111 How the Visual Cortex Recognizes Objects: The Tale of the Standard Model

MAXIMILIAN RIESENHUBER AND TOMASO POGGIO

OBJECT RECOGNITION is fundamental to the behavior of higher primates. It is also the most remarkable achievement of visual cortex and one that probably influences greatly its overall functional architecture.

Object recognition is a computationally hard problem. The visual system rapidly and effortlessly recognizes a large number of diverse objects in cluttered natural scenes. Computer vision systems are still far from passing the Turing test for vision, that is to "see" as well as human beings do. Very recent work in computer vision, however, has achieved impressive results on specific tasks such as object detection/categorization and face identification, suggesting that the problem is solvable with well-understood modern statistical learning algorithms.

The problem of object recognition is even more difficult from the point of view of neuroscience since it involves several levels of understanding, from the information processing or computational level to the level of circuits and of cellular and biophysical mechanisms. After decades of work in striate and extrastriate areas that have produced a significant and rapidly increasing amount of data, the emerging picture of how cortex performs object recognition may be becoming too complex for physiologists (as David Hilbert famously said about quantum mechanics: "Physics is too difficult for physicists"). Now, paralleling the very recent developments in bioinformatics, computational approaches may be riding to the rescue.

In this chapter, we attempt to take a first step. We review progress in the field. We do so by using a computational model as a tool to summarize, organize, and interpret existing data and discuss open questions. We first sketch the basic facts that have been established during the past three decades and that are now broadly accepted. We then describe a Standard Model that emerges from the data and represents in its basic architecture the average belief—often implicit—of many visual physiologists. In this sense, it is definitely not our model. The broad form of the model is suggested by the basic facts; we have made it quantitative and thereby predictive (through computer simulations). From the precise perspective it provides, we will discuss recent data, new experiments, and important open questions. We will also discuss specific alternative hypotheses about modules and mechanisms of the model, as well as evidence that may possibly falsify the whole class of interpretations associated with the Standard Model.

Basic facts

Object recognition in cortex is thought to be mediated by the ventral visual pathway (Ungerleider and Haxby, 1994) running from primary visual cortex, V1, over extrastriate visual areas V2 and V4, to inferotemporal cortex IT. Based on physiological experiments in monkeys, IT has been postulated to play a central role in object recognition. IT, in turn, is a major source of input to prefrontal cortex (PFC), "the center of cognitive control" (Miller, 2000) involved in linking perception to memory.

Over the past decade, several physiological studies in nonhuman primates have established a core of basic facts about cortical mechanisms of recognition that seem to be widely accepted and that confirm and refine older data from neuropsychology. A brief summary of this consensus knowledge begins with the groundbreaking work of Hubel and Wiesel first in the cat (1962, 1965) and then in the macaque (1968). Starting from *simple cells* in primary visual cortex, V1, with small receptive fields that respond preferably to oriented bars, neurons along the ventral stream (Logothetis and Sheinberg, 1996; Perrett and Oram, 1993; Tanaka, 1996) show an increase in receptive field size as well as in the complexity of their preferred stimuli (Kobatake and Tanaka, 1994). At the top of the ventral stream, in anterior inferotemporal cortex (AIT), cells are tuned to complex stimuli such as faces (Desimone, 1991; Desimone et al., 1984; Gross et al., 1972; Perrett et al., 1992). The tuning of the view-tuned and object-tuned cells in AIT depends on visual experience as shown by Logothetis et al. (1995) and supported by Booth and Rolls (1998), DiCarlo and Maunsell (2000), and Kobatake et al. (1998). A hallmark of these IT cells is the robustness of their firing to stimulus transformations such as scale and position changes (Logothetis et al., 1995; Logothetis and Sheinberg, 1996; Perrett and Oram, 1993; Tanaka, 1996). In addition, as other studies have shown (Booth and Rolls, 1998; Hietanen et al., 1992; Logothetis et al., 1995; Perrett and Oram, 1993), most

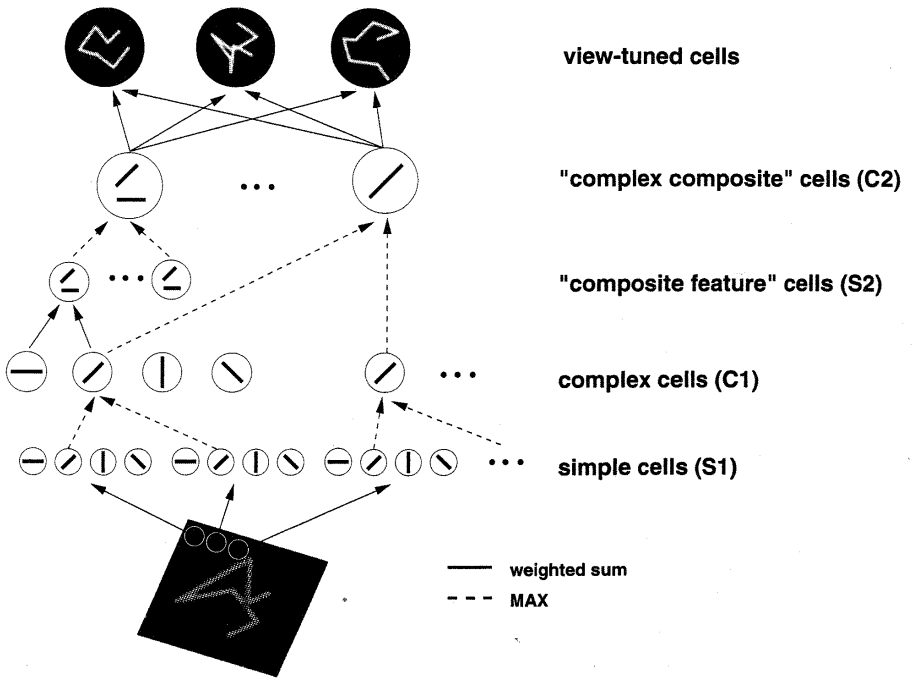

FIGURE 111.1. The figure shows the first part of the Standard Model. It extends several recent models (see especially Fukushima, 1980; see also Perrett and Oram, 1993; Poggio and Edelman, 1990; Riesenhuber and Poggio, 1999b; Wallis and Rolls, 1997). The view-based module, HMAX, shown here is an hierarchical extension of the classical paradigm (Hubel and Wiesel, 1962) of building complex cells from simple cells. The circuitry consists of a hierarchy of layers leading to greater specificity and greater invariance by using two different types of pooling mechanisms. The first layer in V1 represents linear oriented filters followed by input normalization, similar to simple cells (Carandini et al., 1997); each unit in the next layer (C1) pools the outputs of simple cells with the same orientation but at slightly different positions (scales) by using a maximum operation (see text and Riesenhuber and Poggio, 1999b). Each of these units is still orientation selective but more invariant to position (scale), similar to some complex cells. In the next stage, signals from complex cells with different orientations but similar positions are combined (in a weighted sum) to create neurons tuned to a dictionary of more complex features. The next layer (C2) is similar to the C1 cells: by pooling together signals from S2 cells of the same type but at slightly different positions (and scales), the C2 units become more invariant to position (and scale) but preserve feature selectivity. They may correspond roughly to V4 cells. In the model, the C2 cells feed into view-tuned cells, with connection weights that are learned fom exposure to a view of an object. There may be more levels in the hierarchy than are shown in the figure after the C2 layer. The output of the view-based module is represented by view-tuned model units that exhibit tight tuning to rotation in depth (as well as illumination and other object-dependent transformations such as facial expression) but are tolerant to scaling and translation of their preferred object view. Notice that the cells labeled here as view-tuned units encompass, between PIT and AIT, a spectrum of tuning from full views to components or complex features: depending on the synaptic weights determined during learning, each view-tuned cell becomes effectively connected to all or only a few of the units activated by the object view (Riesenhuber and Poggio, 1999a).

