CHAPTER 24

Object Recognition: Physiological and Computational Insights

Doris Y. Tsao, Charles Cadieu and Margaret S. Livingstone

INTRODUCTION

Visual object recognition is the identification of a thing in the outside world based on the sense of vision. Our eyes are bombarded by a wide variety of visual forms, from simple shapes like cups and pens, to complicated shapes like keyboards and saxophones, to amorphous natural forms like flowers and bodies. These forms can appear at multiple poses as they rotate (or as we rotate), can change size as they move closer or farther, and can frequently occlude one another. Yet we have no trouble recognizing them. How does the brain transform the unpredictable retinal array into invariant representations of objects? This problem has two aspects: (1) extracting a stable, compact, and explicit representation of the forms stimulating our eyes at any moment, and (2) comparing the resulting representation to a stored representation in memory. We will focus in this chapter on the first aspect: How does the primate visual system extract stable representations of the objects in the visual world?

Visual object recognition is a function of the inferior temporal lobe, specifically, ventral stream areas V4, TEO, and TE. It has been suggested that the specialization of the inferior temporal lobe for object recognition arose during evolution as an elaboration of the foveal visual representation in V1, while the specialization of the parietal lobe for manipulation of objects in space arose as an outgrowth of the lower field representation of V1 (where the hands would

normally be located) (Maunsell & Van Essen, 1987; Previc, 1990).

Understanding the neural computations underlying object recognition is a uniquely difficult problem for three reasons. First, we don't even know how object recognition might be solved computationally. Even though numerous computer models exist, they don't come close to rivaling the performance of the human visual system. Second, a huge expanse of cortical territory is dedicated to object recognition. Areas TEO and TE, for example, are more than 10 times as large as the much-studied motion area MT (Felleman & Van Essen, 1991). Assuming that multiple levels of organization exist within this vast cortical expanse, it becomes exceedingly difficult to map specific computations onto specific sets of neurons. Thus, single-unit physiologists typically *assume* that the organization of inferotemporal (IT) cortex is homogeneous and that meaningful answers can be obtained from sampling responses of random populations of IT neurons. Finally, the set of possible objects covers a huge parameter space. While it is clear the brain must use some sort of code that is compact, explicit, and stable to identify objects within this space (Connor et al., 2007), it is unclear what this code might be. To summarize, understanding object recognition is difficult due to a lack of effective computer models, the large number of cortical resources used to solve the problem, and a huge parameter space.

Figure 24.1 The challenge of object recognition, exemplified for a face.

Over the past three decades, knowledge has gradually accumulated concerning a system in the temporal lobe that may allow us to overcome these challenges. Both macaques and humans have a specialized system in the temporal lobe for processing faces that produces an extraordinary ability to recognize faces under a variety of conditions (Pinsk et al., 2005; Tsao et al., 2003, 2006, 2008). (Fig. 24.1). This face processing system provides a spatially limited network with readily accessible components representing a parametrically confined set of objects.

Face perception is a microcosm of object recognition processes, and the solution to the particular problem of face recognition will undoubtedly yield insights into the general problem of object recognition. The most difficult challenge in object recognition—distinguishing among similar visual forms despite substantial changes in appearance arising from changes in position, illumination, occlusion, etc.—is something we can do effortlessly for faces (Fig. 24.1). Although face identification is often singled out as demanding particular sensitivity to differences between objects sharing a common basic configuration, in fact, such differences must be represented in the brain for both faces and nonface objects. It is true that most humans can easily identify hundreds of faces (Diamond & Carey, 1986), but, even if one cannot recognize a hundred different bottles by name, one can certainly distinguish them in pairwise discrimination tasks.

In this chapter, we first discuss the functional architecture of the temporal lobe, with a special focus on the architecture of the system of face-selective areas in macaques and humans. We then discuss the physiology of cells in the temporal lobe, with a focus on the response properties of face-selective cells. Finally, we discuss

different computational approaches to object recognition. The central thesis of this chapter is that understanding face processing will illuminate the general problem of visual object recognition.

Functional Architecture of the Inferior Temporal Lobe

The functional architecture of the temporal lobe sets the stage for the neural processes underlying object recognition.

General Architecture of the Macaque Ventral Visual Pathway

Figure 24.2 shows a schematic of the lateral and ventral surfaces of the macaque brain. The three major areas of the ventral form processing pathway are V4, TEO, and TE. Area TE is further subdivided into four parts on the basis of anatomical connection criteria: TEpd (dorso-posterior), TEpv (ventro-posterior), TEad (dorso-anterior), and TEav (ventro-anterior) (Cheng et al., 1997; Saleem & Tanaka, 1996; Yukie & Iwai, 1988). Monkeys with bilateral lesions to V4, TEO, and TE show severe and specific deficits in object recognition tasks (Dean, 1976; Gross, 1973). Both V4 and TEO are retinotopically organized, with TEO containing a coarse but complete representation of the contralateral visual field (Boussaoud et al., 1991); cells in TE have large receptive fields centered on the fovea (Ito et al., 1995; Kobatake & Tanaka, 1994).

The architecture of the inferior temporal lobe has been studied by classic anatomical tracing. V4 sends strong projections to TEO from its central visual field representation (Ungerleider et al., 2008). TEO in turn sends strong feedforward projections to area TE (Saleem et al., 1993).

Figure 24.2 The ventral visual processing pathway. (A) Lateral view of the right hemisphere of the macaque brain, showing the architectonic subdivisions of inferior temporal cortex. The superior temporal sulcus (STS) is opened to show the dorsal bank (d), fundus (f), and ventral bank (v). (B) Ventral view of subdivisions of the inferior temporal cortex. ls, lateral sulcus; STGr, superior temporal gyrus, rostral part; TEad, dorsal subregion of anterior TE; TEav, ventral subregion of anterior TE; TEpd, dorsal subregion of posterior TE; TEpv, ventral subregion of posterior TE; TG, temporal pole; TH, area TH of parahippocampal cortex; TF, area TF of the parahippocampal cortex; TFO, area TFO of the parahippocampal cortex; 28, entorhinal cortex; 36, perirhinal cortex. From Saleem, K. S., Kondo, H., & Price, J. L. (2008). Complementary circuits connecting the orbital and medial prefrontal networks with the temporal, insular, and opercular cortex in the macaque monkey. *Journal of Comparative Neurology*, 506, 659–693. Used with permission.

The projection from TEO to TE shows a precise modularity, with single TEO sites projecting to two to five clusters in TE (Saleem et al., 1993) (Fig. 24.3A). These clusters are columnar, extending across all six cortical layers. Although the functional properties of the TEO injection sites and connected TE columns are unknown, it seems plausible that this precise anatomical circuitry serves a computational purpose. One interesting possibility is that each network of connected sites in TEO and TE is responsible for extracting a specific aspect or class of visual form.

Area TE, the highest purely visual stage of the ventral pathway, sends feedback projections to V4 and TEO, and feedforward projections to several polymodal brain sites including the perirhinal cortex, the frontal cortex, the amygdala, and the striatum (Cheng et al., 1997; Saleem & Tanaka, 1996; Suzuki et al., 2000; Webster et al., 1991, 1993, 1994). The perirhinal projections from TEav and TEad differ, suggesting that these areas constitute distinct processing regions in TE. Focal TEav injections result in a widely distributed labeling in perirhinal cortex, covering around half its total extent, whereas

Figure 24.3 Columnar projections in inferior temporal (IT) cortex. (A) Labeled terminals in TE following injections of the anterograde tracer *Phaseolus vulgaris* leucoagglutinin (PHA) in TEO (indicated by two red spots) form columnar clusters. From Saleem, K. S., Tanaka, K., & Rockland, K. S. (1993). Specific and columnar projection from area TEO to TE in the macaque inferotemporal cortex. *Cerebral Cortex*, 3, 454–464. (B) Labeled terminals of intrinsic horizontal axons show patchy columnar arborization following injection of biocytin into layer 3 in area TE. The injection site is marked by the red asterisk and arrowheads indicate terminal patches. From Fujita, I., & Fujita, T. (1996). Intrinsic connections in the macaque inferior temporal cortex. *Journal of Comparative Neurology*, 368, 467–486. Used with permission.

injections into TEad result in labeling in only a small region of lateral perirhinal cortex (Saleem & Tanaka, 1996). Perirhinal cortex has been implicated in encoding long-term memory of objects (Meunier et al., 1993) and association between familiar objects (Erickson & Desimone, 1999; Erickson et al., 2000; Miyashita et al., 1996). The widespread projections from TEav to perirhinal could thus facilitate association between different shapes (e.g., different views of the same object).

