MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1695
C.B.C.L. Paper No. 190

August 7, 2000

Computational Models of Object Recognition in Cortex: A Review

Maximilian Riesenhuber and Tomaso Poggio

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.

Abstract

Understanding how biological visual systems perform object recognition is one of the ultimate goals in computational neuroscience. Among the biological models of recognition the main distinctions are between feedforward and feedback and between object-centered and view-centered. From a computational viewpoint the different recognition tasks — for instance categorization and identification — are very similar, representing different trade-offs between specificity and invariance. Thus the different tasks do not strictly require different classes of models. The focus of the review is on feedforward, view-based models that are supported by psychophysical and physiological data.

This report describes research done within the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences and in the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. This research is sponsored by a grant from Office of Naval Research under contract No. N00014-93-1-3085, Office of Naval Research under contract No. N00014-95-1-0600, National Science Foundation under contract No. IIS-9800032, and National Science Foundation under contract No. DMS-9872936. Additional support is provided by: AT&T, Central Research Institute of Electric Power Industry, Eastman Kodak Company, DaimlerChrysler Corp., Digital Equipment Corporation, Honda R&D Co., Ltd., NEC Fund, Nippon Telegraph & Telephone, and Siemens Corporate Research, Inc. M.R. is supported by a Merck/MIT Fellowship in Bioinformatics.

1 Introduction

1.1 Object recognition is a difficult computational problem

Imagine waiting for incoming passengers at the arrival gate at the airport. The small camera in the buttonhole of your lapel looking at the incoming crowd suddenly tells you that Dr. Jennings is the third from right, partly occluded by a woman in front. Today his tie — the camera says — shows a pattern of antique cars. Computer vision is well on its way to solve restricted versions of the problem of object recognition — both in identification (recognizing Jenning's specific face) and categorization (recognizing the patterns on the tie as cars). A system, however, that were capable of categorizing all types of objects in complex images, of recognizing individual objects like faces and of providing human-level performance under different illuminations and viewpoints would pass the Turing test for vision. Not surprisingly, such a general and flexible system is still the stuff of science fiction. Object recognition is at the top of a hierarchy of visual tasks. In its general form, it is a very difficult computational problem, which is likely to play a significant role in eventually making intelligent machines. Not surprisingly it is an even more difficult, open and key problem for neuroscience.

1.2 Multiple tasks and strategies in object recognition

As the airport scenario shows, an object can be recognized at different levels of specificity. It can be categorized as a member of a general class, such as "face" or "car". It can also be identified as a unique individual, such as "Jenning's face" or "my car".

Identification and categorization are the two main tasks in recognition.* Which of the two tasks is easier and which comes first? The answers from neuroscience and computer vision are strikingly different. Typically, computer vision techniques found identification relatively easy — as shown by the several companies selling face identification systems — and categorization close to impossible. Psychologists and neuroscientists like to tell the opposite story (see [51] and, for reviews, [24, 49]): in biological visual systems, categorization seems to be the simpler and more immediate stage in the recognition process. In any case it has been common in the past few years — in computer vision and especially in visual neuroscience — to assume that different strategies are required for these different recognition tasks (but see the books by Edelman [10] and Ullman [49]).

*Together with motor-related shape estimation.

1.3 A continuum of recognition tasks along the trade-off between specificity and invariance

In this review we start from a rather different *Ansatz*. From a theoretical standpoint, identification and categorization, rather than two distinct tasks, represent two points in a spectrum of generalization levels [49].† An appropriate theoretical framework for object recognition is computational learning. Within it, the distinction between identification and categorization is mostly irrelevant. The relevant variables are the size of the training set, the universe of distractors, the number of classes and the "legal" transformations allowed for generalization.

We believe that the crux of the problem of object recognition is the trade-off between (object, class) specificity and (transformation) invariance (see below). The distinction in the literature between different flavours of identification and categorization is an idiosyncratic tale of this trade-off.

1.4 The same basic computational strategy can be used from identification to categorization

From this reasoning we expect that the same computational strategies could be adapted to perform either identification or categorization. Thus, the existence of multiple recognition tasks does not require radically different algorithms or representations. Multiple recognition strategies are nevertheless very likely to exist in biological vision, such as the "immediate", perceptual recognition based on similarity of visual appearance as opposed to recognition based on motion dynamics or reasoning-based recognition (*e.g.,* interpreting maps or technical diagrams). Thus we both expect multiple strategies (*e.g.,* algorithms) for the same recognition task and the same basic strategy for multiple tasks (*e.g.,* identification and categorization).

1.5 Models and experiments

The main reason for the lengthy introduction is our belief that models are necessary to make sense of the data and more importantly to plan new experiments. Without quantitative models (properly tested), it may be difficult to ask the right questions.

Our brief review of models of object recognition is far from comprehensive, even within the caveats of the previous discussion. For instance, we will confine ourselves to the recognition of isolated objects (for some ideas on biological object recognition in clutter see [1, 36]), and will focus on recent developments and on

†Notice that even identification tasks can differ widely in difficulties and requirements: consider, for instance, the problem of identifying the image of a specific old high-school friend among all the pictures of faces stored among the terabytes of the World Wide Web versus identifying who among my two siblings is in the picture that my mother has on her coffee table.

those models that can be directly related to experimental physiological data. We show that a mix of models and data has brought us closer to understanding some of the cortical mechanisms of recognition and will discuss key questions that lie ahead.

