Vision: are models of object recognition catching up with the brain?

Tomaso Poggio¹ and Shimon Ullman²

¹Department of Brain and Cognitive Sciences, McGovern Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts; ²Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel

Object recognition has been a central yet elusive goal of computational vision. For many years, computer performance seemed highly deficient and unable to emulate the basic capabilities of the human recognition system. Over the past decade or so, computer scientists and neuroscientists have developed algorithms and systems—and models of visual cortex—that have come much closer to human performance in visual identification and categorization. In this personal perspective, we discuss the ongoing struggle of visual models to catch up with the visual cortex, identify key reasons for the relatively rapid improvement of artificial systems and models, and identify open problems for computational vision in this domain.

Keywords: object recognition; visual models; supervised learning; visual cortex; feedforward; backprojection

Computer models are catching up with the brain

Object recognition is difficult

When you are watching a movie, you instantly recognize the scene and the objects in it, such as people, buildings, and cars. You may identify the location, a specific actor, and the brand of the car she is driving, her clothing, eyeglasses, wristwatch, and the like. Like other natural tasks that our brains perform effortlessly, visual recognition has turned out to be difficult to reproduce in artificial systems. In its general form, it is a highly challenging computational problem that is likely to play a significant role in eventually making intelligent machines. Not surprisingly, it is also an open and key problem for neuroscience.

Within object recognition, it is common to distinguish two main tasks: identification—for instance, recognizing a specific face among other faces—and categorization, for example, recognizing a car among other object classes. We will discuss both of these tasks below, use recognition to include both, and discuss later challenges beyond recognition.

Recognition in computer vision

Early computer vision recognition schemes focused primarily on the recognition of rigid three-dimensional (3-D) objects, such as machine parts, tools, and cars. This is a challenging problem because the same object can have markedly different appearances when viewed from different directions. It proved possible to deal successfully with this difficulty by using detailed 3-D models of the viewed objects, which are compared with the projected 2-D image.¹–³ These methods did not extend, however, to the recognition of more complex objects and object classes, since the variety of images of an object class such as a “dog” cannot be adequately described by the projections of a typical fixed 3-D shape. The advantage of identification over categorization in these recognition models stood in marked contrast to human vision, where general categorization is typically faster and easier than precise identification.⁴

Over the past decade or so, computational models have made significant progress in the task of recognizing natural object categories under realistic, relatively unconstrained viewing conditions. A number...
of early schemes, mainly focusing on the class of human faces, obtained significant improvement over previous methods. The techniques have evolved to reach practical applications, as evidenced by their use in current digital cameras.

The more recent versions of these computational schemes have started to deal successfully with an expanding range of complex object categories such as pedestrians, cars, motorcycles, airplanes, horses, and the like, in unconstrained natural scenes, to deal with a broad range of objects within each class. The algorithms that were refined over the past few years can deal successfully with a large number of different object classes, in complex and highly cluttered scenes. They are being applied to databases of hundreds and even thousands of object classes. Yearly competitions in computer-based recognition, such as the Pascal challenge, witness continuous improvement in the range of classes and scene complexity handled successfully by automatic object categorization algorithms. An example of results is shown in Figure 1 for the class “airplanes,” illustrating the variability of images used in the test set. Comparison with human performance reveals that even the best computational schemes still fall short in most cases. In some specific tasks, however, computers have been trained to achieve high performance, including face identification, face detection, and car and pedestrian detection. Moreover, recent results on the very large ImageNet dataset with thousands of object classes suggest that the performance of convolutional networks, when trained with very large sets of labeled examples, may begin to approach human performance, when humans are limited to short image exposure. No detailed comparisons have been made, but humans are probably still superior in many recognition tasks, especially when dealing with flexible and highly variable objects and when background clutter is large. Reasons for this performance gap and possible directions for bridging the gap are discussed in later sections.

Models of the visual cortex
Over the same time period, some of the best performing recognition systems have come from research at the intersection of computational neuroscience and computer vision. Recent models of visual cortex based directly on known functional anatomy (Fig. 2), building on a range of earlier attempts, were able to account for and predict a number of physiological data from areas of the ventral stream from V1 and V2 to V4 and IT. This family of models was able to mimic human performance in rapid categorization tasks and some of these models of visual cortex were among the best computer vision systems at the time.

