DOI:10.1145/1831407.1831425

**Neuroscience is beginning to inspire a new generation of seeing machines.**

BY THOMAS SERRE AND TOMASO POGGIO

# A Neuromorphic Approach to Computer Vision

IF PHYSICS WAS *the* science of the first half of the 20th century, biology was certainly *the* science of the second half. Neuroscience is now often cited as one of the key scientific focuses of the 21st century and has indeed grown rapidly in recent years, spanning a range of approaches, from molecular neurobiology to neuro-informatics and computational neuroscience. Computer science gave biology powerful new data-analysis tools that yielded bioinformatics and genomics, making possible the sequencing of the human genome. Similarly, computer science techniques are at the heart of brain imaging and other branches of neuroscience.

Computers are critical for the neurosciences, though at a much deeper level, representing the best metaphor for the central mystery of how the brain produces intelligent behavior and intelligence itself. They also provide experimental tools for information processing, effectively testing theories of the brain, particularly those involving aspects of intelligence (such as sensory perception). The contribution of computer science to neuroscience happens at multiple levels and is well recognized. Perhaps less obvious is that neuroscience is beginning to contribute powerful new ideas and approaches to artificial intelligence and computer science as well. Modern computational neuroscience models are no longer toy models but quantitatively detailed while beginning to compete with state-of-the-art computer-vision systems. Here, we explore how computational neuroscience could become a major source of new ideas and approaches in artificial intelligence.

Understanding the processing of information in our cortex is a significant part of understanding how the brain works and understanding intelligence itself. For example, vision is one of our most developed senses. Primates easily categorize images or parts of images, as in, say, an office scene or a face within a scene, identifying specific objects. Our visual capabilities are exceptional, and, despite decades of engineering, no computer algorithm is yet able to match the performance of the primate visual system.

Our visual cortex may serve as a proxy for the rest of the cortex and thus

## » key insights

- **The past century of neuroscience research has begun to answer fundamental questions ranging from the intricate inner workings of individual neurons to understanding the collective behavior of networks of millions of neurons.**
- **A key challenge for the visual cortex is how to deal with the poverty-of-stimulus problem.**
- **A major goal of the visual system is how to adapt to the statistics of its natural environment through visual experience and even evolution.**

