Neuron, Vol. 24, 87–93, September, 1999, Copyright ©1999 by Cell Press

Are Cortical Models Really Bound by the “Binding Problem”?

Review

Maximilian Riesenhuber and Tomaso Poggio*
Department of Brain and Cognitive Sciences
Center for Biological and Computational Learning and
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02142

Processing of visual information in cortex is usually described in terms of an extension of the simple-to-complex hierarchy postulated by Hubel and Wiesel—a feed forward sequence of more and more complex and invariant neuronal representations. The capability of this class of models to perform higher-level visual processing such as viewpoint-invariant object recognition in cluttered scenes has been questioned in recent years by several researchers, who in turn proposed an alternative class of models based on the synchronization of large assemblies of cells, within and across cortical areas. The main implicit argument for this novel and controversial view was the assumption that hierarchical models cannot deal with the computational requirements of high-level vision and suffer from the so-called “binding problem.” Here, we review the present situation and discuss theoretical and experimental evidence showing that the perceived weaknesses of hierarchical models are unsubstantiated. In particular, we show here that recognition of multiple objects in cluttered scenes, arguably among the most difficult tasks in vision, can be done in a hierarchical feedforward model. Two problems in particular make object recognition difficult:

1. The *segmentation problem*. Visual scenes normally contain multiple objects. To recognize individual objects, features must be isolated from the surrounding clutter and extracted from the image, and the feature set must be parsed so that the different features are assigned to the correct object. The latter problem is commonly referred to as the “binding problem” (von der Malsburg, 1995).

2. The *invariance problem*. Objects have to be recognized under varying viewpoints, lighting conditions, etc.

Interestingly, the human brain can solve these problems with ease and *quickly*. Thorpe et al. (1996) report that visual processing in an object detection task in complex visual scenes can be achieved in under 150 ms, which is on the order of the latency of the signal transmission from the retina to inferotemporal cortex (IT), the highest area in the ventral visual stream thought to have a key role in object recognition (Ungerleider and Haxby, 1994; see also Potter, 1975). This impressive processing speed presents a strong constraint for any model of object recognition.

Models of Visual Object Recognition and the Binding Problem

Hubel and Wiesel (1965) were the first to postulate a model of visual object representation and recognition. They recorded from simple and complex cells in the primary visual cortices of cats and monkeys and found that while both types preferentially responded to bars of a certain orientation, the former had small receptive fields with a phase-dependent response while the latter had bigger receptive fields and showed no phase dependence. This observation led them to hypothesize that complex cells receive input from several simple cells. Extending this model in a straightforward fashion, they suggested (Hubel and Wiesel, 1962) that the visual system is composed of a hierarchy of visual areas, from simple cells all the way up to higher-order “hypercomplex” cells.

Later studies (Bruce et al., 1981) of macaque inferotemporal cortex (IT) described neurons tuned to views of complex objects such as faces; i.e., the cells discharged strongly to a face seen from a specific viewpoint but very little or not at all to other objects. A key property of these cells was their scale and translation invariance, i.e., the robustness of their firing to stimulus transformations such as changes in size or position in the visual field.

These findings inspired various models of visual object recognition such as Fukushima’s Neocognitron (1980) or, later, Perrett and Oram’s (1993) outline of a model of shape processing and Wallis and Rolls’ VisNet (1997), all of which share the basic idea of the visual system as a feedforward processing hierarchy where invariance ranges and complexity of preferred features grow as one ascends through the levels.

Models of this type prompted von der Malsburg (1981) to formulate the binding problem. His claim was that visual representations based on spatially invariant feature detectors (to achieve invariant recognition) were ambiguous: “As generalizations are performed independently for each feature, information about neighborhood relations and relative position, size, and orientation is lost. This lack of information can lead to the inability to distinguish between patterns that are composed of the same set of invariant features” (von der Malsburg, 1995). Moreover, as a visual scene containing multiple objects is represented by a set of feature activations, a second problem lies in “singling out appropriate groups from the large background of possible combinations of active neurons” (von der Malsburg, 1995). These problems would manifest themselves in various phenomena such as hallucinations (the feature sets activated by objects actually present in the visual scene combine to yield the activation pattern characteristic of another object) and the figure–ground problem (the inability to correctly assign image features to foreground object and background). These difficulties led von der Malsburg to postulate the necessity of a special mechanism, the synchronous oscillatory firing of ensembles of neurons, to *bind* features belonging to one object together.

