Object recognition in cortex: Neural mechanisms, and possible roles for attention

Maximilian Riesenhuber

Department of Neuroscience
Georgetown University Medical Center
Washington, DC 20007
phone: 202-687-9198
fax: 202-784-3562
Email: mr287@georgetown.edu

Abstract

The primate visual system can rapidly and with great accuracy recognize a large number of diverse objects in cluttered scenes under widely varying viewing conditions. Recent data (Thorpe *et al.*, 1996; Li *et al.*, 2002) have suggested that complex object recognition tasks can be performed in one feedforward pass without the need for attention, providing strong constraints for models of object recognition in cortex. I will review a "Standard Model" that is an extension of the original model of simple and complex cells of Hubel and Wiesel. Despite its simplicity, this feedforward model can already explain a number of experimental findings, and has been shown to be able to perform object detection in natural images. Moreover, the model can be extended in a straightforward way to investigate how "top-down" attention can modulate "bottom-up" processing to improve its performance. This leads to constraints on the scenarios in which attention can aid object recognition, and to experimental predictions on how attention and feedforward processing might interact.

Key words: object recognition, computational modeling, attention, object representation, neuroscience

1 Introduction

Object recognition is a fundamental cognitive task essential for survival, *e.g.,* to detect predators or to discriminate food from non-food. Despite the apparent ease with which the visual system performs object recognition, it is a very complex computational task requiring a quantitative trade-off between invariance to certain object transformations on the one hand, and specificity for individual objects on the other. For instance, object recognition needs to be invariant across huge variations in the appearance of objects such as faces, due to viewpoint, illumination, or occlusions. At the same time, the system needs to maintain specificity, *i.e.,* the ability to discriminate between different faces.

How does the brain perform object recognition, and what is the role of attention in this process? The experimental data paint a complex picture: Even very complicated visual tasks, such as determining whether an arbitrary natural scene contains an animal or not, can be performed in the absence of attention. However, other tasks that would appear to be "simpler" such as discriminating bisected colored discs from their mirror images seem to require attention (Li *et al.*, 2002).

In this chapter, I will first review some basic experimental data on object recognition in cortex which motivate a simple computational model that can be viewed as an extension of the original simple-to-complex cell scheme of Hubel and Wiesel. The model can perform object recognition in cluttered natural scenes in one feedforward pass, in agreement with the experimental data. I will then discuss how the model can be extended to incorporate attentional effects, and under what conditions attention can aid object recognition, leading to predictions for experiments.

2 Object Recognition in Cortex: Some Experimental Results

Object recognition in cortex is thought to be mediated by a hierarchy of brain areas called the "ventral visual stream" (Ungerleider and Haxby, 1994) extending from primary visual cortex (V1) to inferior temporal cortex, IT. IT in turn provides input to prefrontal cortex (PFC) which appears to play a crucial role in linking perception to action. Starting from *simple cells* in primary visual cortex, V1, with small receptive fields that respond preferably to oriented bars, neurons along the ventral stream 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 and other relevant stimuli from the monkey's environment (Logothetis and Sheinberg, 1996). A hallmark of these IT cells is the robustness of their firing to stimulus transformations such as scale and position changes (Logothetis and Sheinberg, 1996). In addition, as these and other studies have shown, most neurons show specificity for a certain object view or lighting condition (so-called view-tuned neurons), while some neurons show view-invariant tuning (view-invariant/object-tuned neurons). The tuning of the view-tuned and object-tuned cells in AIT can be modified by visual experience (for references, see (Riesenhuber and Poggio, 2002)). Recent fMRI data have shown a similar pattern of tuning properties for the Lateral Occipital Cortex (LOC), a brain region in human visual cortex central to object recognition and believed to be the homologue of monkey area IT (Grill-Spector *et al.*, 2001).

ERP experiments have established that the visual system is able to perform even complex recognition tasks such as object detection in natural images within 150ms (Thorpe *et al.*, 1996), which is on the order of the latency of neurons in prefrontal cortex, close to the site of the measured ERP effect. Further experiments have shown that such detection tasks can be performed in the absence of attention (Li *et al.*, 2002), and in parallel for two images (Rousselet *et al.*, 2002). These results point to a feedforward account of object recognition in cortex — at least for object detection tasks in natural images — in which recognition is achieved in one processing pass through the ventral visual stream.

