Models of object recognition

Maximilian Riesenhuber and Tomaso Poggio

Department of Brain and Cognitive Sciences, McGovern Institute for Brain Research, Center for Biological and Computational Learning and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA

Correspondence should be addressed to T.P. (tp@ai.mit.edu)

Understanding how biological visual systems recognize objects is one of the ultimate goals in computational neuroscience. From the computational viewpoint of learning, different recognition tasks, such as categorization and identification, are similar, representing different trade-offs between specificity and invariance. Thus, the different tasks do not require different classes of models. We briefly review some recent trends in computational vision and then focus on feedforward, view-based models that are supported by psychophysical and physiological data.

Imagine waiting for incoming passengers at the arrival gate at the airport. Your visual system can easily find faces and identify whether one of them is your friend's. As with other tasks that our brain does effortlessly, visual recognition has turned out to be difficult for computers. In its general form, it is a very difficult computational problem, which is likely to be significantly involved in eventually making intelligent machines. Not surprisingly, it is also an open and key problem for neuroscience.

The main computational difficulty is the problem of variability. A vision system needs to generalize across huge variations in the appearance of an object such as a face, due for instance to viewpoint, illumination or occlusions. At the same time, the system needs to maintain specificity. It is important here to note that an object can be recognized at a variety of levels of specificity: a cat can be recognized as "my cat" on the individual level, or more broadly on the categorical level as "cat", "mammal", "animal" and so forth.

Within recognition, we can distinguish two main tasks: identification and categorization. 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 achieve identification relatively easily, as shown by the several companies selling face identification systems, and categorization with much more difficulty. For biological visual systems, however, categorization is suggested to be simpler[1]. In any case, it has been common in the past few years, especially in visual neuropsychology, to assume that different strategies are required for these different recognition tasks[2]. Here we take the computational view[3,4] that identification and categorization, rather than being two distinct tasks, represent two points in a spectrum of generalization levels.

Much of the interesting recent computational work in object recognition considers the recognition problem as a supervised learning problem. We start with a very simplified description (Fig. 1a). The learning module's input is an image, its output is a label, for either the class of the object in the image (is it a cat?) or its individual identity (is it my friend's face?). For simplicity, we describe a learning module as a binary classifier that gives an output of either "yes" or "no." The learning module is trained with a set of examples, which are a set of input–output pairs, that is, images previously labeled. Positive and negative examples are usually needed.

In this setting, the distinction between identification and categorization is mostly semantic. In the case of categorization, the range of possible variations seems larger, because the system must generalize not only across different viewing conditions but also across different exemplars of the class (such as different types of dogs). The difficulty of the task, however, does not depend on identification versus categorization but on parameters such as the size and composition of the training set and how much of the variability required for generalization is covered by the training examples. For instance, the simple system described earlier could not identify an individual face from any viewpoint if trained with only a single view of that face. Conversely, the same system may easily categorize, for instance, dog images versus cat images if trained with a large set of examples covering the relevant variability.

Within this learning framework, the key issue is the type of invariance and whether it can in principle be obtained with just one example view. Clearly, the effects of two-dimensional (2D) affine transformations, which consist of combinations of scaling, translation, shearing and rotation in the image plane, can be estimated exactly from just one object view. Generic mechanisms, independent of specific objects and object classes, can be added to the learning system to provide invariance to these transformations. There is no need then to collect examples of one object or object class at all positions in the image to be able to generalize across positions from a single view. To determine the behavior of a specific object under transformations that depend on its 3D shape, such as illumination changes or rotation in depth, however, one view is generally not sufficient. In categorization, invariance also occurs 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, usually require multiple example views during learning. We believe that this distinction along types of invariance is more fundamental that the distinction between categorization and recognition, providing motivation for experiments to dissect the neural mechanisms of object recognition.