In addition to projections to other areas, both areas TEO and TE have local connections. Injections into TE reveal horizontal axons in layers 2 and 3 terminating in patches 0.5 mm wide and cylindrical in shape, spanning layers 1 through 3 or even to layers 4 and 5, with the farthest patches up to 4 mm distant from the

injection site (Fujita & Fujita, 1996) (Fig. 24.3B). TEO shows similar, but slightly smaller, columns of connected patches, with the patches slightly smaller. Again, it seems plausible that each network of locally connected patches forms a system specialized to represent a particular aspect of form.

Electrophysiology and optical imaging studies also suggest a columnar functional architecture in TE. Neighboring cells are responsive to similar visual features of objects (Fujita et al., 1992). Optical imaging reveals that spots ≈ 0.5 mm wide in TE are activated when a monkey views a particular object (Tsunoda et al., 2001; Wang et al., 1996), with different spots activated by different objects. The cortical regions activated by three different objects are shown in Figure 24.4A. Single-unit recordings confirmed

Figure 24.4 Distributed object representations in macaques and humans. (A) Top: Spots activated by different stimuli (1, 2, and 3, right) as revealed by optical imaging, together with numbered electrode sites. Bottom: Responses to the three stimuli at each of 17 electrode sites. Scale bar, 1.0 mm. Tsunoda, K., Yamane, Y., Nishizaki, M., & Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience*, 4, 832–838. (B) Example images from the eight stimulus categories used by Haxby and colleagues (2001) to examine object representations in the human temporal lobe. (C) The pattern of responses to four categories (faces, houses, chairs, shoes) measured separately from data obtained on even-numbered and odd-numbered (two horizontal slices from a single subject are shown). Within-category correlations (0.81, 0.87, 0.45, 0.35) are significantly higher than cross category comparisons (–0.4, –0.47, –0.12, –0.1), allowing prediction of the viewed object category. B and C from Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293, 2425–2430. Used with permission.

the responsiveness of cells within these spots to the particular objects used to activate them (and absence of responsiveness outside these spots). For example, an apple activated spots 1, 2, and 3, and single units in these three sites also responded strongly to the apple. Since multiple spots, distributed over several millimeters of cortex, were active in response to a single image, and since the size of these spots is similar to the size of termination zones observed in tracer experiments (Fig. 24.3), it is conceivable that these spots correspond to connected columns.

Distributed Architecture of the Human Ventral Pathway

Human functional magnetic resonance imaging (fMRI) studies provide a much coarser picture of functional architecture of ventral temporal cortex, but are consistent with results from macaques. The wide field of view of fMRI makes it possible to address the question of how distributed the representation of an object is. Haxby and colleagues (2001) measured response patterns in ventral temporal cortex evoked by viewing of faces, cats, five categories of manmade objects, and nonsense pictures (Fig. 24.4B). They found that object categories are represented by widely distributed and overlapping representations in ventral temporal cortex. Each category elicited a distinct and reproducible distributed response pattern that could be used to identify which category was seen (Fig. 24.4C).

The distinctiveness of the response patterns is apparent even at a scale of centimeters (Fig. 24.4C, compare response to chairs with response to shoes). This suggests that the structure in the response patterns reflects a level of functional organization more macroscopic than 0.5-mm columns. While multivoxel readout techniques might be capable of distinguishing response patterns that differ only at a scale of 0.5-mm columns (by pooling minute but consistent category differences across a large number of voxels) (Norman et al., 2006), such functional distinctions should not be readily apparent using a technique with a resolution on the scale of

centimeters (Boynton, 2005; Kamitani & Tong, 2005). Therefore, the distinct fMRI response patterns to different objects observed by Haxby and colleagues may instead arise from a coarser scale functional organization, possibly, networks of connected, clustered columns.

Architecture of the Face Processing System in Humans

The existence of a face-selective area in the human brain was first suggested by neurological observations. That there is a specialized area for processing upright faces fits with one of the most striking findings from the neuropsychology literature: Patient C.K., who is severely impaired at object recognition, including many basic midlevel visual processes, is nonetheless 100% normal at face recognition (Moscovitch et al., 1997). C.K.'s dissociation is illustrated by his perception of the face made up of vegetables by Arcimbaldo—C.K. sees the face, but not the constituent vegetables. His pattern of deficits indicates that face processing is not simply a final stage tacked onto the end of the nonface object recognition pathway, but rather a different pathway that branches away from object recognition early in the visual hierarchy.

The first direct demonstration of face-selective activation in a human brain area came from positron emission tomography (PET) studies showing activation of the fusiform gyrus in a variety of face perception tasks (Haxby et al., 1991; Sergent et al., 1992) and event-related potential (ERP) studies showing a face-selective event-related potential in the fusiform gyrus (Allison et al., 1994; McCarthy et al., 1999; Puce et al., 1999). Subsequently fMRI revealed more of the specificity of these cortical regions for faces, with demonstrations of fusiform regions that responded more strongly to faces than to letter strings and textures (Puce et al., 1996), flowers (McCarthy et al., 1997), everyday objects, houses, and hands (Kanwisher et al., 1997). Face-specific fMRI activation can be seen in the superior temporal sulcus (STS), in part of the occipital lobe (the “occipital face area,” or OFA), and most robustly, on the lateral side of the right midfusiform gyrus (the “fusiform face

2008), who found that in monkeys, as in humans, face processing, as revealed by functional imaging, is localized to six discrete regions of the temporal lobe (Fig. 24.5B). These six patches are distributed along the anterior-posterior axis of the temporal lobe and organized into one posterior patch on the lateral surface of TEO ("PL," for posterior lateral); two middle face patches in posterior TE, one located in the fundus of the STS ("MF," for middle fundus) and one on the lower lip of the STS ("ML," for middle lateral); and three patches in anterior TE, one located near the fundus of the STS ("AF," for anterior fundus), one on the lower lip of the STS and adjacent gyrus, in TEad ("AL," for anterior lateral), and one more medially on the ventral surface, just lateral and anterior to the anterior-medial temporal sulcus, in TEav ("AM," for anterior medial).

The face-patch system presents us with a new kind of functional organization in TE. The components of this network are coarser (few millimeters in diameter) than feature columns of inferotemporal cortex (≈ 0.5 mm in diameter; Fujita et al., 1992; Wang et al., 1996), yet finer in scale than the coarse partitioning of IT into anatomically defined subregions (Felleman & Van Essen, 1991; Seltzer & Pandya, 1994; Von Bonin & Bailey, 1947). The face-patch system transgresses area boundaries, with face patches located in posterior, middle, and anterior portions of IT (Felleman & Van Essen, 1991).

Tsao and colleagues (2006) further showed that at least two of the macaque face areas, ML and MF, consist almost entirely of face-selective cells (Fig. 24.6). Tsao and colleagues used functional imaging to localize regions in the macaque temporal lobe that were selectively activated by faces, compared to nonface objects, and then they recorded almost 500 single units within ML and MF in two monkeys. They found a remarkable degree of face selectivity within these two regions, with 97% of visually responsive cells being face selective, on average showing almost 20-fold larger responses to faces than to nonface objects. The region where they recorded was quite posterior in the temporal lobe (6 mm anterior to

the interaural canal, corresponding to posterior TE/anterior TEO). The fact that an area consisting almost entirely of face-selective cells exists so early in the ventral stream is consistent with the idea that the face processing pathway is a completely different pathway from the nonface object recognition pathway, and that it branches away early in the visual hierarchy. The existence of an area consisting almost entirely of face-selective cells so early in the ventral stream furthermore implies that the face processing pathway is gated by a face detection stage, at which nonface objects are filtered out. As we propose later, the existence of this detection gate may account for the seemingly special "holistic" aspect of face processing.