3 The "Standard Model"

The data described in the previous section motivate a "Standard Model" of visual processing in cortex, which reflects in its general structure the average belief of many visual physiologists and cognitive scientists. We have provided a quantitative computational implementation (Riesenhuber and Poggio, 1999b) that demonstrates the feasibility of the basic architecture, and allows us to integrate experimental data in a rigorous framework and make quantitative predictions for new experiments. 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 the inset in Fig. 1, leads to model units showing the same scale and position invariance properties as view-tuned IT neurons (Riesenhuber and Poggio, 2002; Logothetis and Sheinberg, 1996; Riesenhuber and Poggio, 1999b). Computationally, this is accomplished by a scheme consisting of a hierarchy of just two operations: i) a "MAX" operation, and ii) a "template match" operation.

In detail, the model proposes that a MAX pooling function, in which a cell's output is determined by its strongest afferent, provides invariance to scaling and translation as well as robustness to clutter while maintaining feature specificity. To illustrate the idea of MAX pooling, consider the example of simple and complex cells in primary visual cortex: Both simple and complex cells respond to bars of a certain orientation, but while simple cells have separate on and off regions (*i.e.,* responding to light and dark bars, resp.) and small receptive fields, complex cells have larger receptive fields with overlapping on/off regions. In the model, complex cells (C1 units in Fig. 1) increase translation invariance by performing a MAX pooling operation over simple cells (S1) tuned to the same feature but at different positions (and phase). Besides increasing translation (and scale) invariance, the MAX pooling function is also advantageous for object recognition in clutter: By design, the MAX operation only selects the strongest input to a cell, and the response is not affected by the presence of other objects that activate other afferents to a lesser degree (but might cause strong activation of the afferents to another cell, *e.g.,* one tuned to a different orientation). Thus, the MAX operation provides a biologically plausible mechanism to perform invariant object recognition in clutter (see (Riesenhuber and Poggio, 1999a)) without the need for a separate segmentation process or special neural circuits to reroute the visual input to a standard reference frame, which are challenged by the timing constraints for object detection imposed by the experimental data, in particular for visual scenes containing more than a single object (but see [DIRECTED VISUAL ATTENTION AND THE DYNAMIC CONTROL OF INFORMATION FLOW]).

While oriented edges are a good model for the preferred features of V1 neurons, neurons in higher areas along the ventral stream are tuned to more complex shapes. This is achieved in the model by the other basic neural mechanism, a "template match" operation in which feature complexity is increased by combining simpler features into more complex ones (*e.g.,* from the C1 to S2 and C2 to VTU levels in Fig. 1). This operation is performed at different levels in the hierarchy to build increasingly complex features while maintaining invariance.

In the second part of the architecture, arbitrary transformations can be learned by interpolating between multiple examples, *i.e.,* different view-tuned neurons, leading to neural circuits performing specific tasks. The key idea here is that interpolation and generalization can be obtained by simple networks that learn from a set of examples, that is input-output pairs (Poggio and Girosi, 1990). In this case, 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 weights from the view-tuned units to the output are learned from the set of examples (see Riesenhuber and Poggio (2002)). In principle, two networks sharing the same VTU input units but with different weights (from the VTUs to the respective output units), could be trained to perform different tasks such as pose estimation, view-invariant recognition, or categorization.

Despite its simplicity, our implementation of this "Standard Model" of object recognition in cortex has turned out to explain a number of experimental results (for a recent review, see Riesenhuber and Poggio (2002)), and make predictions for new experiments. Importantly, there is now evidence from physiology for the MAX pooling prediction, in complex cells in V1 (I. Lampl, M. Riesenhuber, T. Poggio, D. Ferster, *Soc. Neurosci. Abs. 2001*) as well as V4 (Gawne and Martin, 2002). Also, recent data from an experiment in which monkeys were trained to categorize “cat” and “dog” stimuli followed by recordings from the animals’ IT and PFC support the model prediction of a shape-based but object-class specific representation (in this case for “cat”/“dog”-like shapes) that provides input to task-specific circuits, in this case trained on the categorization task (in IT and PFC, resp., see (Freedman *et al.*, 2003)).