One approach to avoid these problems was presented by Olshausen et al. (1993): instead of trying to process all objects simultaneously, processing is limited to one object in a certain part of space at a time, for example by “focusing attention” on a region of interest in the

*To whom correspondence should be addressed (e-mail: tp@ai.mit.edu).

visual field and routing that information through to higher visual areas, while ignoring the remainder of the visual field. The control signal for the input selection in this model is thought to be in the form of the output of a "blob search" system, one that identifies possible candidates in the visual scene for closer examination. While this top-down approach to circumvent the binding problem has intuitive appeal and is compatible with physiological studies that report top-down attentional modulation of receptive field properties (see the review by Reynolds and Desimone, 1999 [this issue of *Neuron*], or the recent study by Connor et al., 1997), such a sequential approach seems difficult to reconcile with the apparent speed with which object recognition can proceed even in very complex scenes containing many objects (Potter, 1975; Thorpe et al., 1996), and it is also incompatible with reports of parallel processing of visual scenes, as observed in pop-out experiments (Treisman and Gelade, 1980). These and other data suggest that object recognition does not seem to depend only on explicit top-down selection in all situations.

A more head-on approach to the binding problem was taken in other studies that have called into question the assumption that representations based on sets of spatially invariant feature detectors are inevitably ambiguous. Starting with Wickelgren (1969) in the context of speech recognition, several studies (Fukushima, 1980; Perrett and Oram, 1993; Wallis and Rolls, 1997) have proposed that coding an object through a set of intermediate features made up of local arrangements of simpler features can sufficiently constrain the representation to uniquely code complex objects without retaining global positional information. Thus, rather than using individual letters to code words, letter pairs or higher-order combinations of letters can be used—i.e., although the word "tomaso" might be confused with the word "somato" if both were coded by the sets of letters they are made up of, this ambiguity is resolved if both are represented through letter pairs (see Mozer, 1991, for an elaboration of this idea). The capabilities of intermediate-level representations based on spatially invariant receptive fields were recently analyzed in detail by Mel and Fiser (1999) for the example domain of English text.

In the visual domain, Mel (1997) presented a model to perform invariant recognition of a high number (100) of objects of different types, using a representation based on a large number of feature channels. While the model performed surprisingly well for a variety of transformations, recognition performance depended strongly on color cues and did not seem as robust to scale changes as experimental neurons (Logothetis et al., 1995). Perrett and Oram (1998) have recently outlined a conceptual model, based on very similar ideas, of how a representation based on feature combinations could in theory avoid the binding problem, for example, by coding a face through a set of detectors for combinations of face parts such as eye–nose or eyebrow–hairline. What has been lacking so far, however, is a computational implementation quantitatively demonstrating that such a model can actually perform "real-world" subordinate visual object recognition to the extent observed in behavioral and physiological experiments (Sato, 1989; Logothetis et al., 1994, 1995; Missal et al., 1997), where effects such as variations in scale and position, occlusion, and overlap pose additional problems not found in an idealized text environment. In particular, unlike in the text domain where the input consists of letter strings, and the extraction of features (letter combinations) from the input is therefore trivial, the crucial task of invariant feature extraction from the image is nontrivial for scenes containing complex shapes, especially when multiple objects are present.

We have developed a hierarchical feedforward model of object recognition in cortex (Riesenhuber and Poggio, 1999b) as a plausibility proof that such a model can account for several properties of IT cells, in particular the invariance properties of IT cells found by Logothetis et al. (1995). In the following section, we will show that such a simple model can perform invariant recognition of complex objects in cluttered scenes and is compatible with recent physiological studies. Thus, this plausibility proof demonstrates that complex oscillation-based mechanisms are not necessarily required for these tasks, and that the binding problem seems to be a problem for only some models of object recognition.

A Hierarchical Model of Object Recognition in Cortex

Studies of receptive field properties along the ventral visual stream in the macaque, from primary visual cortex V1 to anterior IT, report an overall trend of an increase of average feature complexity and receptive field size throughout the stream (Kobatake and Tanaka, 1994). While simple cells in V1 have small localized receptive fields and respond preferentially to simple shapes like bars, cells in anterior IT have been found to respond to views of complex objects while showing great tolerance to scale and position changes. Moreover, some IT cells seem to respond to objects in a view-invariant manner (Perrett et al., 1991; Logothetis et al., 1995; Booth and Rolls, 1998).