What is the functional significance of the anatomical localization of face processing? The cerebral cortex is functionally parcellated: Neurons concerned with similar things are organized into areas and columns, each having extensive interconnections and common inputs and outputs (Mountcastle, 1997). Face processing, an identifiable and discrete form of object recognition, appears also to be organized into anatomically discrete processing centers. Individual neurons connect directly with a small fraction of the rest of the neurons in the brain, usually to nearby cells, because longer axons delay neural transmission, are energetically expensive, and take up space. Furthermore, colocalization of neurons concerned with face processing enables enriched local inhibitory interactions, since inhibitory neurons are always local (Somogyi et al., 1998). Wang and colleagues (2000) recorded responses in anterior IT to a set of complex stimuli before, during, and after applying the γ -aminobutyric acid (GABA) receptor antagonist bicuculline near the recording electrode. In many cases, for both face-selective and non-face-selective cells, blocking local inhibition revealed responses to previously nonactivating stimuli, which were often activating stimuli for neighboring cells. This suggests that neighboring cells refine each other's response selectivity by mutual inhibition.

Figure 24.6 Mapping face and object selectivity in the monkey brain. (A) Five stimulus categories included faces, four nonface object categories (hands, gadgets, fruits, and bodies), and grid scrambled patterns. (B) Responses of 182 neurons from the middle face patch of one monkey to 96 images of faces and nonface objects. (C) Average normalized population response to each image. Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, 311, 670–674. Used with permission.

While tracer studies demonstrate connections between specific nodes in TEO and TE (Saleem et al., 1993), the need to sacrifice the animal to process the tissue prevents assessment of the functional properties of connected nodes. In general, to understand functional architecture, it is necessary to combine connectivity maps with functional topography. The face processing system of macaque monkeys, consisting of six patches that can be identified by fMRI and targeted for anatomical experiments, provides an ideal preparation for dissecting the large-

scale functional anatomy of one discrete aspect of object recognition.

To identify the connectivity of individual face patches, Moeller and colleagues (2008) used electrical microstimulation combined with simultaneous fMRI. Stimulation of each of four targeted face patches produced strong activation specifically within a subset of the other face patches (Fig. 24.7). Stimulation outside the face patches produced an activation pattern that spared the face patches. These results indicate that the

Figure 24.7 Combined functional magnetic resonance imaging and microstimulation demonstrates that the six macaque face patches are strongly and specifically interconnected. (A) Face patches from one macaque shown on a right hemisphere flat map. (B) Regions activated by microstimulation at the region marked by an $\times$ compared to blank. Face patches indicated by green outlines. The animal fixated on a blank screen during the entire experiment. Face region and sulcal abbreviations as in Figure 24.5B. Moeller, S., Freiwald, W. A., & Tsao, D. Y. (2008). Patches with links: a unified system for processing faces in the macaque temporal lobe. *Science*. In press. Used with permission.

AQ2

face patches form a strongly and specifically interconnected network.

The existence of this network creates unprecedented possibilities for gaining a systematic understanding of information flow in IT. For the first time we can access anatomically distinct components of a unified object processing network in vivo and intercept the messages being sent between connected cell pairs within different patches. Thus, in addition to characterizing in detail the properties of cells within each patch, we can sequentially characterize the transformations that occur across patches—transformations that somehow lead to our ability to recognize thousands of different faces effortlessly even under incredible variation (i.e., Fig. 24.1).

Interim Summary

Nonface objects are represented by widely distributed and overlapping representations in ventral temporal cortex. Anatomical tracing, optical imaging, and single-unit studies in macaques suggest

that the neural machinery supporting these non-face object representations has a precise circuitry. The unit element of this circuitry is a column ≈ 0.5 mm wide. Specific subsets of columns within TEO and TE are organized into connected networks. Alongside the machinery for recognizing nonface objects, the temporal lobe of both humans and macaques contains a system of strongly face-selective regions. In macaques, this system comprises six patches of cortex extending from TEO to anterior TE and, in humans, four areas spanning the length of the temporal lobe. The macaque face patches are strongly and specifically connected to each other, and at least two of the macaque face patches consist almost entirely of face-selective cells.

REPRESENTATIONS IN THE INFERIOR TEMPORAL LOBE

In this section we discuss the physiology of cells in IT that support the remarkable capacity to recognize objects. We present a review of the

findings derived from visual stimulation using parametric visual features and general objects. We will also present results of electrophysiological studies of face representations in visual cortex. Finally, we address the issue of holistic facial processing and its relation to electrophysiological findings.

Representation of Objects

Tanaka and colleagues pioneered the study of tuning properties of single cells in IT (Tanaka, 1996; Tanaka et al., 1991). Their work was converged on the concept of a “critical feature,” defined as the simplest feature that elicits the maximum response in a cell. To identify the critical feature of an IT cell, they first showed many three-dimensional plant and animal models to the cell at different views, then made pictures of these objects and systematically simplified them until they were able to identify the minimal feature that elicited maximal activation (Kobatake & Tanaka, 1994). For example, the critical feature of a cell preferring a tiger might be two round black and white gratings arranged like a snowman (Tanaka, 2003). The main conclusion was that most cells in TE are selective for moderately complex feature combinations but not for whole objects. Such cells are already present in V4 and TEO, but at a lower proportion, mixed with cells selective for simple features (such as color or orientation).

Subsequent attempts to decipher the code used by the ventral stream have followed two general lines of attack: (1) a bottom-up approach, starting with a model of structural encoding, and testing the model, and (2) a black box approach, using responses of a random population of cells to various complex objects to derive general principles of IT coding. These two approaches represent two different ways of conquering the huge parameter space occupied by the set of all possible visual inputs. The structural encoding approach assumes that the brain represents most objects by decomposing them into parts and part relations. The black box approach, on the other hand, does not make any concrete assumptions about how objects might be coded, but it does make the

assumption that experimenter-chosen real-world objects will span enough of the coding space in sufficient detail that the approach will yield meaningful insights.

A parts-based representation has intuitive appeal. When asked to describe an object (e.g., a flower), people naturally describe its parts (e.g., petals, leaves, stem). The distributed, combinatorial code created by a parts-based representation permits an extremely rich set of objects to be represented. A parts-based representation of words doubtless in fact underlies your ability to read this page: (1) the geometry of letters may take advantage of the brain’s hardwired ability to distinguish contour parts (Changizi et al., 2006), through their location (“p” vs. “b”), curvature (“v” vs. “u”), and connectivity (“t” vs. “T”), and (2) the possibility of forming infinite meanings from 26 characters illustrates the representational power of a parts-based structural code.

Connor and colleagues have uncovered evidence for parts-based coding in V4 (Pasupathy & Connor, 1999, 2001, 2002) and PIT (Brincat & Connor, 2004, 2006; Haxby et al., 2001) through single-unit recordings examining tuning to parametrically defined contour features. Pasupathy and Connor (2001) created a large set of closed contours that could be parameterized by curvature and angular position (Fig. 24.8A), and then analyzed the responses of V4 cells to these shapes. They found that most cells in V4 were tuned to a particular curvature at a particular location within the cell’s receptive field (e.g., a sharp convexity in the lower left). Furthermore, they showed that by combining these tuning curves, they could reconstruct (approximately) an unknown test shape from the pattern of responses in V4 elicited by the shape (Fig. 24.8B,C) (Pasupathy & Connor, 2002). Extending this work to IT cortex, Connor and colleagues found cells tuned for the presence of multiple parts (specifically, tuning could be described by the sum of two to four subunits with Gaussian tuning in a six-dimensional curvature, orientation, and position space) (Brincat & Connor, 2004). Figure 24.8D illustrates a cell tuned to a combination of concave contour fragments oriented toward the lower right and toward the lower left, and located in the bottom right of the

Figure 24.8 Encoding of contour curvature in V4 and PIT. (A) The white line represents the boundary curvature as a function of angular position of the test shape in the center. (B) Estimated population response across the curvature $\times$ position domain to a test shape with the veridical curvature function superimposed (white line). The response was estimated by weighting each cell's curvature $\times$ position tuning function by its response to the test shape. (C) Reconstruction of the test shape from the population response in B. (D) Response of a posterior inferotemporal cell to its preferred feature: a combination of concave contour fragments oriented toward the lower right and toward the lower left (highlighted in green). The cell was highly sensitive to the position of these contour elements relative to the object. It responded strongly to concavities at the left of object center (bright green), as exemplified by the average response (gray histogram) to the stimuli in the top two rows. The light gray shading indicates the 500-ms stimulus-presentation period. The cell responded only weakly to the same concavities at the right (dark green; bottom two rows). This tuning for object-relative position was consistent across changes in absolute position of the stimuli. Thus, cells in PIT are coding the curvature of contour fragments in object-centered coordinates. A–C from Pasupathy, A., & Connor, C. E. (2002). Population coding of shape in area V4. *Nature Neuroscience*, 5, 1332–1338. D from Connor, C. E., Brincat, S. L., & Pasupathy, A. (2007). Transformation of shape information in the ventral pathway. *Current Opinions in Neurobiology*, 17, 140–147. Used with permission.