While the “Standard Model” is purely feedforward and thus in principle fast enough to explain the data of Thorpe *et al.* (1996), the question is whether such a simple model can indeed perform real-world object recognition tasks. Recent work (Serre *et al.*, 2002; Louie, 2003) has provided some very encouraging results: The feedforward architecture of Fig. 1 can detect objects (in this case, faces) in natural images, at a level comparable or even superior to state-of-the-art machine vision systems (Fig. 2). Key to the model’s success on this difficult task is the learning of a set of object class-specific features, at the S2 level in the model, roughly corresponding to V4 cells in cortex. In combination with the MAX pooling operation, this specialized set of features allows the system to isolate the relevant features from the surrounding clutter, without the need for a separate segmentation step. Moreover, the specialized object class representation also greatly simplifies the complexity of the learning problem, permitting the use of a simple linear classifier (see curve for “Kmeans classifier” in Fig. 2) similar to the architecture shown in Fig. 1.

The observed difference in the attentional demands of different recognition tasks (Li *et al.*, 2002) could thus be related to the “naturalness” of the objects involved: If the visual system has learned optimized features of intermediate complexity for familiar objects (like animals in natural scenes), this would facilitate detection of these objects in a single feedforward processing pass even in the presence of clutter. In the case of unfamiliar or artificial objects, such as bisected colored disks (Li *et al.*, 2002), for which there are no specialized features, however, the higher level of interference caused by the simultaneous presence of other objects is more likely to prevent “attentionless” recognition of the target objects (*i.e.,* in one feedforward pass through the ventral stream), as observed in the experiment (Li *et al.*, 2002). In section 4 we will discuss possible mechanisms of how attention could modulate processing in the recognition system to improve performance in such cases.

3.1 Limitations of the Feedforward Approach

As described in the previous section, the purely feedforward “Standard Model” is already able to perform very complex object recognition tasks. However, the feedforward architecture has some limitations:

- There is experimental evidence that the visual system can exploit informa-

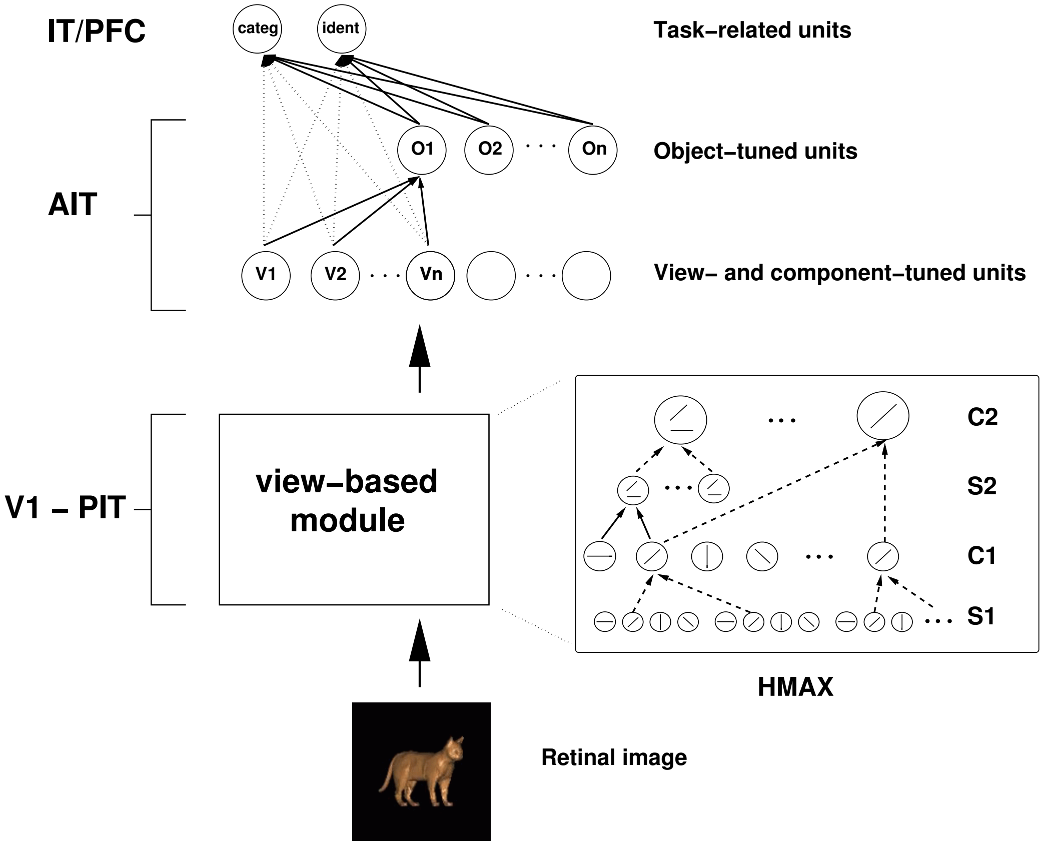

Fig. 1. Sketch of the "Standard Model" of the recognition architecture in cortex. It combines and extends several recent models and effectively summarizes many experimental findings. The view-based module shown in the inset is an hierarchical extension of Hubel and Wiesel's classical paradigm of building 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 mechanisms (a MAX pooling mechanism (dashed lines), to increase invariance, and a template match operation (solid lines), to increase feature specificity, see text). The output of the view-based module is represented by view-tuned model units V_n that exhibit tight tuning to rotation in depth (and illumination, and other object-dependent transformations such as facial expression, *etc.*) but are tolerant to scaling and translation of their preferred object view. Invariance to rotation in depth (or other object-specific transformations) is obtained by combining in a learning module several view-tuned units V_n tuned to different views (or differently transformed versions) 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 same generic learning circuitry but are trained with appropriate sets of examples to perform specific tasks. 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 prefrontal cortex (PFC). For more information on the model, including source code, see `http://riesenhuberlab.neuro.georgetown.edu/hmax`. Modified from (Riesenhuber and Poggio, 2002).

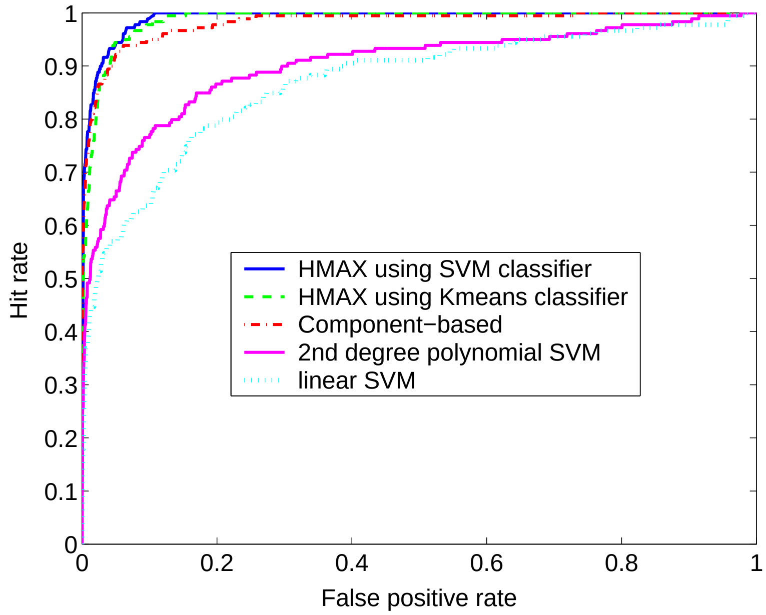

Fig. 2. The feedforward model of object recognition in cortex can perform face detection in natural images (faces were a subset of CMU PIE database, non-faces were selections from natural scenes selected as “face-like” by an LDA classifer, see (Louie, 2003)) at a level comparable to that of one of the best available machine vision face detection systems (Heisele *et al.*, 2002). The figure shows ROC curves of the biological system with feature learning (“HMAX”) and different machine vision face detection systems. The “HMAX” system was like the one shown in Fig. 1, with the difference that S2 features were learned from a set of faces images (Serre *et al.*, 2002). This object-class specific feature set enabled the system to robustly detect faces in cluttered images. For details, see (Louie, 2003).

tion about target location (spatial cues, *e.g.,* (Posner, 1980)) to enhance processing at the location of interest. The model cannot explain such “top-down” effects.

- There are situations where the feedforward system is overwhelmed and cannot correctly detect the object of interest, for instance, when the target appears together with a number of other objects and the surrounding clutter interferes with the representation of the target object (see Tsotsos (1990); Riesenhuber and Poggio (1999a)). This is the case in some visual search tasks where the visual system appears to resort to a serial approach to sequentially process different parts of the visual input.

