II BACKGROUND AND METHODS

17 Introduction to the Anatomy and Function of Visual Cortex

Kendra S. Burbank and Gabriel Kreiman

Summary

We provide here a brief overview of the neuroanatomy and neurophysiology of the primate visual system. We first describe the physical path through the brain that visual information takes as it is undergoing a transformation from an almost pixel-based format to a more abstract representation of behaviorally relevant information. We then describe what is known about the responses of neurons in specific brain areas to different visual stimuli. These responses are researchers' best window into the circuits involved in information transformation. Finally, we describe several computational models of the visual system. Throughout, we mostly focus on the ventral parts of the primate (monkey and human) visual cortex and its role in object recognition.

Introduction

Primates and other species use vision constantly in order to detect motion, estimate distances to objects, and recognize objects. A large part of the primate brain is involved in processing visual information, and it is presumed that visual processing must have been under strong selective pressure to develop a system capable of achieving strong selectivity, robustness to object transformations, high capacity, and high processing speed (Connor, Brincat, and Pasupathy, 2007; Logothetis and Sheinberg, 1996; Serre et al., 2007; Wandell, 1995). The brain must transform incoming visual signals from their very specific (almost pixel-based) form arriving at the eyes to a much more abstract form that is useful for quickly extracting behaviorally relevant information.

Our aim in this chapter is to provide a succinct overview of the architecture and function of the primate visual system. While writing, we have in mind a quantitative student (of math, engineering, physics, computer science, or the like) who first encounters the bewildering and fascinating complexity of visual cortex. The goal of this chapter is *not* to provide an exhaustive account of the visual system, but rather to describe some of the basic insights that are important to understand other chapters in this book and to introduce researchers to the architecture, function, and computational modeling of the visual system. At the same time, we hope that the curious reader will be eager to learn more, and we encourage further reading (of, e.g., Biederman, 1987; Blumberg and Kreiman 2010; Carandini et al., 2005; Connor, Brincat, and Pasupathy, 2007; Dayan and Abbott, 2001; Deco and Rolls, 2004b; Felleman and Van Essen, 1991; Gabbiani and Cox 2010; Gross, 1994; Humphreys and Riddoch, 1993; Koch, 2005; Kreiman, 2004, 2007; Logothetis and Sheinberg, 1996; Riesenhuber and Poggio, 1999; Rolls, 1991; Tanaka, 1996; Ullman, 1996; Wandell, 1995; Wu, David, and Gallant, 2006; as well as other references in this chapter).

Neuroanatomy

It was recognized early on that lesions in the back of the brain tend to produce visual impairments and that the exact nature of the deficit varies with the exact position of the lesion. Subsequent studies identified multiple parts of cortex that are involved in processing visual information. A classic study by Felleman and Van Essen (1991) summarized knowledge about connectivity in the primate visual cortex, organizing visual cortex into an approximate hierarchical system. A subset of that hierarchy, ventral visual cortex, seems most important for visual object recognition. A highly schematic representation of the connectivity in some of the main parts of visual cortex is shown in figure 17.1. Much more is known about the connectivity and anatomy of the nonhuman primate visual cortex than about the human visual cortex; the discussion in this section focuses on the nonhuman primate. Here, we walk through a simplified version of the path that information takes as it makes its way from the eye through the visual cortex.

Early Vision: Retina to Cortex

Information enters the visual system when light reaches the eye. The light is focused by the lens to land on the *retina*, a collection of cells at the back of the eyeball. There, the light excites *photoreceptor* neurons: the *rods*, which are specialized for dim light, and the *cones*, which are specialized for fine detail and color vision. The very center of the retina, the *fovea*, contains only cones and provides higher resolution than the periphery. The signal from the photoreceptors is passed through intermediate types of cells, *horizontal*, *bipolar*, and *amacrine* neurons, before arriving at *retinal ganglion* cells, which are located at the front of the retina. In chapter 2, Sheila Nirenberg describes state-of-the-art methods to quantitatively elucidate how retinal ganglion cells encode visual information.

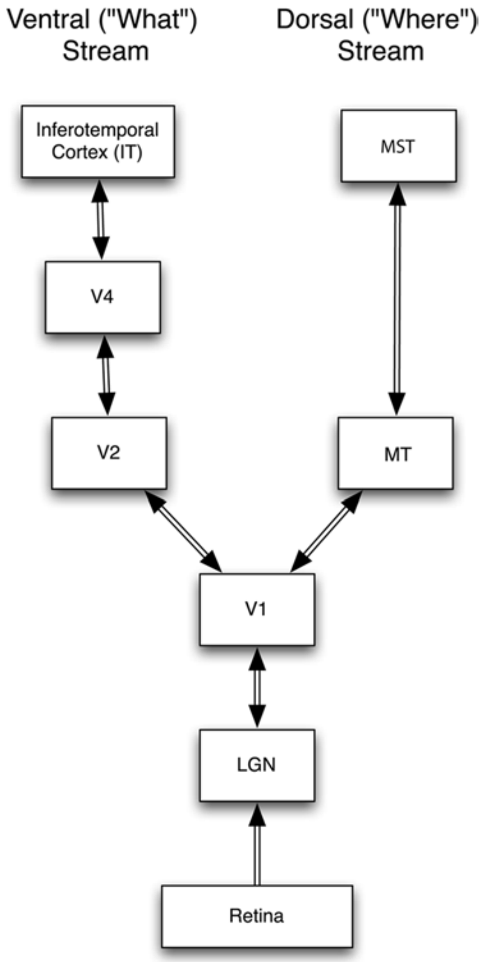

Figure 17.1
A highly oversimplified schematic of the primate visual system. The arrows denote the direction of information flow (see text for details; see also Felleman and Van Essen 1991).

The visual signals pass down the axons of the ganglion cells, which come together in a bundle called the optic nerve and travel out of the eye and back to the rest of the brain. About 90 percent of the retinal ganglion cell axons terminate in the *lateral geniculate nucleus* (LGN), a structure in the thalamus, in the center of the brain. Every cell in the LGN receives input from several retinal ganglion cells.LGN neurons, in turn, have axons that come together in a bundle called the *optic radiation*. From the thalamus, this bundle transmits the visual signals to the first of the visual areas in cortex: *primary visual cortex*, which is at the very back of the brain. For further reading about early vision, see Wandell (1995).