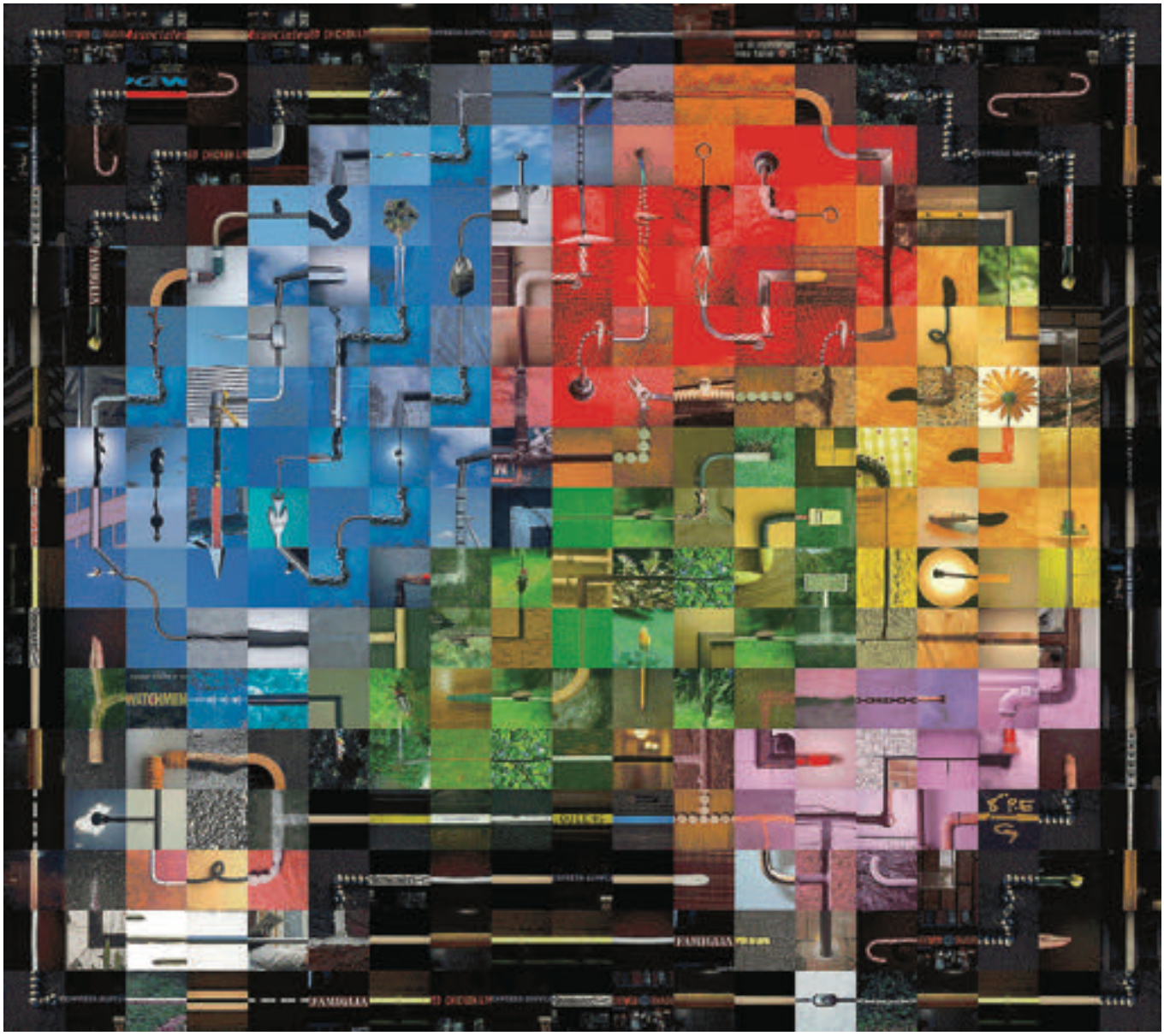


ILLUSTRATION BY KAI SCHREIBER

for intelligence itself. There is little doubt that even a partial solution to the question of which computations are performed by the visual cortex would be a major breakthrough in computational neuroscience and more broadly in neuroscience. It would begin to explain one of the most amazing abilities of the brain and open doors to other aspects of intelligence (such as language and planning). It would also bridge the gap between neurobiology and the various information sciences, making it possible to develop computer algorithms that follow the information-processing principles used by biological organisms and honed by natural evolution.

The past 60 years of experimental work in visual neuroscience has generated a large and rapidly increasing amount of data. Today's quantitative models bridge several levels of understanding, from biophysics to physiology to behavior. Some of these models compete with state-of-the-art computer-vision systems and are close to human-level performance for specific visual tasks.

Here, we describe recent work toward a theory of cortical visual processing. Unlike other models that address the computations in a given brain area (such as primary visual cortex) or attempt to explain a particular phenomenon (such as contrast adaptation and specific visual illusion), we describe a large-scale model that attempts to mimic the main information-processing steps across multiple brain areas and millions of neuron-like units. A first step toward understanding cortical functions may take the form of a detailed, neurobiologically plausible model, accounting for the connectivity, biophysics, and physiology of the cortex.

Models provide a much-needed framework for summarizing and integrating existing data and planning, coordinating, and interpreting new experiments. They can be powerful tools in basic research, integrating knowledge across multiple levels of analysis, from molecular to synaptic, cellular, systems, and complex visual behavior. However, models, as we discuss later, are limited in explanatory power but should, ideally, lead to a deeper and more general theory. Here, we discuss the role of the visual cortex and review key computational principles underlying the processing of information dur-

ing visual recognition, then explore a computational neuroscience model (representative of a class of older models) that implements these principles, including some of the evidence in its favor. When tested with natural images, the model performs robust object recognition on par with computer-vision systems and human performance for a specific class of quick visual-recognition tasks. The initial success of this research represents a case in point for arguing that over the next decade progress in computer vision and artificial intelligence promises to benefit directly from progress in neuroscience.

## Goal of the Visual System

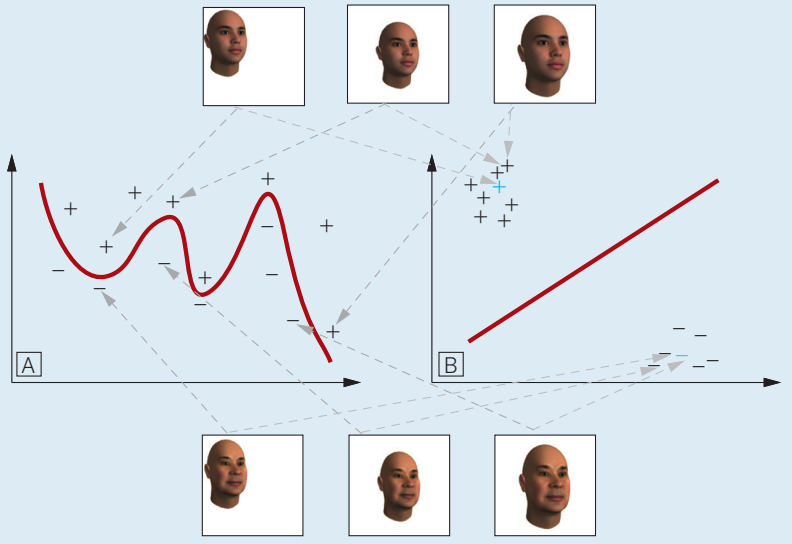
A key computational issue in object recognition[a] is the specificity-invariance trade-off: Recognition must be able to finely discriminate between different objects or object classes (such as the faces in Figure 1) while being tolerant of object transformations (such as scaling, translation, illumination, changes in viewpoint, and clutter), as well as non-rigid transformations (such as variations in shape within a class), as in the change of facial expression in recognizing faces.