Figure 24.8 (Continued)

object. These experiments suggest that object recognition relies to a significant extent on a parts-based code implemented by tuning curves of single neurons. One criticism of this work is that the space of real shapes is much vaster than that tested. But any study of IT must employ an experimenter-limited stimulus set; the finding that cells in both V4 and TEO respond to parts of objects and not to the whole object (albeit in an object-centered reference frame) must be explained by any theory of object representation in the ventral stream.

A structural encoding scheme for representing three-dimensional objects was proposed by Irving Biederman (Biederman, 1995). In the “geon theory” of object recognition, a given view of an object is represented as an arrangement of simple, viewpoint-invariant, volumetric primitives called “geons.” Five examples of geons are shown in Figure 24.9A. According to the geon theory, an object is represented by the geons it contains (out of 24 total), together with pairwise geon relations (e.g., “above”; 81 total) and geon attributes (e.g., “horizontally oriented, narrow relative aspect ratio”; 15 total). These limited parameters could represent 10.5 million different two-geon objects.

The geon theory is supported by experiments with partial line drawings. The effect of priming

with complementary contours that preserve geon structure is just as strong as the effect of priming with identical contours, and significantly stronger than the effect of priming with an abstract object category (Biederman & Cooper, 1991) (Fig. 24.9C). Thus, priming must be attributed to a representation of the parts of the object (and their interrelations) and not to the activation of the image features or abstract categories.

The defining feature of geons is their view invariance. Metric properties of objects (e.g., aspect ratio), in contrast, are not viewpoint invariant. The geon theory predicts that responses to objects differing in geon structure should be easier to distinguish, across changes in view, than responses to objects differing only in metric properties. To test this, Vogels and colleagues (2001) recorded responses of IT cells to six variations of an object: the original object composed of a pair of geons, a “metric property” variant in which one of the geons was metrically varied (e.g., made wider), and a “non-accidental property” variant in which one of the geons was changed (e.g., a cylinder replaced by a block). Rotated versions of these three images were also constructed. The two variants were equated in terms of low-level image statistics. Their results provide partial support of the

geon theory. A population multidimensional scaling (MDS) analysis revealed two dimensions, one corresponding to rotation, and one to geon change (Fig. 24.10D). The population response to the metric variant was very close to that of the original image (i.e., neurons had trouble distinguishing them), while the response to the geon variant was well removed. This supports the idea that cells in IT have a special sensitivity to differences in geon structure over differences in metric structure. However, the large difference between the population response to each object and its rotated counterpart shows that the representation of object shape in IT is *not* view invariant.

The last finding is consistent with an alternative view that the primitives for object recognition in IT are view based (Bulthoff et al., 1995). Experimental evidence for this theory comes

from Logothetis and colleagues, who trained monkeys to match three-dimensional paperclips across changes in view. Subsequently, they recorded in anterior IT and found that many cells responded maximally to a particular view of a paperclip, with the response declining gradually as the object was rotated away from this preferred view. A very small number of cells were found that responded in a view-invariant manner.

While a large number of objects can be represented by parts-based parametric schemes such as curved contours (Pasupathy & Connor, 2001) or geons (Biederman, 1987), there are still many objects that do not fit easily into such parametric shape schemes, yet may be biologically important to distinguish. Thus, a second approach to studying IT cortex has

Figure 24.9 Structural encoding by geons. (A) Examples of five geons. (B) A block and a cylinder, two examples of geons, show particular nonaccidental differences that may be used by the brain to recover their structure from line drawings. (C) Examples of the four stimulus types used by Biederman and Cooper to demonstrate the existence of a geon-level encoding in a psychophysical priming experiment. (D) Neurophysiological evidence for representation of nonaccidental properties. Top: Six classes of images used by Vogels and colleagues to test for a geon representation in macaque IT. Bottom: Multidimensional scaling of inferotemporal responses to these image classes reveals a dimension representing the difference between a view-dependent change and a viewpoint-invariant change (dimension 1), and a dimension representing rotation (dimension 2). A–C from Biederman, I. (1995). Visual object recognition. In: S. M. Kosslyn & D. N. Osherson (Eds.), *Visual cognition* (pp. 121–166). Cambridge, MA: MIT Press. D from Vogels, R., Biederman, I., Bar, M., & Lorincz, A. (2001). Inferior temporal neurons show greater sensitivity to nonaccidental than to metric shape differences. *Journal of Cognitive Neuroscience*, 13, 444–453. Used with permission.

Figure 24.9 (Continued)

been to stimulate cells with random sets of complex images and ask, by means of various readout procedures, what types of information are available in the population of responses (e.g., how fast can information be read out, how many neurons are needed to code certain types of information, etc.). Obviously, a null result does not say much, since the readout algorithm used may simply be inadequate. On the other hand, a positive result from such an approach does reveal what kind of information is present. It should be noted that readout techniques can also be used to examine structural encoding hypotheses—for example, one could use classifier techniques to ask whether information about curvature can be read out from cells in a position-invariant way. The main difference lies in whether the stimulus space consists of parametric combinations of parts or randomly selected real-world objects.

Kiani and colleagues (2007) measured responses of more than 600 neurons in monkey IT cortex to over a thousand different images of natural and manmade objects during passive fixation. Separation of the images into clusters based on response similarity across the population of cells showed that the neuronal population sorted the images into intuitive categories: Animate and inanimate objects created the most distant distinguishable clusters in the population code. The global category of animate objects was divided into bodies, hands, and faces. Faces were divided into primate and nonprimate faces, and the primate-face group was divided into human and monkey faces. Bodies of human, birds, and four-limb animals clustered together, whereas lower animals such as fish, reptiles, and insects made another cluster. Importantly, low-level image

Figure 24.10 Holistic face detection by an inferotemporal cell. Top: Recording site of a face cell. (A–H) Response selectivity. From 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. Used with permission.

characteristics could not account for this categorical classification. Therefore, they concluded that monkey IT specifically extracts complex features for the purpose of object categorization.

Hung and colleagues (2005) used a classifier-based readout technique to analyze responses of 256 IT recording sites to 77 images. They found that the activity of ≈ 100 IT neurons over very short time intervals (as small as 12.5 ms

contained sufficient information to read out object "identity" at 72% (chance = 1/77) and object "category" at 94% (chance = 1/8) accuracy. Importantly, this information generalized over a range of object positions and scales, with less than 10% reduction in performance. These results show that invariant information about object category and identity is available in small populations of neurons even during an early phase of the response. Can object recognition be completely solved by means of units with the invariance and tuning properties of randomly sampled IT cells, as already described in studies such as that of Hung and colleagues? Or are we still waiting to observe a new type of invariance property? Rigorous tests comparing physiology to behavior have not been performed to test this important question.

Representation of Faces

One of the biggest differences between face perception and general object recognition is that arbitrary objects map to different aspects of contour geometry. Faces, on the other hand, share a common template, consisting of eyes, nose, mouth, and face outline (though numerous variations are possible; e.g., depending on the view angle, only one eye may be visible). The vastly reduced template space makes understanding the mechanism of face perception more tractable compared to understanding the detection and recognition of arbitrary objects.

That face cells are truly detecting faces, and not some more abstract basis set in which all possible shapes are represented by different cells, with some cells tuned to particular parameters that happen to fit the faces better than any of the other objects tested, was demonstrated by Foldiak and colleagues (2004). They presented 600 to 1,200 stimuli randomly chosen from several image archives to cells recorded from both the upper and lower bank of the STS and found that the distribution of tuning to these images showed bimodality (i.e., cells were either predominantly face selective or not face selective).