Primary Visual Cortex

Primary visual cortex (V1, also referred to as “striate cortex” and “area 17” in cats) is located at the posterior poles of the left and right occipital cortices. In human adults, the V1 region of each hemisphere is 2mm thick and occupies an area of around 2300 mm^2, or roughly two-thirds the size of a credit card. (In the smaller brain of the macaque monkey, V1 has approximately half the area and a quarter the thickness.) V1 contains around 140 million neurons in each hemisphere; these can be largely classified into two main types: *pyramidal cells* and *interneurons*. The neurons are arranged in six layers that differ in connectivity and function. Primary visual cortex has an additional, columnar organization: the columns are perpendicular to the layer structure, and neurons within a column typically share similar visual preferences (Douglas and Martin, 2004; Nassi and Callaway, 2009). There has been more research investigating area V1 than any other part of visual cortex. Chapter 3, 8, 21, and 22 illustrate the neurophysiological properties of V1 neurons.

The Ventral Stream: V2-ITC

From V1, the visual signal is split into two roughly separate channels, or “streams” (figure 17.1) (Haxby et al., 1991; Mishkin, 1982). The *ventral stream* passes into *secondary visual cortex* (V2), then through area *V4* and into *inferotemporal cortex* (IT). The ventral stream is primarily involved in object recognition, and is sometimes called the “what” stream. The *dorsal stream* projects from V1 to V2 and V3 and also to “*middle temporal cortex*” (MT or V5). The dorsal stream processes spatial locations, stereopsis, and object motion and is known as the “where” or “action” stream. This chapter focuses on the properties of the ventral stream. For a recent overview of the properties of the dorsal stream, see Born and Bradley (2005).

Secondary visual cortex is located just to the front of V1, from which it receives strong feedforward input. Secondary visual cortex has a layered structure and columnar organization similar to that of V1; indeed, these features may be common to all neocortical visual areas. The next area in the ventral stream is area V4, located anterior to V2. The last purely visual cortical area along the ventral stream is Inferior

Temporal Cortex (ITC). Several investigators have in turn divided ITC into multiple subparts such as posterior ITC, central ITC, and anterior ITC. Another nomenclature that is widely used refers to area TEO (roughly corresponding to posterior ITC and central ITC) and area TE (roughly corresponding to anterior ITC). Chapter 7 describes how contour shapes are represented in areas V4 and ITC and chapter 10 describes ultrafast encoding of visual information revealed by decoding the activity of a population of ITC neurons.

Many interareal connections exist beyond those in the feedforward path just described; figure 17.1 is a major oversimplification. There are connections between the dorsal and ventral streams (Felleman and Van Essen, 1991), horizontal connections within each area, "bypass" connections (e.g., LGN projections to extrastriate visual areas beyond V1, V1 projections to V4) and abundant back-projection connections. Indeed, semiquantitative anatomical studies reveal that back-projections are significantly more abundant than feedforward connections (e.g., Binzegger, Douglas, and Martin, 2004; Callaway, 2004; Douglas and Martin, 2004)!

Of course, another important simplification is that each of the boxes in figure 17.1 encompasses millions of neurons. Is it possible to obtain a more detailed picture of the individual connections between neurons? Characterizing neuroanatomical connections at high resolution has traditionally been a daunting task and typically required laborious analysis of the projections of individual neurons (Douglas and Martin, 2004; Rockland and Pandya, 1979; Salin and Bullier, 1995). There has been rapid progress over the last five years in the field of "connectomics," which aims to provide high-resolution connectivity information (at the electron microscopy level) for large neuronal circuits. Yet, it seems that we are still far from obtaining detailed connectivity in neocortex. The availability of such data will eventually enable researchers to move from qualitative description of some connections across areas to a systematic characterization of the key principles governing connectivity in cortex.

Neurophysiological Responses in the Visual System

It is difficult to deduce function exclusively from anatomy, and the presence of connections does not indicate the strength (or sign) of those connections. To describe the function of neuronal circuits during vision, it is necessary to examine the activity of individual neurons and their responses to visual stimuli. The gold standard for measuring the activity of neurons is the use of microwire electrodes to record the action potentials of single neurons at millisecond temporal resolution. In the typical experimental situation, researchers present the subject with a visual stimulus while monitoring the subject's eye movements and recording the evoked response of one or more neurons, as well as behavioral responses in awake experiments (see, e.g., chapters 2, 3, 7, 8, 10, 21, and 22). These experiments have been most frequently

Q

performed in cats and nonhuman primates such as the macaque monkey. There have also been some efforts to examine field potentials and unit activity in the human cortex (Allison et al., 1999; Engel et al., 2005; Kreiman, 2007; Liu et al., 2009). Recently, there has also been an increased and promising resurgence of interest in the rodent visual cortex.

Neurophysiology of the Early Visual System

Significant processing of visual information occurs within the retina itself. The firing of each retinal ganglion cell is affected by light impinging upon a small region of the visual field—this region is termed the cell's *receptive field*. The light can cause either an increase or a decrease in a ganglion cell's firing, depending on exactly where it arrives within the cell's receptive field. For cells called *on-center* cells, light arriving in the center of the receptive field will increase the firing rate, while light arriving in the periphery will instead suppress firing. For *off-center* cells, the opposite pattern is seen. Both cases are examples of *center-surround* receptive field architectures. LGN neurons have receptive fields approximately similar to those of ganglion cells—they also display a center-surround organization.

An important implication of the center-surround architecture is that light impinging upon both the center and the periphery of the receptive field will cause only weak firing. Instead, retinal ganglion cells respond most strongly when illumination is not constant across the receptive field. The necessary nonuniformity could come from the presence of a high-contrast visual feature or from a temporally changing stimulus, such as a dot of light moving through the receptive field. The full response is characterized by *spatiotemporal receptive fields* that are often described by a difference-of-Gaussians model (Dayan and Abbott, 2001; Wandell, 1995; Gabbiani and Cox, 2010). For an on-center cell with a receptive field centered at $x = y = 0$, the structure of the receptive field can be characterized by the filter *F(x,y)*:

$$F(x,y) = \frac{1}{2\pi\sigma_{center}^2}\exp\left(-\frac{x^2+y^2}{2\sigma_{center}^2}\right) - \frac{B}{2\sigma_{surround}^2}\exp\left(-\frac{x^2+y^2}{2\sigma_{surround}^2}\right) \tag{17.1}$$

where σ_{center} and $\sigma_{surround}$ control the size of the center and surround regions respectively and B indicates the relative weight of center and surround responses. In addition to the just-described spatial aspects of the receptive field, the responses of ganglion cells and LGN neurons evolve over time; more elaborate models include this temporal dependency when describing the receptive field properties (Dayan and Abbott, 2001).