A key challenge posed by the visual cortex is how well it deals with the poverty-of-stimulus problem, or simple lack of visual information. Primates are able to learn to recognize an object in quite different images from far fewer labeled examples than are predicted by our present learning theory and algorithms. For instance, discriminative algorithms (such as support vector machines, or SVMs) can learn a complex object-recognition task from a few hundred labeled images. This number is small compared to the apparent dimensionality of the problem (millions of pixels), but a child, even a monkey, is apparently able to learn the same task from a handful of examples. As an example of the prototypical problem in visual recognition, imagine a (naïve) machine is shown an image of a given person and an image of another person. The system's task is to discriminate future images of these two people without seeing other images of them, though it has seen many images of other people and objects and their transformations and may have learned from them in an unsupervised way. Can the system learn to perform the classification task correctly with just two (or a few) labeled examples?

Imagine trying to build such a classifier from the output of two cortical cells, as in Figure 1. Here, the response of the two cells defines a 2D feature space to represent visual stimuli. In a more realistic setting, objects would be represented by the response patterns of thousands of such neurons. In the figure, we denote visual examples from the two people with + and – signs; panels (A) and (B) illustrate what the recognition problem would look like when these two neurons are sensitive vs. invariant to the precise position of the object within their receptive fields.[b] In each case, a separation (the red lines indicate one such possible separation) can be found between the two classes. It has been shown that certain learning algorithms (such as SVMs with Gaussian kernels) can solve any discrimination task with arbitrary difficulty (in the limit of an infinite number of training examples). That is, with certain classes of learning algorithms we are guaranteed to be able to find a separation for the problem at hand irrespective of the difficulty of the recognition task. However, learning to solve the problem may require a prohibitively large number of training examples.

In separating two classes, the two representations in panels (A) and (B) are not equal; the one in (B) is far superior to the one in (A). With no prior assumption on the class of functions to be learned, the "simplest" classifier that can separate the data in (B) is much simpler than the "simplest" classifier that separates the data in (A). The number of wiggles of the separation line (related to the number of parameters to be learned) gives a hand-wavy estimate of the complexity of a classifier. The sample complexity of the problem derived from the invariant representation in (B) is much lower than that of

**Figure 1. Sample complexity.**




A hypothetical 2D (face) classification problem (red) line. One class is represented with + and the other with – symbols. Insets are 2D transformations (translation and scales) applied to examples from the two categories. Panels (A) and (B) are two different representations of the same set of images. (B), which is tolerant with respect to the exact position and scale of the object within the image, leads to a simpler decision function (such as a linear classifier) and requires fewer training examples to achieve similar performance, thus lowering the sample complexity of the classification problem. In the limit, learning in (B) could be done with only two training examples (blue).

a Within recognition, one distinguishes between identification and categorization. From a computational point of view, both involve classification and represent two points on a spectrum of generalization levels.

b The receptive field of a neuron is the part of the visual field that (properly stimulated) could elicit a response from the neuron.

the problem in (A). Learning to categorize the data-points in (B) requires far fewer training examples than in (A) and may be done with as few as two examples. The key problem in vision is thus what can be learned effectively with only a small number of examples.[c]

The main point is not that a low-level representation provided from the retina would not support robust object recognition. Indeed, relatively good computer-vision systems developed in the 1990s were based on simple retina-like representations and rather complex decision functions (such as radial basis function networks). The main problem of these systems is they required a prohibitively large number of training examples compared to humans.

More recent work in computer vision suggests a hierarchical architecture may provide a better solution to the problem; see also Bengio and Le Cun[1] for a related argument. For instance, Heisele et al.[10] designed a hierarchical system for the detection and recognition of faces, an approach based on a hierarchy of "component experts" performing a local search for one facial component (such as an eye or a nose) over a range of positions and scales. Experimental evidence from Heisele et al.[10] suggests such hierarchical systems based exclusively on linear (SVM) classifiers significantly outperform a shallow architecture that tries to classify a face as a whole, albeit by relying on more complex kernels.

The visual system may be using a similar strategy to recognize objects, with the goal of reducing the sample complexity of the classification problem. In this view, the visual cortex transforms the raw image into a position- and scale-tolerant representation through a hierarchy of processing stages, whereby each layer gradually increases the tolerance to position and scale of the image representation. After several layers of such processing stages, the resulting image representation can be used much more efficiently for task-dependent learning and classification by higher brain areas.

**The role of the anatomical back-projections present (in abundance) among almost all areas in visual cortex is a matter of debate.**

These stages can be learned during development from temporal streams of natural images by exploiting the statistics of natural environments in two ways: correlations over images that provide information-rich features at various levels of complexity and sizes; and correlations over time used to learn equivalence classes of these features under transformations (such as shifts in position and changes in scale). The combination of these two learning processes allows efficient sharing of visual features between object categories and makes learning new objects and categories easier, since they inherit the invariance properties of the representation learned from previous experience in the form of basic features common to other objects. In the following sections, we review evidence for this hierarchical architecture and the two correlation mechanisms described earlier.

### Hierarchical Architecture and Invariant Recognition

Several lines of evidence (from both human psychophysics and monkey electrophysiology studies) suggest the primate visual system exhibits at least some invariance to position and scale. While the precise amount of invariance is still under debate, there is general agreement as to the fact that there is at least some generalization to position and scale.