Our model follows this general framework. Previously, Poggio and Edelman (1990) presented a model of how view-invariant cells could arise from view-tuned cells (Figure 1). However, they did not describe any model of how the view-tuned units (VTUs) could come about. We have recently developed a hierarchical model that closes this gap and shows how VTUs tuned to complex features can arise from simple cell–like inputs. A detailed description of our model can be found in Riesenhuber and Poggio (1999b; for preliminary accounts, refer to Riesenhuber and Poggio, 1998a, 1998b, and also to Koch and Poggio, 1999). We briefly review here some of the model's main properties. The central idea of the model is that invariance to scaling and translation, and robustness to clutter on the one hand and feature complexity on the other hand, require different transfer functions, i.e., mechanisms by which a neuron combines its inputs to arrive at an output value. While for feature complexity a weighted sum of different features, which makes the neuron respond preferentially to a specific activity pattern over its afferents, is a suitable transfer function, increasing invariance requires a different transfer function that pools over different afferents tuned to the same feature but transformed to different degrees (for example, at different scales to achieve scale invariance). A suitable pooling function is a so-called MAX function, *where the output of the neuron is determined*

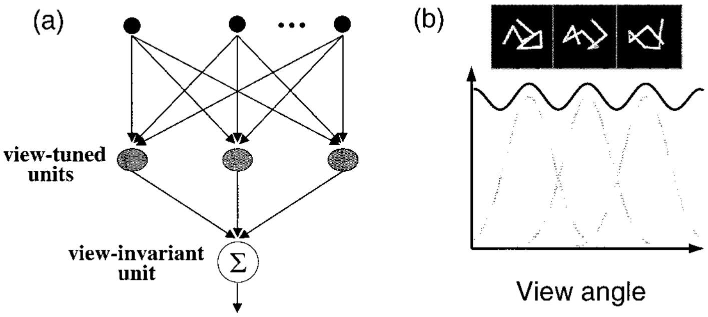

Figure 1. Poggio and Edelman (1990) Model of 3D Rotation-Invariant Object Recognition from Individual Views

(a) Cartoon of the model. The gray ovals correspond to view-tuned units that feed into a view-invariant unit (white circle).
(b) Tuning curves of the view-tuned (gray) and the view-invariant (black) units.

by the strongest afferent, and it thus performs a "selection" (and possibly scanning) operation over afferents tuned to different positions and scales (for a computational justification, see Riesenhuber and Poggio, 1999b). The idea is similar to the original Hubel and Wiesel model of a complex cell receiving input from simple cells at different locations to achieve phase invariance.

In our model of object recognition in cortex (Figure 2), the two types of operations, selection and template matching, are combined in a hierarchical fashion to build up complex, invariant feature detectors from small, localized, simple cell–like receptive fields in the bottom layer. Our model "retina" is composed of 160 × 160 pixels, corresponding to a 5° receptive field size if we set 32 pixels = 1° (Kobatake and Tanaka [1994] report an average V4 receptive field size of 4.4°). Patterns on the model retina are first filtered through layer S1 (adopting Fukushima's [1980] nomenclature referring to feature-building cells as "S" cells and pooling cells as "C" cells) of simple cell–like receptive fields (first derivative of Gaussians, zero-sum, square-normalized to 1, oriented at 0°, 45°, 90°, and 135° with standard deviations of 1.75–7.25 pixels in steps of 0.5 pixels). S1 filter responses are absolute values of the image "filtered" through the units' receptive fields (more precisely, the rectified dot product of the cell's receptive field with the corresponding image patch). Receptive field centers densely sample the input retina. Cells in the next layer (C1) each pool S1 cells of the same orientation over a range of scales and positions using the MAX operation. Filters were grouped in four bands, each spanning roughly 0.5 octaves; sampling over position was done over patches of linear dimensions of 4, 6, 9, and 12 pixels, respectively (starting with the smallest filter band); patches overlapped by half in each direction to obtain more invariant cells responding to the same features as the S1 cells. Different C1 cells were then combined in higher layers—the figure illustrates two possibilities: either combining C1 cells tuned to different features, resulting in S2 cells that respond to coactivations of C1 cells that are tuned to different orientations, or yielding C2 cells that respond to the same feature as the C1 cells but that have bigger receptive fields (i.e., the hierarchy does not have to be a strict alternation of S and C layers). In the version described in this paper, there were no direct C1 to C2 connections, and each S2 cell received input from four neighboring C1 units (in a 2 × 2 arrangement) of arbitrary orientation, yielding a total of $4^4 = 256$ different S2 cell types. S2 transfer functions were Gaussian ($\sigma = 1$, centered at 1). C2 cells then pooled inputs from all S2 cells of the same type, producing invariant feature detectors tuned to complex shapes. Top-level view-tuned units had Gaussian response functions and each VTU received inputs from a subset of C2 cells (see below).