neurons show specificity for a certain object view or lighting condition. In particular, Logothetis et al. (1995) trained monkeys to perform an object recognition task with isolated views of novel three-dimensional objects ("paper clips"; see the objects at the top of Fig. 111.1). When recording from the animals' IT, they found that the great majority of neurons selectively tuned to the training objects were view-tuned (with a half-width of about 20 degrees for rotation in depth) to one of the training objects [about one-tenth of the tuned neurons were view-invariant, in agreement with earlier predictions (Poggio and Edelman, 1990)], but an average translation invariance of 4 degrees (for typical stimulus sizes of 2 degrees) and an average scale invariance of 2 octaves (Riesenhuber and Poggio, 1999b). Thus, whereas view-invariant recognition requires visual experience of the specific novel object, position and scale invariance seems to be immediately present in the view-tuned neurons (Logothetis et al., 1995) without the need of visual experience for views *of the specific object* at different positions and scales. A very recent study (DiCarlo and Maunsell, 2003)—using different stimuli and training paradigm—reports translation invariance from one view of less than 3 degrees, pointing to a possible influence of training history and object shape on invariance ranges. Recent functional magnetic resonance imaging (fMRI) data have shown a similar pattern for the lateral occipital cortex (LOC), a brain region in human visual cortex central to object recognition and believed to be the homolog of monkey area IT (Grill-Spector et al., 2001;

Malach et al., 1995; Tanaka, 1997). Optical recordings in monkeys confirmed the view dependency of several face-tuned neurons (Wang et al., 1996).

A comment about the architecture is important: in its basic initial operation—akin to *immediate recognition*—the hierarchy is likely to be mainly feedforward (though local feedback loops almost certainly have key roles, e.g., possibly in performing a maximum-like pooling; see later) (Perrett and Oram, 1993). Event-related potential data (Thorpe et al., 1996) have shown that the process of object recognition appears to take remarkably little time, on the order of the latency of the ventral visual stream (Perrett et al., 1992), adding to earlier psychophysical studies using a rapid serial visual presentation (RSVP) paradigm (Intraub, 1981; Potter, 1975) that have found that subjects were still able to process images when they were presented as rapidly as eight per second.

In summary, the accumulated evidence points to six mostly accepted properties of the ventral stream architecture:

1. A hierarchical buildup of invariances first to position and scale and then to viewpoint and more complex transformations requiring the interpolation between several different object views
2. In parallel, an increasing size of the receptive fields
3. An increasing complexity of the optimal stimuli for the neurons
4. A basic feedforward processing of information (for "immediate" recognition tasks)
5. Plasticity and learning probably at all stages and certainly at the level of IT
6. Learning specific to an individual object is not required for scale and position invariance (over a restricted range)

The Standard Model

Invariance properties of IT neurons with respect to different transformations can be understood from a computational perspective (Riesenhuber and Poggio, 2000b): mathematically, the effects of two-dimensional (2D) affine transformations, such as scaling and translation in the image plane, can be estimated exactly from just one object view. There is no need then to, for example, collect examples of one object at all positions in the image to be able to generalize across positions from a single view. To determine the behavior of a specific object under transformations that depend on its three-dimensional (3D) shape, such as illumination changes or rotation in depth, however, one view generally is not sufficient. Unlike affine 2D transformations, 3D rotations, as well as illumination changes, usually require multiple example views during learning. Mathematically, the space of the orthographic views of one object is spanned by a single view for affine 2D transformations and by two or more for affine 3D transformations (Poggio, 1990; Ullman and Basri, 1991).

The basic facts summarized earlier, together with the above computational considerations, lead to a Standard Model likely to represent the simplest class of models reflecting the known anatomical and biological constraints.

The model reflects the general organization of visual cortex in a series of layers from V1 to IT to PFC. From the point of view of invariance properties, it consists of a sequence of two main modules based on two key ideas. The first module, shown schematically in Figure 111.1, leads to model units showing the same scale and position invariance properties as the view-tuned IT neurons of Logothetis et al. (1995) using the same stimuli (Fig. 111.3). This is not an independent prediction since the model parameters were chosen to fit Logothetis' data. It is, however, not obvious that an hierarchical architecture using plausible neural mechanisms could account for the measured invariance *and* selectivity. Computationally, this is accomplished by a scheme that can be best explained by taking striate complex cells as an example: invariance to changes in the position of an optimal stimulus (within a range) is obtained in the model by means of a *maximum operation (max)* performed on the simple cell inputs to the complex cells, where the strongest input determines the cell's output. Simple cell afferents to a complex cell are assumed to have the same preferred orientation with their receptive fields located at different positions. Taking the maximum over the simple cell afferent inputs provides position invariance while preserving feature specificity. The key idea is that the step of filtering followed by a max operation is equivalent to a powerful signal processing technique: select the peak of the correlation between the signal and a given matched filter, where the correlation is either over position or over scale. The model alternates layers of units combining simple filters into more complex ones—to increase pattern selectivity—with layers based on the max operation—to build invariance to position and scale while preserving pattern selectivity.

In the second part of the architecture, shown in Figure 111.2, learning from multiple examples, that is, different view-tuned neurons, leads to view-invariant units as well as to neural circuits performing specific tasks. The key idea here is that interpolation and generalization can be obtained by simple networks, similar to Gaussian Radial Basis Function networks (GRBF) (Poggio and Girosi, 1990), that learn from a set of examples, that is, input-output pairs. In this case, the inputs are views and the outputs are the parameters of interest such as the label of the object or its pose or expression (for a face). The GRBF network has a hidden unit for each example view, broadly tuned to the features of an example image (see also opde Beeck et al., 2001). The weights from the hidden units to the output are learned from the set of examples, that is, input-output pairs. In principle,

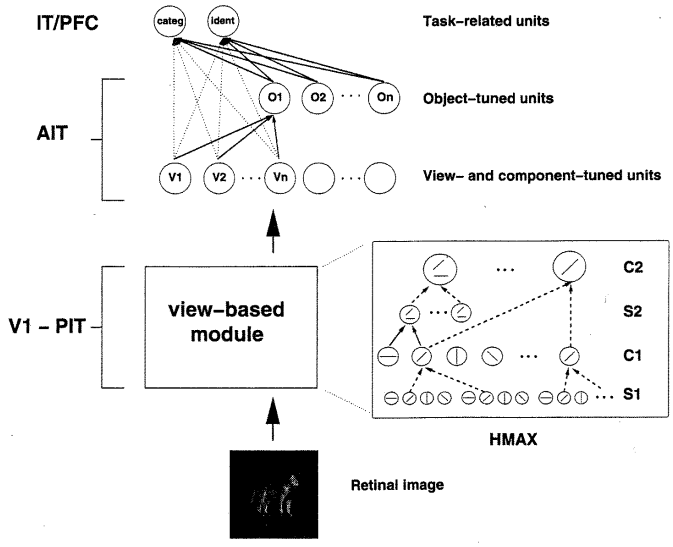

FIGURE 111.2. The figure shows the overall architecture of the Standard Model. The view-tuned module, described in Figure 111.1, is shown here in the inset (up to the C2 layer). The second part of the model starts with the view-tuned cells which represent the output of the view-based module. Invariance to rotation in depth is obtained by combining in a learning module several view-tuned units V_n tuned to different views of the same object (Poggio and Edelman, 1990), creating view-invariant (object-tuned) units O_n. These, as well as the view-tuned units, can then serve as input to task modules that learn to perform different visual tasks such as identification/discrimination or object categorization. They consist of the same generic learning circuitry (similar to an RBF network; see text) but are trained with appropriate sets of examples to perform specific tasks. In addition to the feedforward processing, there are likely feedback pathways (not shown) for top-down modulation of neuronal responses throughout the processing hierarchy and to support the learning phase. The stages up to the object-centered units probably encompass V1 to anterior IT (AIT). The last stage of task-dependent modules may be localized in AIT or PFC. All the units in the model represent single cells modeled as simplified neurons with modifiable synapses. (Modified from Riesenhuber and Poggio, 2000b.)

two networks sharing the same hidden units but with different weights (from the hidden units to the output unit) could be trained to perform different tasks, such as pose estimation or view-invariant recognition. Depending just on the set of training examples, learning networks of this type can learn to categorize across exemplars of a class (Riesenhuber and Poggio, 1999c), as well as to identify an object across different illuminations and different viewpoints. The demonstration (Poggio and Edelman, 1990) that a view-based GRBF model could achieve view-invariant object recognition in fact-motivated psychophysical experiments (Bülthoff and Edelman, 1992; Gauthier and Tarr, 1997). In turn, the psychophysics provided strong support for the view-based hypothesis against alternative theories (for a review, see Tarr and Bülthoff, 1998) and, together with the model, triggered the physiological work of Logothetis et al. (1995).