2 Models: Object-Centered and View-Based, Feedforward and Feedback

The models proposed to explain object recognition can be coarsely divided into two categories: object-centered and view-based (or image-based or appearance-based). In the former group of models, the recognition process consists in extracting a view-invariant structural description of the object that is then matched to stored object descriptions. One of the most prominent models of this type is the "Recognition-by-Components" (RBC) theory of Biederman [3, 19], whose emphasis on representing an object by decomposing it into basic geometrical shapes is reminiscent of the scheme proposed by Marr and Nishihara [25]. RBC predicts that recognition of objects should be viewpoint-invariant as long as the same structural description can be extracted from the different object views.

In contrast to structural description models, the basic tenet of view- or image-based models is that objects are represented as collections of view-specific features, leading to recognition performance that is a function of previously seen object views. In the following, the term "view" is used in the broad sense of "image-based appearance". Thus different views correspond to different appearances, due, for instance, to different viewpoints or different illuminations, or different conditions such as different facial expressions. A view is not restricted to contain just 2D information; it may have 3D information as well, for instance because of stereo or structure-from-motion. A zoo of view-based models of object recognition exists in the literature, each employing very different computational mechanisms. Two major groups of models can be discerned based on whether they employ a purely feedforward model of processing or utilize feedback connections (for the recognition process, *i.e.,* excluding a learning phase, in which top-down teaching signals are likely to be used).

Feedback models include architectures that perform recognition by using an analysis-by-synthesis or hypothesis-and-test approach: the system makes a guess about the object that may be in the image, synthesizes a neural representation of it relying on stored memories, measures the difference between the hallucination and the actual visual input and proceeds to correct the initial hypothesis. The models of Rao & Ballard [35], or of Mumford [30], and in part Ullman's [49] belong to this category. Other models use feedback control to "renormalize" the input image in position and scale before attempting to match it to a database of stored objects (as in the "shifter" circuit [2, 31]), or to conversely tune the recognition system depending on the object's transformed state (for instance by matching filter size to object size [15]).

While feedback processing is essential for object recognition in the previous group of models, other image-based models rely on just feedforward processing. One of the earliest representatives of this class of models is the "Neocognitron" of Fukushima [12], a hierarchical network in which feature complexity and (translation) invariance were alternatingly increased in different ("S" and "C", resp.) layers of a processing hierarchy by a template match, and a pooling operation over units tuned to the same feature but at different positions, respectively. The concept of pooling of units tuned to transformed versions of the same object or feature was subsequently proposed by Perrett & Oram to explain invariance also to non-affine transformations such as invariance to rotation in depth or illumination changes [33]. Indeed, Poggio & Edelman [34] had shown earlier that view-invariant recognition of an object was possible by interpolating between a small number of stored views of that object.

The strategy of using different computational mechanisms to attain the twin goals of invariance and specificity (as opposed to a homogeneous architecture as used in, for example, Wallis' and Rolls' VisNet [53]) has been employed successfully in later models, among them Mel's SEEMORE system [26] that represented objects by histograms over various feature channels, and the HMAX model by Riesenhuber and Poggio [36, 37, 45], whose structure is similar to Fukushima's Neocognitron with its feature complexity-increasing "S" layers and invariance-increasing "C" layers. HMAX, however, uses a new pooling mechanism (a MAX operation) to increase invariance in the "C" layers, in which the most strongly activated afferent determines the response of the pooling unit, endowing the system with the ability to isolate the feature of interest from non-relevant background and thus build feature detectors robust to translation and scale changes, and even clutter [36]. More complex features in higher levels of HMAX are thus built from simpler features with tolerance to deformations in their local arrangement due to the invariance properties of the lower level units. In this respect, HMAX is similar to (so far non-biological) recognition architectures based on feature trees which emphasize compositionality [1].

2.1 A basic module: feedforward and view-based

We are thus left with two major fault lines running through the landscape of models of object recognition: object-centered *vs.* image-based, and, within the latter group, feedforward *vs.* feedback models. How well do these different model classes hold up to constraints derived from neurophysiological data?