The neural mechanisms underlying such invariant visual recognition have been the subject of much computational and experimental work since the early 1990s. One general class of computational models postulates that the hierarchical organization of the visual cortex is key to this process; see also Hegdé and Felleman[9] for an alternative view. The processing of shape information in the visual cortex follows a series of stages, starting with the retina and proceeding through the lateral geniculate nucleus (LGN) of the thalamus to primary visual cortex (V1) and extrastriate visual areas, V2, V4, and the inferotemporal (IT) cortex. In turn, IT provides a major source of input to the prefrontal cortex (PFC) involved in linking perception to memory and action; see Serre et al.[29] for references.

As one progresses along the ventral

c The idea of sample complexity is related to the point made by DiCarlo and Cox[4] about the main goal of processing information from the retina to higher visual areas to be "untangling object representations," so a simple linear classifier can discriminate between any two classes of objects.

stream of the visual cortex, neurons become selective for stimuli that are increasingly complex—from simple oriented bars and edges in early visual area V1 to moderately complex features in intermediate areas (such as a combination of orientations) to complex objects and faces in higher visual areas (such as IT). Along with this increase in complexity of the preferred stimulus, the invariance properties of neurons seem to also increase. Neurons become more and more tolerant with respect to the exact position and scale of the stimulus within their receptive fields. As a result, the receptive field size of neurons increases from about one degree or less in V1 to several degrees in IT.

Compelling evidence suggests that IT, which has been critically linked with a monkey's ability to recognize objects, provides a representation of the image that facilitates recognition tolerant of image transformations. For instance, Logothetis et al.[16] showed that monkeys can be trained to recognize paperclip-like wireframe objects at a specific location and scale. After training, recordings in their IT cortex revealed significant selectivity for the trained objects. Because monkeys were unlikely to have been in contact with the specific paperclip prior to training, this experiment provides indirect evidence of learning. More important, Logothetis et al.[16] found selective neurons also exhibited a range of invariance with respect to the exact position (two to four degrees) and scale (around two octaves) of the stimulus, which was never presented before testing at these new positions and scales. In 2005, Hung et al.[12] showed it was possible to train a (linear) classifier to robustly read out from a population of IT neurons the category information of a briefly flashed stimulus. Hung et al. also showed the classifier was able to generalize to a range of positions and scales (similar to Logothetis et al.'s data) not presented during the training of the classifier. This generalization suggests the observed tolerance to 2D transformation is a property of the population of neurons learned from visual experience but available for a novel object without object-specific learning, depending on task difficulty.

## Computational Models of Object Recognition in Cortex

We developed[26,29] (in close cooperation with experimental labs) an initial quantitative model of feedforward hierarchical processing in the ventral stream of the visual cortex (see Figure 2). The resulting model effectively integrates the large body of neuroscience data (summarized earlier) characterizing the properties of neurons along the object-recognition processing hierarchy. The model also mimics human performance in difficult visual-recognition tasks[28] while performing at least as well as most current computer-vision systems.[27]

Feedforward hierarchical models have a long history, beginning in the 1970s with Marko and Giebel's homogeneous multilayered architecture[17] and later Fukushima's Neocognitron.[6] One of their key computational mechanisms originates from the pioneering physiological studies and models of Hubel and Wiesel (http://serre-lab.clps.brown.edu/resources/ACM2010). The basic idea is to build an increasingly complex and invariant object representation in a hierarchy of stages by progressively integrating, or pooling, convergent inputs from lower levels. Building on existing models (see supplementary notes http://serre-lab.clps.brown.edu/resources/ACM2010), we have been developing[24,29] a similar computational theory that attempts to quantitatively account for a host of recent anatomical and physiological data; see also Mutch and Lowe[19] and Masquelier et al.[18]

The feedforward hierarchical model in Figure 2 assumes two classes of functional units: simple and complex. Simple act as local template-matching operators, increasing the complexity of the image representation by pooling over local afferent units with selectivity for different image features (such as edges at different orientations). Complex increase the tolerance of the representation with respect to 2D transformations by pooling over afferent units with similar selectivity but slightly different positions and scales.

**Learning and plasticity.** How the organization of the visual cortex is influenced by development vs. genetics is a matter of debate. An fMRI study[21] showed the patterns of neural activity elicited by certain ecologically important classes of objects (such as faces and places in monozygotic twins) are significantly more similar than in dizygotic twins. These results suggest that genes may play a significant role in the way the visual cortex is wired to process certain object classes. Meanwhile, several electrophysiological studies have demonstrated learning and plasticity in the adult monkey; see, for instance, Li and DiCarlo.[15] Learning is likely to be both faster and easier to elicit in higher visually responsive areas (such as PFC and IT[15]) than in lower areas.

This learning result makes intuitive sense. For the visual system to remain stable, the time scale for learning should increase ascending the ventral stream.[d] In the Figure 2 model, we assumed unsupervised learning from V1 to IT happens during development in a sequence starting with the lower areas. In reality, learning might continue throughout adulthood, certainly at the level of IT and perhaps in intermediate and lower areas as well.

*Unsupervised learning in the ventral stream of the visual cortex.* With the exception of the task-specific units at the top of the hierarchy ("visual routines"), learning in the model in Figure 2 is unsupervised, thus closely mimicking a developmental learning stage.