Thus the two key ideas in the model are (1) the max operation to provide invariance at several steps of the hierarchy and (2) the RBF-like learning network to learn a specific task based on a set of cells tuned to example views.

Interpreting experimental data

The Standard Model summarizing the basic facts about the ventral pathways predicts additional experimental results and provides interesting perspectives on still other data. For instance, the model accounts (see Riesenhuber and Poggio, 1999a, 1999c, 2000b) for the response of tuned IT cells to scrambled objects (Vogels, 1999) (see Fig. 5 in Riesenhuber and Poggio, 1999b), clutter (Missal et al., 1997) (Fig. 111.4), and mirror views (Logothetis and Sheinberg, 1996) (see Fig. 4 in Riesenhuber and Poggio, 1999b). It also shows a degree of performance roughly in agreement with physiological and psychophysical data in specific tasks (the same stimuli are used for simulations and experiments) such as the cat versus dog categorization task described by Freedman et al. (2001a) (Fig. 111.6), object identification (in separate experiments faces, cars, paper clips were used; see Figs. 111.5 and 111.7), gender classification, and possibly the face habituation effect of (Leopold et al., 2001).

A key function of models is to clarify key issues. Here we use the Standard Model to discuss several central topics concerning the neural mechanisms of object recognition.

SELECTIVITY Invariance is only one requirement for object recognition, the other one being selectivity. Several studies have established that IT neurons can become tuned to task-relevant objects and their full or partial (Vetter et al., 1995) views (DiCarlo and Maunsell, 2000; Kobatake et al., 1998; Logothetis et al., 1995) or to objects in the monkey's envi-

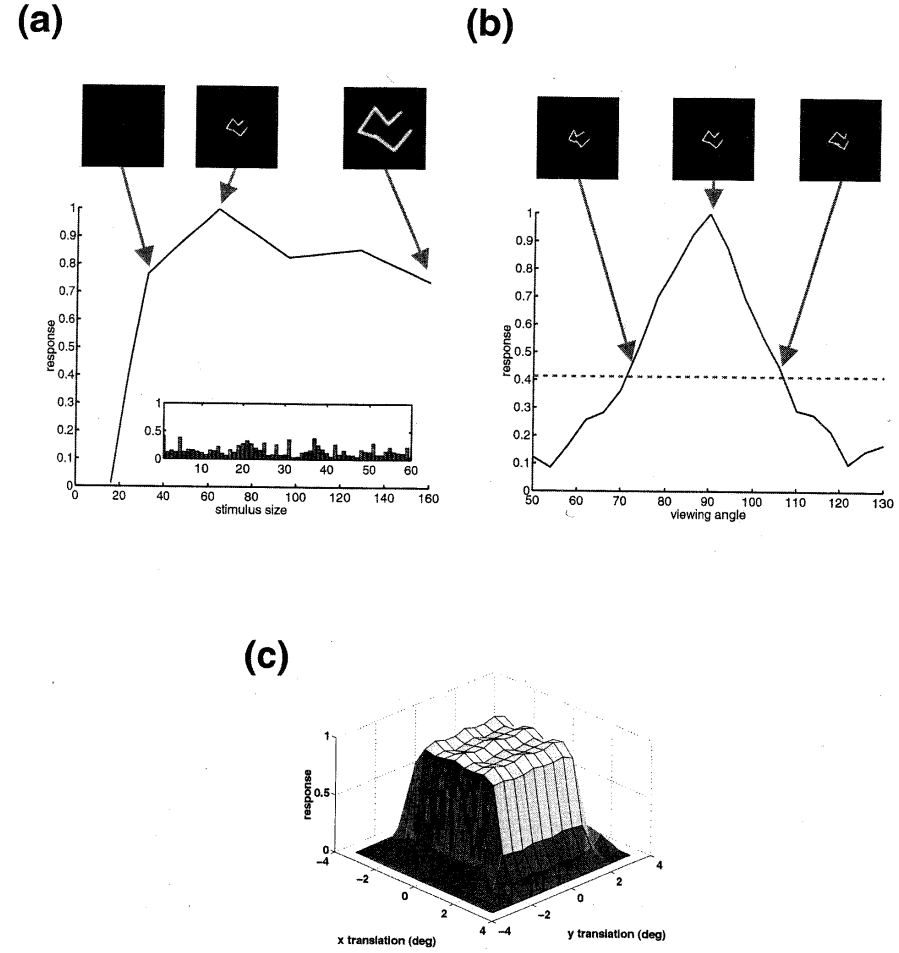

FIGURE 111.3. Responses of a sample model neuron to different transformations of its preferred stimulus. The different panels show the same neuron's response to (*a*) varying stimulus sizes [the inset shows the response to 60 distractor objects, selected randomly from the paper clips used in the physiology study (Logothetis et al., 1995)], (*b*) experiments, rotation in depth, and (*c*) translation. The training size was 64 × 64 pixels corresponding to 2 degrees of visual angle. The simulation results shown here are in quantitative agreement with recordings in IT by Logothetis et al. (1995) (see Fig. 8 in that article), using the same stimuli.

ronment (Booth and Rolls, 1998), suggesting that these neurons provide a representation of objects occurring in an animal's environment. The preferred stimuli of neurons in intermediate stages of the ventral stream are less clear, partly owing to the difficulty of knowing which stimuli to use to probe the neural selectivity. Reports of preferred features of neurons in V4, the visual area preceding IT in the ventral pathway, vary, depending on the set of stimuli used to probe responses, including Cartesian gratings (Desimone and Schein, 1987) polar and hyperbolic sinusoidal gratings (Gallant et al., 1996), and contour features (Pasupathy and Connor, 1999). In V2, a recent study (Hegde and Van Essen, 2000) has reported preferences to complex stimuli such as arcs, intersecting lines, and non-Cartesian gratings. Instead of probing neuronal tuning with a fixed set of stimuli, a set of studies (Fujita et al., 1992; Kobatake and Tanaka, 1994; Tanaka, 1993, 1996) used a heuristic "simplification procedure" in an effort to arrive at the features crucial to activate a neuron. In this approach, a complex natural stimulus (such as a face) to which the neuron under study responds is progressively "simplified" (e.g., by removing color or texture or by reducing complex shapes to simpler geometric primitives) in a way that preserves or increases the neuronal firing. The stimulus that cannot be simplified further without significantly decreasing the firing rate is then labeled the *effective stimulus* for that cell. A study using this paradigm (Kobatake and Tanaka, 1994) has reported an increase in feature complexity from area V2 to anterior IT. However, a recent IT optical imaging study (Tsunoda et al., 2001), supported by single-cell recordings, demonstrates the fundamental difficulty of determining a neuron's preferred feature in higher visual areas. The authors report that simplifying a stimulus often causes additional IT neurons to respond relative to the more complex stimulus. Interestingly, the Standard Model