Psychophysical data from humans [5, 44, 47] as well as monkeys [22] point to a view-dependence of object recognition (for reviews, see [24, 46]). Interestingly, data from physiology also support a view-based theory: several studies have previously shown that cells in the inferotemporal cortex (IT) of macaque monkeys (an area thought to be crucial for object recognition [24, 43]) respond to views of complex objects, such as faces [4, 8]. Logothetis and co-workers [23] systematically studied the tuning properties of IT cells by training a monkey to perform recognition of "paperclip" objects, strictly controlling the object views the monkeys had been exposed to during training. Even though the monkeys had access to the full 3D shape description of the object (by presenting the object as rotating in depth by $\pm 10°$), psychophysical experiments [22] showed (in agreement with human studies [5]) that recognition was based around the views seen during training. Even more intriguing, when Logothetis *et al.* recorded from IT neurons of trained monkeys [23], they found cells tuned to *views* of the training objects, along with a much smaller number of view-invariant neurons tuned to objects the monkey had been trained to recognize from any viewpoint, as predicted by the model of Poggio and Edelman [34]. Moreover, psychophysical recognition performance and neuronal tuning seemed to be intimately related. Further constraining computational models of object recognition are findings from EEG studies [48] that have shown that humans appear to be able to perform object detection tasks (such as determining whether an image contains an animal or not) in natural images within 150 ms, which is on the order of the latency of visual signals from primary visual cortex to inferotemporal cortex [13, 40]. This does not rule out the use of feedback processing but strongly constrains its role in "immediate" object recognition.

In summary, the combined weight of experimental data and theoretical work strongly suggests that feedforward view-based models describe well one of the basic strategies used by the brain for "immediate" recognition of 3D objects. In the rest of the review we will focus on this class of models.

2.2 Invariance and specificity

The different approaches reviewed in the previous section illuminate a central issue in object recognition, namely the invariance–specificity trade-off: Recognition should be tolerant to object transformations such as scaling, translation, or viewpoint changes (and, for the case of categorization, also to shape variations within a class), *i.e.,* generalize over a variety of image changes, while at the same time being able to finely discriminate between different objects (for identification) or different object classes (for categorization). The visual system seems able to satisfy both goals of specificity and invariance simultaneously, but with different degrees of success depending on the transformation in question, as shown in the object identification experiment by Logothetis *et al.* [23]: while their view-tuned IT units (VTUs) generally showed only narrow (in terms of image similarity as measured, for instance, by correlation) invariance for rotation in depth, they show relatively great tolerance to changes in stimulus position and scale changes ([23], see caption of Fig. 1).

Thus, not all object transformations appear to be treated equally, in agreement with computational considerations. The effects of affine transformations in the image plane, such as scaling or 2D translation, can be estimated exactly from just one object view. To determine the behavior of a specific object under transformations that depend on its 3D shape, such as illumination changes or rotation in depth, however, one view generally is not sufficient. These fundamental limitations are borne out by the observed invariance properties of the view-tuned IT neurons as described above: while it is possible to construct a translation- and scaling-invariant set of features that allows the system to perform position- and size-invariant recognition of novel objects, invariance to 3D-based transformations does not transfer as freely but has to be learned anew for each paperclip ([5, 34]) individually (for "nice" object classes in which the objects have a similar 3D shape and behave similarly under the transformation in question [51], [52], invariance might in part transfer to other class members, see below). In categorization, generalization is across members of the class. Thus, multiple example views are also needed to capture the appearance of multiple objects. Unlike affine 2D transformations, 3D rotations, as well as illumination changes and shape variations within a class, may require multiple example views during learning.

3 Models of Object Recognition: A Summary

Figure 2 summarizes, in an oversimplified and cartoonish way, the discussion above, putting together and extending models such as HMAX [36], Poggio & Edelman [34], Perrett & Oram [33], the Neocognitron [12], and even VisNet [53]. An initial "view-based module" stage takes care of the invariance to image-based transformations leading to view-tuned cells — several for each object. In the following, with view-tuned cells we mean cells tuned to a full or a partial view (*i.e.,* connected only to a few of the feature units activated by the object view [36]) of an object. At higher stages, invariance to rotation in depth (illumination, facial expression, *etc.*) is achieved by pooling together the view-tuned cells for each object. Finally, categorization and identification tasks, up to the motor response if necessary, are performed by circuits looking at the activities of the object-specific and view-invariant cells (or, in the absence of relevant view-invariant units, *e.g.,* when the subject has only experienced an object from a certain viewpoint as in the experiments on paperclip recognition [5, 22, 23], directly at the view-tuned units, as indicated by the dashed lines in Fig. 2) . In general, a particular object, say a specific face, will elicit different activity in the O_n cells tuned to a small number of "prototypical" faces [39]. Thus the memory of the particular face is represented in the identification circuit in an implicit way (*i.e.*, without dedicated "grandmother" cells) by a population code. In a similar way, a categorization module — say, for dogs *vs.* cats — uses as inputs the activities of a number of cells tuned to various animals, with weights set so that the unit responds differently to animals from different classes [38].