Neurophysiology in V1

The first systematic description of V1 neurons' responses to visual stimuli was given by Hubel and Wiesel (1959, 1962). Neurons in primary visual cortex have small receptive fields near the center of the visual field. On average, the receptive fields in V1 comprise less than 1° of visual angle. The neurons in V1 are arranged so that their receptive fields tile visual space in a *retinotopic map*. That is, nearby neurons in primary visual cortex represent nearby locations in the visual field. This tiling is most dense for visual input coming from the foveal region. The *cortical magnification factor* describes the nonlinear representation of the visual field in cortex.

To a first approximation, V1 pyramidal neurons fall into two general classes. *Simple cells* have elongated receptive fields that contain specific excitatory and inhibitory regions. Simple cell responses are well modeled by linear summation of the stimulus present in excitatory and inhibitory regions of their receptive field. An effective stimulus for a simple cell might be an oriented bar, exactly positioned so that its edge matches the border between the excitatory and inhibitory regions in the cell's receptive field. A slight shift in the stimulus location can greatly decrease, or even eliminate, the simple cell's response. The spatial structure of the responses of a V1 simple cell with a receptive field centered at $x = y = 0$ can often be well described by a Gabor function (product of a Gaussian and cosine):

$$F(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} exp\left(-\frac{x}{2\sigma_x^2} - \frac{y}{2\sigma_y^2}\right)\cos(kx-\phi)\,, \tag{17.2}$$

where σ_x and σ_y determine the spatial extent in x and y, k is the preferred spatial frequency, and ϕ indicates the preferred spatial phase. If the spatial and temporal aspects of the receptive field are separable, an additional multiplicative term is added to characterize the temporal dynamics of the responses (for a discussion of the separability of spatial and temporal aspects of V1 responses, see Dayan and Abbott, 2001; Ringach, Hawken, and Shapley, 1997).

A second class of V1 pyramidal cells, complex cells, have receptive fields that do not show simply defined excitatory and inhibitory regions. Instead, complex cells respond to particular features—generally oriented bars—with considerable tolerance of the stimulus' position within the receptive field. Other properties have been described in V1 responses. Particularly important are end-stopped cells, cells that respond best when the oriented bar ends within the receptive field.

Hubel and Wiesel proposed a simple and elegant model of how orientation tuning at the level of V1 could arise from the combination of LGN center-surround units with receptive fields aligned according to the orientation preference of the V1 unit. Several other models have been proposed to describe the origin of V1 receptive fields (for a recent overview and discussion, see Carandini et al., 2005).

Although only stimuli within a neuron's receptive field can directly drive its activity, the neuron's activity can be enhanced or suppressed depending on context in nearby regions. One common type of contextual modulation is *surround* suppression (Allman, Miezin, and McGuinness, 1985; Angelucci and Bressloff, 2006), where Q the presence of a parallel oriented bar just outside of a neuron's receptive field can suppress the response of that neuron to an oriented bar within the receptive field. The mechanisms for contextual modulation are a matter of current research; they may arise from feedforward connections, from lateral connections within a brain area, or from feedback from higher areas. There is also evidence that V1 neuron responses can be modulated by attention (Desimone and Duncan, 1995; for a recent review, see Reynolds and Chelazzi, 2004). Chapter 3 shows recent evidence that illustrates how the activity of neuronal populations in V1 can be influenced by task demands and attention. While there has been significant progress toward describing the responses of V1 neurons, much remains to be done to fully and quantitatively characterize the V1 neuronal circuitry (Carandini et al., 2005).

Neurophysiology beyond V1

Beyond primary visual cortex, in spite of a large body of work by multiple talented investigators, what neurons "prefer" remains largely terra incognita. Part of the challenge is the large multidimensional space in which possible visual inputs reside combined with the relatively short recording times. In typical neurophysiological experiments, it is possible to sample only a small fraction of the conceivable set of visual stimuli. It is therefore very difficult to estimate the joint probability distribution of visual stimuli and neuronal responses. To make matters even more complicated, neurons' responses are modulated by context from outside the receptive field; an exhaustive response characterization would require also varying the contextual conditions. Such an approach is clearly unfeasible with current techniques. Instead, researchers make educated guesses about which stimulus characteristics are likely to be important to the neurons' responses, and they vary only these characteristics. This approach has been quite successful in early brain areas such as V1, where a few simple characteristics such as orientation and contrast can be shown to determine much of a neuron's response (see, however, Carandini et al., 2005). However, in extrastriate visual areas (those outside V1), the complex selectivities that neurons display makes it difficult to determine a set of simplified stimulus characteristics to sample. Indeed, it is entirely possible that even if the subset of important stimulus characteristics were known, the resulting space of possible stimuli would still be too large to sample experimentally. Although there have been multiple studies examining the responses of neurons along the ventral visual stream from V2 to ITC, we lack a clear quantitative understanding of feature preferences, let alone the mechanisms by which these feature preferences originate. A promising line of research involves using algorithms that aim to iteratively refine the stimuli presented to neurons to converge on the preferred features (e.g., Connor, Brincat, and Pasupathy, 2007; see also chapter 7). This is an area of active research, and the field will benefit from the systematic interplay of theoretical predictions and neurophysiological recordings.

In the following paragraphs, we provide an overview of several studies that illustrate the type of responses encountered in extrastriate visual cortex to different types of stimuli, but we emphasize that a systematic, quantitative, and theory-based understanding of neurophysiological responses remains an important open question in the field.

