# *A Neuromorphic Approach to Computer Vision*

**Citation:** Thomas Serre and Tomaso Poggio. 2010. A neuromorphic approach to computer vision. Commun. ACM 53, 10 (October 2010), 54-61

**As Published:** http://dx.doi.org/10.1145/1831407.1831425

**Publisher:** Association for Computing Machinery

**Persistent URL:** http://hdl.handle.net/1721.1/64483

**Version:** Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

# A neuromorphic approach to computer vision (Solicited Paper)

Thomas Serre
McGovern Institute for Brain Research
Department of Brain & Cognitive Sciences
Massachusetts Institute of Technology
77 Massachusetts Avenue,
Bldg 46-5155B
+1 (617) 253 0548
serre@mit.edu

Tomaso Poggio
McGovern Institute for Brain Research
Department of Brain & Cognitive Sciences
Massachusetts Institute of Technology
77 Massachusetts Avenue,
Bldg 46-5177B
+1 (617) 253 5230
tp@ai.mit.edu

## ABSTRACT

If physics was “the” science of the first half of last century, biology was certainly the science of the second half. Neuroscience is often mentioned as the focus of the present century. The field of neuroscience has indeed grown very rapidly over the last several years, spanning a broad range of approaches from molecular neurobiology to neuro-informatics and computational neuroscience. Computer science provided to biology powerful new data analysis tools which created bioinformatics and genomics: they made possible the sequencing of the human genome. In a similar way, computer science techniques are at the heart of brain imaging and other branches of neuroscience. Computers are critical for the Neurosciences, however, at a much deeper level: they represent the best metaphor for the central mystery of how the brain produces intelligent behavior and intelligence itself. They also provide experimental tools for performing experiments in information processing, effectively testing theories of the brain, in particular theories of aspects of intelligence such as sensory perception. The contribution of computer science to neuroscience happens at a variety of 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. Modern computational neuroscience models are no longer toy models: they are quantitatively detailed and at the same time, they are starting to compete with state-of-the-art computer vision systems. In fact we will argue in this review that in the next decades computational neuroscience may be a major source of new ideas and approaches in artificial intelligence.

## Keywords

Computational neuroscience, neurobiology, models, cortex, theory, computer vision, artificial intelligence

## 1. INTRODUCTION

Understanding the processing of information in our cortex is a significant part of understanding how the brain works and, in a sense, understanding intelligence itself. One of our most developed senses is vision. Primates can easily categorize images or parts of them, for instance as an office scene or as a face within that scene, and identify a specific object. Our visual capabilities are exceptional and despite decades of efforts in engineering, no computer algorithm has been able to match the level of performance of the primate visual system.

It has been argued that vision is a form of intelligence: it is suggestive that the sentence ‘*I see*’ is often used to mean ‘*I understand*’! Our visual cortex may serve as a proxy for the rest of the cortex and thus 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 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 information sciences making it possible to develop computer algorithms following the information processing principles used by biological organisms and honed by natural evolution.

The past fifty 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 and behavior. Some of these models already compete with state-of-the-art computer vision systems and are close to human level performance for specific visual tasks.

In this review, we will describe recent work in our group towards a theory of cortical visual processing. In contrast to other models that address the computations in any one given brain area (such as primary visual cortex) or attempt to explain a particular phenomenon (such as contrast adaptation or a specific visual illusion), we will describe a large-scale model that attempts to mimic the main information processing steps across multiple brain areas and millions of neuron-like units. We believe that a first step towards understanding cortical functions may take the form of a detailed, neurobiologically plausible model taking into account the connectivity, the biophysics and the physiology of cortex.

Models can provide a much-needed framework for summarizing and integrating existing data and for planning, coordinating and interpreting new experiments. Models can be powerful tools in basic research, integrating knowledge across several levels of analysis – from molecular to synaptic, cellular, systems and to complex visual behavior. Models, however, as we will discuss at the end of the paper, are limited in their explanatory power; ideally they should eventually lead to a deeper and more general theory.

We first argue about the role of the visual cortex and review some of the key computational principles underlying the processing of information during visual recognition. We then describe a computational neuroscience model – representative of a whole class of older models – that implements those principles. We also discuss some of the evidence in its favor. When tested with natural images the model is able to perform robust object recognition on par with then current computer vision systems and at the level of human performance for a specific class of rapid 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 may benefit directly from progress in neuroscience.

## 2. GOAL OF THE VISUAL SYSTEM

One key computational issue in object recognition[1] is the specificity-invariance trade-off. On the one hand, recognition must be able to finely discriminate between different objects or object classes (such as the faces illustrated in insets A and B of Figure 1). At the same time, recognition must be tolerant to object transformations such as scaling, translation, illumination, changes in viewpoint, clutter, as well as non-rigid transformations such as variations in shape within a class (for instance change of facial expression for the recognition of faces). Though the tolerance shown by our visual system is not *complete*, it is still significant.