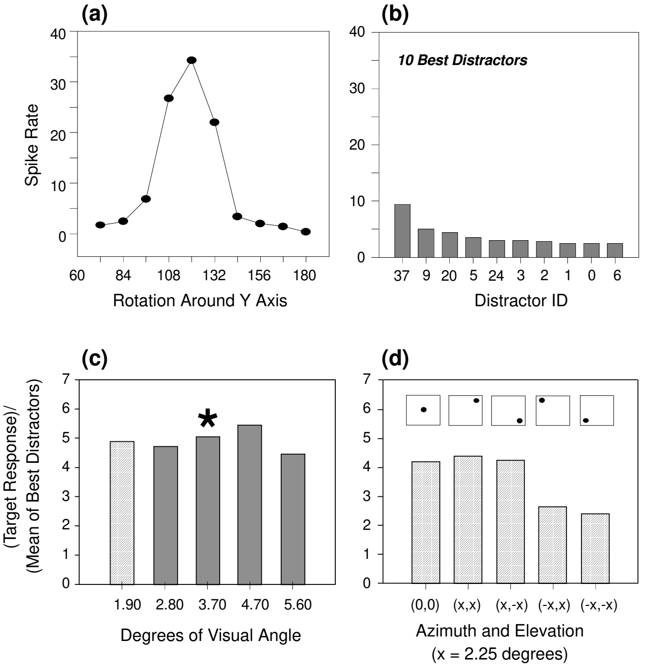

Figure 1: Invariance properties of one neuron (modified from Logothetis *et al.* [23]). The figure shows the response of a single cell found in anterior IT after training the monkey to recognize paperclip-like objects. The cell responded selectively to one view of a paperclip and showed limited invariance around the training view to rotation in depth, along with significant invariance to translation and size changes, even though the monkey had only seen the stimulus at one position and scale during training. **(a)** shows the response of the cell to rotation in depth around the preferred view. **(b)** shows the cell's response to the 10 distractor objects (other paperclips) that evoked the strongest responses. The lower plots show the cell's response to changes in stimulus size, **(c)** (asterisk shows the size of the training view), and position, **(d)** (using the 1.9° size), resp., relative to the mean of the 10 best distractors. Defining "invariance" as yielding a higher response to transformed views of the preferred stimulus than to distractor objects, neurons exhibit an average rotation invariance of 42° (during training, stimuli were actually rotated by $\pm 15°$ in depth to provide full 3D information to the monkey; therefore, the invariance obtained from a single view is likely to be smaller), translation and scale invariance on the order of $\pm 2°$ and ± 1 octave around the training view, resp. (J. Pauls, personal communication).

4 Beyond the Recognition of Specific Objects: Object Classes

4.1 A basic module for identification and categorization: sparse population codes

Most experimental studies of object recognition have focussed on testing the recognition performance on the same (small number of) objects used during training [5, 14, 22, 44]. However, in everyday object recognition, the ability to generalize from previously seen objects of a class to novel representatives of the same class, such as in the case of faces, is essential. The difference in the object recognition tasks — no generalization over shape in the first case in favor of high specificity *vs.* the ability to also discriminate between novel objects in the second case — appears also to be reflected in the neuronal tuning of object-tuned neurons: while Logothetis *et al.* [23] found neurons that were tightly shape-tuned ("grandmother"-like neurons), with a neuron responding to (a view of) just to a single object from the training set, recent studies of face cells in IT have argued for a distributed representation of this object class where the identity of a face is jointly encoded by the activation pattern over a *group* of face units [54, 55], corresponding in the model of Fig. 2 to an activation pattern over view-tuned (and object-tuned) units belonging to *different* objects (none of which in general is identical to the input object, unlike in the "grandmother" case). Discrimination (or memorization of specific objects, Fig. 3a) can then proceed by comparing activation patterns over the relevant (*i.e.*, the strongly activated) object- or view-tuned units — with the advantage that for a certain level of specificity, only the activations of a small number of units have to be remembered, forming a sparse code (in contrast to activation patterns on lower levels where units are less specific and hence activation patterns tend to be more distributed). Recent computational studies in our lab [39] have provided evidence for the feasibility of such a representation, with a discrimination performance comparable to that achieved by dedicated "grandmother" units. An interesting and non-trivial conjecture (supported by several