4 Extending the Feedforward System: Roles for Top-Down Attentional and Task-Dependent Modulations

4.1 Spatial Cueing

It is straightforward to incorporate task-relevant information in form of a spatial cue into the framework of the feedforward model by appropriately modulating the pooling range of units performing a MAX operation. In this way, spatial attention could enhance signals from the region of interest and suppress input from nonrelevant parts of the visual field (see Fig. 3). This is compatible with reports from physiology that show that the receptive fields of neurons in V4 can constrict around the location of interest (Luck *et al.*, 1997), and similar observations of enhancement of processing for the region of interest and suppression elsewhere in fMRI experiments [BIASING COMPETITION IN HUMAN VISUAL CORTEX,SPATIALLY-SPECIFIC ATTENTIONAL MODULATION REVEALED BY FMRI]. Regarding the underlying neural mechanisms, recent data suggest that deploying spatial attention to a region that includes the receptive field of a particular neuron causes a leftward shift of that neuron's contrast response curve [VISUAL CORTICAL CIRCUITS AND SPATIAL ATTENTION]. Thus, focusing attention on a particular region in space would be equivalent to raising the effective contrast of that part of the input (and conversely, non-attended regions would be expected to show a lowered effective contrast). In the framework of the model, such a modulation of effective contrast directly reduces the interference caused by non-attended regions, as high-contrast stimuli cause higher responses which are more likely to win the MAX competition and thus determine the response of pooling units along the pathway and ultimately of view-tuned IT neurons (Riesenhuber and Poggio, 1999a). This parallel between attentional modulation and contrast is also very appealing since it directly relates to the notion of "salience" at the heart of popular models of attentional selection Itti and Koch (2000) [MODELS OF BOTTOM-UP ATTENTION AND SALIENCY].

4.2 Nonspatial Cueing

The case of nonspatial cueing is not as straightforward as the spatial case, however. While a spatial signal can be translated into a modulatory signal for cells at all levels of the processing hierarchy depending on the overlap of a neuron's receptive field with the extent of the "spotlight" of attention, it is not clear how a nonspatial, *e.g.,* object-level cue (such as "look for a face") can be translated into response modulations of neurons tuned to different features along the ventral pathway to selectively improve detection of the object of interest. For instance, if the goal is to detect "a face", there is a multitude of potential target objects and it is not clear which neurons should be modulated and in which way to increase the detectability of any face *vs.* non-faces. Consider the simplest case: Assume the target is a particular face in a particular pose (lighting condition, expression *etc.*), and that further there is a particular view-tuned cell in IT tuned to this exact face (a so-called "grandmother cell"), and the target face is declared "detected" if the activation of this VTU exceeds a certain threshold. How should afferent neurons tuned to simpler features in lower processing levels, *e.g.,* in V4, be modulated to improve the system's selectivity (*i.e.,* to improve detection without an increase in false alarms) for the target object? If the VTU is tuned to a characteristic distributed activation pattern over its afferents (with high and low activations, depending on which features are present in the face and to what degree), then how should those afferents be modulated, in particular in the absence of information about target contrast? Increasing the afferents' gain might change their response to the target object in such a way that the resulting activation pattern over the afferents could actually be less optimal to activate the VTU than the unmodulated activation pattern (for supporting simulation results, see (Schneider and Riesenhuber, 2004)). The situation is even more problematic in the more general case where object identity is encoded by a population of view-tuned units tuned to *e.g.,* different faces (Young and Yamane, 1992). Here, the same V4 neuron can provide input to different VTUs and conceivably receives top-down signals from more than one IT cell. How should the possibly different top-down inputs be combined to modulate the V4 neuron?

These computational arguments concerning the conceptual simplicity of spatial cueing on the one hand and the difficulties associated with nonspatial cueing on the other are compatible with reports from electrophysiology that suggest that featural and spatial attentional effects are qualitatively different. In particular, while spatial effects appear to occur close to response onset, nonspatial modulations appear to have a latency of at least 150ms (Motter, 1994; Hillyard and Anllo-Vento, 1998). A possible interpretation of these data could be that nonspatial attention only sets in *after* an initial feedforward pass through the visual system, when the precise shape, contrast *etc.* of the target and the activation it evokes at the different processing stages would be known. However, this would suggest very different roles for spatial and featural attention, at least for the case of rapidly presented images (when processing is limited mainly to a feedforward pass): While spatial cues could aid recognition as it is possible to "tune" the system before stimulus onset, information about features would not be able to enhance performance of the initial feedforward processing pass, but might serve to, *e.g.,* "highlight" instances of the target object in the visual field, as in a saliency map (Motter, 1994; Mazer and Gallant, 2003) to inform other processes such as eye movements to potential targets (see also section 5).