The receptive fields of V2 neurons form a retinotopic map, like that in V1, but are roughly 2–3 times larger (Burkhalter and Van Essen, 1986; Gattass, Gross, and Sandell, 1981). Neurons in V2 can be excited by simple stimuli, in a similar fashion to V1 neurons. But at least some V2 neurons appear to be specialized for detecting more complex features. Some authors have proposed that V2 neurons detect curvature or angles (Hegde and Van Essen, 2003; Ito and Komatsu, 2004). The responses of V2 neurons can be modulated by abstract features of the stimulus—even features present outside the neurons' receptive fields. Such modulatory influences include the presence of illusory contours (Peterhans and von der Heydt, 1991; von der Heydt, Friedman, and Zhou, 1999) and spatial attention (Desimone and Duncan, 1995). Although such modulation is also seen partly in area V1, the effects are stronger and more frequent in V2 (von der Heydt, Peterhans, and Baumgartner, 1984).

V4 neurons have receptive fields around 4–7 times as large as V1 neurons (Desimone and Schein, 1987) The tuning properties of V4 neurons are more complex than those of V2 neurons, with some appearing to be tuned for simple geometric shapes (Cadieu et al., 2007; David, Hayden, and Gallant, 2006; Desimone and Schein, 1987; Pasupathy and Connor, 2001). V4 is more strongly affected by attentional modulation than areas V1 and V2 (Moran and Desimone, 1985). Neuronal activity in area V4 plays an important role in analyzing color (Zeki, 1983).

Finally, neurons in ITC have significantly larger receptive fields than those in earlier areas, but reports vary widely in terms of their exact magnitudes from neurons with receptive fields of a few degrees (DiCarlo and Maunsell, 2004) all the way to neurons with receptive fields spanning several tens of degrees (Rolls, 1991; Tanaka, 1996). Neurons respond preferentially to complex shapes. A large variety of visual stimuli have been shown to elicit enhanced responses in ITC neurons including faces, objects (including shapes such as paperclips), natural images, but also artificial shapes and fractal patterns (Desimone et al., 1984; Hung et al., 2005; Logothetis and Sheinberg, 1996; Tanaka, 1996). The most parsimonious explanation of this apparently bewildering complexity in neuronal preferences seems to be that neurons may be tuned to complex parametric shape features that are present in many of these shapes but are not defined by the arbitrary choices made by the investigators. We illustrate this possibility in figure 17.2 by comparing neuronal responses recorded in monkey ITC with the responses of a simulated neuron that

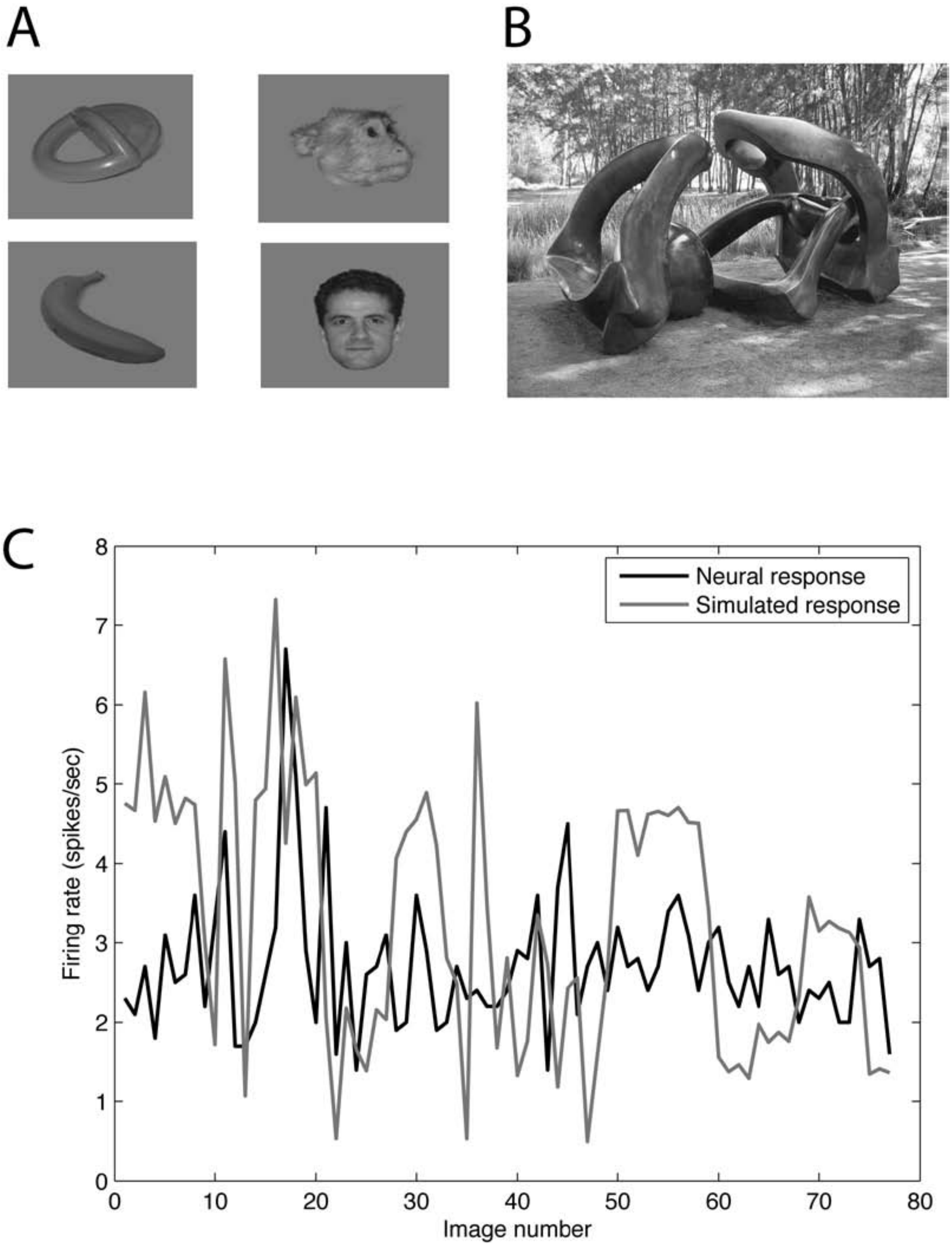

Figure 17.2
Responses of a simulated neuron, artificially tuned to prefer images similar to a photograph of a Henry Moore sculpture, show similar variability to the multiunit activity recorded in monkey area inferior temporal cortex in response to the same images. (A) Four of the 77 grayscale images presented to the

was tuned to prefer one particular complex and arbitrary shape, that of a sculpture by English artist Henry Moore. In this toy example, the variability across different images in the actual neuronal responses from ITC is comparable to the corresponding variability for the artificial sculpture-tuned unit.