Experiments with cartoon faces show that in general face cells require an intact face and are

not just selective for individual features (Bruce et al., 1981; Desimone et al., 1984; Kobatake & Tanaka, 1994; Leonard et al., 1985; Oram & Perrett, 1992; Perrett et al., 1982, 1984; Scaldidhe et al., 1999; Tsao et al., 2006). Figure 24.10 shows nonlinear combinatorial response properties of a face-selective cell recorded in IT (most likely in the face patch ML, judging by the recording location) by Kobatake and Tanaka (1994). Out of a large number of three-dimensional objects, this cell responded best to the face of a toy monkey (A), and by testing various simplified two-dimensional paper stimuli, they determined that the cell would also respond to a configuration of two black dots over a horizontal line within a disk (B), but not in the absence of either the spots or the line (C and D) or the circular outline (E). The contrast between the inside and the outside of the circle was not critical (G), but the spots and the bar had to be darker than the disk (H). In other words, the cell only responded when the stimulus looked like a face, no matter how simplified. How face cells detect a facial Gestalt is still unknown, and will likely require identifying the inputs to the face patches.

Once a face has been detected, it needs to be identified. Recordings in the face patches ML and MF have begun to shed light on the neural mechanism for distinguishing between different faces. Freiwald and colleagues (2009) used cartoon faces to study the neural basis of face measurement. Dense parametric mapping was used to measure responses of cells in the middle face patch to a cartoon stimulus in which all face parameters were independently varied. Cells were found to be tuned to the geometry of facial features, with different cells tuned to different feature subsets. Tuning was strikingly ramp shaped, with a one-to-one mapping of feature magnitude to firing rate (Fig. 24.11). These extreme values extended to or even transgressed the limits of realistic face space. For example, intereye distances ranged from almost cyclopean to abutting the edges of the face, and the most extreme face aspect ratios were outside those of any known primate. Monotonic tuning allows for simple readout (Guigon, 2003) and may be a general principle for high-level coding

Figure 24.11 Tuning of face cells to a cartoon face space. (A) Six example dimensions (hair width, eyebrow slant, eyebrow height, intereye distance, iris size, mouth top shape) of the 19-dimensional cartoon space. Each row shows example values for one parameter, with all other parameters fixed at their mean. (B) Tuning curves of three example cells to each of the 19 feature dimensions. In gray maximal, minimal, and mean values from shift predictor are shown. Stars mark significant modulation. From Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2008). A face feature space in the macaque temporal lobe. Under review . Used with permission.

of visual shapes (Kayaert et al., 2005). It may also aid in emphasizing what makes an individual face unique (i.e., separates it from the average face), and could be the neural basis for the power of caricatures. The breadth of tuning underscores the fact that cells in the middle face patches, ML and MF, encode axes and not individual faces.

Recordings made in the highest stage of the form processing hierarchy, the medial temporal lobe, reveal the existence of cells that respond to specific familiar individuals in a representation-invariant manner (Quiroga et al., 2005), as expected of a “grandmother cell.” For example, some cells each responded to multiple pictures of a well-known individual as well as to a letter string of their name, but were unresponsive to all other images. Such individual-specific cells have not been found in the lateral inferior temporal lobe, where most face cells in monkeys have been recorded, although, as a population, cells in the anterior inferior temporal gyrus of the macaque (in the general vicinity of AM) can support view-invariant identification (Eifuku et al., 2004).

Holistic Processing

Since the thesis of this chapter is that understanding face areas will help us understand object recognition, it is appropriate to discuss the claim that face processing is supported by a unique set of mechanisms that do not generalize to nonface objects. Face processing obviously is unique in being housed in a set of specialized cortical regions, but this doesn’t imply that the underlying mechanisms are unique.

Specifically, face processing is claimed to be distinct from nonface object processing in that it is “holistic”; that is, faces are represented as nondecomposed wholes, rather than as a combination of independently represented component parts (eyes, nose, mouth) and the relations between them (Farah et al., 1998). Evidence for holistic processing of faces comes from a number of behavioral paradigms, of which the two most cited are the part-whole effect (Tanaka & Farah, 1993) and the composite effect (Young et al., 1987). In the part-whole effect, subjects are better able to identify two

face parts when the parts are presented in the context of a whole face than in isolation. In the composite effect, subjects are slower to identify half of a chimeric face aligned with an inconsistent other half-face than if the two half-faces are misaligned (Young et al., 1987). As with the part-whole effect, the composite effect indicates that even when subjects attempt to process only part of a face, they suffer interference from the other parts, suggesting an inability to access parts of the face and mandatory processing of the whole face.

Holistic face processing could be explained by the existence of an obligatory detection stage that uses a coarse upright template to detect whole faces (Tsao & Livingstone, 2008). An aligned chimera would be *obligatorily detected* as a whole face and therefore processed as a unit by subsequent measurement and classification stages. The key evidence favoring this early detection gating hypothesis comes from the finding of six face-selective areas in the macaque (Tsao et al., 2006, 2008) and the finding that the middle face patches, located early in this hierarchy, already consist entirely of face-selective cells (Tsao et al., 2006). According to this viewpoint, holistic psychological markers do not indicate that faces are processed by a mysterious and unique mechanism; they simply indicate that faces are processed by a system that is employed only for stimuli that are first detected as faces, but this system may be parts based, and is also likely optimized for analyzing face parts. An as yet unexplored question is whether other object categories are first detected and then analyzed by specialized modules. Other object categories that may fit this scheme include bodies and words.

Interim Summary

Perhaps the central theme that has emerged from IT research over the past two decades is that apart from biologically special categories (most conspicuously faces), objects are coded in IT in terms of their parts and not wholes. Early experiments studying the responses of IT cells to decomposed complex objects found that most cells are selective for moderately complex

features but not for whole objects. Consistent with a parts-based code, cells in V4 and PIT show tuning to the curvature and location of contour elements in object-centered coordinates. Biederman and colleagues have argued for explicit structural encoding of view-invariant volumetric primitives, but evidence suggests that the responses of most cells in TEO and TE are view dependent. Although IT cells do not generally appear to be detectors for complex objects, there are consistently observed populations of cells selectively responsive to animate objects such as faces, bodies, and hands, suggesting that animate objects may be treated differently from other types of complex patterns. A major distinction between face processing and nonface object processing is that early on (by the middle face patches at the latest), the form of a face has been *obligatorily detected as a whole*. This observation may explain holistic psychophysical phenomena associated with face perception. Cells in the middle face patch encode facial identity through ramp-shaped tuning to subsets of facial features, consistent with a parts-based code of face identity. At the highest stages of form processing, in the medial temporal lobe, cells are found that respond to specific familiar individuals in a representation-invariant manner.

COMPUTATIONAL APPROACHES TO OBJECT RECOGNITION

Theories of visual object recognition have received attention from both the neuroscience community and the machine vision community. Both approaches have contributed to our understanding of how visual object recognition might work in biological systems. We will summarize key work in both of these areas. While such a survey does not do justice to the individual contributions of each theory or system, it does allow us to see the commonalities that might be useful for producing a consensus view of computational object recognition and for applying this interpretation to processing in the primate brain, especially in relation to face processing. In this section we begin by providing a cursory survey of models of biological object

recognition. We then turn to a class of models from machine vision that produces state-of-the-art recognition abilities. We also examine a different class of models that uses feedback or contextual information during inference and object recognition. Finally, we relate these works to our view of face processing in human and non-human primates.

Theories of Biological Object Recognition

One of the central problems of object recognition is that effective systems must deal with the variations present in the natural world: variation in position, size, rotation, illumination, and even nonrigid motion (i.e., movement of limbs). This problem materializes when a visual system must generalize from its previous experience with an object under specific conditions to conditions in which the object has never been viewed before. This is a computationally difficult problem. It is made even more difficult by the fact that the novel object view may differ from previously experienced views in several of these variations at the same time. In order to deal with these variations, researchers have sought representations that will be invariant, or unchanging, under transformations that produce image variation. With respect to this problem, most models of biological object recognition take one of two approaches: models that directly compute invariant responses and those that correct for transformations to produce invariant responses. See Wiskott (2004) for an additional review and analysis of these two approaches.

Models that directly compute invariant responses take many approaches. However, there are strong commonalities among a group of models that seek to explain invariant object recognition in biological systems. This class of models can be traced back to the hypothesis originally proposed by Hubel and Wiesel that complex cell responses in primary visual cortex are formed by combining the responses of spatially shifted simple cell responses (Hubel & Wiesel, 1962, 1965). One of the earliest computational instantiations of this framework was the Neocognitron model of Fukushima (Fukushima et al., 1988). The Neocognitron is a hierarchical

model consisting of layers that alternately perform feature extraction (similar to simple cells) and layers that build invariance to position or small deformations (similar to complex cells). The weights in each layer determine the type of computation performed. The hierarchy alternately builds up selectivity and invariance until at the top layer the responses are selective to complex shapes and invariant to large changes in size and position. Recent theoretical work and simulations have extended this approach in several directions: VisNet (Wallis & Rolls, 1997), SeeMore (Mel, 1997), and HMAX (Riesenhuber & Poggio, 1999). Of note are the extensions and elaboration of the HMAX model, which have demonstrated correspondence to human level object recognition performance during brief presentations (Serre et al., 2007b), explanations of selectivity and invariance in inferotemporal cortex (Hung et al., 2005) and visual area V4 (Cadieu et al., 2007), and models of biophysical computation (Kouh & Poggio).