4.3 Visual Search

If the target object appears together with a number of similar distractor objects that interfere with the representation of the target object so much that it cannot be recognized in a feedforward pass anymore (see Fig. 3(b), dark bars) one computational strategy for the visual system is to "divide and conquer" to reduce the influence of clutter and detect the target by sequentially analyzing parts of the image. This piecemeal approach could use the same mechanism underlying spatial attention described above, *i.e.,* spatial modulation of the pooling range of neurons along the processing hierarchy, but now controlled not by external spatial cues but by, for instance, a "saliency map" (Itti and Koch, 2000) [MODELS OF BOTTOM-UP ATTENTION AND SALIENCY] as shown in Fig. 3 (Walther *et al.*, 2002). It is an interesting question whether this saliency map is based solely on "bottom-up" factors such as orientation or intensity contrast, or whether there are task-specific components to saliency (that would serve to, *e.g.,* increase the salience of purple regions when looking for Barney, the Dinosaur). Also, given the feature-based modulations of V4 neurons and their possible role in object recognition described in section 4.2, it is interesting to ask whether more complex features, like those represented by V4 neurons (Kobatake and Tanaka, 1994) are integrated into the saliency map. Clearly, investigating this link between attention and recognition should be a priority for future studies.

5 Summary and Conclusions

"Basic" object recognition tasks such as object detection in natural images can be understood to a first approximation as resulting from a single feedforward pass through the processing hierarchy of the ventral visual stream in cortex from primary visual cortex, V1, to inferotemporal cortex, IT. A hierarchical computational model of the ventral stream based on just two operations, a MAX pooling function to increase tolerance to stimulus translation and scaling and a template match operation to increase feature complexity, can provide an explanation for the shape tuning and invariance properties of view-tuned cells in IT, and explain how the ventral stream can perform object detection in complex natural scenes. The model makes very few assumptions. For instance, model unit responses have no dynamics, and there are no lateral interactions, oscillations or synchronous ensembles of units. This does not mean such mechanisms do not play a role in vision. However, the simulation results show that they are not necessary to explain the relevant data on object recognition.