A particularly interesting aspect of the ITC responses is that these neurons are often somewhat tolerant to perturbations of the stimuli: a neuron responding to a face, for instance, could respond similarly if the face is presented at different scales or positions (Hung et al., 2005; Ito et al., 1995). Investigators have examined the degree of tolerance in ITC responses to changes in scale, position, rotation, illumination, clutter, occlusion, color, and many other transformations (Logothetis and Sheinberg, 1996; Tanaka, 1996). Maintaining selectivity in the presence of object transformations is arguably one of the key challenges that the progression of computations along ventral visual cortex needs to solve. The responses of ITC neurons are strongly affected by contextual influences, including task demands and attention.

Models of Neurons and Networks of Neurons

A fundamental goal of science is to be able to formulate quantitative and predictive theories that explain the observed phenomena. The accumulation of empirical observations about neuroanatomical connectivity and neurophysiology beg for a theoretical formulation to account for the selective, robust, and rapid aspects of visual recognition. Quantitative models force us to formalize the assumptions and the hypothesis in the experiments. They can also provide quantitative bounds and constraints that can change the interpretation of the problem. Good models can integrate and summarize observations across different experiments, across different spatial and temporal resolutions, across different laboratories. Additionally, a good model can lead to nonintuitive experimental predictions. The models need to be guided and constrained by experimental findings and they can also inspire new experiments and novel ways of thinking about old problems. The model can also point to important missing data or critical information. Finally, quantitative models implemented through simulations can be useful from an engineering viewpoint. A machine that could solve visual recognition at human performance levels would find

Figure 17.2 (continued)
monkey (Hung et al., 2005) and to the simulated neuron. (B) The photograph used as the "preferred stimulus" for the simulated neuron. The simulated neuron's response to each image was chosen to be $A\exp\left(-d^2/\sigma^2\right)$, where d was the pixel-by-pixel Euclidian distance between the presented image and the sculpture photograph. σ and A were chosen to set the response strengths in the correct range. (C) Average multiunit activity recorded from a single electrode in inferior temporal cortex in response to each of the 77 images, counting spikes from 100 to 300 ms after stimulus presentation (black squares) and response of the simulated neuron to the same images (gray circles).

nearly infinite uses. Here we provide a succinct overview of some of the theoretical efforts to explain visual recognition and some of the computational models that have been implemented based on these theories.

One of the first questions to consider when developing a computational model pertains to the level of abstraction to represent neurons or the basic units in the model. From the extremely large to the very small, one could start by considering "boxes" that represent the average activity over seconds and over millions of neurons (see, for example, chapter 16). This type of box model matches the poor spatial and temporal resolution of current noninvasive neuroimaging techniques. At the other end of the spectrum, some computational efforts aim to consider the detailed biophysics of individual neurons (Markram, 2006). Detailed biophysical models have provided fundamental insights about the computations performed by individual neurons. However, it is not easy to scale up to large networks, a process that requires the use of supercomputers, many free parameters, and extensive simulations. Moreover, it is not clear what aspects of neuronal biophysics are central to understand a complex cognitive process such as object recognition (Anderson and Kreiman, 2011). Do we need to incorporate the detailed 3D geometry of every neuron? Do models need to consider the distribution of synapses along each dendrite? Do models need to incorporate the wide variety of different types of interneurons? How about the 3D shape of every protein within the neurons?

Single Neuron Models

In between "box models" and highly realistic simulations, several investigators have developed different models of single neurons (Koch, 1999; Gabbiani and Cox 2010). *Filter models* assume that each neuron is performing a filter operation on its input and typically provide a scalar output that is often interpreted as a firing rate. Moving up in complexity, a simple and widely used model of single neurons that incorporates dynamics and produces spike outputs is the *integrate-and-fire model*. The simplest instantiation of this model is equivalent to an RC circuit. The model contains a capacitor (C) and a leak resistance (R). The circuit integrates the incoming input current ($I(t)$). Whenever the voltage ($V(t)$) reaches a threshold, a spike is generated and the voltage is reset to 0. The subthreshold voltage dynamics are governed by:

$$C\frac{dV}{dt} = -\frac{V(t)}{R} + I(t) \tag{17.3}$$

This model has been extensively studied and there are several variations including adding a refractory period and variable thresholds. Still more detailed is the *Hodgkin-Huxley model* (Hodgkin and Huxley, 1952). This model describes voltage-dependent ion currents into and out of the neuron and how these lead to action potentials. Hodgkin and Huxley provided a nice example of the integration of powerful empirical measurements and quantitative work. Their models continue to be extensively used. The most detailed models of individual neurons incorporate the idea of multiple compartments and typically separate the axon, soma, and dendrites (even more detailed models can have multiple compartments and try to incorporate more realistic geometries). Simpler models are faster and, in some cases, have analytical solutions. More complex models quickly run into regimes that require simulations and increasing computational costs.

Models of Individual Brain Areas

One of the tests of a theory of visual recognition involves asking how well we can predict neuronal responses throughout the visual system. Eventually, we would like to be able to take an arbitrary visual stimulus and predict the elicited response of neurons at every level of the visual system. How close are we to being able to predict neuronal responses through modeling? Phenomenological models have been proposed to fairly accurately predict the responses of certain types of retinal ganglion cells as well as LGN cells. We have models that can predict the responses of some V1 cells (e.g., Keat et al., 2001), but even here there are many aspects of the responses that are not well understood (Carandini et al., 2005). When we characterize the responses of V1 neurons to simple stimuli, those results do not generalize well to allow us to predict the responses to complex natural stimuli. In particular, the role of lateral connections and feedback from other visual areas is poorly understood. As emphasized earlier, at levels V2 and higher, we suspect that cells' preferred stimuli become ever more complex, but we lack a way to systematically probe these preferred stimuli (see chapter 7 for an example of prediction of neuronal responses outside V1).