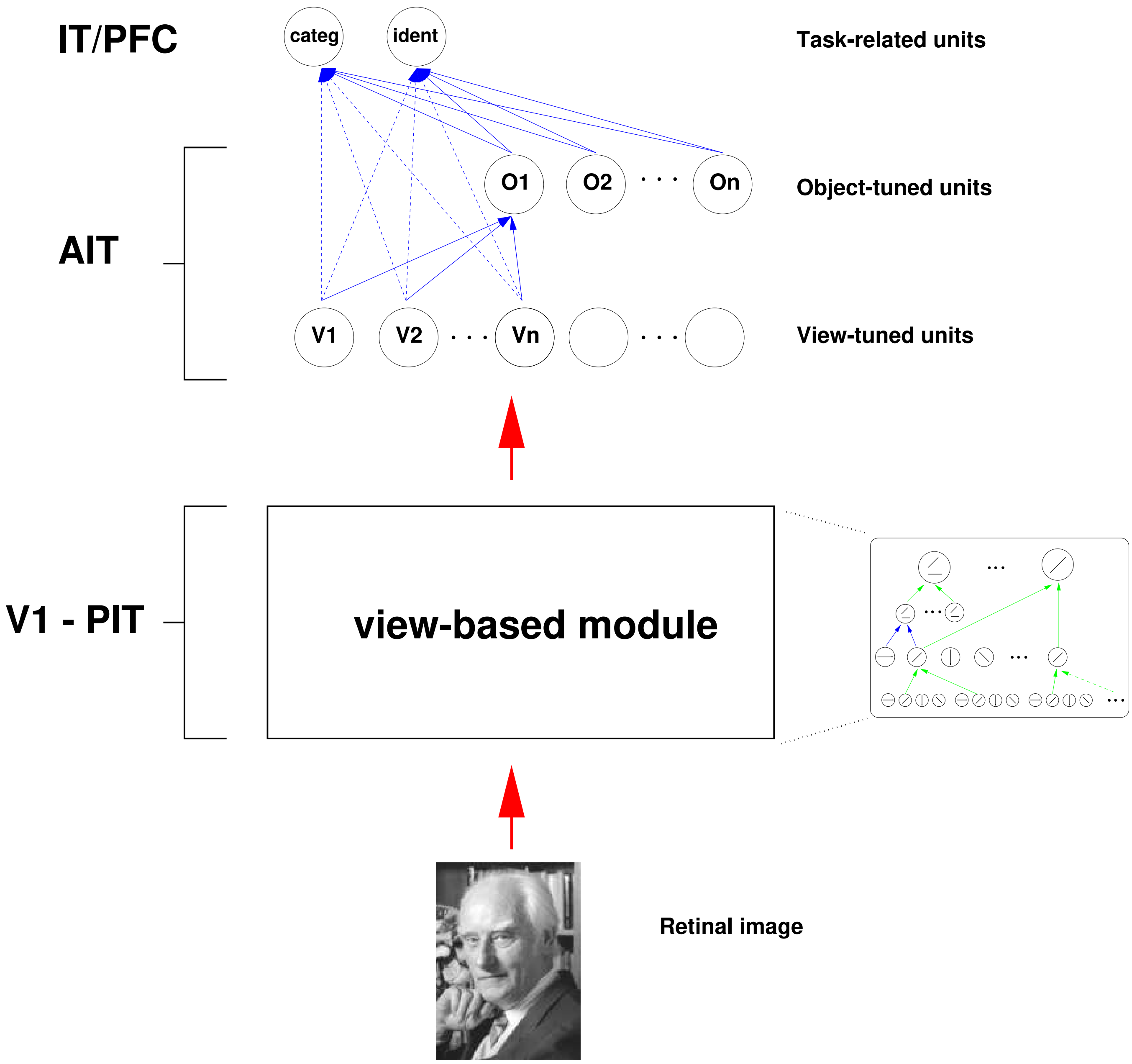

Figure 2: Sketch of a class of models of object recognition, combining and extending models such as Fukushima [12], Poggio & Edelman [34], Perrett & Oram [33], VisNet [53], and HMAX [37]. View-tuned units (V_n) (on top of a view-based module, as shown in the inset [37]) in the model exhibit tight tuning 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. Notice that the cells labeled here and in Fig. 3 as view-tuned units may be tuned to full or partial views. Invariance to rotation in depth (as an example of an object-dependent transformation) can then be significantly increased by interpolating between several view-tuned units tuned to different views of the same object [34], creating view-invariant (or object-tuned) units (O_n). These, as well as the view-tuned units, can then serve as input to task modules performing visual tasks such as identification/discrimination or object categorization (see below). Categorization could be supported even just by connections with the cells in the last layer of the model in the inset. For most object classes, the object-tuned units will be much fewer than the objects that can be identified. The stages up to the object-centered unit probably encompass V1 to anterior IT (AIT). The last stage of task dependent modules may be localized in prefrontal cortex (PFC) and beyond (see D.J. Freedman *et al.*, *Soc. Neurosci. Abs.*, 1999).

experiments [9, 28, 39], see also [10]) of this population-based representation is that it should be capable of generalizing from a single view of a new object of a nice [51] class — such as a specific face — to other views with a higher performance than for non-nice objects — such as paperclips.

The same substrate, a population-based class representation, has the nice property that it can also support categorization, as sketched in Fig. 3b: for instance, a "cat/dog categorization unit" can be connected to units responding to "cat" and "dog" prototypes[‡] in such a way that it shows different response levels for cats and dogs, respectively, as we have recently demonstrated [38].

4.2 A unified view

Thus, we see that the task-related "black boxes" at the top of Fig. 2 can possibly be realized as straightforward extensions of the previously proposed architecture: using template match operations, units in higher layers can learn to perform categorization tasks (Fig. 3b), or to identify individual objects ("my car", Fig. 3a). In all these cases, inputs to the top level units can be tuned to full or partial views (as illustrated in Fig. 3) or originate from object-tuned neurons, as described before. Thus, there is a dissociation of the tuning to individual stimuli and their labels in different recognition tasks, allowing the system, for instance, to implement different categorization schemes [38] and hierarchies of categories on the same stimuli.

An intriguing possibility is that these "top-level" units of Fig. 3a-b might themselves serve as inputs to other task-related units, as shown in Fig. 3c, *e.g.*, when learning additional hierarchy levels in a categorization scheme.