What are the roots of computer models’ recent progress?
In our view, the qualitative improvement in the performance of recognition models can be attributed to three main components. The first is the use of extensive learning in constructing recognition models. In this framework, rather than specifying a particular model, the scheme starts with a large family of possible models and uses observed examples to guide the construction of a specific model that is best suited to the observed data. The second component was the development of new forms of object representation for the purpose of categorization, based on both computational considerations and guidelines from...
known properties of the visual cortex. These two components, representation and learning, are interrelated: initially, the class representation provides a family of plausible models, and effective learning methods are then used to construct a particular model for a novel class such as “dog” or “airplane” based on observed examples. The third component was the use of new statistical learning techniques, such as regularization classifiers (SVM and others) and Bayesian inference (such as graphical models). We next discuss each of these advances in more detail.

**Learning instead of design**

A conceptual advance that facilitated recent progress in object recognition was the idea of learning the solution to a specific classification problem from examples, rather than focusing on the classifier design. This was a marked departure from the dominant practices at the time: instead of an expert program with a predetermined set of logical rules, the appropriate model was learned and selected from a possibly infinite set of models, based on a set of examples. The techniques used in the 1990s originated in the area of supervised learning, where image examples are provided together with the appropriate class labels (e.g., “face” or “nonface”). A comprehensive theory of the foundations of supervised learning has been developed, with roots in functional analysis and probability theory. The formal analysis of learning continues to evolve and to contribute to our understanding of the role of learning in visual recognition.

**New image representations**

During learning, a recognition scheme typically extracts a set of measurements, or features, and uses them to construct new object representations. Objects are then classified and recognized based on their feature representation. Feature selection and object representation are crucial, because they facilitate the identification of elements that are shared by objects in the same class and support discrimination between similar objects and categories. Different types of visual features have been used in computational models in the past, ranging from simple local-image patterns such as wavelets, edges, blobs, or local-edge combinations to abstract 3-D shape primitives, such as cylinders, spheres, cubes, and the like.

A common aspect of most earlier recognition schemes is that they used a fixed small generic set of feature types to represent all objects and classes. In contrast, recent recognition schemes use pictorial features extracted from examples, such as object fragments or patches, together with their spatial arrangement. Unlike generic parts, these schemes use a large set of features, extracted from different classes of objects. The use of large feature sets is also connected to an interesting new trend in signal processing, related to overcomplete representations. 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). This type of representation may be mirrored by the several thousands of complex shapes.
to which different single neurons in the posterior inferotemporal cortex of the macaque were found to be tuned.\textsuperscript{53} Representations using such features have been used successfully in recent computer vision recognition systems for two reasons. First, these representations can be learned and used efficiently; second, they proved to effectively capture the broad range of variability in appearance within a visual class.

Two additional comments are appropriate. First, the representations described above are view-based, as opposed to object-centered models, but this should not be confused with the use of 3-D information.\textsuperscript{54,55} A representation based on image appearance can include not only 2-D image properties, but also 3-D aspects such as local depth variations or 3-D curvature, as supported by some physiological evidence.\textsuperscript{56–58} Second, we think that multiple object representations are likely to be used by the brain, probably for different tasks. In particular, an object-based representation using invariant 3-D properties is likely to be useful for tasks involving planning as well as perceiving object manipulation, or for slower recognition processes (e.g., those involved in mental rotation experiments, see Shepard and Metzler\textsuperscript{59}).

**New statistical learning methods**

Over the past few years, the mathematics of learning has become the lingua franca of large areas of computer science and, in particular, of computer vision. As we discussed, the use of a learning framework enabled a qualitative jump in object recognition. Whereas the initial techniques used to construct useful classification models from data were quite simple, there are now more efficient algorithms originally introduced in the area of learning in the 1990s such as regularization algorithms (also called kernel machines), which include SVM\textsuperscript{44,60,61} and boosting.\textsuperscript{62} In addition to discriminative algorithms, the area of learning has grown to include probabilistic approaches with the goal of providing full probability distributions as solutions to object recognition tasks. These techniques are mostly Bayesian and range from graphical models\textsuperscript{63,64} to hierarchical Bayesian models.\textsuperscript{65–68} At the same time, the focus of research is shifting from supervised to unsupervised and semisupervised learning problems, using techniques such as manifold learning.\textsuperscript{68} The inclusion of semisupervised problems, in which the training set consists of a large number of unlabeled examples and a small number of labeled ones, is beginning to formulate the learning problem in a more biologically plausible way, as biological organisms seem to be able to learn from experience with a surprisingly small amount of supervision. We will discuss such an approach in the context of learning invariant representations.