In computer vision applications, the learning module of Fig. 1a has been implemented in various ways. In one simple approach (Fig. 1b), each unit stores one of the example views and measures the similarity of the input image with the stored

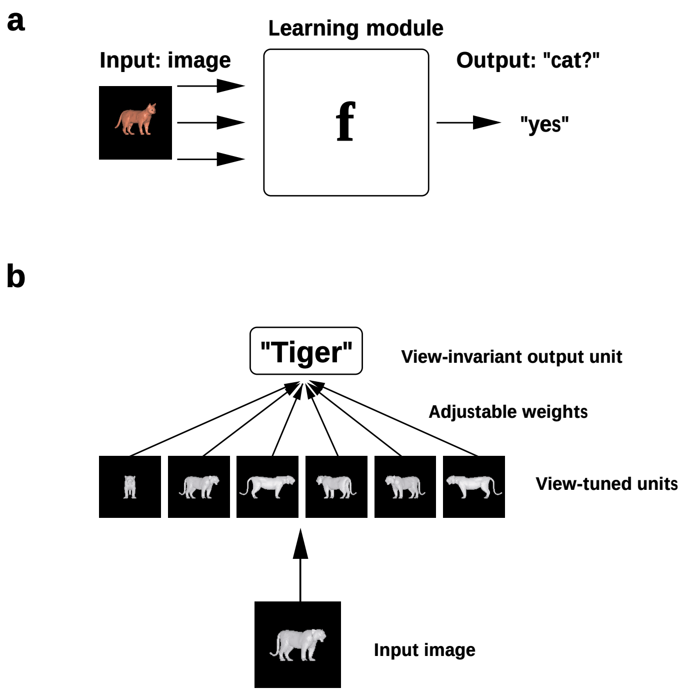

Fig. 1. Learning module schematics. (**a**) The general learning module. (**b**) A specific learning module: a classifier, trained to respond in a view-invariant manner to a certain object.

example. The weighted outputs of all units are then added. If the sum is above a threshold, then the system's output is 1; otherwise it is 0. During learning, weights and threshold are adjusted to optimize correct classification of examples. One of the earliest versions of this scheme was a model[5] for identification of an individual object irrespective of viewpoint. Note that this approach is feedforward and view-based in the sense that there is no 3D model of the object that is mentally rotated for recognition, but rather novel views are recognized by interpolation between (a small number of) stored views (Fig 1b).

More elaborate schemes have been developed, especially for the problem of object categorization in complex real-world scenes. They focus on classifying an image region for a particular viewpoint and then combine classifiers trained on different viewpoints. The main difference between the approaches lies in the features with which the examples are represented. Typically, a set of *n* measurements or filters are applied to the image, resulting in an *n*-dimensional feature vector. Various measures have been proposed as features, from the raw pixel values themselves[6–8] to overcomplete measurements (see below), such as the ones obtained through a set of overcomplete wavelet filters[9]. Wavelets can be regarded as localized Fourier filters, with the shape of the simplest two-dimensional wavelets being suggestive of receptive fields in primary visual cortex.

The use of overcomplete dictionaries of features is an interesting new trend in signal processing[10]. Instead of representing a signal in terms of a traditional complete representation, such as Fourier components, one uses a redundant basis, such as the combination of several complete bases. It is then possible to find sparse representations of a given signal in this large dictionary, that is, representations that are very compact because any given signal can be represented as the combination of a small number of features. Mallat makes the point by analogy: a complete representation is like a small English dictionary of just a few thousand words. Any concept can be described using the vocabulary but at the expense of long sentences. With a very large dictionary—say 100,000 words—concepts can be described with much shorter sentences, sometimes with a single word. In a similar way, overcomplete dictionaries of visual features allow for compact representations. Single neurons in the macaque posterior inferotemporal cortex may be tuned to such a dictionary of thousands of complex shapes[11].

Newer algorithms add a hierarchical approach in which non-overlapping[12,13] or overlapping components[8,14] are first detected and then combined to represent a full view. As we discuss below, in models for object recognition in the brain, hierarchies arise naturally because of the need to obtain both specificity and invariance of position and scale in a biologically plausible way.

The performance of these computer vision algorithms is now very impressive in tasks such as detecting faces, people and cars in real-world images[8,13]. In addition, there is convincing evidence from computer vision[15] that faces—and other objects—can be reliably detected in a view-invariant fashion over 180° of rotation in depth by combining just three detectors (Fig. 1b), one trained with and tuned to frontal faces, one to left profiles, and one to right profiles.

All these computer vision schemes lack a natural implementation in terms of plausible neural mechanisms. Some of the basic ideas, however, are relevant for biological models.