This model had originally been developed to account

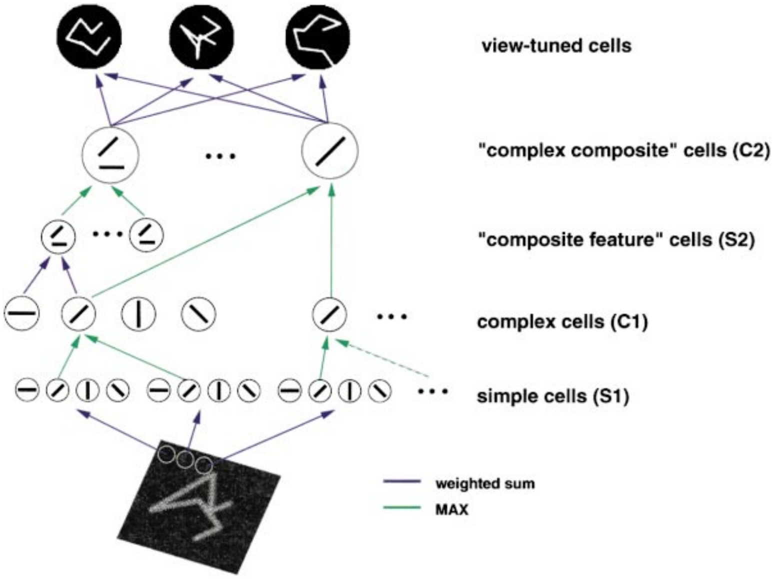

Figure 2. Diagram of Our Hierarchical Model of Object Recognition in Cortex

The model consists of layers of linear units that perform a template match over their afferents (blue arrows) and of nonlinear units that perform a MAX operation over their inputs, where the output is determined by the strongest afferent (green arrows). While the former operation serves to increase feature complexity, the latter increases invariance by effectively scanning over afferents tuned to the same feature but at different positions (to increase translation invariance) or scale (to increase scale invariance; not shown). In the version described in this paper, learning only occured at the connections from the C2 units to the top-level view-tuned units. From Riesenhuber and Poggio (1999b).

for the transformation tolerance of view-tuned units in IT as recorded by Logothetis et al. (1995). It turns out, however, that the model also has interesting implications for the binding problem.

Binding without a Problem

To correctly recognize multiple objects in clutter, two problems must be solved: (1) features must be robustly extracted, and (2) based on these features, a decision has to be made about which objects are present in the visual scene. The MAX operation can perform robust feature extraction (cf. Riesenhuber and Poggio, 1999b): a MAX pooling cell that receives inputs from cells tuned to the same feature at, e.g., different locations, will select the most strongly activated afferent; i.e., its response will be determined by the afferent with the closest match to its preferred feature in its receptive field. Thus, the MAX mechanism effectively isolates the feature of interest from the surrounding clutter. Hence, to achieve robustness to clutter, a VTU should only receive input from cells that are strongly activated by the VTU's preferred stimulus (i.e., those features that are relevant to the definition of the object) and thus less affected by clutter (which will tend to activate the afferents less and will therefore be ignored by the MAX response function). Also, in such a scheme, two view-tuned neurons receiving input from a common afferent feature detector will tend to both have strong connections to this feature detector. Thus, there will be little interference even if the common feature detector only responded to one (the stronger) of the two stimuli in its receptive field due to its MAX response function. Note that the situation would be hopeless for a response function that pools over all afferents through, for example, a linear sum function: the response would *always* change when another object is introduced in the visual field, making it impossible to disentangle the activations caused by the individual stimuli without an additional mechanism—such as, for instance, an attentional sculpting of the receptive field or some kind of segmentation process.

In the following two sections, we will show simulations that support these theoretical considerations, and we will compare them to recent physiological experiments.