An interesting research direction is the formulation of learning that mirrors the combination of existing genetic information and learning from experience that is likely used by biological organisms. In this context, the range of neurobiologically plausible strategies for the computation of probabilistic graphical models capable of taking into account specific priors is quite open. As an example, it has been suggested that sampling techniques, such as MCMC, could be implemented by circuits of noisy neurons. Probabilistic models of biological vision are likely to grow in importance, at least as phenomenological and convenient descriptions of biological information processing. These models formalize in mathematical terms how established information is combined with learning from data, even if the brain does not directly use or compute probabilities.

**Computer vision and the visual cortex: fundamental differences**

The recent past has shown convergence of computational schemes and brain modeling. There still are, however, major differences between models and the cortex, as well as large differences in the performance between models and the brain. We will discuss below two examples of prominent features of cortical structure that have only a minor role in current computational models.

**Why hierarchies?**

The organization of visual cortex is hierarchical, with features of increasing complexity represented at successive layers. Models of the visual cortex have naturally adopted hierarchical structures. In contrast, in computer vision, the large majority of current schemes are nonhierarchical. Some recent models\textsuperscript{28} have adopted a hierarchical structure and obtained high recognition performance. In addition, some computational schemes are implicitly hierarchical and possibly derive some of their power from their hierarchical
While scale and position invariance can be achieved quite readily in computer vision systems by sequentially scanning the image at different positions and scales, such a strategy appears unlikely to be realized in neural hardware. When properly measured, scale and position tolerance for new objects are less than originally claimed, but still substantial: for at least some of the cells in A1, position tolerance is on the order of 2–4° in the fovea and scale invariance is on the order of a factor of 2–4, which is remarkably large. We will discuss the issue of invariant representations in the next section.

A second possible advantage of hierarchical representations has to do with efficiency: computational speed and use of computational resources. For instance, hierarchy may increase the efficiency of dealing with multiple classes in parallel, by allowing the use of shared features at multiple levels. An increase in efficiency may also be related to the issue of sample complexity. Hierarchical architectures in which each layer is adapted through learning to properties of the visual world may reduce the complexity of the learning task and thus the overall number of labeled examples required for training. Finally, hierarchies also offer an advantage in not only obtaining recognition of the object as a whole, but also in recognizing and localizing parts and subparts at multiple levels, such as a face together with the eyes, nose, mouth, eyebrow, nostril, upper lip, and the like.

Learning from very few labeled examples

One of the striking differences between current machine learning algorithms and learning in biological organisms is that the former require large amounts of labeled training data, whereas biological organisms seem to work well with a much more limited set of labeled examples. We believe that this is due in part to the ability of primate visual systems to compute representations of images that are invariant to the most common image transformations such as translations, changes of scale, rotations, changes in illumination, and, in some cases, changes in pose. One of us has proposed that the main computational goal of the ventral stream in the visual cortex is to learn during development how objects transform and then to compute signatures of new images that are automatically invariant to the same transformations. A biologically plausible way to learn and compute invariant signatures leads to a theory of the ventral stream and of similar hierarchical architectures that achieves the dual goals of learning representations that are invariant to transformations learned in an unsupervised way and considerably reducing the sample complexity of the recognition problem, allowing categorization of new images with a small number of supervised examples. The theory suggests that invariances shape the architecture of the ventral stream and the tuning properties of its neurons. It also formalizes the main properties of neural network architectures such as the hierarchical model and X (HMAX) and convolutional networks, proposing in addition how such an architecture can learn transformations from unsupervised visual experience, instead of being hardwired for some of them.