5 Challenges Ahead

5.1 Top-down and role of feedback

In this review we have taken the view that basic recognition processes take place in a bottom-up way; it is, however, very likely that top-down signals play an essential role in controlling the learning phase of recognition [18] and in some top-down effects (for instance in detection tasks, to bias recognition towards the features of interest, as suggested by physiological studies [6, 16, 27, 29]). There is an obvious anatomical substrate for top-down processing: the massive descending projections in the visual cortex that tend to reciprocate the forward connections. Ullman [49] suggested a role in top-down processing for matching models to inputs that is symmetric and as important as the bottom-up matching of inputs to models (see also the Helmholtz Machine [7]). Other roles for top-down processing have been proposed such as controlling attention [21], and grouping and synchronization of neural groups [41].

[‡]Of course they can also be connected directly to earlier "components" units: categorization is known to be sensitive to partial matches.

5.2 Learning

We can learn to recognize a specific object (such as a new face) immediately after a brief exposure. The models we described predict that only the last stages need to change their synaptic connections over a fast time scale. Current psychophysical, physiological and fMRI evidence, however, suggests that learning takes place throughout the cortex from V1 to IT and beyond. It is natural to assume that modifications of earlier layers take place over longer times and are thus experience-dependent but less object-specific. A challenge lies in finding a learning scheme that describes how input stimuli drive the development of features at lower levels, while at the same time assuring that features of the same type are pooled over in an appropriate fashion by the pooling units. Hyvärinen and Hoyer [20] have recently presented a learning rule whose aim is to decompose an image into *independent feature subspaces.* The learning rule is similar to the independence maximization rule with a sparsity prior used by Olshausen and Field [32] with the difference that here the independence between the norms of projections on linear subspaces is maximized. With this learning rule, Hyvärinen and Hoyer are able to learn shift- and phase-invariant features similar to complex cells. It remains to be seen whether a hierarchical version of such a scheme to construct an increasingly complex set of features is also feasible. Wallis and Rolls [53] have studied learning at all levels in a model of object recognition (using a variant of Földiák's Trace Rule [11]) capable of recognizing simple configurations of bars, and even faces. Ullman has suggested a computational scheme in which features are learned though the selection of significant components common to different examples of a class of objects [50]. It would be interesting to translate the main aspects of his proposal in a biologically plausible circuitry.

5.3 The time dimension

The models reviewed here do not take into explicit account the fact that the retinal input usually has a time component to it: objects move and the eyes move, too. In addition, any neural circuit will have its own time dynamics. However, most of the models so far are not sufficiently detailed for making any reasonable prediction. Measured neuronal responses are functions of time and even for an image presented in a flash different types of information may be carried over time [33, 42], or in the time structure of the neuronal response. Incorporating the time dimension in neuronal models of recognition is a challenge that is just beginning to be tackled (see M. Giese, ARVO 2000).