Object recognition in cortex

View-based models have also been proposed to explain object recognition in cortex. As described above, in this class of models, objects are represented as collections of view-specific features, leading to recognition performance that is a function of previously seen object views, in contrast to so-called 'object-centered' or 'structural description' models, which propose that objects are represented as descriptions of spatial arrangements among parts in a three-dimensional coordinate system that is centered on the object itself[16]. One of the most prominent models of this type is the 'recognition by components' (RBC) theory[17,18], in which the recognition process consists of extracting a view-invariant structural description of the object in terms of spatial relationships among volumetric primitives, 'geons', that is then matched to stored object descriptions. 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.

The question of whether the visual system uses a view-based or an object-centered representation has been the subject of much controversy[19,20] (for reviews, see refs. 2, 21). Psychophysical[22,23] and physiological data[24,25] support a view-based approach, and we will not discuss these data further here. In this paper, we focus on view-based models of object recognition and show how they provide a common framework for identification and categorization.

Based on physiological experiments in monkeys[2,11], object recognition in cortex is thought to be mediated by the ventral visual pathway[26] from primary visual cortex, V1, through extrastriate visual areas V2 and V4 to inferotemporal cortex, IT (Fig. 2). Neuropsychological and fMRI studies point to a crucial role of inferotemporal cortex for object recognition also in human vision[2,26]. As one proceeds along the ventral stream, neurons seem to show increasing receptive field sizes, along with a preference for increasingly complex stimuli[27]. Whereas neurons in V1 have small receptive fields and respond to simple bar-like stimuli, cells in IT show large receptive fields and prefer complex stimuli such as faces[2,11,25]. Tuning properties of IT cells seem to be shaped by task learning[24,28,29]. For instance, after monkeys were trained to discriminate between individual and highly similar 'paperclip' stimuli composed of bar segments[24], neurons in IT were found that were tightly shape-tuned to the training objects. The great majority of these neurons responded only to a single view of the object, with a much smaller number responding to a single object-invariant viewpoint. In addition to this punctate representation, more distributed representations are likely also used in IT. Studies of 'face cells' (that is, neurons responding preferentially to faces) in IT argue for a distributed representation of this object class with the identity of a face being jointly encoded by the activation pattern over a group of face neurons[30,31]. Interestingly, view-tuned face cells are much more prevalent than view-invariant face cells[25]. In either case, the activation of neurons in IT can then serve as input to higher cortical areas such as prefrontal cortex, a brain region central for the control of complex behavior[32].

A zoo of view-based models of object recognition in cortex exists in the literature. Two major groups of models can be discerned based on whether they use a purely feedforward model of processing or use feedback connections. A first question to be asked of all models is how they deal with the 2D affine transformations described earlier. Although scale and position invariance can be achieved very easily in computer vision systems by serial scanning approaches, in which the whole image is searched for the object of interest sequentially at different positions and scales, such a strategy seems unlikely to be realized in neural hardware.

Feedback models include architectures that perform recognition by an analysis-by-synthesis approach: the system makes a guess about what object may be in the image and its position and scale, 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[3,33,34]. Other models use top-down control to 'renormalize' the input image in position and scale before attempting to match it to a database of stored objects[35,36], or conversely to tune the recognition system depending on the object's transformed state, for instance by matching filter size to object size[37].

Interestingly, EEG studies[38] show that the human visual system can solve an object detection task within 150 ms, which is on the order of the latency of view- and object-tuned cells in inferotemporal cortex[25]. This does not rule out the use of feedback processing but strongly constrains its role in perceptual recognition based on similarity of visual appearance.

Indeed, this speed of processing is compatible with a class of view-based models that rely only on feedforward processing, similar to the computer vision algorithms described above. However, in these models, image-based invariances are not achieved by an unbiological scanning operation but rather are gradually built up in a hierarchy of units of increasing receptive field size and feature complexity, as found in the ventral visual stream. One of the earliest representatives of this class of models is the 'Neocognitron'[39], a hierarchical network in which feature complexity and translation invariance are alternatingly increased in different layers of a processing hierarchy. Feature complexity is increased by a 'template match' operation in which higher-level neurons only fire if their afferents show a particular activation pattern; invariance is increased by pooling over units tuned to the same feature but at different positions. The concept of pooling units tuned to transformed versions of the same feature or object was subsequently proposed[40] to explain invariance also to non-affine transformations, such as to rotation in depth or illumination changes, in agreement with the shorter latency of view-tuned cells relative to view-invariant cells observed experimentally[25]. Indeed, a biologically plausible model[5] had motivated physiological experiments[24], showing 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 has been used successfully in later models, among them the SEEMORE system[41] and the HMAX model[42]. The latter, using a new pooling operation, demonstrated how scale and translation invariance could be achieved in view-tuned cells by a hierarchical model, in quantitative agreement with experimental IT neuron data[24].