We suggest, based on this analysis, that present high-performance computer vision systems require millions of labeled examples to achieve good performance because they do not use invariant representations. It seems that at the moment even convolutional networks take only partial advantage of invariance to transformations. It is likely, however, that invariant representations are not the only reason for reducing the number of labeled examples required for a certain level of performance. A second mechanism is sometimes called by a number of different terms, including cross-class generalization, learning to learn, and transfer learning. The basic idea in this approach is to use experience obtained during the learning of a given class to subsequently accelerate the learning of related classes. A third mechanism, more difficult to define precisely at this point but probably important, consists of priors (we use the term prior in a general way here, without necessarily implying a probabilistic interpretation)
incorporated by evolution, which constrain the hypothesis space that has to be explored during recognition.

**Feedforward versus backprojections**

A feedforward architecture from V1 to prefrontal cortex, in the spirit of the Hubel and Wiesel simple–complex hierarchy, seems to account for several properties of visual cells. In particular, recent read-out experiments measuring information that could be read from populations of IT cells\(^7^3\) confirm previous estimates that after about \(\sim 100\) ms from onset of the stimulus, performance of the readout classifier was essentially at its asymptotic performance during passive viewing. In addition, feedforward models also appear to account for recognition performance of human subjects for images flashed briefly and followed by a mask.\(^3^1,^7^7\) The evidence suggests therefore that a feedforward process is sufficient for a fast initial recognition phase, during which primates can already complete difficult recognition tasks involving what the image is. What, then, is the role of the extensive anatomical backprojections in the primate visual system?

Their role may be restricted to learning, but we believe that it is broader. We suggest that even when the feedforward projections by themselves may be capable of answering the **what** question of vision, the backprojections must be added to answer other questions that may be asked in a visual task, including **what is where** (as described by Marr\(^7^8\)). This proposal is similar to previous ideas suggesting that visual cortex follows a hypothesis-and-verification strategy\(^7^9,^8^0\) or a Bayesian inference procedure in which top-down priors are used to compute a set of mutually consistent conditional probabilities at various stages of the visual pathway.\(^8^1\) A recent model\(^2^5\) along these lines demonstrated the use of initial classification using a bottom-up sweep, followed by precise localization of the object and its parts and subparts by a top-down pass.

Top-down pathways in the visual cortex also include the dorsal stream and connections between the dorsal and the ventral stream that are likely to be involved in attentional effects. A Bayesian model\(^8^2\) that takes into account these bottom-up and top-down signals performs well in recognition tasks and predicts some of the characteristic psychophysical and physiological properties of attention. For natural images, the top-down signal improves object recognition performance and predicts human eye fixations, and probably attentional shifts. It is likely to be important for recognition in significant clutter, since the performance of feedforward models decreases significantly when there are more than three to five objects.\(^3^1\) The top-down flow of information combined with a hierarchical representation allows the system to answer not only the what question—that is, perform object identification and categorization—but to answer the what is where question—that is, identification and localization at multiple levels. Of course, we do not believe that the process of vision can be fully characterized in terms of answering, what is where?\(^7^8\) For example, humans can recognize subtle aspects of actions, goals, and social interactions at a level which is far beyond the capabilities of our present algorithms. They can also answer essentially any reasonable question, beyond what and where, on any given image, in a kind of Turing test for vision. The top-down pathway is likely to play an important role for this broader range of visual tasks. We think therefore that while the forward pathway may be described in terms of feedforward architectures in the spirit of Hubel and Wiesel, HMAX, and convolutional networks, the full process of vision also requires verification and interpretation stages that may be best described in terms of top-down specialized routines.

**Discussion: future**

The more we learn about vision, the more questions appear. In this final discussion, we briefly consider two problem domains for future studies. The first focuses on how to close the gap between computer and human vision in the tasks considered above—object categorization and identification. The second part considers broader aspects of vision and its roots in evolution.