As emphasized by several authors, statistical regularities in natural visual scenes may provide critical cues to the visual system for learning with very limited or no supervision. A key goal of the visual system may be to adapt to the statistics of its natural environment through visual experience and perhaps evolution, too. In the Figure 2 model, the selectivity of simple and complex units can be learned from natural video sequences (see supplementary ma-

d In the hierarchical model in Figure 1, learning proceeds layer by layer, starting at the bottom, a process similar to recent work by Hinton[11] but that is quite different from the original neural networks that used back-propagation and simultaneously learned all layers at the same time. Our implementation includes the unsupervised learning of features from natural images but assumes the learning of position and scale tolerance, thus hardwired in the model; see Masquelier et al.[18] for an initial attempt at learning position and scale tolerance in the model.

terial http://serre-lab.clps.brown.edu/resources/ACM2010 for details).

*Supervised learning in higher areas.* After this initial developmental stage, learning a new object category requires training only of task-specific circuits at the top of the ventral-stream hierarchy, thus providing a position and scale-invariant representation to task-specific circuits beyond IT to learn to generalize over transformations other than image-plane transformations (such as 3D rotation) that must be learned anew for each object or category. For instance, pose-invariant face categorization circuits may be built, possibly in PFC, by combining several units tuned to different face examples, including different people, views, and lighting conditions (possibly in IT).

A default routine may be running in a default state (no specific visual task), perhaps the routine `What is there?` As an example of a simple routine consider a classifier that receives the activity of a few hundred IT-like units, tuned to examples of the target object and distractors. While learning in the model from the layers below is stimulus-driven, the PFC-like classification units are trained in a supervised way following a perceptron-like learning rule.

## Immediate Recognition

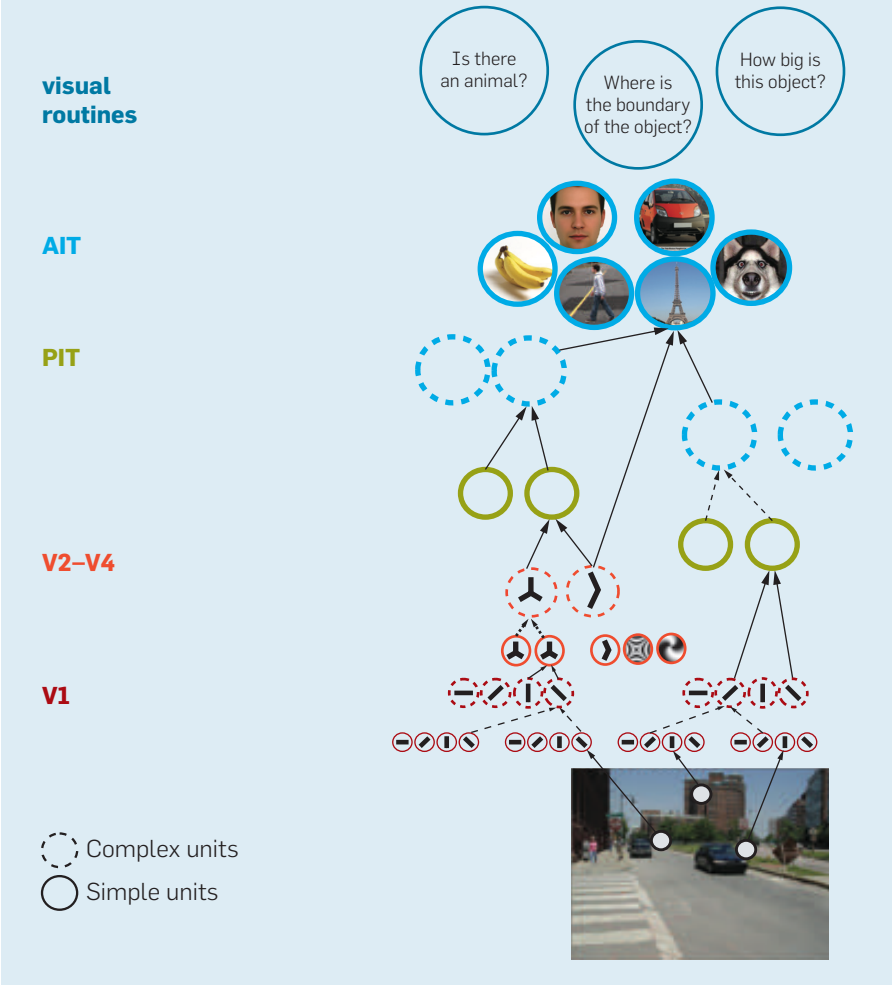
The role of the anatomical back-projections present (in abundance) among almost all areas in the visual cortex is a matter of debate. A commonly accepted hypothesis is that the basic processing of information is feedforward,[30] supported most directly by the short times required for a selective response to appear in cells at all stages of the hierarchy. Neural recordings from IT in a monkey[12] show the activity of small neuronal populations over very short time intervals (as short as 12.5ms and about 100ms after stimulus onset) contains surprisingly accurate and robust information supporting a variety of recognition tasks. While this data does not rule out local feedback loops within an area, it does suggest that a core hierarchical feedforward architecture (like the one described here) may be a reasonable starting point for a theory of the visual cortex, aiming to explain immediate recognition, the initial phase of recognition before eye movement and high-level processes take place.