Two features are key to the success of hierarchical models[39,42,43]. First, the gradual and parallel increase of feature complexity and receptive field size, as found in the visual system, is crucial in avoiding a combinatorial explosion of the number of units in the system on one hand or insufficient discriminatory ability on the other hand. Although the invariance range is low at lower levels, thus requiring many cells to cover the required range of scales and positions, only a small set of simple features must be represented. Conversely, in higher layers, where neurons are tuned to a greater number of more complex features, neurons show a greater degree of invariance, thus requiring fewer cells tuned to the same feature at different positions and scales[44]. Second, in hierarchical models, a redundant set of more complex features in higher levels of the system is built from simpler features. These complex features are tolerant to local deformations as a result of the invariance properties of their afferents[39,42,43]. In this respect, they are related to (so far non-biological) recognition architectures based on feature trees that emphasize compositionality[45]. The end result is an overcomplete dictionary similar to the computer vision approaches reviewed earlier.

Models of object recognition

We propose a model (Fig. 3) that extends several existing models[5,39,40,42,43]. A view-based module, whose final stage consists of units tuned to specific views of specific objects, takes care

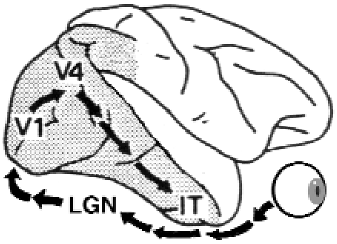

Fig. 2. The ventral visual stream of the macaque (modified from ref. 26).

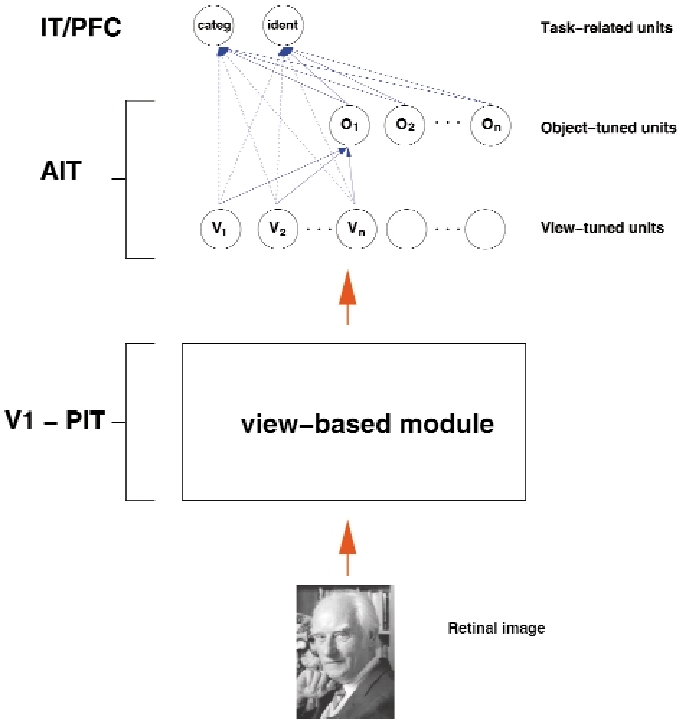

Fig. 3. A class of models of object recognition. This sketch combines and extends several recent models[5,39,40,42,43]. On top of a view-based module, view-tuned model units (V_n) show tight tuning to rotation in depth (and illumination, and other object-dependent transformations such as facial expression and so forth) but are tolerant to scaling and translation of their preferred object view. Note that the cells labeled here as view-tuned units may be tuned to full or partial views, that is, connected to only a few of the feature units activated by the object view[44]. All the units in the model represent single cells modeled as simplified neurons with modifiable synapses. Invariance, for instance to rotation in depth, can then be increased by combining in a learning module several view-tuned units tuned to different views of the same object[5], creating view-invariant 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. They can be the same generic learning modules (**Fig. 1**) but trained to perform different 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 located in the prefrontal cortex (PFC).