**Closing the performance gap**

One general question regarding possible improvements in visual recognition is whether recognition is obtained by multiple specialized mechanisms, or by a uniform scheme applied to different recognition tasks. For example, suggestions have been made that general categorization and individual recognition may be subserved by different mechanisms, or that face recognition may depend on special mechanisms not used for other object categories. It appears to us that the underlying computational
problems in different recognition tasks are similar, and can therefore be approached by the same general scheme, applied to different training sets (and possibly implemented by more than a single neuronal mechanism). The basic recognition scheme could be augmented, however, by specialized mechanisms, dealing with special cases and exceptions. The full system could then be a combination of a scheme that may be characterized as rule based, which can capture the main properties of a category and generalize broadly to novel examples, and a memory-based recognition scheme, which can deal with atypical cases and exceptions to the rule-based scheme.

We next consider future directions that we think could play a useful role in bringing the performance of artificial recognition models closer to the performance level of human vision. These are not the only possible routes for closing the performance gap, but they provide examples of promising general directions motivated by human perception that could usefully be incorporated into artificial systems.

**Continuous learning of rich models**

In current computational schemes, a model for the object or category of interest is constructed during a learning stage and then used for recognition. In contrast, the primate visual system exhibits continuous plasticity and can continue to learn when confronted with new examples. The disadvantage of a fixed limited training stage is that the resulting object model may remain too simple. A visual category often contains a core of typical examples, but also a large number of possible variations, atypical members, and counter-examples. An object can be recognized by its overall shape, but also by small distinguishing parts, and both aspects need to be included in its representation. To achieve human-level performance, it appears therefore that it will be necessary to construct rich object models, learned continuously from a large number of examples.

Such use of continuous learning raises interesting computational challenges: new methods will be required to learn from errors and to continuously modify an existing representation based on new incoming information, possibly combining the rule-based and memory-based mechanisms mentioned above.

**Integrating segmentation and recognition**

Recognition and segmentation are related tasks in the perception of objects: we can usually recognize the object and at the same time identify in the image the precise region containing the object of interest. Historically, segmentation and recognition were treated in computer vision as sequential processes: figure-ground segmentation first identifies in the image a region likely to correspond to a single object; recognition processes are subsequently applied to the selected region to identify the segmented object. More recently, computational models have started to treat the two tasks together, performing object segmentation not only in a bottom-up manner based on image properties, but also in a top-down manner based on object representations stored in memory, as illustrated in Figure 3. This led to substantial progress in object segmentation; however, most current recognition systems do not include segmentation as an integral part of the recognition process. It seems to us that recognition and segmentation are closely linked tasks, and their solutions constrain each other. This
Are models of object recognition catching up with the brain? Poggio & Ullman

Figure 4. Challenges to future visual recognition include (A) actions performed by agents and (B) complex scene interpretation. These broader aspects of visual recognition cannot be efficiently handled by simple extension of existing recognition methods.

Integration appears to be supported by considerable physiological and psychophysical evidence. A closer integration of recognition and segmentation at both the object and part levels is likely therefore to improve the recognition of objects and their parts.

A greater challenge: vision and evolution

The brain uses vision, together with other senses, to obtain knowledge about the world and to act upon it. This knowledge goes beyond object recognition and categorization: vision is also used, for example, to recognize actions performed by agents in the surrounding environment as well as their goals and social interactions, and to complete scene understanding (Fig. 4).

It seems to us that these broader aspects of visual recognition cannot be efficiently handled by simple extension of existing recognition methods. It is likely that in addition to the general learning mechanisms currently used in object recognition models, the brain also uses specialized mechanisms, which have evolved to focus on and extract information required for making judgments about actions, goals, social interactions, and the like.
Innate structures and circuits in the brain do not by themselves incorporate full solutions to these challenging problems, but are more likely to provide useful constraints and initial biases which later lead, guided by learning from the environment, to powerful specific mechanisms.  

A general broad question for future studies is therefore the nature of the innate machinery used by the visual system, its genetic encoding, and how the combination of innate machinery and learning from the environment leads to our understanding of the visual world.