Recognition of Multiple Objects

The ability of the model neurons to perform recognition of multiple, nonoverlapping objects was investigated in the following experiment: 21 model neurons were each tuned to a view of a randomly selected paperclip object, as used in theoretical (Poggio and Edelman, 1990), psychophysical (Bülthoff and Edelman, 1992; Logothetis et al., 1994), and physiological (Logothetis et al., 1995) studies on object recognition. Model neurons were each presented with 21 displays consisting of that neuron's preferred clip combined with each of the 21 clip stimuli (in the upper left and lower right corner of the model retina, respectively; see Figure 3a) yielding $21^2 = 441$ two-clip displays. Recognition performance was evaluated by comparing the neuron's response to these displays with its responses to 60 other, randomly chosen "distractor" paperclip objects (Figure 3). Following the studies on view-invariant object recognition (Bülthoff and Edelman, 1992; Logothetis et al., 1994, 1995; Riesenhuber and Poggio, 1999b), an object is said to be recognized if the neuron's response to the two-clip displays (containing its preferred stimulus) is greater than to any of the distractor objects. For 40 afferents to each view-tuned cell (i.e., the 40 C2 units excited most strongly by the neuron's preferred stimulus—this choice produced top-level neurons with tuning curves similar to the experimental neurons; Riesenhuber and Poggio, 1999b), we find that on average, in 90% of the cases, recognition of the neuron's preferred clip is still possible, indicating that there is little interference between the activations caused by the two stimuli in the visual field. The maximum recognition rate is 94% for 18 afferents, dropping to 55% if each neuron is connected to all 256 afferents. Figure 3c plots the recognition rate as a function of the number of afferents to each VTU: the rate climbs in the beginning as discriminability of different clips increases with the number of afferents, and then falls again as the presence of the second object in the visual field increasingly interferes with the input to the VTU caused by the first object. Interference occurs because the probability that another object activates a feature detector connected to the VTU more strongly than the preferred object increases as the VTU also receives input from feature detectors that are excited only weakly by its preferred object.

These simulation results have an interesting experimental counterpart in the work of Sato (1989), who studied the responses of neurons in macaque IT to displays consisting of one or two simultaneously appearing stimuli within the IT cell's receptive field. He defines a "summation index," *SmI*, as

$$SmI = \frac{R_{A+B} - \max(R_A, R_B)}{\min(R_A, R_B)}$$

with R_A the IT neuron's response to stimulus A, R_B the neuron's response to another stimulus B, and R_{A+B} the neuron's response to both stimuli presented simultaneously in its receptive field. Neurons performing a linear summation would have an *SmI* of 1, while MAX neurons would show an *SmI* of 0. For a fixation task, Sato reports a mean *SmI* of -0.18 ($\sigma = 0.5$, $N = 70$, both stimuli in the same hemifield). From these data, the response of real IT neurons appears to have strong MAX characteristics. In fact, a reduction of the response to the two-stimulus display compared to the response to the stronger stimulus alone, implied by the negative *SmI*, and also found in an experiment by Rolls and Tovee (1995), is compatible with the response reduction observed in the two-clip simulation shown in Figure 3b. Interestingly, for a visual discrimination task, Sato (1989) reports very similar average *SmI* values, suggesting that the same bottom-up-driven MAX response mechanism might be operating in both cases.

Recognition in Clutter

So far, we have examined the model's performance for two well-separated objects in the visual field. What about the case of two overlapping stimuli, e.g., when the object of interest is in front of a background object?

This stimulus configuration was used in a physiology

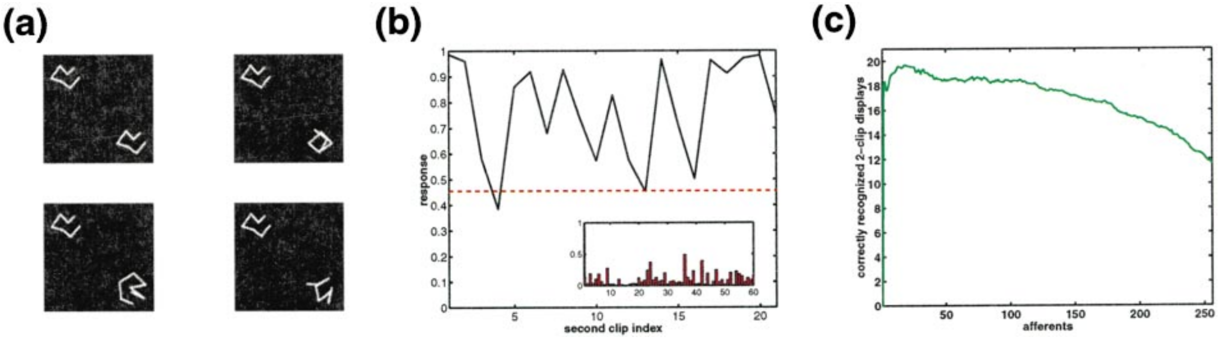

Figure 3. Model Neuron Responses and Average Recognition Rate for the Case of Two Objects in the Visual Field

(a) Example stimuli.

(b) Response curve of one neuron to its 21 two-clip stimuli (the first four being the stimuli shown in 3a), with the dashed line showing the response to the best distractor and the inset showing the response to 60 distractors.