*Agreement with experimental data.* Since its original development in the late 1990s,[24,29] the model in Figure 2 has been able to explain a number of new experimental results, including data not used to derive or fit model parameters. The model seems to be qualitatively and quantitatively consistent with (and in some cases predicts[29]) several properties of subpopulations of cells in V1, V4, IT, and PFC, as well as fMRI and psychophysical data (see the sidebar "Quantitative Data Compatible with the Model" for a complete list of findings).

We compared the performance of the model against the performance of human observers in a rapid animal vs. non-animal recognition task[28] for which recognition is quick and cortical back-projections may be less relevant. Results indicate the model predicts human performance quite well during such a task, suggesting the model may indeed provide a satisfactory description of the feedforward path. In particular, for this experiment, we broke down the performance of the model and human observers into four image categories with varying amounts of clutter. Interestingly, the performance of both the model and the human observers was most accurate (~90% correct for both human participants and the model) on images for which the amount of information is maximal and clutter minimal and decreases monotically as the clutter in the image increases. This decrease in performance with increasing clutter likely reflects a key limitation of this type of feedforward architecture. This result is in agreement with the reduced selectivity of neurons in V4 and IT when presented with multiple stimuli within their receptive fields for which the model provides a good quantitative fit[29] with neurophysiology data (see the sidebar).

**Application to computer vision.**

**Figure 2. Hierarchical feedforward models of the visual cortex.**



How does the model[29] perform real-world recognition tasks? And how does it compare to state-of-the-art artificial-intelligence systems? Given the specific biological constraints the theory must satisfy (such as using only biophysically plausible operations, receptive field sizes, and a range of invariances), it was not clear how well the model implementation would perform compared to systems heuristically engineered for these complex tasks.

Several years ago, we were surprised to find the model capable of recognizing complex images,[27] performing at a level comparable to some of the best existing systems on the CalTech-101 image database of 101 object categories with a recognition rate of about 55% (chance level < 1%); see Serre et al.[27] and Mutch and Lowe.[19] A related system with fewer layers, less invariance, and more units had an even better recognition rate on the CalTech data set.[20]

We also developed an automated system for parsing street-scene images[27] based in part on the class of models described earlier. The system recognizes seven different object categories—cars, pedestrians, bikes, skies, roads, buildings, trees—from natural images of street scenes despite very large variations in shape (such as trees in summer and winter and SUVs and compact cars from any point of view).

Content-based recognition and search in videos is an emerging application of computer vision, whereby neuroscience may again suggest an avenue for approaching the problem. In 2007, we developed an initial model for recognizing biological motion and actions from video sequences based on the organization of the dorsal stream of the visual cortex,[13] which is critically linked to the processing of motion information, from V1 and MT to higher motion-selective areas MST/FST and STS. The system relies on computational principles similar to those in the model of the ventral stream described earlier but that start with spatio-temporal filters modeled after motion-sensitive cells in the primary visual cortex.

We evaluated system performance for recognizing actions (human and animal) in real-world video sequences,[13] finding that the model of the dorsal stream competed with a state-of-the-art action-recognition system (that outperformed many other systems) on all three data sets.[13] A direct extension of this approach led to a computer system for the automated monitoring and analysis of rodent behavior for behavioral phenotyping applications that perform on par with human manual scoring. We also found the learning in this model produced a large dictionary of optic-flow patterns that seems consistent with the response properties of cells in the medial temporal (MT) area in response to both isolated gratings and plaids, or two gratings superimposed on one another.

## Quantitative Data Compatible with the Model

**Black corresponds to data used to derive the parameters of the model, red to data consistent with the model (not used to fit model parameters), and blue to actual correct predictions by the model. Notations: PFC (prefrontal cortex), V1 (visual area I or primary visual cortex), V4 (visual area IV), and IT (inferotemporal cortex). Data from these areas corresponds to monkey electrophysiology studies. LOC (Lateral Occipital Complex) involves fMRI with humans. The psychological studies are psychophysics on human subjects.**

| Area | Type of data | Ref. biol. data | Ref. model data |
|---|---|---|---|
| **Psych.** | Rapid animal categorization | (1) | (1) |
| | Face inversion effect | (2) | (2) |
| **LOC** | Face processing (fMRI) | (3) | (3) |
| **PFC** | Differential role of IT and PFC in categorization | (4) | (5) |
| **IT** | Tuning and invariance properties | (6) | (5) |
| | Read out for object category | (7) | (8,9) |
| | Average effect in IT | (10) | (10) |
| **V4** | MAX operation | (11) | (5) |
| | Tuning for two-bar stimuli | (12) | (8,9) |
| | Two-spot interaction | (13) | (8) |
| | Tuning for boundary conformation | (14) | (8,15) |
| | Tuning for Cartesian and non-Cartesian gratings | (16) | (8) |
| **V1** | Simple and complex cells tuning properties | (17–19) | (8) |
| | MAX operation in subset of complex cells | (20) | (5) |