Acknowledgments

We thank J. Leibo and J. Mutch for reading the manuscript and for discussions. S.U. was supported by ERC Advanced Grant 269627 “Digital Baby.” Additional support was provided by SkyTech, ASFOR, and the Eugene McDermott Foundation. We would like to thank the McGovern Institute for Brain Research for their support.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Lowe, D.G. 1987. Three-dimensional object recognition from single two-dimensional images. Artif. Intell. 31: 355–395.
2. Grimson, W.E.L. 1990. Object Recognition by Computer: The Role of Geometric Constraints. Cambridge, MA: MIT Press.
3. Ullman, S. & R. Basri. 1991. Recognition by linear combinations of models. IEEE Trans. Pattern Anal. 13: 992–1006.
4. Rosch, E., C.B. Mervis, W.D. Gray, et al. 1976. Basic objects in natural categories. Cogn. Psychol. 8: 382–439.
5. Turk, M. & A. Pentland. 1991. Eigenfaces for recognition. J.Cogn. Neurosci. 3: 71–86.
6. Brunelli, R. & T. Poggio. 1993. Face recognition: features versus templates. IEEE Trans. Pattern Anal. 15: 1042–1052.
7. Sung, K.K. & T. Poggio. 1998. Example-based learning for view-based human face detection. IEEE Trans. Pattern Anal. 20: 39–51.
8. Viola, P. & M. Jones. 2001. Robust real-time object detection. Int. J. Comput. Vis. 57: 137–154.
9. Rowley, H.A., S. Baluja & T. Kanade. 1998. Neural network-based face detection. IEEE 20: 23–38.
10. Papageorgiou, C. & T. Poggio. 1999. Trainable pedestrian detection. IEEE 4: 35–39.
11. Papageorgiou, C.P., M. Oren & T. Poggio. 1998. A general framework for object detection. Proc. CVPR IEEE 6: 555–562.
12. Sali, E. & S. Ullman. 1999. Combining class-specific fragments for object classification. BMVC 1: 203–213.
13. Weber, M., M. Welling & P. Perona. 2000. Towards automatic discovery of object categories. IEEE Comput. Soc. 2: 101–108.
14. Mohan, A., C. Papageorgiou & T. Poggio. 2001. Example-based object detection in images by components. IEEE Trans. Pattern Anal. 23: 349–361.
15. Agarwal, S. & D. Roth. 2002. Learning a sparse representation for object detection. Computer 4: 1–15.
16. Ullman, S., M. Vidal-Naquet & E. Sali. 2002. Visual features of intermediate complexity and their use in classification. Nat. Neurosci. 5: 682–687.
17. Leibe, B. & B. Schiele. 2003. Interleaved object categorization and segmentation. BMVC 2: 759–768.
18. Fei-fei, L., R. Fergus & P. Perona. 2003. A Bayesian approach to unsupervised one-shot learning of object categories. Proc. CVPR IEEE 2: 1134–1141.
19. Lowe, D.G. 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60: 91–110.
20. Zhang, J. & A. Zisserman. 2006. “Dataset issues in object recognition.” In Toward Category-Level Object Recognition. J. Ponce, M. Hebert, C. Schmid & A. Zisserman, Eds. New York: Springer.
21. Fei-Fei, L., R. Fergus & P. Perona 2004. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. Proceedings of Workshop on Generative-Model Based Vision.
22. Deng, J., W. Dong, R. Socher, et al. 2009. ImageNet: a large-scale hierarchical image database. Proc. CVPR IEEE 1:248–255.
23. Ponce, J. 2006. Toward Category-Level Object Recognition. New York: Springer.
24. The PASCAL visual object classes homepage. [Online]. http://pascal.inria.fr/challenges/VOC. Accessed January 18 2013.
25. Felzenszwalb, P., D. McAllester & D. Ramanan. 2008. A discriminatively trained, multiscale, deformable part model. Proc. CVPR IEEE 8: 1–8.
26. Vedaldi, A., V. Gulshan, M. Varma & A. Zisserman. 2009. Multiple kernels for object detection. In Proceedings of IEEE 12th International Conference on Computer Vision 2009 (ICCV), vol. 1, pp. 606–613.
27. http://www.mobileye.com/. Accessed September 1, 2012.
28. LeCun, Y., B. Boser, J.S. Denker, et al. 1989. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1: 541–551.
29. Krizhevsky, A., I. Sutskever & G. Hinton. 2012. ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. 25: 1106–1114.
30. ArXiv: Computer Vision and Pattern Recognition, [Online]. http://www.science.io/paper/1388351/the-neural-representation-benchmark-and-its. Accessed January 18, 2013.
31. Serre, T., A. Oliva & T. Poggio. 2007. A feedforward architecture accounts for rapid categorization. Proc. Natl. Acad. Sci. USA 104: 6424–6429.
32. Serre, T., M. Kouh, C. Cadieu, et al. 2005. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. http://dspace.mit.edu/handle/1721.1/2040. Accessed September 1, 2012.
33. Fukushima, K. 1975. Cognitron: a self-organizing multilayered neural network. *Biol. Cybern.* **20**: 121–136.
34. Wallis, G. & E.T. Rolls. 1997. A model of invariant object recognition in the visual system. *Prog. Neurobiol.* **51**: 167–194.
35. Mel, B.W. 1997. SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comput.* **9**: 777–804.
36. Riesenhuber, M. & T. Poggio. 1999. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**: 1019–1025.
37. Thorpe, S. 2002. Ultra-rapid scene categorization with a wave of spikes. *Lect. Notes Comput. SC.* **2525**: 1–15.
