learn that there is independent coding of texture, color, and shape. Certainly we can look for our car in a parking lot or a garment in the laundry basket on the basis of its color, ignoring the variations in shape. It has been known for some time that people can demonstrate perfectly efficient selective attention to shape and ignore surface color or luminance and vice versa. For example, Fitts and Biederman [10] showed that the speed and accuracy in discriminating all black or all white circles and squares by human subjects was unaffected when
the shapes varied irrelevantly in luminance. The discrimination of luminance was similarly unaffected by irrelevant variation in shape. Biederman [1] demonstrated selective attention between dimensions of color, size, and orientation. Combinations of stimulus dimensions that can be selectively attended, such as luminance and shape, are said to be analyzable. In contrast, some dimensions, such as hue and saturation, cannot be efficiently selectively attended. Such dimensions are said to be integral. The neuroimaging work of Cant and Goodale suggest that efficient (i.e., fast and accurate) selective attention to a particular stimulus attribute in the presence of other varying attributes may be dependent on the coding these attributes at different cortical loci.

26.2.4 How Efficient are Line Drawings for Object Recognition?

Biederman and Ju [7] investigated the speed and accuracy of naming, in one experiment, and verification, in another, of briefly presented line drawings and photographs of common objects. In the naming tasks, an image with a common basic-level name is briefly presented followed by a mask and the subject has to name the object as quickly as possible. The time from the onset of the presentation to the onset of naming is measured. In a verification task, following the presentation of a target name, for example, “chair,” on half the trials a subsequently presented image matches the target and the subject is to press a key that indicates a match. In the other half of the trials, the image is not of a chair and the subject presses a non-match key. Both basic-level naming and verification were as fast and as accurate for line drawings as they were for color photography of the same instances. That is, the availability of surface cues—color, texture, luminosity gradients—added nothing to the recognition speed and accuracy of a line drawing of an object that specified well its orientation and depth discontinuities.

Some additional results confirmed the dominance of shape over surface for object classification/recognition. Some objects have diagnostic colors, such as a fish, a banana, or a fork, all familiar to human subjects. Other objects, such as a chair, pen, or a mitten do not. In both naming and verification the objects with the diagnostic colors showed the same equivalence with their line drawing counterparts as did the objects with non-diagnostic colors.

The equivalence of line drawings and colored photos only holds for concrete objects with definite boundaries. Linguistically, these tend to be count nouns for which, as the term implies, we can apply number and the indefinite article, so we can say three chairs or a chicken. Visual entities specified by mass nouns, such as sand or water, tend to be identified through their surface properties, for which we cannot apply number or the indefinite article. Perhaps predictably, the identification of objects dependent on their surface properties, including the case where an object’s characteristic shape is altered, as occurs with a balled up shirt in the laundry basket, is markedly slower and more error prone than object classes with well-defined discontinuities that can be conveyed by a line drawing [6].
26.3 What Aspects of Shape are Coded in LOC?

To pose the above question in a principled manner, we have to first understand what aspects of shape are apparent in object recognition behavior. There are six results (Table 26.1) motivated from Geon Theory (also termed Recognition-by-Components), Biederman's [2] account of human object recognition that have received strong behavioral and neural support but are not expressed by appearance-based theories of object recognition.

We have previously considered the sufficiency of a line-drawing representation (#2) Space precludes an extensive account of all six characteristics of human object recognition so I will just provide a brief overview of the evidence for a parts-based representation (#3) as well as the evidence for explicit coding of the relations between object parts (and objects) (#1).

26.3.1 Evidence for the Representation of Objects in Terms of Parts Rather than Local Features, Templates or Concepts

An object's shape is coded in terms of its simple parts—its geons—rather than templates of the whole object or the local features. There are considerable advantages in coding an object in terms of its parts. Under the most common causes of image variation, namely partial occlusion or rotation in depth, a template (such as the silhouette) can change markedly but only minimal costs in recognition are observed as long as two or three of the parts remain in view. If an object is missing a part, human subjects—even young children—can readily describe what it is that is missing.

It is important to note that the invariance to rotation in depth, as well as to changes in size, position, reflection, and direction of illumination can be witnessed with a "snap shot" brief presentation of a novel object [3]. That is, a single 100 msec view
of an object never seen previously is sufficient to achieve an invariant representation of that object so that the gain in speed and accuracy in identifying a subsequent presentation of the object is unaffected by a change in the viewing conditions. There is, however, an episodic memory, of the viewing parameters of the object so that we can remember where the object was, its size, and its orientation (e.g., [4]).