(c) Dependence of average recognition rate (over 21 model neurons) on the number of afferents to each VTU.

experiment by Missal et al. (1997). They trained a monkey on a paired-associate task involving 30 polygonal shapes, followed by recordings of shape-selective cells in IT to the training stimuli and, among others, to displays consisting of the training stimuli superimposed on randomly generated background (other polygons in outline), which were selected so as not to drive the cells (Figure 4a). In this condition, the monkeys behavioral performance decreased slightly (from 98% to 89%), but the average neuronal response dropped precipitously to 25%. How could the monkey still do the task so well in the face of such a drastic change in neuronal response? Furthermore, do we see a similar behavior in the model?

Simulation of the experimental paradigm with our model is straightforward. Foreground stimuli were the 21 clips used in the simulations described previously; backgrounds were randomly generated polygons consisting of eight edges, chosen so that each corner was at a distance from the center of at least 45% of the stimulus size (Figure 4b). Following Missal et al. (1997), we only chose backgrounds that did not drive the model cells, here defined as generating an input to the VTU more than two standard deviations away from the preferred stimulus. Taking the 21 view-tuned cells described above (with 40 afferents out of 256 C2 cells each), and testing each neuron's response to an input image consisting of that neuron's preferred stimulus superimposed on a randomly generated polygonal background, responses on average (over 10 trials and 21 model neurons each) drop to 49% of the response to the stimulus alone (Figure 5). However, average responses to the best distractor (out of 60) are even lower (42%). Note that the response level of the neurons (but not the recognition rates) depends on the standard deviation σ of their Gaussian response function, which is a free parameter and was set to 0.16 in all simulations, producing tuning curves qualitatively similar to those observed experimentally (Riesenhuber and Poggio, 1999b): $\sigma = 0.12$, for instance, would give average responses of 33% to the stimulus–background combination and 23% to the best distractor. This leads to an average recognition rate of 65% in this condition (unlike in the Missal et al. [1997] experiment, using no color cues—if features are color selected, performance is likely to increase). The maximum average recognition rate was 74% for 100 afferents; the maximum average rate for one trial (over 21 neurons) was 90% with 105 afferents. Model parameters were not specially tuned in any way to achieve this performance, so higher recognition rates (for instance, through pooling the responses of several neurons tuned to the same object but receiving inputs from different afferents) are very likely achievable. This simulation thus demonstrates that the ability of the MAX response function to ignore nonrelevant information (in this case, the background figure), together with an object definition based on its salient components, is sufficient to perform recognition in clutter.

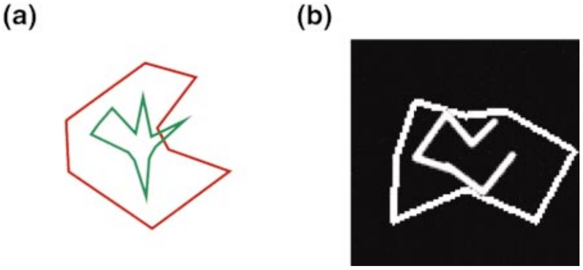

Figure 4. Recognition in Clutter

(a) Example stimulus (green) and outline background (red) used in the physiology experiments and redrawn from Missal et al. (1997).

(b) Example stimulus for the corresponding experiment with the model (see text). The foreground clip was correctly recognized by the corresponding model neuron (which was the same as in Figure 3).

Discussion

As with most existing theories of the brain, our model is likely to be incomplete at best and quite possibly wrong altogether. It provides, however, a plausibility proof that biologically plausible models do in fact exist that do not suffer from the binding problem in performing difficult recognition tasks. This is of course just an explicit demonstration of a known but often neglected fact—that the binding problem is not a fundamental