In approaches that deal with transformations explicitly, the model tries to correct for, or undo, the transformations so that the image can be matched against a stored canonical representation. In the model of Olshausen and colleagues (1993), the visual input is dynamically routed by control neurons through the visual hierarchy to rescale and shift the image. At the top of the hierarchy, the shifted and rescaled image can then be compared to stored templates of canonical object views, thus producing invariant recognition. The control neurons, which explicitly model the transformation, engage in an active process to selectively turn on and off connections between layers of the visual hierarchy. Later work by Arathorn (2005) suggests a computationally tractable method for determining the appropriate transformations for a given input image. It is interesting that theories that deal explicitly with transformations often use feedback: Memory representations guide the search for the correct transformation, while theories that directly compute invariant representations are constructed in a primarily feedforward network.

Bilinear models are a class of mathematical models closely related to the dynamic routing

theory of invariant recognition. In recent work, Tenenbaum and Freeman (2000) describe a bilinear model that separates the “content” and “style” in various data types. These models extend traditional linear models and are referred to as bilinear because they are linear when one set of the variables is held fixed. This model allows each set of variables to explicitly represent different types of information. For example, in text, the character identity, or content, is separated from the font, or style. In relation to the dynamic routing theory, the style variables can be considered as dynamically routing the content variables to produce an image with specific style and content. Therefore, in the context of invariant object recognition, the content could be ascribed to the identity of the object and the style to the visual variations or transformations. In one promising example, Tenenbaum and Freeman use a bilinear model on a dataset of face images containing a number of individuals viewed at a range of poses. The bilinear model was able to learn from this dataset to separate the information into a set of variables indicating the identity of a person and a set of variables indicating the pose of the person.

Computer Vision and Object Recognition

While the models we have addressed so far have been directed at explaining biological object recognition or perception, a number of techniques from computer vision have interesting implications for biological object recognition and face processing in the brain. Importantly, several computer vision models have demonstrated impressive performance on nontrivial object recognition problems and have demonstrated performance comparable to humans for specific tasks. For example, in an impressive demonstration, Sivic and Zisserman (2003) developed a system that is capable of searching through a Hollywood movie for a specific object, aptly named “Video Google.” Related to the HMAX model we discussed previously, Serre and coauthors (2007a) demonstrated that a feedforward computer vision algorithm performed at levels comparable to human subjects during a rapid presentation of a visual stimulus

in which they attempted to determine if an animal was present in the image. Furthermore, the computer vision algorithm and the human subjects exhibited similar patterns of error on individual images and image categories, such as far versus close views of animals.

Many of the most effective computer vision algorithms can be described as consisting of two stages: The first stage produces a set of local features, while the second stage performs classification or matching to determine the presence of an object or the identity of an exemplar (Berg et al., 2005; Lowe, 2004; Ranzato et al., 2007; Serre et al., 2007b). In the work of Serre and colleagues (2007b), a feature hierarchy produces a dictionary of visual features that are invariant to object variation. Standard classification techniques are used on the outputs of the hierarchy for a variety of visual recognition tasks, such as identification and classification. Interestingly, this model is inspired by a theory of biological object recognition (Serre et al., 2005) and falls into the class of biological object recognition models that seek to directly compute invariant responses. Additional work of Ranzato and colleagues (2007) takes a similar approach and develops an unsupervised learning algorithm to specify the selectivity of the intermediate layers. Lowe's influential SIFT algorithm (2004) also produces a set of invariant features, but instead of using a classifier to detect or identify object, the SIFT algorithm determines a geometric correspondence between the features. This increases the accuracy of the algorithm for object recognition because it ensures a consistent geometric interpretation of the features (e.g., the eye features, nose features, and mouth features must be in the appropriate geometric relationship). Related work by Berg and colleagues (2005) uses a different set of features, but also tries to estimate the geometrical transformation between a novel object and stored object representations.

Another interesting, and relevant, class of models uses feedback, or context, to aid in the interpretation of an image and the recognition of objects within a scene. Jin and Geman (2006) describe a hierarchy of parts for a license plate detection system that uses context to

disambiguate low-level information. In their system, license plates are modeled as a hierarchy of parts: whole license plates, groups of numbers, individual numbers, and number parts. When presented with an image, the system produces an interpretation of the visual scene by using both bottom-up cues and top-down hypotheses from prior knowledge of what license plates look like. The authors show that using their contextual hierarchy greatly improves recognition accuracy in a real-world dataset. Another example of the efficacy of contextual information is presented by Torralba and colleagues (2003). In their system, context is used to prime the location of a visual object. For example, in an indoor office scene there is a much higher probability of seeing a computer monitor than there is in an outdoor forest scene.

Insights for Biological Face Recognition

These results can be used to interpret findings related to face processing in primates, and allow us to develop mathematical models that make explicit the architecture of biological face processing. There are three main connections we want to point out. First, a rapid feedforward computation that produces invariant responses may be central to the detection of faces. Second, complicated facial discriminations and inferences may be mediated by a more complicated series of interconnected processing stages that explicitly model the visual transformations. Third, the representations in these later stages may represent the independent aspects of facial appearance and pose.

Because it is advantageous for facial detection to be computed quickly, approaches that directly compute invariant responses may be suitable for face detection. While direct computation of invariant responses often suffers from the combinatorial explosion of possible inputs (and therefore requires more processing to overcome the combinations of variations for a specific object), face detection is unique because of its ecologically critical role, and because faces constitute a relatively constrained set of visual inputs compared to all visual objects. The early stages in the ventral stream that exclusively

respond to faces may serve as the culmination of this rapid process of directly computing invariant responses to faces.

We may also gain some insights about biological face processing from bilinear models, which seek to explicitly and separately represent different aspects of the visual input. In the example examined by Tenenbaum and Freeman, the pose of a face is represented independently of the facial identity (Fig. 24.12). In face processing it is often the case that the relevant information is not only the presence or the identity of a face but also what it is doing. For example, in complex social interactions it is important to infer intentions from subtle cues about head pose and gaze direction. In these cases the transformations that facial forms undergo are the goal of representation and not just variations in the input that must be

overcome. Therefore, it seems sensible that explicitly representing these transformations would be advantageous. Could the different nodes of the interconnected system of face patches be representing different, largely independent aspects of facial form and variation? Such a representation would make specific information independently available for complex social inferences of ecological value.

Interestingly, the two main theoretical approaches to biological object recognition, directly computing invariant responses and explicitly accounting for transformations, may both be employed for face recognition and processing: Direct computation of invariant responses may mediate face detection, while computing transformations may mediate further inferences about facial characteristics. The evidence for holistic face processing seems

Figure 24.12 A bilinear model from Tenenbaum and Freeman. In this model, face images are represented by two sets of variables, one encoding pose, “pose coefficients,” and the other encoding identity, “person coefficients.” The basis images (top left) are weighted by the joint activity of the person coefficients (top right) and the pose coefficients (bottom left) to generate reconstructed images of faces (bottom right). From Tenenbaum, J. B., & Freeman, W. T. (2000). Separating style and content with bilinear models. *Neural Computation*, 12, 1247–1283. Used with permission.

Figure 24.13 A model of face processing. In this model of face recognition a fast, feedforward detection stage provides feedback to the face model. The face model, which is only activated upon face detection, computes a variety of parameters specific to faces, such as pose, identity, and expression (among possibly others, such as gaze direction). The face detection stage may bias the face model to certain regions of the image, and certain sizes of faces from the information it infers from the input. Such information aids the face model in inferring more specific face parameters such as identity or pose. (The downward arrow from the face model to the image indicates the generative formulation of bilinear models as the model parameters seek to provide an interpretation of the image.)

to indicate that face detection gates this further processing.