Not surprisingly, simulations show that there are situations where such a sim-

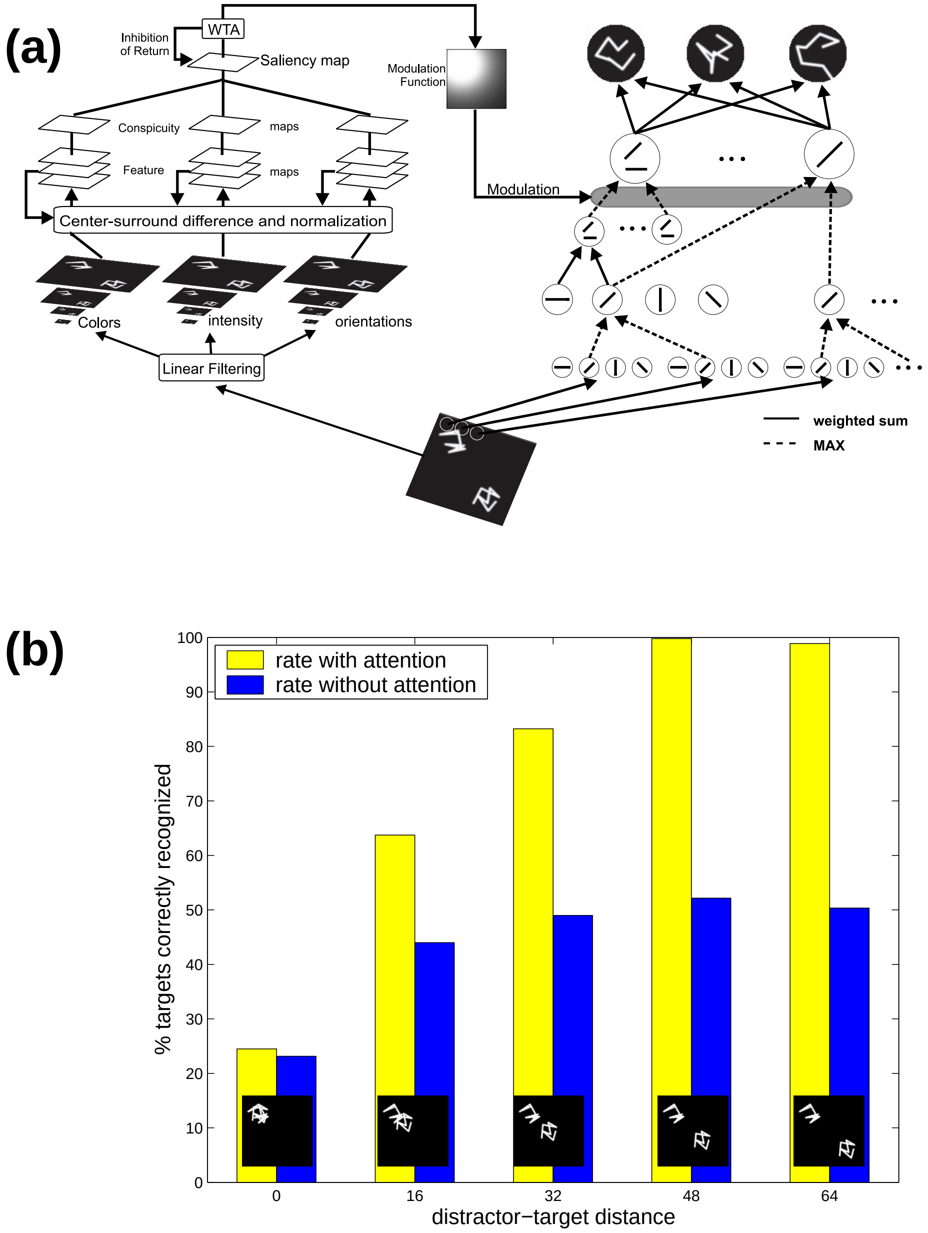

Fig. 3. Coupling of the Saliency Map model of attention (Itti and Koch, 2000) and the model of object recognition (Riesenhuber and Poggio, 1999b). **(a)** Sketch of the integrated system: The saliency map (left) provides a modulatory signal to C2 model units (corresponding to V4 neurons in cortex) to modulate their pooling range, causing their receptive fields to focus around the location of interest selected by the saliency map. **(b)** Recognition results with and without attentional modulation, as a function of target and distractor stimulus separation. While a purely feedforward analysis of the image yields only poor performance, iterative piecewise analysis of the image through attentional modulation of the spatial extent of receptive fields dramatically improves recognition performance. From (Walther *et al.*, 2002).

ple feedforward system breaks down, for instance, in the case of visual clutter when the receptive field of a model IT unit contains several interfering objects. Similar effects are observed in natural vision, and it will be interesting to compare the conditions under which feedforward vision fails in the model and in the experiment. As the modeling studies on face detection demonstrate (Serre *et al.*, 2002; Louie, 2003), familiarity with objects from the target class is expected to play a crucial role: If a subject is well trained on a certain object class such that specific intermediate representations have been learned, then interference caused by simultaneously presented distractor objects and thus attentional demands for this task should decrease. Reports that some initally "serial" tasks can become "parallel" with practice (Sireteanu and Rettenbach, 1995), and that well-practiced object recognition tasks (such as animal detection in natural scenes) do not seem to require attention whereas more artificial ones (like discriminating bisected colored disks) do (Li *et al.*, 2002) are compatible with this hypothesis.

The challenge to perform object recognition also in more difficult situations when the feedforward system fails, together with experimental data that show spatial cueing effects in behavior as well as attention-related modulations of processing observed in physiology and fMRI, has motivated extensions of the basic model to incorporate spatial modulations of processing to reduce the effect of clutter. These modulations can be based on explicit information about the target (for instance in form of a spatial cue), or could possibly be driven more indirectly by stimulus saliency (Itti and Koch, 2000; Walther *et al.*, 2002) as in serial visual search. In the case of spatial attention, information about target location permits a "tuning" of the visual system prior to stimulus exposure that can improve the performance of the feedforward system.