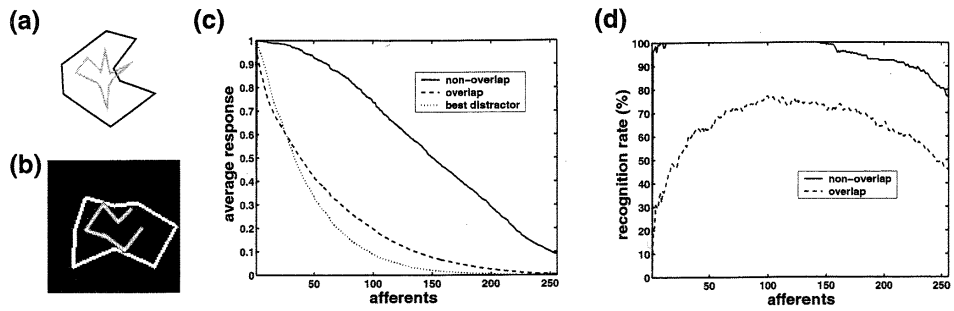

FIGURE 111.4. Recognition in clutter in the Standard Model. *a*, Example stimulus (light gray) and outline background (black) from the experiment by Missal et al. (1997) (redrawn from that source), in which a monkey was trained to recognize polygonal shapes in isolation and on a polygonal background. In the experiment, foreground and background stimuli were of different colors. The study found that while the monkeys' performance was only slightly impaired by the introduction of a background stimulus (from 98% to 89% correct), firing rates of IT neurons were affected much more strongly. *b*, Example stimulus for the corresponding experiment with the model, in which paper clip stimuli were superimposed on polygonal backgrounds (for details, see Riesenhuber and Poggio, 1999a). Unlike in the experiment (Missal et al., 1997), the stimuli and background were of the same color, increasing task difficulty. The foreground clip in *b* was correctly recognized by the corresponding model IT unit (which was the same as in Fig. 111.3). *c*, Average response levels (over 21 model IT units tuned to different paper clips, 10 runs each with different randomly selected backgrounds) for a nonoverlap condition (*solid line*), in which the background polygon and the foreground paper clip were in opposite corners of the model IT unit's receptive field, and an overlap condition (*dashed line*), in which stimuli overlapped, as shown in *b*, compared to the average response to the best distractor (*dotted line*), as a function of the number of afferents (ordered by decreasing connection strength) to each view-tuned model neuron. *d*, Average recognition rates in the two conditions, again plotted as a function of the number of afferents to each view-tuned unit. Decreasing the number of afferents increases robustness to clutter; see Riesenhuber and Poggio (1999a). Note that while firing rates are strongly affected, most stimuli can be recognized successfully, as in the experiment (Missal et al., 1997).

does in fact predict qualitatively what is observed—neurons tuned to a dictionary of features at different levels of complexity. Moreover, preliminary simulations (Knoblich and Riesenhuber, 2002) suggest that for IT model units the effect of the simplification procedure may well lead to the observations of Tsunoda et al. (2001). In any case, it is important to emphasize that features are defined not only by the specific image but also by the system looking at it. A pattern that seems simple to us may activate more filters in a vision system looking at it than another stimulus that is apparently more complex, depending on the filters used by the system. Thus the Standard Model can provide specific hypotheses to guide experiments regarding how more complex features (and ultimately object-tuned cells) are built from simpler ones.

REPRESENTATION Related to the question of neuronal tuning is the question of the precise nature of object representation in cortex. It has recently been put forward, based on a set of human fMRI studies, that some object classes—faces (Kanwisher et al., 1997), places (Epstein and Kanwisher, 1998), body parts (Downing et al., 2001)—are processed by distinct modules in cortex. Interestingly, another fMRI study (Haxby et al., 2001) has shown that objects of a certain class (e.g., faces) evoke a distributed pattern of activity that is not confined to the aforementioned specialized modules [e.g., the fusiform face area (FFA), a brain area that is strongly activated by face stimuli (Kanwisher et al., 1997)], and that the part of the activation pattern *outside* the specific module is sufficient for object categorization. Thus, some data appear to argue for a *modular* framework of object representation in cortex, where specific brain areas perform computations unique to the object class at hand, while others support a model in which the same computation is performed for different objects and represented in a distributed way. The latter claim is supported by the Standard Model, which also helps to reconcile the two sets of data. Model IT units have preferred C2 activation vectors that can be full or partial view of an object (Riesenhuber and Poggio, 1999a), depending on their connectivity. The distinction between *complex features* and *object* is largely semantic, since during training a cell can become tuned to a feature that is diagnostic for the object rather than to a full view (Pauls, 1997). What is relevant for object recognition is that the objects to be discriminated produce distinct activation patterns. From the point of view of the model, groups of neurons responding to representatives from the same object class do not have to be segregated but are expected to be interdigitated. Since the activity of one fMRI voxel is the average over typically hundreds of thousands of neurons,[1] strong activation of the FFA for faces would argue for a

[1]For cautionary notes about the interpretation of fMRI images see Logothetis et al. (2001).

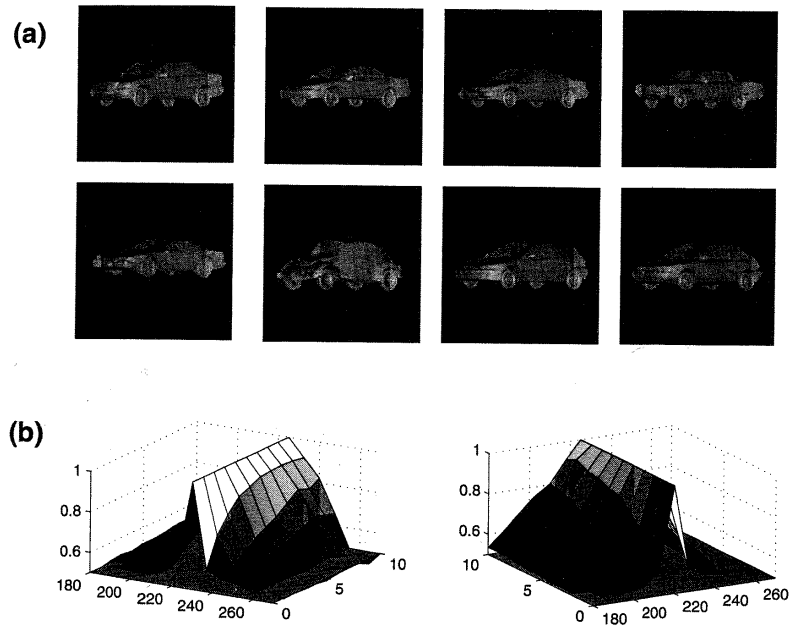

FIGURE 111.5. Recognition of non–paper clip objects (morphed cars) in HMAX. *a*, The eight prototype cars spanning the car morph space (Riesenhuber and Poggio, 2000a). *b*, Recognition performance of the model on the eight car morph space for a discrimination task in which first a target car was shown from a viewpoint φ_0 = 225 degrees. This car, presented at viewpoint φ (varying between 180 and 270 degrees), then had to be discriminated from a distractor car presented at the same viewpoint φ. The similarity of the target and distractor cars was controlled by varying the separation of their corresponding parameter vectors in morph space. In particular, the parameter vectors were chosen to lie on lines connecting two prototypes from *a*. The *x*-axis shows viewpoint φ of the nonmatch object, *y*-axis target/distractor morph distance *d* (in steps along the morph line that the sample object lies on, a distance of 10 corresponds to the easiest case of the two cars being different prototypes), and *z*-axis model discrimination performance for all (φ, *d*) stimulus pairs in the sample set. Stimulus identity was represented by a population code over 16 "car neurons." The two subplots show the same graph from two different viewpoints to show positive rotations (i.e., toward the front, so that the front of the car is turning toward the viewer, as used in the psychophysical experiments), left plot, and negative rotations (i.e., toward the back, so that the side of the car faces the viewer), right plot. Note that the invariance range for the class of car stimuli is comparable to the results obtained with the paper clip stimuli in psychophysics and modeling (Bülthoff and Edelman, 1992; Logothetis et al., 1994; Poggio and Edelman, 1990) (see Fig. 111.3), demonstrating the generality of the model (over different object classes and representations).

higher density of face neurons in that part of cortex, possibly reflecting the great cognitive importance of face neurons. However, subjects with great expertise for other object classes might show significant activation of parts of cortex for objects from their field of expertise. Indeed, in bird and car experts, brain areas, overlapping with but not limited to the FFA, have been found to be specifically activated by birds and cars, respectively (Gauthier et al., 2000).