We can make the connections between computational theories and biological evidence more precise by describing a simple model of face processing in primates. Presented in Figure 24.13, the system incorporates these insights. First, a fast, feedforward processing stage computes visual features from the visual input, and faces are detected from these features in a face detection area. Second, the detection of a face in the face detection module guides and gates further face-specific processing. Third, a

face-specific processing stage, or face model, produces a rich description of the face. The face model may be similar to a bilinear model and mimic the dynamic routing theory of invariant object recognition.

SUMMARY

Hubel and Wiesel showed that primary visual cortex represents images in a space of localized, oriented edges. This was stunning because it showed that the brain performs a simple mathematical transformation that condenses the

information present in the visual world. Primates possess an entire temporal lobe to further condense the visual form of objects. The thesis of this chapter is that the system of face patches in the macaque brain may be a “turtle’s underbelly” (Medawar, 1981)—if we can understand it, we will be able to pry open the general problem of invariant pattern recognition.

What mathematical transformations are being performed in the face patches? An unprecedented opportunity now exists to understand the successive stages of face processing at a mechanistic level. The face-patch system offers a set of dedicated, connected, yet anatomically distinct components, inviting us to analyze for the first time the messages being sent within a hierarchical system for high-level object recognition. If we can figure out how the brain recognizes faces, then we will have gone a long way toward understanding how the brain represents a complex object made of multiple parts.

ACKNOWLEDGMENTS

We are grateful to Winrich Freiwald for discussions and comments on the manuscript.

REFERENCES

- Allison, T., McCarthy, G., Nobre, A., Puce, A., & Belger, A. (1994). Human extrastriate visual cortex and the perception of faces, words, numbers, and colors. *Cerebral Cortex*, 4, 544–554.
- Arathorn, D. W. (2005). A cortically-plausible inverse problem solving method applied to recognizing static and kinematic 3D objects. In: *Advances in neural information processing systems*. Cambridge, MA: MIT Press.
- Berg, A., Berg, T., & Malik, J. (2005). Shape matching and object recognition using low distortion correspondence. *Proceedings of CVPR*, 1, 26–33.
- Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94, 115–147.
- Biederman, I. (1995). Visual object recognition. In: S. M. Kosslyn & D. N. Osherson (Eds.), *Visual cognition* (pp. 121–166). Cambridge, MA: MIT Press.
- Biederman, I., & Cooper, E. E. (1991). Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology*, 23, 393–419.
- Boussaoud, D., Desimone, R., & Ungerleider, L. G. (1991). Visual topography of area TEO in the macaque. *Journal of Comparative Neurology*, 306, 554–575.
- Boynton, G. M. (2005). Imaging orientation selectivity: Decoding conscious perception in V1. *Nature Neuroscience*, 8, 541–542.
- Brincat, S. L., & Connor, C. E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature Neuroscience*, 7, 880–886.
- Brincat, S. L., & Connor, C. E. (2006). Dynamic shape synthesis in posterior inferotemporal cortex. *Neuron*, 49, 17–24.
- 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.
- Bulthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex*, 5, 247–260.
- Cadiou, C., Kouh, M., Pasupathy, A., Connor, C. E., Riesenhuber, M., & Poggio, T. (2007). A model of V4 shape selectivity and invariance. *Journal of Neurophysiology*, 98, 1733.
- Changizi, M. A., Zhang, Q., Ye, H., & Shimojo, S. (2006). The structures of letters and symbols throughout human history are selected to match those found in objects in natural scenes. *American Naturalist*, 167, E117–139.
- Cheng, K., Saleem, K. S., & Tanaka, K. (1997). Organization of corticostriatal and corticoamygdalar projections arising from the anterior inferotemporal area TE of the macaque monkey: A Phaseolus vulgaris leucoagglutinin study. *Journal of Neuroscience*, 17, 7902–7925.
- Connor, C. E., Brincat, S. L., & Pasupathy, A. (2007). Transformation of shape information in the ventral pathway. *Current Opinions in Neurobiology*, 17, 140–147.
- Dean, P. (1976). Effects of inferotemporal lesions on the behavior of monkeys. *Psychological Bulletin*, 83, 41–71.
- Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties

AQ4

- of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 4, 2051–2062.
- Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, 115, 107–117.
- 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.
- Erickson, C. A., & Desimone, R. (1999). Responses of macaque perirhinal neurons during and after visual stimulus association learning. *Journal of Neuroscience*, 19, 10404–10416.
- Erickson, C. A., Jagadeesh, B., & Desimone, R. (2000). Clustering of perirhinal neurons with similar properties following visual experience in adult monkeys. *Nature Neuroscience*, 3, 1143–1148.
- Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is “special” about face perception? *Psychological Review*, 105, 482–498.
- Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1, 1–47.
- Foldiak, P., Xiao, D., Keyser, C., Edwards, R., & Perrett, D. I. (2004). Rapid serial visual presentation for the determination of neural selectivity in area STSa. *Progress in Brain Research*, 144, 107–116.
- Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. *Nature Neuroscience*, in press.
- Fujita, I., & Fujita, T. (1996). Intrinsic connections in the macaque inferior temporal cortex. *Journal of Comparative Neurology*, 368, 467–486.
- Fujita, I., Tanaka, K., Ito, M., & Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, 360, 343–346.
- Fukushima, K., Miyake, S., & Ito, T. (1988). Neocognitron: A neural network model for a mechanism of visual pattern recognition. In: *Computer Society Press technology series neural networks* (pp. 136–144). Los Alamitos, CA: IEEE Computer Society Press.
- Gross, C. G. (1973). Visual functions of inferotemporal cortex. In: R. Jung (Ed.), *Handbook of sensory physiology* (p. 451–482). Berlin: Springer.
- Guigon, E. (2003). Computing with populations of monotonically tuned neurons. *Neural Computation*, 15, 2115–2127.
- Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293, 2425–2430.
- Haxby, J. V., Grady, C. L., Horowitz, B., Ungerleider, L. G., Mishkin, M., Carson, R. E., et al. (1991). Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proceedings of the National Academy of Sciences*, 88, 1621–1625.
- Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *Journal of Physiology*, 160, 106.
- Hubel, D., & Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas of the cat. *Journal of Neurophysiology*, 229–289.
- Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310, 863–866.
- Ito, M., Tamura, H., Fujita, I., & Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, 73, 218–226.
- Jin, Y., & Geman, S. (2006). Context and hierarchy in a probabilistic image model. *IEEE Conference on Computer Vision and Pattern Recognition*.
- Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8, 679–685.
- Kanwisher, N., McDermott, J., & Chun, M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17, 4302–4311.
- Kayaert, G., Biederman, I., Op de Beeck, H. P., & Vogels, R. (2005). Tuning for shape dimensions in macaque inferior temporal cortex. *European Journal of Neuroscience*, 22, 212–224.
- Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, 97, 4296–4309.
- 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.
- Kouh, M., & Poggio, T. A canonical neural circuit for cortical nonlinear operations. *Neural Computation*.