Computational Models of the Visual System

Researchers in the field of machine learning have been working for half a century to build computer programs that are capable of performing visual tasks such as object recognition. The difficulties they have encountered, and the strategies they have developed to overcome these difficulties, are of interest to neuroscientists because the brain itself must solve some of the same problems. Conversely, biophysically inspired models of visual cortex can also inspire and help develop machines that can generalize and perform complex visual recognition tasks.

One of the early approaches to machine object recognition was to implement a "brute force" template matching approach. Imagine that our task is to recognize a handwritten letter on a piece of paper. We do not know the exact position, size, or shape of the letter. We can try a template-matching approach where we sequentially examine each letter from A to Z (uppercase and lowercase, and perhaps even in different fonts). Because we do not know the position, we can scan the entire paper by shifting the template. Because we do not know the exact scale, we can try different scales (for each font, letter, and position). Given 26 letters, 4 possible fonts, a 600 × 800 pixel position matrix scanned every 5 pixels, and 10 possible scales, we have to make about 20 million comparisons. This does not include many of the possible sources of variation for each letter. This approach requires a large storage space for each object, there is no extrapolation and no intelligent learning, and we need to learn about each object in each possible rendering. Consider recognizing a face under different possible sizes, positions, colors, illuminations, rotations, gestures, makeup, beard, and so on. The problem is that any object can cast an infinite number of projections on the retina.

Several strategies have been proposed to overcome the challenges in the "brute force" approach to vision. The different models can be generically described as neural networks consisting of layers of artificial "neurons," with connections between neurons in adjacent layers. Inputs to the networks are in the form of patterns of activation of the neurons in the first layer. In a network for visual recognition, the activity of each first-layer neuron might represent the value of a single pixel in an image to be identified or categorized. The activity of each second-layer neuron is then determined by the joint activity of all the first-layer neurons to which it is connected, and in this way an input pattern propagates through the network. Additionally, there could be back-projections as well as recurrent connections within each layer. The strengths of the individual neuronal connections determine the computations performed by the overall network. For example, in neural networks for visual categorization, the goal is to transform input patterns so that those belonging to different categories can be more easily separated. Typically, this involves a series of nonlinear calculations that eventually enable the transformed patterns to be separated by a simple linear classifier. As one of the first examples of work in the field of neural networks, the *perceptron* is a type of artificial neural network proposed in 1958 (Bishop, 1995). It is composed of two layers of binary artificial neurons with unidirectional connections between the layers. The perceptron could be trained to perform classification tasks, but it worked only in the simplest cases, where the inputs were already linearly separable. However, two later modifications allowed the perceptrons to perform nonlinear classification. First, the network was expanded to more than two layers. Second, the artificial neurons were made to respond as a nonlinear function of their inputs. Frequently, the effect of the nonlinear calculations is to make important discriminative features become more explicit in the transformed patterns (for reviews on computational models of visual recognition, see Bishop, 1995; Deco and Rolls, 2004a; Riesenhuber and Poggio, 2000; Serre et al., 2007).

The "neocognitron" was proposed in 1980 by Fukushima (1980). Like models that had come before, it consists of a multilayered hierarchical neural network designed for visual pattern recognition. The neocognitron's key innovation was its explicit incorporation of alternating layers that were designed to produce invariance to small translations in the input stimulus. The architecture of the network was inspired by the neurophysiological studies of Hubel and Wiesel. In parallel to the "simple" and "complex" cells described in primary visual cortex, the model consisted of "S" and "C" units. The "C" units perform an "OR" operation over a local set of "S" units with identical tuning to provide increased robustness to position changes. Extending the neurophysiology in V1, the model alternated "S" and "C" units throughout a multilayered hierarchy. The neocognitron was able to classify simple digits and characters even when the inputs were slightly distorted.

The neocognitron was but the first example of the class of neural network models called "convolutional networks." These networks share three architectural concepts that make them ideal for visual pattern recognition. First, cells in convolutional networks have "local receptive fields," which means that their responses are determined only by stimulus features in some small and connected region of space. Second, cells in convolutional networks have "shared weights": the network is trained so that each cell in a specific position has many counterpart cells in different positions that all respond identically to identical but spatially translated stimuli. A set of cells that all share the same stimulus selectivities forms a "feature map." Third, convolutional networks include spatial or temporal subsampling. This subsampling allows the network to combine closely related inputs to produce invariance, as with the complex cell layers in the neocognitron. Not all multilayer feedforward networks for visual recognition can be characterized as convolutional: some have feature extractors that are not describable in terms of a convolution kernel. For instance, some networks calculate histogram type features, which are useful for texture representation (LeCun et al., 1998).

The use of neural networks for pattern recognition exploded after the advent of a network training method called backpropagation. Backpropagation is a general algorithm that trains multilayer networks so as to minimize an error function such as pattern classification error. Convolutional networks trained with backpropagation have been quite successful in visual recognition. However, several models of the visual system have used methods other than backpropagation to train the networks. For instance, models designed to explore properties of more biologically realistic systems frequently hard-wire the early layers of a network so that their responses resemble those of the early visual system, typically by using Gabor filters. Higher layers may also be set manually: for instance, Serre et al. (2007) chose the feature maps for their higher layers by choosing portions of images seen during training to use as templates for a convolutional filter. A machine learning classifier (e.g., Bishop, 1995; Vapnik, 1995) can be used as a final classification layer acting on the output of the hierarchical network (see chapters 18 and 19 for more information about machine learning). Interestingly, the performance of this biologically inspired architecture is comparable to the performance of computer vision approaches that are not guided or constrained by neurobiological principles.

All of the models described so far are purely feedforward: information flows in a single direction from input to output (see also the discussion in chapter 1). However, a number of authors have proposed visual system models that incorporate feedback. With a feedback model, information from higher layers can influence the activity of neurons at lower layers. This higher-layer information might include preliminary classifications or prior expectations, each of which could help with the interpretation of ambiguous low-level signals (Mumford, 1992; Rao, 2005a). Indeed, image recognition can be viewed as a Bayesian inference problem, and networks have been designed that probabilistically combine feedforward and feedback information to compute the most probable interpretation of the data (Lee and Mumford, 2003; Rao, 2005b; Chikkerur et al 2009).