Thus the Standard Model suggests the following object representation in cortex [for the appropriate computational simulations, see Riesenhuber and Poggio (1999c, 2000a)], a related proposal can be found in Edelman (1999). A particular object, say a specific face, will elicit different activity in the view-specific V_n and object-specific O_n cells of Figure 111.2 (an example of which is shown in Fig. 111.7). Thus, the memory of the particular face is represented in the identification circuit in an implicit way by a sparse population code through the activation pattern over the coarsely tuned V_n and O_n cells, generally without cells dedicated to represent individual objects (*grandmother cells*).[2] Discrimination, or memorization of specific objects, can then proceed by comparing activation patterns over the strongly activated object- or view-tuned units (Riesenhuber and Poggio, 2000a) tuned to a small number of "prototypical" faces (Young and Yamane, 1992). For a certain level of specificity, only the activations of a small number of units have to be stored, forming a sparse code—in contrast to activation patterns on lower levels, where units are less specific and hence activation patterns tend to involve more neurons. In a similar fashion, categorization neurons, located putatively in the PFC, can be trained (Riesenhuber and Poggio, 1999c) to receive input from relevant object-tuned units.

[2]For special cases in which a small, fixed number of objects had to be discriminated [as, for example, it Logothetis et al. (1995)], the representations found in the model during learning turned out to be more grandmother-like.

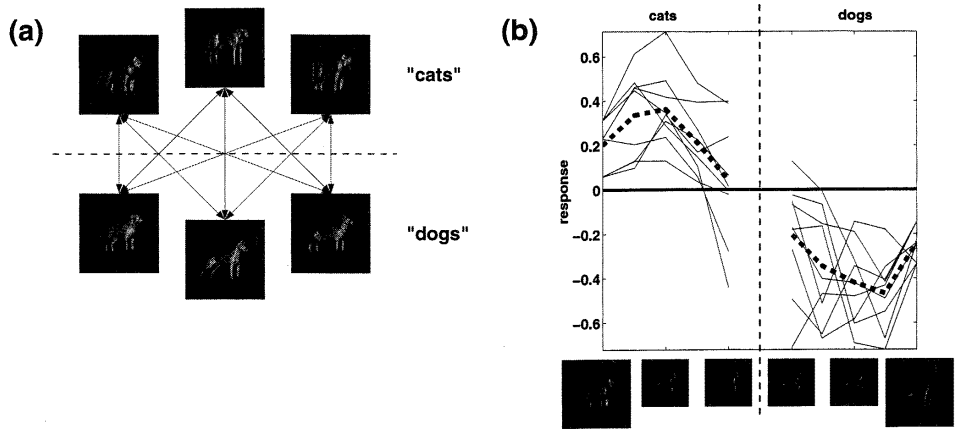

FIGURE 111.6. Example of categorization in the Standard Model. *a*, Illustration of the cat/dog stimulus space [used in the physiology experiments of Freedman et al. (2001a)]. The morph space is spanned by the pictures of three cats shown on top ("House Cat," "Cheetah," and "Tiger") and the three dogs below ("House Dog," "Doberman," and "German Shepherd"). All prototypes have been normalized with respect to viewing angle, lighting parameters, size, and color. *b*, Response of a cat/dog categorization unit along the nine class boundary-crossing morph lines (*thin lines*). The unit receives input from 144 view-tuned units tuned to cat/dog training stimuli, as used in Freedman et al. (2001a), and was trained to respond differently to cats and dogs (cf. Fig. 111.7). The dashed line shows the average over all morph lines. The solid horizontal line shows the class boundary in response space, and the dashed vertical line shows the category boundary in morph space the unit was trained on. The images on the bottom are taken from one morph line (from the House Cat to the Doberman) to illustrate the fine shape changes involved in categorization. All stimuli in the left half of the plot are "cats" and all stimuli in the right half are "dogs" (the class boundary is at the morph line center). The model unit correctly categorizes 94% of all stimuli [as good as the monkey trained on the same categorization task (Freedman et al., 2001a)], with errors occuring at the class boundary where categorization is most difficult (Riesenhuber and Poggio, 1999c).

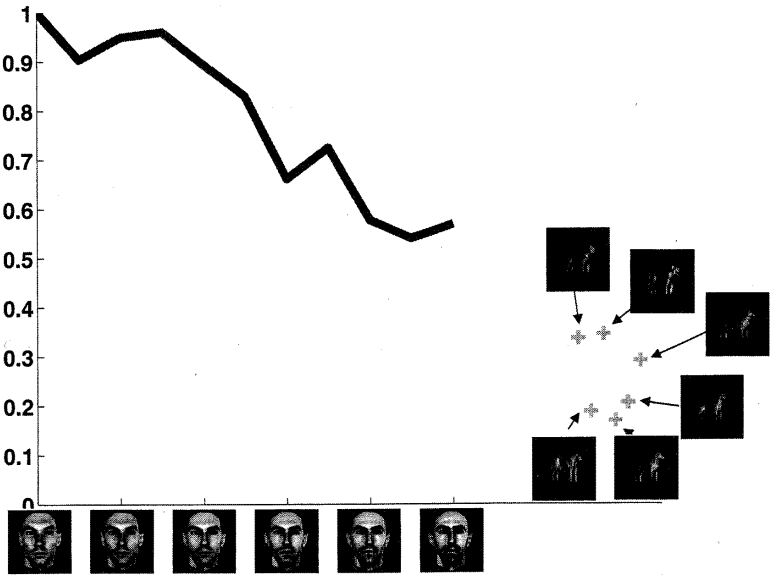

FIGURE 111.7. Tuning of a model face unit. The unit is a view-tuned unit, as shown in Figure 111.2, tuned to the face shown in the lower left. The solid line shows how the unit's response changes as the stimulus is gradually morphed away from the preferred stimulus to another face (faces shown at lower edge of plot. The unit's response changes gradually with changes in the stimulus, permitting subordinate-level discrimination [especially when using a population code consisting of several units tuned to different representatives of the class (Riesenhuber and Poggio, 2000a)]. The same unit also responds to animal stimuli [cf. Freedman et al. (2001a); responses shown by crosses], but at a lower level than to the faces, permitting a coarse categorization of a stimulus as an animal stimulus based on the face unit's firing (cf. Haxby et al., 2001). As for subordinate-level recognition, combining several such units in a sparse population code (Young and Yamane, 1992), improves recognition performance (Riesenhuber and Poggio, 1999c). [Faces and morphing software courtesy of Thomas Vetter (Blanz and Vetter, 1999).]

Categorization and Identification An object can be recognized at different levels—a face can be recognized as a face, but also more specifically as a "male face," "Tommy Poggio's face," or "Tommy Poggio's smiling face." It has been common in cognitive science to assume that recognition of an object at different levels is based on different computational mechanisms (Murphy and Brownell, 1985; Tversky and Hemenway, 1984). In particular, it has been proposed that *subordinate-level* recognition (identification) is based on *configurational* judgments (i.e., based on the spatial arrangement of features), whereas *basic-level* categorization (a face? a dog? a car?) is based on a qualitative representation based on the presence or absence of features. However, as Figure 111.2 makes clear and as we pointed out earlier (Riesenhuber and Poggio, 2000b), all supervised recognition tasks—in which the subject is trained with labeled examples—are identical from a computational point of view: they all involve a classification based on positive and negative exemplars. Indeed, it is not clear why different computations should be required in recognizing a face on the subordinate level and in, for example, determining its gender. Rather, in the simple framework of the Standard Model, the same learning algorithm and architecture can support a variety of object recognition tasks. In particular, identification and categorization circuits (possibly located in PFC) should receive signals from the same or equivalent cells tuned to specific objects or prototypes (in IT). This prediction is supported by recent results from physiology (opde Beeck et al., 2001), where different monkeys were trained on a discrimination and a categorization task on the same stimuli. Subsequent recordings from neurons in the animals' IT revealed no systematic differences in the representation of stimuli in the different animals. Further support comes from combined IT and PFC recordings (Freedman et al., in press) of monkeys trained on a cat/dog categorization task (Freedman et al., 2001a), where PFC neurons showed stronger category tuning than IT neurons. Simulation results (Knoblich et al., 2002) suggest that the tuning properties of these IT neurons could in fact arise without any explicit category information during training. Recent human fMRI data which indicate that the FFA is involved not just in subordinate-level face recognition but also in face detection (Grill-Spector et al., 2001) also argue against a specialization of brain areas for recognition tasks such as subordinate-level recognition independent of object class.