of the invariance to image-based transformations. This module, of which HMAX[42] is a specific example, probably comprises neurons from primary visual cortex (V1) up to, for instance, posterior IT (PIT). At higher stages such as in anterior IT, invariance to object-based transformations, such as rotation in depth, illumination and so forth, is achieved by pooling together the appropriate view-tuned cells for each object. Note that view-tuned models[5] predict the existence of view-tuned as well as view-invariant units, whereas structural description models strictly predict only the latter. Finally, categorization and identification tasks, up to the motor response, are performed by circuits, possibly in prefrontal cortex (D.J. Freedman *et al.*, *Soc. Neurosci. Abstr*., 25, 355.8, 1999), receiving inputs from object-specific and view-invariant cells. Without relevant view-invariant units, such as when the subject has only experienced an object from a certain viewpoint, as in the experiments on paperclip recognition[22,24,46], task units could receive direct input from the view-tuned units (Fig. 3, dashed lines).

In general, a particular object, say a specific face, will elicit different activity in the object-specific O_n cells of Fig. 3 tuned to a small number of 'prototypical' faces, as observed experimentally[31]. Thus, the memory of the particular face is represented in the identification circuit implicitly by a population code through the activation pattern over the coarsely tuned O_n cells, without dedicated 'grandmother' cells. Discrimination, or memorization of specific objects, can then proceed by comparing activation patterns over the strongly activated object- or view-tuned units. For a certain level of specificity, only the activations of a small number of units have to be stored, forming a sparse code—in contrast to activation patterns on lower levels, where units are less specific and hence activation patterns tend to involve more neurons. Computational studies in our laboratory[47] provide evidence for the feasibility of such a representation. An interesting and non-trivial conjecture (supported by several experiments[47–49]) of this population-based representation is that it should be able to generalize from a single view of a new object belonging to a class of objects sharing a common 3D structure—such as a specific face—to other views. Generalization is expected to be better than for other object classes in which members of the same class can have very different 3D structure, such as the 'paperclip' objects[46]. Similarly to identification, a categorization module (say, for dogs versus cats) uses as inputs the activities of a number of cells tuned to various animals, with weights set during learning so that the unit responds differently to animals from different classes[50].

This simple framework illustrates how the same learning algorithm and architecture can support a variety of object recognition tasks such as categorization and identification (for a related proposal, see ref. 4). It can easily be extended to include inputs to task units from lower-level feature units and even other task units, with interesting implications for the learning of object class hierarchies or phenomena such as categorical perception[50]. The model can successfully perform recognition tasks such as identification[47] and categorization[50], with performance similar to human psychophysics[47] and in qualitative agreement with monkey physiology (D.J. Freedman *et al.*, *Soc. Neurosci. Abstr*., 25, 355.8, 1999).

Several predictions for physiology follow from this proposed architecture (Fig. 3). For instance, objects sharing a similar 3D structure, such as faces, would be expected to be represented in terms of a sparse population code, as activity in a small group of cells tuned to prototypes of the class. Objects that do not belong to such a class (paperclips) should need to be represented for unique identification in terms of a more punctate representation, similar to a look-up table and requiring, in the extreme limit, the activity of a single 'grandmother' cell. Further, identification and categorization circuits should receive signals from the same or equivalent cells tuned to specific objects or prototypes.

Challenges ahead

Here we have taken the view that basic recognition processes occur in a bottom-up way; it is, however, very likely that top-down signals are essential in controlling the learning phase of recognition[51] and in some attentional effects, for instance in detection tasks, to bias recognition toward features of interest, as suggested by physiological studies[52–55]. The massive descending projections in the visual cortex are an obvious candidate for an anatomical substrate for top-down processing. One of the main challenges for future models is to integrate such top-down influences with bottom-up processing.