Computational models today can perform very well on relatively simple tasks such as character recognition. On one popular test dataset of handwritten digits, networks can correctly identify more than 99 percent of the characters. However, recognition of natural images is considerably more difficult; for example, state-of-the-art performance on the CALTECH 101 database is only around 80 percent correct (Mutch and Lowe, 2006). Current models can also require a very large number of examples during training. Their performance also degrades rapidly in the presence of clutter or occlusion. No current models begin to approach the abilities of the human visual system yet!

As we begin to apply computational models of pattern recognition to the biological visual system, we need to evaluate them using different criteria. Are the mechanisms they describe biologically plausible? Do the models make falsifiable predictions? Biology offers tight constraints, and understanding these may help us exclude certain types of models. On the other hand, the primate visual system is the product of millions of years of evolution. It is conceivable that the type of solution to the visual recognition problem implemented by the ventral visual cortex is a highly efficient and accurate one. Computer vision algorithms may benefit also from an understanding of the neuronal circuitry involved in biological vision.

Perhaps the most difficult constraint is that of speed: multiple experimental protocols show that visual recognition occurs incredibly quickly, within 100–150ms after presentation of a visual stimulus. After this short period, scalp EEG signals in human cortex can correlate with recognition in a complex task (Thorpe et al., 1996) and neural activity in IT is selective for complex shapes (Hung et al., 2005; Liu et al., 2009). Such fast processing allows us to process a large amount of visual input very quickly. These times sharply constrain the number of computational steps that the brain could be using for initial recognition (Oram and Perrett, 1992; Serre et al., 2007; Thorpe, Fize, and Marlot, 1996). Of course, the initial "fast" recognition is not the entire story. With more processing time, human performance at recognizing images is much improved. This is unsurprising because, given many seconds, people can move their eyes, shift attention, recall information, and compare different parts of an image. Much research to date has attempted to reduce the influence of these complicating factors by focusing on the fast initial stages of recognition.

A Final Word

As emphasized at the beginning, this chapter does not pretend to provide a comprehensive account of the visual system (how could it anyway?). Studying the visual system is a highly active area of research that involves multidisciplinary approaches including computational and theoretical modeling, neurophysiological recordings, functional neuroimaging, cognitive psychology, neurology, and neuroanatomy, among many others. We hope that aficionados in this field will forgive the highly succinct nature of this chapter and the multiple omissions of large fields of research. At the same time, we naively hope that newcomers will share our enthusiasm and we encourage them to read further and, eventually, to contribute to the field.

References

Allison T, Puce A, Spencer D, McCarthy G. 1999. Electrophysiological studies of human face perception. I: Potentials generated in occipitotemporal cortex by face and non-face stimuli. *Cereb Cortex* 9: 415–430.

Allman J, Miezin F, McGuinness E. 1985. Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annu Rev Neurosci* 8: 407–430.

Anderson WS, Kreiman G. 2011. What we cannot model, we do not understand. *Curr Biol* 21: R123–R125.

Angelucci A, Bressloff PC. 2006. Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate V1 neurons. *Prog Brain Res* 154: 93–120.

Biederman I. 1987. Recognition-by-components: A theory of human image understanding. *Psychol Rev* 24: 115–147.

Binzegger T, Douglas RJ, Martin KA. 2004. A quantitative map of the circuit of cat primary visual cortex. *J Neurosci* 24: 8441–8453.

Bishop CM. 1995. *Neural networks for pattern recognition*. Oxford: Clarendon Press.

Blumberg J, Kreiman G. 2010. How cortical neurons help us see: visual recognition in the human brain. *J. Clin Invest* 120: 3054–3063.

Born RT, Bradley DC. 2005. Structure and function of visual area MT. *Annu Rev Neurosci* 28: 157–189.

Burkhalter A, Van Essen DC. 1986. Processing of color, form and disparity information in visual areas VP and V2 of ventral extrastriate cortex in the macaque monkey. *J Neurosci* 6: 2327–2351.

Q

Cadieu C, Kouh M, Pasupathy A, Connor C, Riesenhuber M, Poggio T. 2007. A model of V4 shape selectivity and invariance. *J Neurophysiol* 98: 1733–1750.

Callaway EM. 2004. Feedforward, feedback and inhibitory connections in primate visual cortex. *Neural Netw* 17: 625–632.

Carandini M, Demb JB, Mante V, Tolhurst DJ, Dan Y, Olshausen BA, Gallant JL, Rust NC. 2005. Do we know what the early visual system does? *J Neurosci* 25: 10577–10597.

Chikkerur S, Serre T, Tan C, Poggio T. 2010. What is where: A Bayesian inference theory of attention *Vis Res* 50: 2233–2247.

Connor CE, Brincat SL, Pasupathy A. 2007. Transformation of shape information in the ventral pathway. *Curr Opin Neurobiol* 17: 140–147.

David SV, Hayden BY, Gallant JL. 2006. Spectral receptive field properties explain shape selectivity in area V4. *J Neurophysiol* 96: 3492–3505.

Dayan P, Abbott L. 2001. *Theoretical neuroscience*. Cambridge, MA: MIT Press.

Deco G, Rolls ET. 2004a. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Res* 44: 621–642.

Deco G, Rolls ET. 2004b. *Computational neuroscience of vision*. Oxford: Oxford University Press.

Desimone R, Albright T, Gross C, Bruce C. 1984. Stimulus-selective properties of inferior temporal neurons in the macaque. *J Neurosci* 4: 2051–2062.

Desimone R, Duncan J. 1995. Neural mechanisms of selective visual attention. *Annu Rev Neurosci* 18: 193–222.

Desimone R, Schein SJ. 1987. Visual properties of neurons in area V4 of the macaque: sensitivity to stimulus form. *J Neurophysiol* 57: 835–868.

DiCarlo J, Maunsell J. 2004. Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *J Neurophysiol* 89: 3264–3278.

Douglas RJ, Martin KA. 2004. Neuronal circuits of the neocortex. *Annu Rev Neurosci* 27: 419–451.

Engel AK, Moll CK, Fried I, Ojemann GA. 2005. Invasive recordings from the human brain: clinical insights and beyond. *Nat Rev Neurosci* 6: 35–47.