The problem with many experiments investigating the relationship between categorization and identification which claim that there is a time advantage of basic-level recognition versus subordinate-level recognition (Jolicoeur et al., 1984; Rosch et al., 1976) is that the tasks used for the different recognition levels are of different difficulty: discriminating a face from a chair (in categorization) is a much easier task than discriminating the faces of the two authors (in identification), as the latter two are more similar to each other. Assuming that physically similar stimuli produce similar neuronal activation patterns, and that the ability to discriminate between two stimuli requires a certain level of evidence, finer discrimination would require the accumulation of evidence (in the form of firing rate differences) over a longer time period than when the activation patterns are very different.

Pooling and Max Mechanism A key component of hierarchical models of cortical processing, such as Fukushima's and Hubel and Wiesel's, is pooling of afferents with similar tuning to increase the neuronal response invariance of certain stimulus transformations. Our version of the Standard Model—the HMAX model of Figure 111.1—assumes that this pooling is done—at some but not all levels in cortical processing—by a *maximum-like operation* rather than by a linear sum.

Invariance for translation and scaling is achieved by pooling the responses from noninvariant neural detectors over multiple spatial positions or scales using a maximum-like operation. Pooling by a maximum operation, as opposed to linear summation, ensures that the invariant response is robust against background clutter and does not lose the selectivity for the original feature. In neural terms, the firing rate of a maxinum circuit corresponds to the firing rate of the strongest input pooled over some set of synaptic inputs. The maximum operation can be realized with biologically plausible neural circuits.

It will be critical to evaluate whether there are sets of neurons along the ventral pathway that implement such an operation. The model predicts (Riesenhuber and Poggio, 1999b) that the first stage showing a maximum-operation is a subset of complex cells in V1. Preliminary data (Lampl et al., 2001), obtained by intracellular recordings in simple and complex cells, are encouraging.

Extensions of the Standard Model and open issues

The Standard Model can serve as a conceptual tool to guide experiments and as the integrative "glue" between single-cell physiology and behavior. It also incorporates a large body of experimental data into one coherent framework. However, this framework needs to be extended in a number of directions. For instance, it has to take into account top-down and bottom-up attentional effects (Itti and Koch, 2000), learning at most stages of the hierarchy (Riesenhuber and Poggio, 2000b), and recognition of moving objects (Giese and Poggio, 2003).

Attention Attention permeates most natural object recognition tasks. Thus, all the issues discussed earlier have also to be addressed in the context of task dependency,

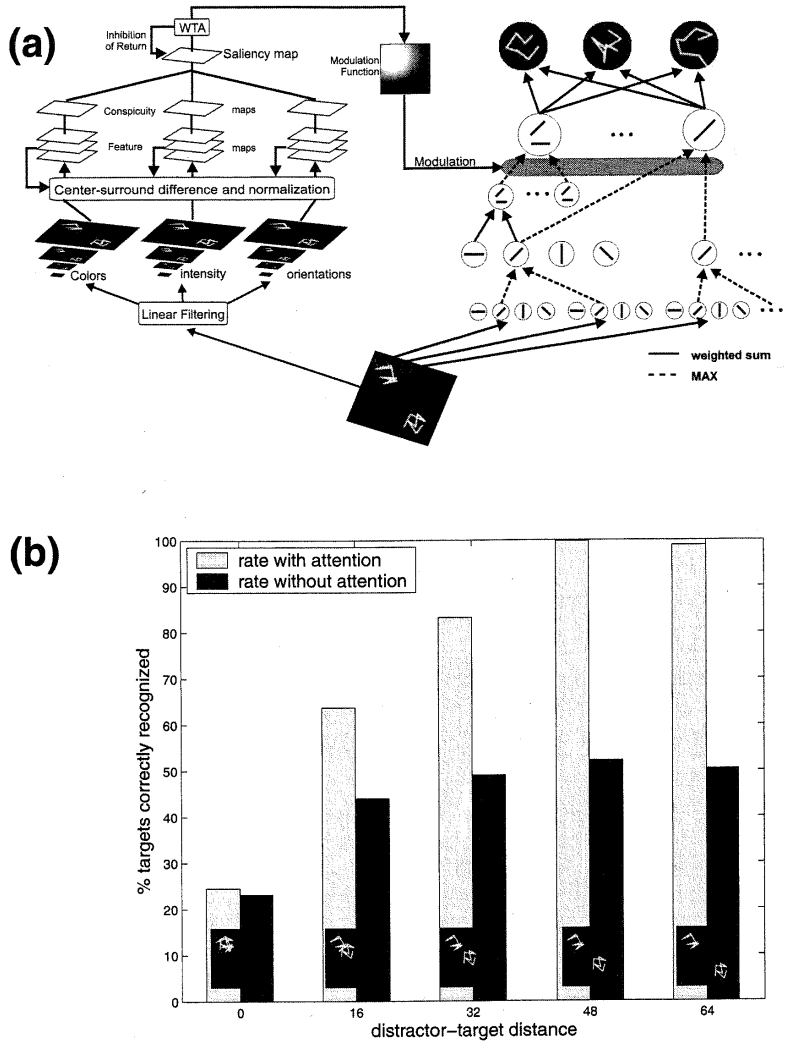

FIGURE 111.8. Coupling of the saliency map (Itti and Koch, 2000) and HMAX (Walther et al., 2002). *a*, Sketch of the integrated system. Targets in the saliency map (*left*) cause a modulation of receptive field size of C2 neurons in HMAX (*right*) by modulating their spatial pooling range (*center*). *b*, Recognition results in HMAX with and without attentional modulation (cf. Fig. 111.4), depending on the separation of stimuli. Attentional modulation of the spatial extent of receptive fields dramatically improves recognition performance.

context, and attentional control. The experimental paradigms used so far in object recognition simplify these issues to a minimum. The Standard Model does not yet incorporate any attentional bias. This is justified by the fact that recognition is possible for scenes viewed in rapid visual presentation that do not allow sufficient time for shifts of attention (Potter, 1975). Furthermore, electroencephalographic studies (Thorpe et al., 1996) provide evidence that the human visual system is able to perform an object detection task—which includes categorization and search—within 150 msec, which is on the order of the latency of view- and object-tuned cells in IT (Perrett et al., 1992). Recent studies indeed suggest that such object detection tasks can be performed in the near absence of attention (Li et al., 2002) and in parallel (Rousselet et al., 2002). None of this rules out, of course, the use of feedback processing, but it suggests a hierarchical feedforward architecture as the core circuitry underlying immediate recognition, with recursion and higher-level interactions playing a role only over longer time periods.

What should happen, however, in the model when the object to be detected is embedded in a large image? Saliency computation is likely to drive eye movement and to interact with the recognition machinery in this process (Fig. 111.8). Task-dependent priming may affect tuning of cells at different levels of the model.

This calls for investigating the relationship between object recognition and visual attention. The required quantitative theory will assume that object recognition derives its key properties of invariance and specificity from two cortical pathways working in parallel. An attentional system (which may be mapped in part onto the dorsal stream of primate cerebral cortex in addition to certain subcortical structures) selects salient or task-relevant candidate locations in the visual field. Attentional biases include both rapid, bottom-up, saliency-driven components as well as slower, top-down, task-dependent components. Attended locations are processed in a mainly feedforward (ventral) stream from primary visual cortex, to IT cortex, and PFC in order to recognize objects. We expect this cooperation between attentional and recognition systems to be especially relevant in object recognition situations too complex—due to, for instance, visual clutter or ambiguous displays—to be parsed by one feedforward pass through the object recognition pathway in isolation.

Learning In the Standard Model simulations described so far, *learning and adaptation* of synapses occur only at the last stages—from the dictionary of features produced by the view-based module in V4 and PIT to the view-tuned, object-tuned, and identification/categorization units. We believe, however, that experience-dependent learning occurs at all stages, possibly starting in V1. There is physiological evidence that the dictionary of complex features in PIT is affected by visual experience (Kobatake et al., 1998). Computationally, hard recognition tasks involving background and clutter require the selection of appropriate complex features, depending on the class of object [early simulations of feature selection used a HyperBF model (Bricolo et al., 1997; Vetter et al., 1995)]. It will be important to investigate with computer experiments how feature sets could be learned using plausible cortical mechanisms and how they may affect object recognition, in particular object detection performance. Simulation results have shown that the model's object detection performance in natural images can be significantly improved by learning target object class-specific intermediate features at the S2 level (Serre et al., 2002). Learning effects may also explain intriguing, recent data on limitations of position invariance (DiCarlo and Maunsell, 2003). It is likely that learning at different stages in cortex takes place over different time scales, the shortest in IT and PFC and the longer in V1.