1. Serre, T., Oliva, A., and Poggio, T. *Proc. Natl. Acad. Sci.104*, 6424 (Apr. 2007).
2. Riesenhuber, M. et al. *Proc. Biol. Sci. 271*, S448 (2004).
3. Jiang, X. et al. *Neuron 50*, 159 (2006).
4. Freedman, D.J., Riesenhuber, M., Poggio, T., and Miller, E.K. *Journ. Neurosci. 23*, 5235 (2003).
5. Riesenhuber, M. and Poggio, T. *Nature Neuroscience 2*, 1019 (1999).
6. Logothetis, N.K., Pauls, J., and Poggio, T. *Curr. Biol. 5*, 552 (May 1995).
7. Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, *J.J. Science 310*, 863 (Nov. 2005).
8. Serre, T. et al. *MIT AI Memo 2005-036 / CBCL Memo 259* (2005).
9. Serre, T. et al. *Prog. Brain Res. 165*, 33 (2007).
10. Zoccolan, D., Kouh, M., Poggio, T., and DiCarlo, *J.J. Journ. Neurosci. 27*, 12292 (2007).
11. Gawne, T.J. and Martin, J.M. *Journ. Neurophysiol. 88*, 1128 (2002).
12. Reynolds, J.H., Chelazzi, L., and Desimone, R. *Journ. Neurosci.19*, 1736 (Mar. 1999).
13. Taylor, K., Mandon, S., Freiwald, W.A., and Kreiter, A.K. *Cereb. Cortex 15*, 1424 (2005).
14. Pasupathy, A. and Connor, C. *Journ. Neurophysiol. 82*, 2490 (1999).
15. Cadieu, C. et al. *Journ. Neurophysiol. 98*, 1733 (2007).
16. Gallant, J.L. et al. *Journ. Neurophysiol. 76*, 2718 (1996).
17. Schiller, P.H., Finlay, B.L., and Volman, S.F. *Journ. Neurophysiol. 39*, 1288 (1976).
18. Hubel, D.H. and Wiesel, T.N. *Journ. Physiol. 160*, 106 (1962).
19. De Valois, R.L., Albrecht, D.G., and Thorell, L.G. *Vision Res. 22*, 545 (1982).
20. Lampl, I., Ferster, D., Poggio, T., and Riesenhuber, M. *Journ. Neurophysiol. 92*, 2704 (2004).


### Conclusion

Demonstrating that a model designed to mimic known anatomy and physiol-

ogy of the primate visual system leads to good performance with respect to computer-vision benchmarks may suggest neuroscience is on the verge of providing novel and useful paradigms to computer vision and perhaps to other areas of computer science as well. The feedforward model described here can be modified and improved by taking into account new experimental data (such as more detailed properties of specific visual areas like V1[25]), implementing some of its implicit assumptions (such as learning invariances from sequences of natural images), taking into account additional sources of visual information (such as binocular disparity and color), and extention to describe the detailed dynamics of neural responses. Meanwhile, the recognition performance of models of this general type can be improved by exploring parameters (such as receptive field sizes and connectivity) by, say, using computer-intensive iterations of a mutation-and-test cycle.

However, it is important to realize the intrinsic limitations of the specific computational framework we have described and why it is at best a first step toward understanding the visual cortex. First, from the anatomical and physiological point of view the class of feedforward models we've described here is incomplete, as it does not account for the massive back-projections found in the cortex. To date, the role of cortical feedback remains poorly understood. It is likely that feedback underlies top-down signals related to attention, task-dependent biases, and memory. Back-projections must also be taken into account in order to describe visual perception beyond the first 100msec–200msec.

Given enough time, humans use eye movement to scan images, and performance in many object-recognition tasks improves significantly over that obtained during quick presentations. Extensions of the model to incorporate feedback are possible and under way.[2] Feedforward models may well turn out to be approximate descriptions of the first 100msec–200msec of the processing required by more complex theories of vision based on back-projections.[3,5,7,8,14,22,31] However, the computations involved in the initial phase are nontrivial but essential for any scheme involving feedback. A related point is that normal visual perception is much more than classification, as it involves interpreting and parsing visual scenes. In this sense, the class of models we describe is limited, since it deals only with classification tasks. More complex architectures are needed; see Serre et al.[26] for a discussion.

Finally, we described a class of models, *not* a theory. Computational models are not sufficient on their own. Our model, despite describing (quantitatively) aspects of monkey physiology and human recognition, does not yield a good understanding of the computational principles of the cortex and their power. What is yet needed is a mathematical theory to explain the hierarchical organization of the cortex.

## Acknowledgments

We thank Jake Bouvrie for his useful feedback on the manuscript, as well as the referees for their valuable comments.