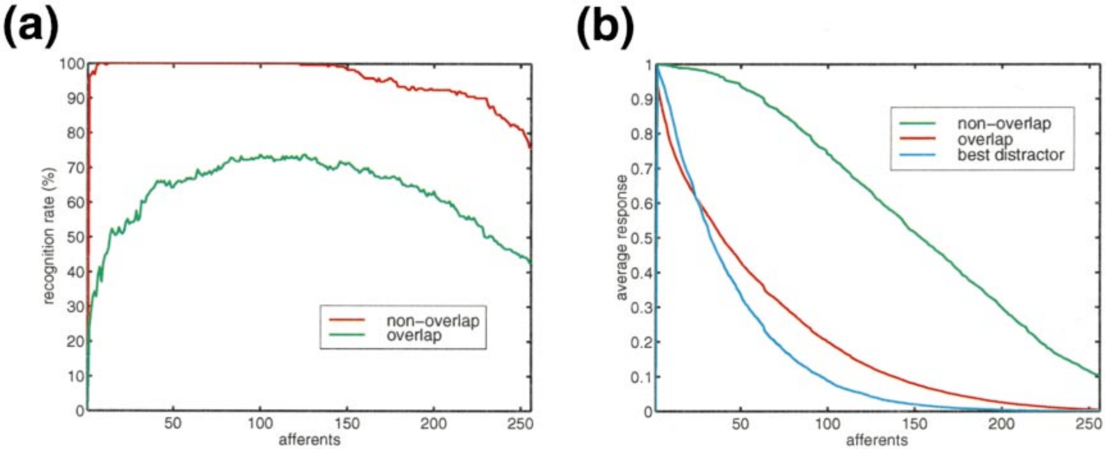

Figure 5. Model Performance for Recognition in Clutter

(a) Average recognition rates (over 21 cells; 10 runs each with different, randomly selected backgrounds) for nonoverlap and overlap conditions (cf. text).

(b) Average response levels in the two conditions, compared to the average response to the best distractor, as a function of the number of afferents to each view-tuned model neuron.

computational problem, like, for instance, the correspondence problem in vision. Instead, the binding problem arises only from the limitations of certain specific computational architectures. Our model shows that there are natural, old-fashioned cortical models that do agree with available data, do not suffer from the binding problems, and do not need oscillations or synchronization mechanisms.

In models like ours, recognition *can* take place without an explicit segmentation stage. The key is to ignore nonrelevant information. At the level of the C cells, this is done through the MAX response function that allows a unit to scan over the image and pick best matches. At the level of the final view-tuned cells (for instance), this is achieved by restricting the afferents to the VTU to those that correspond to the relevant/salient features for the object. This in turn requires an earlier, overcomplete set of "feature"-selective cells that may roughly correspond to the dictionary of shapes described by Tanaka (1993). Subsets of this dictionary are inputs to each of several VTU units.

Many approaches to solving the binding problems do not use oscillation or synchronization mechanisms but instead rely on top-down attentional mechanisms. In fact, it has been argued that top-down control might help in "binding" features together by focusing attention on a region of interest (see Reynolds and Desimone, 1999; Wolfe and Cave, 1999 [both in this issue of *Neuron*]). However, we can perfectly well perform very complex object recognition tasks (e.g., determining whether an image contains a certain object) without focusing attention on a specific part of space (cf. Thorpe et al., 1996). Our model is bottom-up and does not require an explicit top-down signal but is consistent with its use in certain situations. To explain the latter point, we will briefly describe a possible approximate implementation of the MAX operation in terms of cortical microcircuits of lateral, possibly recurrent, inhibition between neurons in a cortical layer. A specific example is a circuit based on feedforward (or recurrent) shunting presynaptic (or postsynaptic) inhibition by "pool" cells (Poggio et al., 1981). The circuit performs a gain control operation and, for certain values of the parameters, a "softmax" operation (an approximation to the MAX operation in which the degree of nonlinearity is controlled by a parameter): each of the N signals x_i (the activation of the afferents) undergoes a softmax operation as

$$y_i = \frac{x_i^{\,p}}{C + \Sigma_j \, x_j^q}$$

Thus, for large p and for $q = p - 1$, we have $y_i = x_i$ if $x = \max_j x_j$ and $y_i = 0$ otherwise. Softmax circuits have been proposed by Nowlan and Sejnowski (1995) and others (Heeger, 1992; Lee et al., 1999) to account for several cortical functions. Circuits of this type may perform an operation ranging between a simple sum and a MAX on the inputs of a layer of cells under the control of a single variable, and thus may form the basis in cortex for normalization of signals at one extreme and for a MAX-like operation at the other (Chance et al., 1999). Thus, in the context of this hypothetical circuitry for the MAX operation, an intriguing possibility is that the same softmax mechanism might be used in both situations, either predominantly driven by bottom-up information or using top-down signals that may control a parameter (equivalent to locally raising q or C) that switches off the "competition" between inputs in locations outside the "focus of attention." Several experiments suggest that the visual system uses a MAX or softmax operation to select bottom-up among different inputs: for instance, there is evidence that a MAX-like operation is used in tasks involving object recognition in context (Sato, 1989). As discussed by Nowlan and Sejnowski (1995), the same active selection mechanism underlying preattentive perceptual phenomena may also be used by top-down overt attentional signals—for instance, when focusing attention to a specific part of visual space (Lee et al., 1999; Reynolds et al., 1999).