Recognizing Objects in Motion Initial work on extending the Standard Model to the time domain for the recognition of biological movements and actions has begun (Giese and Poggio, 2003). The key idea in this extension of the model is that action recognition is based on learned prototypical complex motion patterns. Simulations consistently summarize many existing results and simultaneously provide a plausibility proof that the recognition of complex biological motion patterns might be based on relatively simple, well-established neural mechanisms. In addition, the model shows that biological motion recognition can be based on a relatively limited number of learned prototypical motion patterns. This representational principle is analogous to the encoding of complex stationary three-dimensional shapes in terms of learned *prototypical views*.

In this model, *view-tuned neurons* similar to the units that have been found in area IT of macaques can learn to respond selectively to configurations of the human body that are characteristic of biological movements or actions. The model further assumes that a sequence of such "snapshots" represents a limited number of prototypical example movement patterns. The highest level of the form pathway consists of neurons that respond selectively to whole motion patterns, like walking and running. These neurons integrate the activity of the view-tuned neurons at the previous level that code for the same motion pattern over time. Since recognition of complex movement patterns must be selective for temporal order, the model assumes a simple mechanism that is based on asymmetric lateral connections between the view-tuned, pattern-selective neurons. Lateral connections lead to a network dynamics that stabilizes a traveling activity pulse only if the stimulus frames are presented in the right temporal order. The effectiveness of this mechanism for sequence selectivity has been shown in simulations. Scrambling of the temporal order of the stimulus leads to competition between the stimulus input and the intrinsic dynamics of the network, resulting in a strong reduction of neural activity.

Falsifying the Standard Model

There is no doubt that existing and future data will force modifications of the Standard Model. But what kinds of experiments would falsify it completely and provide no-go results for the whole class of neural architectures associated with it? Data showing that the basic operation is *intrinsically* not feedforward are prime candidates. Dependency of recognition on different computations and computational modules for every object class would be difficult to reconcile with the Standard Model. Finding that supervised forms of categorization and identification do not use the same basic machinery to provide input signals would require a major revision of the Standard Model but not of its view-based HMAX module. The simplest interpretation of the model predicts that both categorization and identification rely on IT, but direct neural correlates of categorization are not in IT. Refutation of this prediction, however, would not falsify the Standard Model (cf. Fig. 111.2).

The road ahead

The architecture of the Standard Model allows us to quantitatively summarize and structure knowledge from experimental data and plan new experiments. It is an implicit map to future experimental work in terms of both its predictions and the questions that are left open. Clearly, the road ahead will require a very close interaction of experiments and computational work.

Acknowledgments

Supported by grants from ONR, Darpa, NSF, ATR, and Honda. M.R. is supported by a McDonnell-Pew Award in Cognitive Neuroscience. T.P. is supported by the Uncas and Helen Whitaker Chair at the Whitaker College, Massachusetts Institute of Technology. We are grateful to Nikos Logothetis for useful comments and suggestions, to Ulf Knoblich for creating Figure 111.6, and to Dirk Walther for creating Figure 111.8.

REFERENCES

Blanz, V., and T. Vetter, 1999. A morphable model for the synthesis of 3D faces, in *SIGGRAPH '99 Proceedings*, ACM Computer Society Press, pp. 187–194.

Booth, M. C., and E. T. Rolls, 1998. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex, *Cereb. Cortex*, 8:510–523.

Bricolo, E., T. Poggio, and N. K. Logothetis, 1997. 3D object recognition: a model of view-tuned neurons, in *Advances in Neural Information Processing Systems*, vol. 9, Cambridge, MA: MIT Press, pp. 41–47.

Bülthoff, H., and S. Edelman, 1992. Psychophysical support for a two-dimensional view interpolation theory of object recognition, *Proc. Natl. Acad. Sci. USA*, 89:60–64.

Carandini, M., D. J. Heeger, and J. A. Movshon, 1997. Linearity and normalization in simple cells of the macaque primary visual cortex, *J. Neurosci.*, 17:8621–8644.

Desimone, R., 1991. Face-selective cells in the temporal cortex of monkeys, *J. Cogn. Neurosci.*, 3:1–8.

Desimone, R., T. D. Albright, C. G. Gross, and C. Bruce, 1984. Stimulus-selective properties of inferior temporal neurons in the macaque, *J. Neurosci.*, 4(8):2051–2062.

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

DiCarlo, J. J., and J. H. R. Maunsell, 2000. Form representation in monkey inferotemporal cortex is virtually unaltered by free viewing, *Nat. Neurosci.*, 3:814–821.

DiCarlo, J. J., and J. H. R. Maunsell, 2003, Anterior intertemporal recognition can be highly sensitive to object retinal position, *J. Neurophysiol.*, 89:3264–3278.

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

Edelman, S., 1999. *Representation and Recognition in Vision*, Cambridge, MA: MIT Press.

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

Freedman, D., M. Riesenhuber, T. Poggio, and E. Miller, 2001a. Categorical representation of visual stimuli in the primate prefrontal cortex, *Science*, 291:312–316.

Freedman, D. J., M. Riesenhuber, T. Poggio, and E. Miller, in press. Comparison of primate prefrontal and inferior temporal cortices during visual categorization, *J. Neurosci.*

Fujita, I., K. Tanaka, M. Ito, and K. Cheng, 1992. Columns for visual features of objects in monkey inferotemporal cortex, *Nature*, 360:343–346.

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.

Gallant, J. L., C. E. Connor, S. Rakshit, J. W. Lewis, and D. C. Van Essen, 1996. Neural responses to polar, hyperbolic, and cartesian gratings in area V4 of the macaque monkey, *J. Neurophysiol.*, 76:2718–2739.

Gauthier, I., P. Skudlarski, J. C. Gore, and A. W. Anderson, 2000. Expertise for cars and birds recruits brain areas involved in face recognition, *Nat. Neurosci.*, 3:191–197.

Gauthier, I., and M. J. Tarr, 1997. Becoming a "Greeble" expert: exploring mechanisms for face recognition, *Vis. Res.*, 37:1673–1682.

Giese, M. A., and T. Poggio, 2003. Biological movement recognition, *Nature Reviews Neuroscience*, 4:179–192.

Grill-Spector, K., and N. G. Kanwisher, 2001. The functional organization of human ventral temporal cortex is based on stimulus selectivity not recognition task, *Soc. Neurosci. Abstr.*, 27:122.10.

Grill-Spector, K., Z. Kourtzi, and N. Kanwisher, 2001. The lateral occipital complex and its role in object recognition, *Vis. Res.*, 41:1409–1422.

Gross, C. G., C. E. Rocha-Miranda, and D. B. Bender, 1972. Visual properties of neurons in inferotemporal cortex of the macaque, *J. Neurophysiol.*, 35:96–111.

Haxby, J. V., M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini, 2001. Distributed and overlapping representations of faces and objects in the ventral temporal cortex, *Science*, 293:2425–2430.

Hegde, J., and D. C. Van Essen, 2000. Selectivity for complex shapes in primate visual area V2, *J. Neurosci.*, 20(R61):1–6.

Hietanen, J. K., D. I. Perrett, P. J. Benson, and W. H. Dittrich, 1992. The effects of lighting conditions on responses of cells selective for face views in the macaque temporal cortex, *Exp. Brain Res.*, 89:157–171.

Hubel, D. H., and T. N. Wiesel, 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, *J. Physiol.*, 160:106–154.

Hubel, D. H., and T. N. Wiesel, 1965. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat, *J. Neurophysiol.*, 28:229–289.

Hubel, D. H., and T. N. Wiesel, 1968. Receptive fields and functional architecture of monkey striate cortex, *J. Physiol.*, 195: 215–243.