We can learn to recognize a specific object (such as a new face) immediately after a brief exposure. In the model we described in Fig. 3, 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. A challenge lies in finding a learning scheme that describes how visual experience drives the development of features at lower levels, while assuring that features of the same type are pooled over in an appropriate fashion by the pooling units. Schemes for learning overcomplete representations have been proposed[56], with extensions to the learning of invariances[57]. It remains to be seen whether a hierarchical version of such a scheme to construct an increasingly complex set of features is also feasible. Learning at all levels has been studied in a model of object recognition capable of recognizing simple configurations of bars, and even faces independent of position[43], by exploiting temporal associations during the learning period[58]. In one proposal[14] (see also refs. 13, 59), features are learned through the selection of significant components common to different examples of a class of objects. It would be interesting to translate the main aspects of this approach into a biologically plausible circuit.

Acknowledgements

Supported by grants from ONR, DARPA, NSF, ATR, Honda, a Merck/MIT Fellowship in Bioinformatics, and a McDonnell Pcw award (M.R.). T.P. is supported by the Uncas and Helen Whitaker Chair at the Whitaker College, MIT. For comments and suggestions, we are grateful to Heinrich Bülthoff, Peter Dayan, Shimon Edelman, David Freedman, Christof Koch, Earl Miller, David Perrett, Pawan Sinha and Francis Crick (also for the picture in Fig. 3).

Received 2 June; accepted 6 october 2000

1. Rosch, E., Mervis, C., Gray, W., Johnson, D. & Boyes-Braem, P. Basic objects in natural categories. *Cogn. Psychol.* 8, 382–439 (1976).
2. Logothetis, N. & Sheinberg, D. Visual object recognition. *Annu. Rev. Neurosci.* 19, 577–621 (1996).
3. Ullman, S. *High-Level Vision: Object Recognition and Visual Cognition* (MIT Press, Cambridge, Massachusetts, 1996).
4. Edelman, S. *Representation and Recognition in Vision* (MIT Press, Cambridge, Massachusetts, 1999).
5. Poggio, T. & Edelman, S. A network that learns to recognize 3D objects. *Nature* 343, 263–266 (1990).
6. Brunelli, R. & Poggio, T. Face recognition: Features versus templates. *IEEE PAMI* 15, 1042–1052 (1993).
7. Yang, M.-H., Roth, D. & Ahuja, N. A. in *Advances in Neural Information Processing Systems* Vol. 12 (eds. Solla, S.A., Leen, T.K. & Müller, K.-K.) 855–861 (MIT Press, Cambridge, Massachusetts, 1999).
8. Schneiderman, H. & Kanade, T. in *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 45–51 (IEEE, Los Alamitos, California, 1998).
9. Oren, M. Papageorgiou, C., Sinha, P., Osuna, E. & Poggio, T. in *IEEE Conference on Computer Vision and Pattern Recognition* 193–199 (IEEE, Los Alamitos, CA, 1997).
10. Chen, S., Donoho, D. & Saunders, M. Atomic decomposition by basis pursuit. Technical Report 479 (Dept. of Statistics, Stanford University, 1995).
11. Tanaka, K. Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19, 109–139 (1996).
12. Mohan, A. Object detection in images by components. AI Memo 1664 (CBCL and AI Lab, MIT, Cambridge, Massachusetts, 1999).
13. Heisele, B., Poggio, T. & Pontil, M. Face detection in still gray images. AI Memo 1687 (CBCL and AI Lab, MIT, Cambridge, Massachusetts, 2000).
14. Ullman, S. & Sali, E. in *Proceedings of BMCV2000*, Vol. 1811 of *Lecture Notes in Computer Science* (eds. Lee, S.-W., Bülthoff, H. & Poggio, T.) 73–87 (Springer, New York, 2000).
15. Schneiderman, H. & Kanade, T. A statistical method for 3D object detection applied to faces and cars. in *IEEE Conference on Computer Vision and Pattern Recognition* (in press).
16. Marr, D. & Nishihara, H. K. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond. B Biol. Sci.* 200, 269–294 (1978).
17. Biederman, I. Recognition-by-components: A theory of human image understanding. *Psychol. Rev*. 94, 115–147 (1987).
18. Hummel, J. & Biederman, I. Dynamic binding in a neural network for shape recognition. *Psychol. Rev*. 99, 480–517 (1992).
19. Biederman, I. & Gerhardstein, P. Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *J. Exp. Psychol. Hum. Percept. Perform*. 19, 1162–1182 (1993).
20. Tarr, M. & Bülthoff, H. Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993). *J. Exp. Psychol. Hum. Percept. Perform*. 21, 1494–1505 (1995).
21. Tarr, M. & Bülthoff, H. Image-based object recognition in man, monkey and machine. *Cognition* 67, 1–20 (1998).
22. Logothetis, N., Pauls, J., Bülthoff, H. & Poggio, T. View-dependent object recognition by monkeys. *Curr. Biol*. 4, 401–414 (1994).
23. Tarr, M., Williams, P., Hayward, W. & Gauthier, I. Three-dimensional object recognition is viewpoint-dependent. *Nat. Neurosci*. 1, 275–277 (1998).
24. Logothetis, N., Pauls, J. & Poggio, T. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol*. 5, 552–563 (1995).
25. Perrett, D., Hietanen, J., Oram, M. & Benson, P. Organization and functions of cells responsive to faces in the temporal cortex. *Phil. Trans. R. Soc. Lond. B Biol. Sci.* 335, 23–30 (1992).
26. Ungerleider, L. & Haxby, J. 'What' and 'where' in the human brain. *Curr. Opin. Neurobiol*. 4, 157–165 (1994).
27. Kobatake, E. & Tanaka, K. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol*. 71, 856–867 (1994).
28. Booth, M. & Rolls, E. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb. Cortex* 8, 510–523 (1998).
29. Kobatake, E., Wang, G. & Tanaka, K. Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *J. Neurophysiol*. 80, 324–330 (1998).
30. Wang, G., Tanaka, K. & Tanifuji, M. Optical imaging of functional organization in the monkey inferotemporal cortex. *Science* 272, 1665–1668 (1996).
31. Young, M. & Yamane, S. Sparse population coding of faces in the inferotemporal cortex. *Science* 256, 1327–1331 (1992).
32. Miller, E. The prefrontal cortex and cognitive control. *Nat. Rev. Neurosci.* 1, 59–65 (2000).
33. Mumford, D. On the computational architecture of the neocortex. II. The role of corticocortical loops. *Biol. Cybern*. 66, 241–251 (1992).
34. Rao, R. & Ballard, D. Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Comput*. 9, 721–763 (1997).
35. Anderson, C. & van Essen, D. Shifter circuits: a computational strategy for dynamic aspects of visual processing. *Proc. Natl. Acad. Sci. USA* 84, 6297–6301 (1987).
36. Olshausen, B., Anderson, C. & van Essen, D. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci*. 13, 4700–4719 (1993).
37. Gochin, P. Properties of simulated neurons from a model of primate inferior temporal cortex. *Cereb. Cortex* 5, 532–543 (1994).
38. Thorpe, S., Fize, D. & Marlot, C. Speed of processing in the human visual system. *Nature* 381, 520–522 (1996).
39. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern*. 36, 193–202 (1980).
40. Perrett, D. & Oram, M. Neurophysiology of shape processing. *Image Vis. Comput*. 11, 317–333 (1993).
41. Mel, B. SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comput*. 9, 777–804 (1997).
42. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci*. 2, 1019–1025 (1999).