**References**

1. Bengio, J. and Le Cun, Y. Scaling learning algorithms towards AI. In *Large-Scale Kernel Machines*, L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, J., Eds. MIT Press, Cambridge, MA, 2007, 321–360.
2. Chikkerur, S., Serre, T., Tan, C., and Poggio, T. *What and Where: A Bayesian Inference Theory of Attention* (in press). Vision Research, 2010.
3. Dean, T. A computational model of the cerebral cortex. In *Proceedings of the 20th National Conference on Artificial Intelligence* (Pittsburgh, PA, July 9–13, 2005), 938–943.
4. DiCarlo, J.J. and Cox, D.D. Untangling invariant object recognition. *Trends in Cognitive Science 11*, 8 (Aug. 2007), 333–341.
5. Epshtein, B., Lifshitz, I., and Ullman, S. Image interpretation by a single bottom-up top-down cycle. *Proceedings of the National Academy of Sciences 105*, 38 (Sept. 2008), 14298–14303.
6. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics 36*, 4 (Apr. 1980), 193–202.
7. George, D. and Hawkins, J. A hierarchical Bayesian model of invariant pattern recognition in the visual cortex. In *Proceedings of the International Joint Conference on Neural Networks 3*, (Montréal, July 31–Aug. 4). IEEE Press, 2005, 1812–1817.
8. Grossberg, S. Towards a unified theory of neocortex: Laminar cortical circuits for vision and cognition. *Progress in Brain Research 165* (2007), 79–104.
9. Hegdé, H. and Felleman, D.J. Reappraising the functional implications of the primate visual anatomical hierarchy. *The Neuroscientist 13*, 5 (2007), 416–421.
10. Heisele, B., Serre, T., and Poggio, T. A component-based framework for face detection and identification. *International Journal of Computer Vision 74*, 2 (Jan. 1, 2007), 167–181.
11. Hinton, G.E. Learning multiple layers of representation. *Trends in Cognitive Sciences 11*, 10 (Oct. 2007), 428–434.
12. Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. Fast read-out of object identity from macaque inferior temporal cortex. *Science 310*, 5749 (Nov. 4, 2005), 863–866.
13. Jhuang, H., Serre, T., Wolf, L., and Poggio, T. A biologically inspired system for action recognition. In *Proceedings of the 11th IEEE International Conference on Computer Vision* (Rio de Janeiro, Brazil, Oct. 14–20). IEEE Press, 2007.
14. Lee, T.S. and Mumford, D. Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America 20*, 7 (July 2003), 1434–1448.
15. Li, N. and DiCarlo, J.J. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science 321*, 5895 (Sept. 12, 2008), 1502–1507.
16. Logothetis, N.K., Pauls, J., and Poggio, T. Shape representation in the inferior temporal cortex of monkeys. *Current Biology 5*, 5 (May 1, 1995), 552–563.
17. Marko, H. and Giebel, H. Recognition of handwritten characters with a system of homogeneous layers. *Nachrichtentechnische Zeitschrift 23* (1970), 455–459.
18. Masquelier, T., Serre, T., Thorpe, S., and Poggio, T. *Learning Complex Cell Invariance from Natural Videos: A Plausibility Proof*. MIT Center for Biological & Computational Learning Paper #269/MIT-CSAIL-TR #2007-060, Cambridge, MA, 2007.
19. Mutch, J. and Lowe, D. Multiclass object recognition using sparse, localized features. In *Proceedings of the Computer Vision and Pattern Recognition* (New York, June 17–22, 2006).
20. Pinto, N., Cox, D.D., and DiCarlo, J.J. Why is real-world visual object recognition hard? *PLoS Computational Biology 4*, 1 (Jan. 1, 2008), e27.
21. Polk, T.A., Park, J.E., Smith, M.R., and Park, D.C. Nature versus nurture in ventral visual cortex: A functional magnetic resonance imaging study of twins. *Journal of Neuroscience 27*, 51 (2007), 13921–13925.
22. Rao, R.P. and Ballard, D.H. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience 2*, 1 (1999), 79–87.
23. Reynolds, J.H., Chelazzi, L., and Desimone, R. Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience 19*, 5 (Mar. 1, 1999), 1736–2753.
24. Riesenhuber, M. and Poggio, T. Hierarchical models of object recognition in cortex. *Nature Neuroscience 2*, 11 (1999), 1019–1025.
25. Rolls, E.T. and Deco, G. *Computational Neuroscience of Vision*. Oxford University Press, Oxford, U.K., 2002.
26. Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., and Poggio, T. A quantitative theory of immediate visual recognition. *Progress in Brain Research 165* (2007), 33–56.
27. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. Object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*, 3 (2007), 411–426.
28. Serre, T., Oliva, A., and Poggio, T. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences 104*, 15 (Apr. 10, 2007), 6424–6429.
29. Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., and Poggio, T. A *Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex*. MIT AI Memo 2005-036 / CBCL Memo 259, AI Memo 2005-036 / CBCL Memo 259 2005. Cambridge, MA, 2005.
30. Thorpe, S., Fize, D., and Marlot, C. Speed of processing in the human visual system. *Nature 381*, 6582 (1996), 520–522.
31. Yuille, A. and Kersten, D. Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Science 10*, 7 (July 2006), 301–308.

**Thomas Serre** (thomas_serre@brown.edu) is an assistant professor in the Department of Cognitive, Linguistic & Psychological Sciences at Brown University, Providence, RI.

**Tomaso Poggio** (tp@ai.mit.edu) is the Eugene McDermott Professor in the Department of Brain and Cognitive Sciences in the McGovern Institute for Brain Research at the Massachusetts Institute of Technology, Cambridge, MA.