Intraub, H., 1981. Rapid conceptual identification of sequentially presented pictures, *J. Exp. Psych. Hum. Percept. Perform.*, 7:604–610.

Itti, L., and C. Koch, 2000. A saliency-based search mechanism for overt and covert shifts of visual attention, *Vis. Res.*, 40:1489–1506.

Jolicoeur, P., M. A. Gluck, and S. M. Kosslyn, 1984. Pictures and names: making the connection, *Cogn. Psychol.*, 16:243–275.

Kanwisher, N., J. McDermott, and M. M. Chun, 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception, *J. Neurosci.*, 17:4302–4311.

Knoblich, U., D. J. Freedman, and M. Riesenhuber, 2002. Categorization in IT and PFC: Model and experiments, CBCL paper #2110/AI Memo #2002-007. Cambridge, MA: Massachusetts Institute of Technology.

Knoblich, U., and M. Riesenhuber, 2002. Stimulus simplification and object representation: a modeling study, CBCL Paper #215/AI Memo #2002-004, Massachusetts Institute of Technology, Cambridge, MA.

Kobatake, E., and K. Tanaka, 1994. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex, *J. Neurophysiol.*, 71:856–867.

Kobatake, E., G. Wang, and K. Tanaka, 1998. Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys, *J. Neurophysiol.*, 80:324–330.

Lampl, I., T. Poggio, D. Ferster, and M. Riesenhuber, 2001. Spatial integration of complex cells of the cat primary visual cortex.

Leopold, D. A., A. J. O'Toole, T. Vetter, and V. Blanz, 2001. Prototype-references shape encoding revealed by high-level aftereffects, *Nat. Neurosci.*, 4:3–5.

Li, F. F., R. van Rulien, C. Koch, and P. Perona, 2002. Rapid natural scene categorization in the near absence of attention. *Proc. Nat. Acad. Sci. USA*, 99:9596–9601.

Logothetis, N. K., J. Pauls, M. Augath, T. Trinath, and A. Oeltermann, 2001. Neurophysiological investigation of the basis of the fMRI signal, *Nature*, 412:150–157.

Logothetis, N. K., J. Pauls, H. H. Bülthoff, and T. Poggio, 1994. View-dependent object recognition by monkeys, *Curr. Biol.*, 4:401–414.

Logothetis, N. K., J. Pauls, and T. Poggio, 1995. Shape representation in the inferior temporal cortex of monkeys, *Curr. Biol.*, 5:552–563.

Logothetis, N. K., and D. L. Sheinberg, 1996. Visual object recognition, *Annu. Rev. Neurosci.*, 19:577–621.

Malach, R., J. B. Reppas, R. R. Benson, K. K. Kwong, H. Jiang, W. A. Kennedy, P. J. Ledden, T. J. Brady, B. R. Rosen, and R. B. Tootell, 1995. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex, *Proc. Natl. Acad. Sci. USA*, 92:8135–8139.

Miller, E., 2000. The prefrontal cortex and cognitive control, *Nat. Rev. Neurosci.*, 1:59–65.

Missal, M., R. Vogels, and G. A. Orban, 1997. Responses of macaque inferior temporal neurons to overlapping shapes, *Cereb. Cortex*, 7:758–767.

Murphy, G. L., and H. H. Brownell, 1985. Category differentiation in object recognition: typicality constraints on the basic category advantage, *J. Exp. Psychol. Learn. Mem. Cogn.*, 11:70–84.

opde Beeck, H., J. Wagemans, and R. Vogels, 2001. Inferotemporal neurons represent low-dimensional configurations of parametrized shapes, *Nat. Neurosci.*, 4:1244–1252.

Pasupathy, A., and C. E. Connor, 1999. Responses to contour features in macaque area V4, *J. Neurophysiol.*, 82:2490–2502.

Pauls, J., 1997. The Representation of 3-Dimensional Objects in the Primate Visual System. Ph.D. thesis, Baylor College of Medicine, Houston, TX.

Perrett, D. I., J. K. Hietanen, M. W. Oram, and P. J. Benson, 1992. Organization and functions of cells responsive to faces in the temporal cortex, *Philos. Trans. R. Soc. B*, 335:23–30.

Perrett, D. I., and M. Oram, 1993. Neurophysiology of shape processing. *Img. Vis. Comput.*, 11:317–333.

Poggio, T., 1990. A theory of how the brain might work. *Cold Spring Harbor Symp. Quant. Biol.*, 55:899–910.

Poggio, T., and S. Edelman, 1990. A network that learns to recognize 3D objects, *Nature*, 343:263–266.

Poggio, T., and F. Girosi, 1990. Networks for approximation and learning, *Proc. IEEE*, 78(9):1481–1497.

Potter, M. C., 1975. Meaning in visual search, *Science*, 187:565–566.

Riesenhuber, M., and T. Poggio, 1999a. Are cortical models really bound by the "Binding Problem"? *Neuron*, 24:87–93.

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

Riesenhuber, M., and T. Poggio, 1999c. A note on object class representation and categorical perception. Technical Report AI Memo 1679, CBCL Paper 183, Cambridge, MA: MIT AI Lab and CBCL.

Riesenhuber, M., and T. Poggio, 2000a. The individual is nothing, the class everything: psychophysics and modeling of recognition in object classes, Technical Report AI Memo 1682, CBCL Paper 185, Cambridge, MA: MIT AI Lab and CBCL.

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

Rosch, E., C. B. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem, 1976. Basic objects in natural categories, *Cogn. Psychol.*, 8: 382–439.

Rousselet, G. A., M. Fabre-Thorpe, and S. J. Thorpe, 2002. Parallel processing in high-level categorization of natural images, *Nat. Neurosci.*, 5:629–630.

Serre, T., M. Riesenhuber, J. Louie, and T. Poggio, 2002. On the role of object-specific features for real world object recognition in biological vision. In: Biologically Motivated Computer Vision, Second International workshop (BMCV 2002), H. H. Bülthoff, S.-W. Lee, T. Poggio, and C. Wallraven (eds.), Tübingen, Germany.

Tanaka, K., 1993. Neuronal mechanisms of object recognition, *Science*, 262:685–688.

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

Tanaka, K., 1997. Mechanisms of visual object recognition: monkey and human studies, *Curr. Opin. Neurobiol.*, 7:523–529.

Tarr, M. J., and H. H. Bülthoff, 1998. Image-based object recognition in man, monkey and machine, *Cognition*, 67:1–20.

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

Tsunoda, K., Y. Yamane, M. Nishizaki, and M. Tanifuji, 2001. Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns, *Nat. Neurosci.*, 4:832–838.

Tversky, B., and K. Hemenway, 1984. Objects, parts, and categories, *J. Exp. Psych. Gen.*, 113:169–197.

Ullman, S., and R. Basri, 1991. Recognition by linear combinations of models, *IEEE Trans. Patt. Anal. Mach. Intell.*, 13:992–1006.

Ungerleider, L. G., and J. V. Haxby, 1994. "What" and "where" in the human brain, *Curr. Opin. Neurobiol.*, 4:157–165.

Vetter, T., A. Hurlbert, and T. Poggio, 1995. View-based models of 3D object recognition: invariance to imaging transformations, *Cereb. Cortex*, 3:261–269.

Vogels, R., 1999. Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study, *Eur. J. Neurosci.*, 11:1239–1255.

Wallis, G., and E. T. Rolls, 1997. A model of invariant object recognition in the visual system, *Prog. Neurobiol.*, 51:167–194.

Walther, D., L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, 2002. Attentional selection for object recognition—a gentle way. In: Biologically Motivated Computer Vision, Second International Workshop (BMCV 2002), H. H. Bülthoff, S.-W. Lee, T. Poggio, and C. Wallraven (eds.). Tübingen, Germany, pp. 472–479.

Wang, G., K. Tanaka, and M. Tanifuji, 1996. Optical imaging of functional organization in the monkey inferotemporal cortex, *Science*, 272:1665–1668.

Young, M. P., and S. Yamane, 1992. Sparse population coding of faces in the inferotemporal cortex, *Science*, 256:1327–1331.