AQ5

AQ6

AQ7

- Leonard, C. M., Rolls, E. T., Wilson, F. A., & Baylis, G. C. (1985). Neurons in the amygdala of the monkey with responses selective for faces. *Behavioral Brain Research*, 15, 159–176.
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60, 91–110.
- Maunsell, J. H., & Van Essen, D. C. (1987). Topographic organization of the middle temporal visual area in the macaque monkey: Representational biases and the relationship to callosal connections and myeloarchitectonic boundaries. *Journal of Comparative Neurology*, 266, 535–555.
- McCarthy, G., Puce, A., Belger, A., & Allison, T. (1999). Electrophysiological studies of human face perception. II: Response properties of face-specific potentials generated in occipitotemporal cortex. *Cerebral Cortex*, 9, 431–444.
- McCarthy, G., Puce, A., Gore, J. C., & Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience*, 9, 605–610.
- Medawar, P. B. (1981). *Advice to a young scientist*. Perseus Publishing.
- Mel, B. W. (1997). SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation*, 9, 777–804.
- Meunier, M., Bachevalier, J., Mishkin, M., & Murray, E. A. (1993). Effects on visual recognition of combined and separate ablations of the entorhinal and perirhinal cortex in rhesus monkeys. *Journal of Neuroscience*, 13, 5418–5432.
- Miyashita, Y., Okuno, H., Tokuyama, W., Ihara, T., & Nakajima, K. (1996). Feedback signal from medial temporal lobe mediates visual associative mnemonic codes of inferotemporal neurons. *Brain Research: Cognitive Brain Research*, 5, 81–86.
- Moeller, S., Freiwald, W. A., & Tsao, D. Y. (2008). Patches with links: A unified system for processing faces in the macaque temporal lobe. *Science*, 320, 1355–1359.
- Moscovitch, M., Winocur, G., & Behrmann, M. (1997). What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of Cognitive Neuroscience*, 9, 555–604.
- Mountcastle, V. B. (1997). The columnar organization of the neocortex. *Brain*, 120 (Pt 4), 701–722.
- Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Science*, 10, 424–430.
- Olshausen, B., Anderson, C., & Van Essen, D. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13, 4700.
- Oram, M. W., & Perrett, D. I. (1992). Time course of neural responses discriminating different views of the face and head. *Journal of Neurophysiology*, 68, 70–84.
- Pasupathy, A., & Connor, C. E. (1999). Responses to contour features in macaque area V4. *Journal of Neurophysiology*, 82, 2490–2502.
- Pasupathy, A., & Connor, C. E. (2001). Shape representation in area V4: Position-specific tuning for boundary conformation. *Journal of Neurophysiology*, 86, 2505–2519.
- Pasupathy, A., & Connor, C. E. (2002). Population coding of shape in area V4. *Nature Neuroscience*, 5, 1332–1338.
- 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., et al. (1984). Neurones responsive to faces in the temporal cortex: Studies of functional organization, sensitivity to identity and relation to perception. *Human Neurobiology*, 3, 197–208.
- 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*, 102, 6996–7001.
- Previc, F. (1990). Functional specialization in the lower and upper visual fields in humans: Its ecological origins and neurophysiological implications. *Behavior and Brain Sciences*, 13, 519–575.
- Puce, A., Allison, T., Asgari, M., Gore, J. C., & McCarthy, G. (1996). Differential sensitivity of human visual cortex to faces, letterstrings, and textures: A functional magnetic resonance imaging study. *Journal of Neuroscience*, 16, 5205–5215.
- Puce, A., Allison, T., & McCarthy, G. (1999). Electrophysiological studies of human face perception. III: Effects of top-down processing

AQ8

- on face-specific potentials. *Cerebral Cortex*, 9, 445–458.
- Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435, 1102–1107.
- Ranzato, M., Huang, F., Boureau, Y., & LeCun, Y. (2007). Unsupervised learning of invariant feature hierarchies with applications to object recognition. *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR'07)*.
- Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2, 1019–1025.
- Saleem, K. S., Kondo, H., & Price, J. L. (2008). Complementary circuits connecting the orbital and medial prefrontal networks with the temporal, insular, and opercular cortex in the macaque monkey. *Journal of Comparative Neurology*, 506, 659–693.
- Saleem, K. S., & Tanaka, K. (1996). Divergent projections from the anterior inferotemporal area TE to the perirhinal and entorhinal cortices in the macaque monkey. *Journal of Neuroscience*, 16, 4757–4775.
- Saleem, K. S., Tanaka, K., & Rockland, K. S. (1993). Specific and columnar projection from area TEO to TE in the macaque inferotemporal cortex. *Cerebral Cortex*, 3, 454–464.
- Scalaidhe, S. P., Wilson, F. A., & Goldman-Rakic, P. S. (1999). Face-selective neurons during passive viewing and working memory performance of rhesus monkeys: Evidence for intrinsic specialization of neuronal coding. *Cerebral Cortex*, 9, 459–475.
- Seltzer, B., & Pandya, D. N. (1994). Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: A retrograde tracer study. *Journal of Comparative Neurology*, 343, 445–463.
- Sergent, J., Ohta, S., & MacDonald, B. (1992). Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain, Part 1*, 15–36.
- Serre, T., Oliva, A., & Poggio, T. (2007a). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, 104, 6424.
- Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007b). Robust object recognition with cortex-like mechanisms. In: *IEEE transactions on pattern analysis and machine intelligence* (pp. 411–426). IEEE Computer Society.
- Serre, T., Wolf, L., & Poggio, T. (2005). Object recognition with features inspired by visual cortex. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Sivic, J., & Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. *Computer vision, 2003. Proceedings of the ninth IEEE international conference* (pp. 1470–1477).
- Somogyi, P., Tamas, G., Lujan, R., & Buhl, E. H. (1998). Salient features of synaptic organisation in the cerebral cortex. *Brain Research: Brain Research Reviews*, 26, 113–135.
- Suzuki, W., Saleem, K. S., & Tanaka, K. (2000). Divergent backward projections from the anterior part of the inferotemporal cortex (area TE) in the macaque. *Journal of Comparative Neurology*, 422, 206–228.
- Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19, 109–139.
- Tanaka, K. (2003). Columns for complex visual object features in the inferotemporal cortex: Clustering of cells with similar but slightly different stimulus selectivities. *Cerebral Cortex*, 13, 90–99.
- Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology A*, 46, 225–245.
- Tanaka, K., Saito, H., Fukada, Y., & Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, 66, 170–189.
- Tenenbaum, J. B., & Freeman, W. T. (2000). Separating style and content with bilinear models. *Neural Computation*, 12, 1247–1283.
- Torralba, A., Murphy, K., Freeman, W., & Rubin, M. (2003). Context-based vision system for place and object recognition. *Computer Vision, 2003. Proceedings of Ninth IEEE International Conference* (pp. 273–280).
- 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. H., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, 311, 670–674.

AQ9

- 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. *PNAS*, 105, 19514–19519.
- Tsunoda, K., Yamane, Y., Nishizaki, M., & Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience*, 4, 832–838.
- Ungerleider, L. G., Galkin, T. W., Desimone, R., & Gattass, R. (2008). Cortical connections of area V4 in the macaque. *Cerebral Cortex*, 18, 477–499.
- Vogels, R., Biederman, I., Bar, M., & Lorincz, A. (2001). Inferior temporal neurons show greater sensitivity to nonaccidental than to metric shape differences. *Journal of Cognitive Neuroscience*, 13, 444–453.
- Von Bonin, G., & Bailey, P. (1947). *The neocortex of Mucoca mulutru*. Urbana, IL: University of Illinois Press.
- Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, 51, 167–194.
- Wang, G., Tanaka, K., & Tanifuji, M. (1996). Optical imaging of functional organization in the monkey inferotemporal cortex. *Science*, 272, 1665–1668.
- Wang Y, Fujita I, Murayama Y (2000) Neuronal mechanisms of selectivity for object features revealed by blocking inhibition in inferotemporal cortex. *Nature Neuroscience* 3, 807–813.
- Webster, M. J., Bachevalier, J., & Ungerleider, L. G. (1993). Subcortical connections of inferior temporal areas TE and TEO in macaque monkeys. *Journal of Comparative Neurology*, 335, 73–91.
- Webster, M. J., Bachevalier, J., & Ungerleider, L. G. (1994). Connections of inferior temporal areas TEO and TE with parietal and frontal cortex in macaque monkeys. *Cerebral Cortex*, 4, 470–483.
- Webster, M. J., Ungerleider, L. G., & Bachevalier, J. (1991). Connections of inferior temporal areas TE and TEO with medial temporal-lobe structures in infant and adult monkeys. *Journal of Neuroscience*, 11, 1095–1116.
- Wiskott, L. (2004). How does our visual system achieve shift and size invariance? In: J. L. van Hemmen & T. J. Sejnowski (Eds.), *23 Problems in systems neuroscience*. New York: Oxford University Press.
- Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, 16, 747–759.
- Yukie, M., & Iwai, E. (1988). Direct projections from the ventral TE area of the inferotemporal cortex to hippocampal field CA1 in the monkey. *Neuroscience Letters*, 88, 6–10.

QUERIES TO BE ANSWERED BY AUTHOR (SEE MANUAL MARKS)

IMPORTANT NOTE: Please mark your corrections and answers to these queries directly onto the proof at the relevant place. Do NOT mark your corrections on this query sheet.

Chapter 24

Q. No.	Pg No.	Query
AQ1	476	Please update reference. Also, "Used with permission" line needed?
AQ2	479	Please update reference.
AQ3	487	Please update reference.
AQ4	494	Please check that all journal names are correct as spelled out.
AQ5	495	Please update reference.
AQ6	495	Please provide volume number.
AQ7	495	Please provide volume number and page range for this journal reference.
AQ8	496	Please provide location of publisher.
AQ9	497	Please provide location of publisher.