Felleman DJ, Van Essen DC. 1991. Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex* 1: 1–47.

Fukushima K. 1980. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern* 36: 193–202.

Gabbiani F, Cox SJ. 2010. *Mathematics for Neuroscientists*. London, UK Academic Press.

Gattass R, Gross CG, Sandell JH. 1981. Visual topography of V2 in the macaque. *J Comp Neurol* 201: 519–539.

Gross CG. 1994. How inferior temporal cortex became a visual area. *Cereb Cortex* 5: 455–469.

Haxby J, Grady C, Horwitz B, Ungerleider L, Mishkin M, Carson R, Herscovitch P, Schapiro M, Rapoport S. 1991. Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proc Natl Acad Sci USA* 88: 1621–1625.

Hegde J, Van Essen DC. 2003. Strategies of shape representation in macaque visual area V2. *Vis Neurosci* 20: 313–328.

Hodgkin AL, Huxley AF. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol* 117: 500–544.

Hubel D, Wiesel T. 1959. Receptive fields of single neurons in the cat's striate cortex. *J Physiol* 148: 574–591.

Hubel DH, Wiesel TN. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol* 160: 106–154.

Humphreys G, Riddoch M. 1993. Object agnosias. *Baillieres Clin Neurol* 2: 339–359.

Hung C, Kreiman G, Poggio T, DiCarlo J. 2005. Fast read-out of object identity from macaque inferior temporal cortex. *Science* 310: 863–866.

Ito M, Komatsu H. 2004. Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *J Neurosci* 24: 3313–3324.

Ito M, Tamura H, Fujita I, Tanaka K. 1995. Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J Neurophysiol* 73: 218–226.

Keat J, Reinagel P, Reid RC, Meister M. 2001. Predicting every spike: a model for the responses of visual neurons. *Neuron* 30: 803–817.

Koch C. 1999. *Biophysics of computation*. New York: Oxford University Press.

Koch C. 2005. *The quest for consciousness*. Los Angeles: Roberts & Company.

Kreiman G. 2004. Neural coding: computational and biophysical perspectives. *Phys Life Rev* 2: 71–102.

Kreiman G. 2007. Single neuron approaches to human vision and memories. *Curr Opin Neurobiol* 17: 471–475.

LeCun Y, Bottou L, Bengio Y, Haffner P. 1998. Gradient-based learning applied to document recognition. *Proc IEEE* 86: 2278–2324.

Lee TS, Mumford D. 2003. Hierarchical Bayesian inference in the visual cortex. *J Opt Soc Am A Opt Image Sci Vis* 20: 1434–1448.

Liu H, Agam Y, Madsen JR, Kreiman G. 2009. Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron* 62: 281–290.

Logothetis NK, Sheinberg DL. 1996. Visual object recognition. *Annu Rev Neurosci* 19: 577–621.

Markram H. 2006. The blue brain project. *Nat Rev Neurosci* 7: 153–160.

Mishkin M. 1982. A memory system in the monkey. *Philos Trans Roy Soc Lond Series B* 298: 85.

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

Mumford D. 1992. On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol Cybern* 66: 241–251.

Mutch J, Lowe D. 2006. Multiclass object recognition with sparse, localized features. In *Proc CVPR'06*, 11–18. New York: IEEE.

Nassi JJ, Callaway EM. 2009. Parallel processing strategies of the primate visual system. *Nat Rev Neurosci* 10: 360–372.

Oram MW, Perrett DI. 1992. Time course of neural responses discriminating different views of the face and head. *J Neurophysiol* 68: 70–84.

Pasupathy A, Connor CE. 2001. Shape representation in area V4: position-specific tuning for boundary conformation. *J Neurophysiol* 86: 2505–2519.

Peterhans E, von der Heydt R. 1991. Subjective contours—bridging the gap between psychophysics and physiology. *Trends Neurosci* 14: 112–119.

Rao RP. 2005a. Bayesian inference and attentional modulation in the visual cortex. *Neuroreport* 16: 1843–1848.

Rao RPN. 2005b. Bayesian inference and attentional modulation in the visual cortex. *Neuroreport* 16: 1843–1848.

Reynolds JH, Chelazzi L. 2004. Attentional modulation of visual processing. *Annu Rev Neurosci* 27: 611–647.

Riesenhuber M, Poggio T. 1999. Hierarchical models of object recognition in cortex. *Nat Neurosci* 2: 1019–1025.

Riesenhuber M, Poggio T. 2000. Models of object recognition. *Nat Neurosci* 3(Suppl): 1199–1204.

Ringach DL, Hawken MJ, Shapley R. 1997. Dynamics of orientation tuning in macaque primary visual cortex. *Nature* 387: 281–284.

Rockland KS, Pandya DN. 1979. Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Res* 179: 3–20.

Rolls E. 1991. Neural organization of higher visual functions. *Curr Opin Neurobiol* 1: 274–278.

Q

Salin PA, Bullier J. 1995. Corticocortical connections in the visual system: structure and function. *Physiol Rev* 75: 107–154.

Serre T, Kreiman G, Kouh M, Cadieu C, Knoblich U, Poggio T. 2007. A quantitative theory of immediate visual recognition. *Prog Brain Res* 165C: 33–56.

Tanaka K. 1996. Inferotemporal cortex and object vision. *Annu Rev Neurosci* 19: 109–139.

Thorpe S, Fize D, Marlot C. 1996. Speed of processing in the human visual system. *Nature* 381: 520–522.

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

Vapnik V. 1995. *The nature of statistical learning theory*. New York: Springer.

von der Heydt R, Friedman HS, Zhou H. 1999. The neural representation of color stimuli during perceptual filling-in. *Invest Ophthalmol Vis Sci* S639.

von der Heydt R, Peterhans E, Baumgartner G. 1984. Illusory contours and cortical neuron responses. *Science* 224: 1260–1262.

Wandell BA. 1995. *Foundations of vision*. Sunderland, MA: Sinauer Associates.

Wu MC, David SV, Gallant JL. 2006. Complete functional characterization of sensory neurons by system identification. *Annu Rev Neurosci* 29: 477–505.

Zeki S. 1983. Color coding in the cerebral cortex—The reaction of cells in monkey visual cortex to wavelengths and colors. *Neuroscience* 9: 741–765.