43. Wallis, G. & Rolls, E. A model of invariant object recognition in the visual system. *Prog. Neurobiol.* 51, 167–194 (1997).
44. Riesenhuber, M. & Poggio, T. Are cortical models really bound by the "binding problem"? *Neuron* 24, 87–93 (1999).
45. Amit, Y. & Geman, D. A computational model for visual selection. *Neural Comput*. 11, 1691–1715 (1999).
46. Bülthoff, H. & Edelman, S. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Natl. Acad. Sci. USA* 89, 60–64 (1992).
47. Riesenhuber, M. & Poggio, T. 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, Massachusetts, 2000).
48. Edelman, S. Class similarity and viewpoint invariance in the recognition of 3D objects. *Biol. Cybern*. 72, 207–220 (1995).
49. Moses, Y., Ullman, S. & Edelman, S. Generalization to novel images in upright and inverted faces. *Perception* 25, 443–462 (1996).
50. Riesenhuber, M. & Poggio, T. A note on object class representation and categorical perception. AI Memo 1679, CBCL Paper 183 (MIT AI Lab and CBCL, Cambridge, Massachusetts, 1999).
51. Hinton, G., Dayan, P., Frey, B. & Neal, R. The wake-sleep algorithm for unsupervised neural networks. *Science* 268, 1158–1160 (1995).
52. Chelazzi, L., Duncan, J., Miller, E. & Desimone, R. Responses of neurons in inferior temporal cortex during memory-guided visual search. *J. Neurophysiol*. 80, 2918–2940 (1998).
53. Haenny, P., Maunsell, J. & Schiller, P. State dependent activity in monkey visual cortex. II. Retinal and extraretinal factors in V4. *Exp. Brain Res*. 69, 245–259 (1988).
54. Miller, E., Erickson, C. & Desimone, R. Neural mechanism of visual working memory in prefrontal cortex of the macaque. *J. Neurosci*. 16, 5154–5167 (1996).
55. Motter, B. Neural correlates of feature selective memory and pop-out in extrastriate area V4. *J. Neurosci*. 14, 2190–2199 (1994).
56. Olshausen, B. & Field, D. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609 (1996).
57. Hyvärinen, A. & Hoyer, P. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Comput*. 12, 1705–1720 (2000).
58. Földiák, P. Learning invariance from transformation sequences. *Neural Comput*. 3, 194–200 (1991).
59. Weber, M., Welling, W. & Perona, P. Towards automatic discovery of object categories. in *IEEE Conference on Computer Vision and Pattern Recognition* (in press).