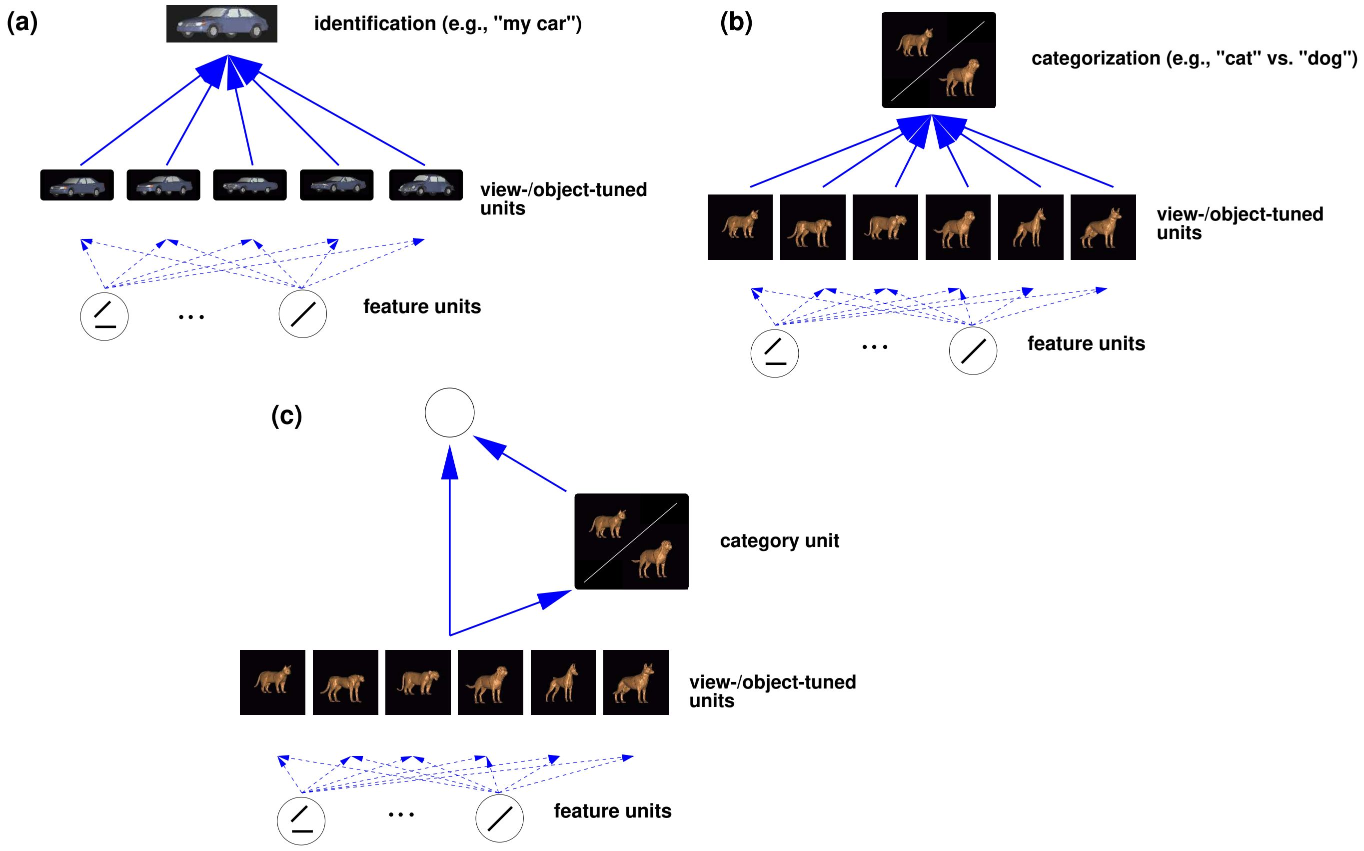

Figure 3: Possible implementations of different recognition tasks in a common computational framework. **(a)** Memorizing an individual object by storing the activation pattern of a population of object-/view-tuned units. **(b)** Learning a categorization task. Note that in principle both identification and categorization could also be learned using the feature unit inputs directly (*e.g.*, the C2 units in HMAX [37]). This is especially important for many categorization tasks which are based on components of the object. Using a more specialized representation, however, such as units tuned to (partial or full views of) the relevant objects, simplifies the learning task (in addition to other computational advantages such as increased robustness to noise [39]). Similarly, combining several of these modules in a hierarchy allows the system to exploit prior knowledge in the learning of new tasks **(c)**, *e.g.*, when using a "cat/dog" categorization unit as input to a "wild cat" unit. In another situation, if the activation of such a "cat/dog" unit was used in a discrimination task along with the activity pattern over the relevant object-tuned cells, discrimination of stimulus pairs straddling the boundary would be expected to be facilitated [38] — the classical "Categorical Perception" effect [17].

5.4 Some key predictions

We label some of the predictions as critical (**) if their falsification will show that the whole class of models described here (and summarized in figure 1) is a "no-go". Experimental evidence against others would falsify specific models (*).

1**) Several "immediate" recognition tasks (identification and categorization) mostly use feed-forward connections during the task itself (possibly not in the learning phase).

2*) Objects of a "nice" class [51] (*e.g.,* objects roughly sharing a similar 3D structure; faces are the best example) are represented in terms of a sparse population code, as activity in a small (hundreds to thousands) set of cells tuned to prototypes of the class. Objects that do not belong to a nice class (*e.g.,* paperclips) may need to be represented for unique identification in terms of a more punctate representation, similar to a look-up table and requiring, in the limit, the activity of just a few "grandmother-like" cells.

3*) Identification and categorization circuits may receive signals from the same or equivalent cells tuned to specific objects or prototypes. Identification of specific objects should be more susceptible to damage (for instance by lesions) than categorization, as identification requires a more specific discrimination (cf. above).

4*) For objects that are members of a nice class, generalization from a single view may be better than for other objects (for non-image plane transformations such as different illuminations or different viewpoints [28]).

Acknowledgments

We are grateful to Francis Crick, Peter Dayan, Shimon Edelman, Christof Koch and Pawan Sinha for useful comments and suggestions.

References

[1] Amit, Y. and Geman, D. (1999). A computational model for visual selection. *Neural Comp.* **11**, 1691–1715.

[2] Anderson, C. and van Essen, D. (1987). Shifter circuits: a computational strategy for dynamic aspects of visual processing. *Proc. Nat. Acad. Sci. USA* **84**, 6297–6301.

[3] Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psych. Rev.* **94**, 115–147.

[4] Bruce, C., Desimone, R., and Gross, C. (1981). Visual properties of neurons in a polysensory area in the superior temporal sulcus of the macaque. *J. Neurophys.* **46**, 369–384.

[5] Bülthoff, H. and Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Nat. Acad. Sci. USA* **89**, 60–64.

[6] Chelazzi, L., Duncan, J., Miller, E., and Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *J. Neurophys.* **80**, 2918–2940.

[7] Dayan, P., Hinton, G., and Neal, R. (1995). The Helmholtz Machine. *Neural Comp.* **7**, 889–904.

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

[9] Edelman, S. (1995). Class similarity and viewpoint invariance in the recognition of 3D objects. *Biol. Cyb.* **72**, 207–220.

[10] Edelman, S. (1999). *Representation and Recognition in Vision*. MIT Press, Cambridge, MA.

[11] Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Comp.* **3**, 194–200.

[12] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cyb.* **36**, 193–202.