The MAX mechanism performs an input-driven selection (and possibly scanning) operation over its inputs that might have interesting implications for the pop-out effect (Treisman and Gelade, 1980; cf. the review by Wolfe and Cave, 1999): as the MAX operation is performed in parallel over many neurons, detection of stimuli does not require an attention-controlled “focused” search, as described above, if surrounding stimuli do not interfere with the VTU’s preferred object. Therefore, for objects that activate different features (such as a square amid circles), recognition is possible without sequential search—the stimuli “pop out.” However, in the case of interference, as in a display consisting of many similar paperclips, detection might require “focusing attention” (as discussed by Reynolds and Desimone, 1999) to reduce the influence of competing stimuli. In this case, there would be no pop-out, but rather sequential search would be required to perform successful recognition.

The observed invariance ranges of IT cells after training with one view are reflected in the architecture used in our model: one of its underlying ideas is that invariance and feature specificity have to grow hierarchically so that view-tuned cells at higher levels show sizable invariance ranges even after training with only one view, as a result of the invariance properties of the afferent units. The key concept is to start with simple localized features—since the discriminatory power of simple features is low, the invariance range has to be kept correspondingly low to avoid the cells being activated indiscriminately. As feature complexity and thus discriminatory power grows, the invariance range, i.e., the size of the receptive field, can be increased as well. Thus, loosely speaking, feature specificity and invariance range are inversely related, which is one of the reasons the model avoids a combinatorial explosion in the number of cells: while there is a larger number of different features in higher layers, there do not have to be as many neurons responding to these features as in lower layers, since higher-layer neurons have bigger receptive fields and respond to a greater range of scales. Notice also that the cells in the model are not binary but have continuous response functions, greatly increasing the representational power of the system (which is why “hallucinations” do not occur).

This hierarchical buildup of invariance and feature specificity greatly reduces the overall number of cells required to represent additional objects in the model: the first layer contains a little more than one million cells (160 × 160 pixels, at four orientations and 12 scales each—for simplicity, dense sampling was used at all scales). Connections in higher levels are in principle subject to learning, driven by the input ensemble and the requirements of the recognition task at hand. As described, we did not investigate learning in the model but rather focused on demonstrating that invariant recognition in clutter is possible using a simple hierarchical feedforward architecture. Hence, except for the C2 to VTU connections, which are learned, all connections were preset by picking a simple pooling scheme in C1 (described above, resulting in 46,000 C1 cells) and a combinatorial rule to create S2 features from C1 inputs (yielding close to three million S2 cells), which were then pooled over to yield the final 256 complex composite feature dectectors in C2. The actual number of cells required to perform the tasks is likely to be lower—in fact, a model with bigger pooling ranges in C1 resulting in about one-fourth the number of S2 cells has been shown to have very similar recognition rates. The crucial observation is that if additional objects are to be recognized irrespective of scale and position, the addition of only one unit in the top layer, with sparse connections to the (256) C2 units, is required. Furthermore, in the case of a distributed code, where individual neurons participate in the coding of several objects, requirements are likely to be even less. This does not appear to be specific to the class of paperclip objects: the exact same model described in this paper has already been applied successfully (with the only difference being the appropriate setting of the weights from the C2 units to the VTUs) to the recognition of computer-rendered images of cars (Riesenhuber and Poggio, 1996b). Thus, the recognition of different classes of objects would only require the addition of more view-tuned units in the top layer of the network.

How could stimulus qualities other than shape, such as color, be added to the model? There are two straightforward options, both of which have some experimental support: (1) to make cells in the first layer color selective, i.e., to have additional sets of S1 cells, at each orientation and scale, for different colors; or (2) to have “blob” cells in S1 that respond to a certain color, in addition to the present noncolor-selective S1 cells. Driven by the demands of the task in the same way composite features would be learned in a shape-only model, higher S layers would combine units selective to different colors, or units tuned to a certain orientation and blobs responding to a certain color, respectively. So far, there are no systematic physiological studies investigating the extent of color tuning of IT cells to a degree similar to the studies on shape tuning: current studies, such as Missal et al. (1997), have mainly limited themselves to global changes of object color. Once more detailed data are available, it will be interesting to see which of the two schemes, or if a combination of both, can yield the required color specificity.

Although clearly further work is required to determine whether the generality and power of a representational scheme such as the one outlined here could ultimately account for the vast representational abilities of the brain, we believe that the model provides evidence that “the binding problem,” as commonly conceived, may only be a problem in the eye of the beholder, but it is not necessarily a problem for all object recognition devices and perhaps may not be one for the brain.

References

A comprehensive reference list for all reviews can be found on pages 111–125.