38. Amit, Y. & M. Mascal. 2003. An integrated network for invariant visual detection and recognition. *Vis. Res.* **43**: 2073–2088.
39. Wersing, H. & E. Körner. 2003. Learning optimized features for hierarchical models of invariant object recognition. *Neural Comput.* **15**: 1559–1588.
40. Serre, T., G. Kreiman, M. Koulh, et al. 2007. A quantitative theory of immediate visual recognition. *Prog. Brain Res.* **165**: 33–56.
41. Mutch, J. & D.G. Lowe. 2006. Multiclass object recognition with sparse, localized features. *Proc. CVPR IEEE* **1**: 11–18.
42. Mutch, J. & D.G. Lowe. 2008. Object class recognition and localization using sparse features with limited receptive fields. *Int. J. Comput. Vis.* **80**: 45–57.
43. Pinto, N., D. Doukhan, J.J. DiCarlo & D.D. Cox. 2009. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Comput. Biol.* **5**: 12.
44. Vapnik, V.N. 1998. *Statistical Learning Theory*. New York: Wiley.
45. Devroye, L., L. Györfi & G. Lugosi. 1996. *A Probabilistic Theory of Pattern Recognition*. New York: Springer.
46. Cucker, F. & S. Smale. 2001. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.* **39**: 1–50.
47. Poggio, T. & S. Smale. 2003. The mathematics of learning: dealing with data. *Not. Am. Math. Soc.* **50**: 537–544.
48. Bousquet, O. 2004. Introduction to statistical learning theory. *Biol. Cybern.* **3176**: 169–207.
49. Poggio, T., R. Rifkin, S. Mukherjee & P. Niyogi. 2004. General conditions for predictivity in learning theory. *Nature* **428**: 419–422.
50. Marr, D. & H.K. Nishihara. 1978. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. B* **200**: 269–294.
51. Biederman, I. 1985. Human image understanding: recent research and a theory. *Comput. Vis. Graph.* **32**: 29–73.
52. Belongie, S., J. Malik & J. Puzicha. 2002. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. 24*: 509–522.
53. Kobatake, E. & K. Tanaka. 1994. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.* **71**: 856–867.
54. Bülthoff, H.H., S.Y. Edelman & M.J. Tarr. 1994. How are three-dimensional objects represented in the brain? *Cereb. Cortex* **5**: 247–260.
55. Tarr, M.J. & H.H. Bülthoff. 1998. Image-based object recognition in man, monkey and machine. *Cognition* **67**: 1–20.
56. Janssen, P., R. Vogels & G.A. Orban. 2000. Selectivity for 3D shape that reveals distinct areas within macaque inferior temporal cortex. *Science* **288**: 2054–2056.
57. Tanaka, K. 2000. Curvature in depth for object representation. *Neuron* **27**: 195–196.
58. Yamane, Y., E.T. Carlson, K.C. Bowman, et al. 2008. A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nat. Neurosci.* **11**: 1352–1360.
59. Shepard, R.N. & J. Metzler. 1971. Mental rotation of three-dimensional objects. *Science* **171**: 701–703.
60. Wahba, G. 1990. *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.
61. Poggio, T. & F. Girosi. 1990. Networks for approximation and learning. *Proc. IEEE* **78**: 1481–1497.
62. Freund, Y. & R.E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**: 119–139.
63. Jordan, M.I. 2004. Graphical models. *Stat. Sci.* **19**: 140–155.
64. Bienenstock, E., S. Geman & D. Potter. 1998. Compositionality, MDL Priors, and Object Recognition. *In Advances in Neural Information Processing Systems* **9**: 838–844.
65. Teh, Y.W., M.I. Jordan, M.J. Beal & D.M. Blei. 2006. Hierarchical dirichlet processes. *J. Amer. Stat. Assoc.* **101**: 1566–1587.
66. Kemp, C. & J.B. Tenenbaum. 2008. The discovery of structural form. *Proc. Natl. Acad. Sci. USA* **105**: 10687–10692.
67. Sudderth, E.B. & M.I. Jordan. 2008. Shared segmentation of natural scenes using dependent pitman-yor processes. *Electr. Eng.* **21**: 1–10.
68. Belkin, M. & P. Niyogi. 2004. Semi-supervised learning on riemannian manifolds. *Mach. Learn.* **56**: 209–239.
69. Logothetis, N.K., J. Pauls, H.H. Bülthoff & T. Poggio. 1994. View-dependent object recognition by monkeys. *Curr. Biol.* **4**: 401–414.
70. Logothetis, N.K., J. Pauls & T. Poggio. 1995. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* **5**: 552–563.
71. Logothetis, N.K. & D.L. Sheinberg. 1996. Visual object recognition. *Ann. Rev. Neurosci.* **19**: 577–621.
72. Bruce, C., R. Desimone & C.G. Gross. 1981. Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *J. Neurophysiol.* **46**: 369–384.
73. Hung, C.P., G. Kreiman, T. Poggio & J.J. DiCarlo. 2005. Fast readout of object identity from macaque inferior temporal cortex. *Science* **310**: 863–866.
74. Smale, S., L. Rosasco, J. Bouvrie, et al. 2009. Mathematics of the neural response. *Found. Comput. Math.* **10**: 67–91.
75. Epstein, B., I. Lifshitz & S. Ullman. 2008. Image interpretation by a single bottom-up top-down cycle. *Proc. Natl. Acad. Sci. USA* **105**: 14298–14303.
76. Poggio, T. 2011. The computational magic of the ventral stream: towards a theory. *Nature Precedings*. http://dx.doi.org/10.1038/npre.2011.6117.1 Accessed January 18, 2013.
77. Thorpe, S., D. Fize & C. Marlot. 1996. Speed of processing in the human visual system. *Nature* **381**: 520–522.
Are models of object recognition catching up with the brain?

78. Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W.H. Freeman.

79. Carpenter, G.A. & S. Grossberg. 1987. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Comput. Vis. Graph.* 37: 54–115.

80. Hawkins, J. & S. Blakeslee. 2004. *On Intelligence*. New York: Times Books.

81. Lee, T.S. & D. Mumford. 2003. Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A* 20: 1434–1448.

82. Chokkerur, S., T. Serre & T. Poggio. 2009. Computer science and artificial intelligence laboratory technical report a Bayesian inference theory of attention: neuroscience and algorithms. http://dspace.mit.edu/handle/1721.1/49416. Accessed September 1, 2012.

83. Sharon, E., M. Galun, D. Sharon, et al. 2006. Hierarchy and adaptivity in segmenting visual scenes. *Nature* 442: 810–813.

84. Borenstein, E. & S. Ullman. 2008. Combined bottom-up and top-down segmentation. *IEEE Trans. Pattern Anal.* 30: 1–17.

85. Zhou, H., H.S. Friedman & R. Von Der Heydt. 2000. Coding of border ownership in monkey visual cortex. *J. Neurosci.* 20: 6594–6611.

86. Zemel, R.S., M. Behrmann, M.C. Mozer & D. Bavelier. 2002. Experience-dependent perceptual grouping and object-based attention. *J. Exp. Psychol. Hum.* 28: 202–217.

87. Ullman, S., D. Harari & N. Dorfman. 2012. From simple innate biases to complex visual concepts. *Proc. Natl. Acad. Sci. USA.* 109: 18215–18220.