[13] Fuster, J. (1990). Inferotemporal units in selective visual attention and short-term memory. *J. Neurophys.* **64**, 681–697.

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

[15] Gochin, P. (1994). Properties of simulated neurons from a model of primate inferior temporal cortex. *Cereb. Cortex* **5**, 532–543.

[16] Haenny, P., Maunsell, J., and Schiller, P. (1988). State dependent activity in monkey visual cortex. II. Retinal and extraretinal factors in V4. *Exp. Brain Res.* **69**, 245–259.

[17] Harnad, S. (1987). *Categorical perception: The groundwork of cognition*. Cambridge University Press, Cambridge, UK.

[18] Hinton, G., Dayan, P., Frey, B., and Neal, R. (1995). The wake-sleep algorithm for unsupervised neural networks. *Science* **268**, 1158–1160.

[19] Hummel, J. and Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psych. Rev.* **99**, 480–517.

[20] Hyvärinen, A. and Hoyer, P. (2000). Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Comp.* **12**, 1705–1720.

[21] Koch, C. and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* **4**, 219–227.

[22] Logothetis, N., Pauls, J., Bülthoff, H., and Poggio, T. (1994). View-dependent object recognition by monkeys. *Curr. Biol.* **4**, 401–414.

[23] Logothetis, N., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* **5**, 552–563.

[24] Logothetis, N. and Sheinberg, D. (1996). Visual object recognition. *Ann. Rev. Neurosci.* **19**, 577–621.

[25] Marr, D. and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond. B Biol. Sci.* **200**, 269–294.

[26] Mel, B. (1997). SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comp.* **9**, 777–804.

[27] Miller, E., Erickson, C., and Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J. Neurosci.* **16**, 5154–5167.

[28] Moses, Y., Ullman, S., and Edelman, S. (1996). Generalization to novel images in upright and inverted faces. *Perception* **25**, 443–462.

[29] Motter, B. (1994). Neural correlates of feature selective memory and pop-out in extrastriate area V4. *J. Neurosci.* **14**, 2190–2199.

[30] Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol. Cyb.* **66**, 241–251.

[31] Olshausen, B., Anderson, C., and van Essen, D. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.* **13**, 4700–4719.

[32] Olshausen, B. and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609.

[33] Perrett, D. and Oram, M. (1993). Neurophysiology of shape processing. *Img. Vis. Comput.* **11**, 317–333.

[34] Poggio, T. and Edelman, S. (1990). A network that learns to recognize 3D objects. *Nature* **343**, 263–266.

[35] Rao, R. and Ballard, D. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Comp.* **9**, 721–763.

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

[37] Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025.

[38] Riesenhuber, M. and Poggio, T. (1999). A note on object class representation and categorical perception. AI Memo 1679, CBCL Paper 183, MIT AI Lab and CBCL, Cambridge, MA.

[39] Riesenhuber, M. and Poggio, T. (2000). The individual is nothing, the class everything: Psychophysics and modeling of recognition in object classes. AI Memo 1682, CBCL Paper 185, MIT AI Lab and CBCL, Cambridge, MA.

[40] Rolls, E., Judge, S., and Sanghera, M. (1977). Activity of neurones in the inferotemporal cortex of the alert monkey. *Brain Res.* **130**, 229–238.

[41] Sporns, O., Tononi, G., and Edelman, G. (1991). Modeling perceptual grouping and figure-ground segregation by means of active reentrant connections. *Proc. Nat. Acad. Sci. USA* **88**, 129–133.

[42] Sugase, Y., Yamane, S., Ueno, S., and Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature* **400**, 869–873.

[43] Tanaka, K. (1996). Inferotemporal cortex and object vision. *Ann. Rev. Neurosci.* **19**, 109–139.

[44] Tarr, M. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonom. Bull. & Rev.* **2**, 55–82.

[45] Tarr, M. (1999). News on views: pandemonium revisited. *Nat. Neurosci.* **2**, 932–935.

[46] Tarr, M. and Bülthoff, H. (1998). Image-based object recognition in man, monkey and machine. *Cognition* **67**, 1–20.

[47] Tarr, M., Williams, P., Hayward, W., and Gauthier, I. (1998). Three-dimensional object recognition is viewpoint-dependent. *Nat. Neurosci.* **1**, 275–277.

[48] Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* **381**, 520–522.

[49] Ullman, S. (1996). *High-level vision: Object recognition and visual cognition*. MIT Press, Cambridge, MA.

[50] Ullman, S. and Sali, E. (2000). Object classification using a fragment-based representation. In *Proceedings of BMCV2000*, Lee, S.-W., Bülthoff, H., and Poggio, T., editors, volume 1811 of *Lecture Notes in Computer Science*, 73–87 (Springer, New York).

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

[52] Vetter, T. and Poggio, T. (1997). Linear object classes and image synthesis from a single example

image. In *Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, volume 19, 733–742.

[53] Wallis, G. and Rolls, E. (1997). A model of invariant object recognition in the visual system. *Prog. Neurobiol.* **51**, 167–194.

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

[55] Young, M. and Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science* **256**, 1327–1331.