From a computational point of view, featural attention, where the system is given information about the shape of a target but not its location, seems to be fundamentally different (Schneider and Riesenhuber, 2004): While the translation of information about target location into a neuronal modulation pattern is straightforward based on the match of the location of a cell's receptive field with the region of interest, and can be applied to any cell irrespective of its position in the processing hierarchy, information about complex target objects is difficult to translate into appropriate attentional modulations of simpler feature detectors in lower levels of the hierarchy. Nevertheless, object recognition tasks that leave sufficient time to perform iterations of feedforward and feedback processing might profit from featural attention. In these cases, an initial feedforward pass could provide hypotheses about possible targets which could then provide specific top-down signals to influence lower levels of processing, possibly explaining observed effects of featural attention on task performance in some experiments (Rossi and Paradiso, 1995; Blaser *et al.*, 1999; Lee *et al.*, 1999). Alternatively, these results might suggest the intriguing and computationally feasible hypothesis that pre-stimulus "featural tuning" of feedforward processing is possible for the basic visual features that neurons at the lower processing levels are tuned to, such as color and orientation. Such modulations of processing could consist in a sharpening of tuning curves (possibly paired with an increase in gain akin to the aforementioned increase in effective contrast in the spatial case) (Lee *et al.*, 1999) of those neurons that are directly tuned to the target stimulus. Opportunities abound for interesting hypothesis-driven experiments seeking to clarify the role of attention in object recognition.

6 Acknowledgments

Thanks to Tommy Poggio for discussions, to Thomas Serre for creating Fig. 2, and to Dirk Walther for creating Fig. 3.

References

E. Blaser, G. Sperling, and Z.L. Lu. Measuring the amplification of attention. *Proc. Nat. Acad. Sci. USA*, 96:11681–11686, 1999.

D.J. Freedman, M. Riesenhuber, T. Poggio, and E.K. Miller. Comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J. Neurosci.*, 23:5235–5246, 2003.

T.J. Gawne and J.M. Martin. Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. *J. Neurophys.*, 88:1128–1135, 2002.

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

B. Heisele, T. Serre, M. Pontil, T. Vetter, and T. Poggio. Categorization by learning and combining object parts. In *Advances in Neural Information Processing Systems*, volume 14, 2002.

S.A. Hillyard and L. Anllo-Vento. Event-related brain potentials in the study of visual selective attention. *Proc. Nat. Acad. Sci. USA*, 95:781–787, 1998.

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

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

D.K. Lee, L. Itti, C. Koch, and J. Braun. Attention activates winner-take-all competition among visual filters. *Nat. Neurosci.*, 2:375–381, 1999.

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

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

J. Louie. A biological model of object recognition with feature learning. Master's thesis, MIT, Cambridge, MA, 2003.

S.J. Luck, L. Chelazzi, S.A. Hillyard, and R. Desimone. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J. Neurophys.*, 77:24–42, 1997.

J. A. Mazer and J. L. Gallant. Goal-related activity in v4 during free viewing visual search. evidence for a ventral stream visual salience map. *Neuron*, 40:1241–1250, 2003.

B.C. Motter. Neural correlates of feature selective memory and pop-out in extrastriate area V4. *J. Neurosci.*, 14:2190–2199, 1994.

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

T. Poggio and F. Girosi. Networks for approximation and learning. *Proc. IEEE*, 78(9):1481–97, 1990.

M.I. Posner. Orienting of attention. *Quart. J. Exp. Psych.*, 32:3–25, 1980.

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

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

M. Riesenhuber and T. Poggio. Neural mechanisms of object recognition. *Curr. Op. Neurobiol.*, 12:162–168, 2002.

A.F. Rossi and M.A. Paradiso. Feature-specific effects of selective visual attention. *Vis. Res.*, 35:621–634, 1995.

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

R. Schneider and M. Riesenhuber. On the difficulty of feature-based attentional modulations in visual object recognition: A modeling study. Technical Report AI Memo 2004-004, CBCL paper 235, MIT AI Lab and CBCL, Cambridge, MA, 2004.

T. Serre, M. Riesenhuber, J. Louie, and T. Poggio. On the role of object-specific features for real world object recognition. In H.H. Buelthoff, S.-W. Lee, T. Poggio, and C. Wallraven, editors, *Proceedings of BMCV2002*, volume 2525 of *Lecture Notes in Computer Science*, New York, 2002. Springer.

R. Sireteanu and R. Rettenbach. Perceptual learning in visual search: fast, enduring, but non-specific. *Vis. Res.*, 35:2037–2043, 1995.

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

J.K. Tsotsos. Analyzing vision at the complexity level. *Behav. Brain Sci.*, 13:423–469, 1990.

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

D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional selection for object recognition — a gentle way. In H.H. Buelthoff, S.-W. Lee,

T. Poggio, and C. Wallraven, editors, *Proceedings of BMCV2002*, volume 2525 of *Lecture Notes in Computer Science*, New York, 2002. Springer.

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