Viewpoint • On theorists and data in computational neuroscience

A diversity of activities in neuroscience are labeled 'theory'. Developing Bayesian spike sorting algorithms, making a theory of consciousness, attractor neural network dynamics, constructing multi-compartment simulations of neurons, these and many other activities have a theoretical component. So of course there is a role for theory in neuroscience.

A question about the future of computational neuroscience can be bluntly put. Is understanding how the brain works going to be an enterprise in which pure theorists, scientists without experimental laboratories and not mere subsidiary parts of an experimentalist's laboratory, make essential contributions? Are independent theorists important to neuroscience? Important enough, say, to merit independent faculty positions in universities? Or will researchers doing experiments (or at least controlling experimental laboratories) make all the significant contributions, and be the only appropriate occupants of professorial positions in neuroscience?

The history of chemistry is the closest parallel. It is a subject in which both qualitative theory (the periodic table, the chemical bond) and quantitative theory (statistical mechanics, quantum mechanics) have been important. Modern quantitative theory and its impact on chemistry was brought forward by people who did not themselves do experiments, such as chemistry Nobelists Onsager and Kohn, whose ability in mathematics was key to understanding how to make new predictions and how to ground in understanding concepts that came qualitatively from experiments (in the areas of chemical bonding and irreversible thermodynamics).

Physics, geology, chemistry and astronomy have developed independent theorists when the breadth of these subjects exceeded the span of talents of a single individual. Within neuroscience I know no one who is both outstandingly able to perform inventive rat brain surgery and able to cogently describe modern artificial intelligence theories of learning and learnability. These are such different dimensions of expertise! Having both the talent and the time to span such a range is now impossible. Computational neuroscience is therefore in the process of bifurcating into theorists and experimentalists.

Sensible theory in science is rooted in facts, be they general or specific, so theory and experiment must interact. In physical science the development of a theoretical branch was at the time made easier because the relatively small number of essential experimental facts were all available in scientific journals. Now, in the more complex parts of these subjects, large data sets are only summarized in publications, and sharing of the extensive data sets themselves has become commonplace. Two forces have pushed this accessibility. One is the genuine wish to advance science rapidly. The other is pragmatic: doing experimental science is expensive. Science is chiefly paid for from the public purse, either directly by government or indirectly by the tax-free subsidization of charitable foundations. In appealing for publicly based support for a science, it is important that resources are seen to be used effectively.

Good experimentalists excel in the art of knowing which parts of their own unpublished data should be ignored, so not *all* data ought be shared. But certain sharing should become common practice. For example, neuroscientists understand that the (partial) publication of data only through summaries such as post-stimulus time histograms can conceal what is actually happening. In these days of web sites, it would be trivial to make available all spike rasters from which summaries are published.

Some of my friends lament "we will fail to get credit for our work." But most scientists know that it was the careful measurements of Tycho Brahe that led Kepler to his three laws of planetary motion. Reputations of experimentalists are only enhanced by having their data cited as significant by others in the motivation or testing of ideas.

J. J. Hopfield
Princeton University, Princeton, New Jersey 08544, USA
e-mail: jhopfield@watson.princeton.edu