Biederman and Cooper [5] tested whether the representation was indeed part-based by employing complementary, contour-deleted line drawings of familiar objects (Fig. 26.4) to assess the presentation that mediated repetition priming of objects. Repetition priming, as the name implies, is the increase in the speed and accuracy of recognizing briefly presented, masked images of objects on their second presentation compared to their first. In their experiment, subjects named the objects, all of which had a common basic-level name, for example, “elephant,” “piano,” as quickly and as accurately as possible. In the first of two experiments, the contour deletion was of every other line and vertex of each part (Fig. 26.4) so that when both members of a complementary pair were combined they would make for the original intact object without any overlap in contour. Subjects named the objects in two blocks separated by approximately 10 minutes. The images in the second block could be classified into three conditions: (a) images that were identical to those on the first block, (b) complementary images which had the deleted contour of those for that object on the first block, and (c) a same name, different exemplar, such as an upright piano on the first block and a grand on the second.

The Different Exemplar condition was designed to rule out non-visual sources of facilitation, such as lexical access or basic-level priming. That is, if all the facilitation was just from repetition of the basic-level name or concept, then the Different
Exemplar condition should have been equivalent to the Identical condition. This was not the case. The different exemplar condition showed minimal facilitation from the first to the second presentation compared to the identical condition and most, if not all, that facilitation was because different exemplars of the same basic level class share some of the parts and relations, for example, airplanes will have wings emerging from their bodies and dogs will have four legs. Thus almost all the facilitation on this task could be considered visual priming and not conceptual or lexical priming. The remarkable result was the equivalence in the Identical and Complementary conditions. This indicates that none of the visual priming can be attributed to local features, as there was zero overlap of the contours and vertices between members of the same exemplar.

Although the results excluded that repetition of local features or basic level concepts or names accounted for the priming, the possibility remained that the priming was accounted for by repetition of subordinate-level concepts, that is, that it was the idea that one had seen a grand piano and not an upright piano or an elephant in a particular pose that mediated the priming. A complementary-priming experiment was then executed in which the complements were composed of different parts. Here complex objects were used each requiring at least six parts to look complete. The contour deletion was not of local features but of complete parts so that each member of a complementary pair had half the parts (and about half the total contour) of the other member. In this experiment there was absolutely no visual priming; performance in the Complementary condition was now equivalent to the Different Exemplar condition with the Identical condition showing a sizable advantage (lower RTs and error rates) compared to the Complementary Condition. These behavioral results have received confirmation in fast, event-related fMRI-adaptation experiments [13]. In these experiments, subjects viewed two images of contour-deleted objects, each presented for 300 msec separated by a blank frame. When the two images were identical, the BOLD response in LOC was minimal, an indicator of adaptation. When the images were of local feature-deleted complementary versions of the same objects, there also was no release from adaptation. However, when the two images were complements with different parts, there was a significant release of adaptation—a higher BOLD response—indicating that they were coded as different representations in LOC even though they were of the same exemplar. In brief, the fMRI-adaptation experiment was completely consistent with the behavioral priming experiment in indicating that objects are coded in terms of their parts rather than in their local features or subordinate level concepts.

26.3.2 Relations Between Parts and Between Objects

Just as different orderings of the same set of letters can form different words, different relations among the same set of parts can form different objects [2]. People have ready access to these relations and not only can describe them verbally but can quickly judge whether different novel objects composed of different geons, say, are
in the same or different arrangements in terms of their medial axis structure. Moreover, even when classifying which set of geons makes up a novel object while presumably, ignoring axis structure, the pattern of fMRI activation nevertheless shows sensitivity to the axis structure [19]. This neural sensitivity to the relations among object parts is also witnessed when viewing minimal scenes composed of different objects [14, 17]. When subjects view two frames of a novel scene, say a bus above a turtle, separated by a brief interval, there is a marked release from adaptation (greater BOLD response) if the second frame is that of a turtle above a bus, and virtually no release if the objects are simply translated by an equal extent but maintain the same relation. Appearance models posit that local features between the objects specify the relations but direct tests of such a proposal found no evidence for such coding [14]. A structural description representation explicitly specifying parts and relations can be argued to be a necessary prerequisite to support true understanding of visual structure. By rendering neither parts nor relations explicit, this prerequisite is not achieved by “bag of features” appearance models, e.g., [27].

To incorporate explicit relations into representations of objects and scenes, it is necessary to solve the binding problem, minimally, that the bus is ABOVE the turtle. So the relation of ABOVE is bound to bus and BELOW is bound to turtle. Although some have claimed that vision can solved without solution to the binding problem, e.g., [26], it is not clear that image understanding (vs. classification into familiar categories) can be so achieved, e.g., [15].

26.4 Coda

Probably the greatest challenge to machine vision efforts to achieve object recognition is the extraction of orientation and depth discontinuities and distinguishing such discontinuities from shadows, reflection edges, texture and color differences, etc. A secondary challenge is whether machine-based attempts at object recognition that do not achieve a structural description, as well as capturing the other characteristics listed in Table 26.1, can rival recognition performance readily evidenced by humans.

References