A key challenge posed by the visual cortex is how well it deals with the *poverty of stimulus problem*: Primates can learn to recognize an object in quite different images from far fewer labeled examples than our present learning theory and learning algorithms predict. For instance, discriminative algorithms such as Support Vector Machines (SVMs) can learn a complex object recognition task from a few hundred labeled images. This is a small number compared with the apparent dimensionality of the problem (millions of pixels), but a child, or even a monkey, can apparently learn the same task from just a handful of examples. As an example of the prototypical problem in visual recognition, imagine that a (naïve) machine is shown one image of a given person and one image of another person. The system's task is to discriminate future images of these two people. The system did not see other images of these two people though it has seen many images of other people and other 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 very few) labeled examples?

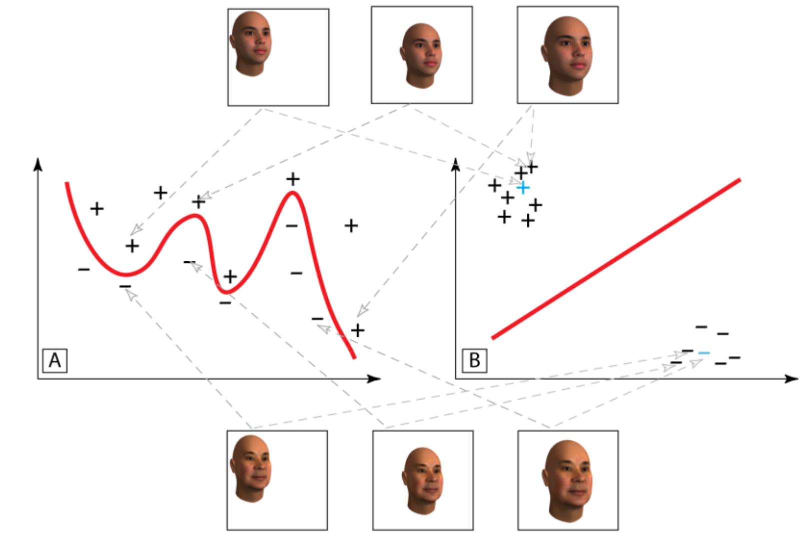


**Figure 1: The problem of *sample complexity*.** A hypothetical 2-dimensional (face) classification problem (red) line: One category is represented with "+" and the other with "–". Insets show 2D transformations (translation and scales) applied to examples from the two classes. Illustrated in panel (A) and (B) are two different representations for the same set of images. The representation in (B), which is tolerant with respect to the exact position and scale of the object within the image, leads to a simpler decision function (e.g., a linear classifier) and will require less training examples to achieve a similar level of performance thus lowering the *sample complexity* of the classification problem. In the limit, learning in panel (B) could be done with only two training examples (illustrated in blue).

For simplicity, imagine trying to build such classifier from the output of two cortical cells (as illustrated in Fig. 1). Here the response of these 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. Here 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[2]. In both cases it is possible to find a separation (the red lines indicate one such possible separation) between the two classes. In fact 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). In other words, 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 that respect, the two representations in panels (A) and (B) are not equal: The representation in panel (B) is far superior to the one in panel (A). With no prior assumption on the class of functions to be learned, the "simplest" classifier that can separate the data in panel (B) is much *simpler* than the "simplest" classifier that can separate the data in panel (A). The number of wiggles of the separation line gives a hand-wavy

[1] Within recognition, one can distinguish between identification and categorization. From the computational point of view, both of these tasks are classification tasks and represent two points in a spectrum of generalization levels.

[2] The receptive field of a neuron is the part of the visual field, which if properly stimulated with the right stimulus, may elicit a response from the neuron.

estimate of the complexity of a classifier, which is related to the number of parameters to be learned. The *sample complexity* of the problem derived from the invariant representation in panel (B) is much lower than that of the problem in panel (A). Learning to categorize the data-points in panel (B) will require far fewer training examples than in panel (A), and it may be done with as few as two examples. Thus the key problem in vision is what can be learned with a small number of examples and how.[3]

Our main argument is not that a low-level representation as provided from the retina would not be able to support robust object recognition. Indeed relatively good computer vision systems developed in the 90's were based on simple retina-like representations and on rather complex decision functions (such as Radial Basis Function (RBF) networks, etc). The main problem of these systems is that they required a prohibitively large number of training examples compared to humans.

More recent work in computer vision suggests that a hierarchical architecture may provide a better solution to this problem (see also [2] for a related argument). For instance Heisele et al. (see [3] for a recent review) designed a hierarchical system for the detection and recognition of faces. The approach is based on a hierarchy of "component experts" performing a local search for one facial component (e.g., an eye, a nose) over a range of positions and scales. Experimental evidence from [3] suggests that such hierarchical system based exclusively on linear (SVM) classifiers outperformed significantly a shallow architecture that tries to classify a face as a whole albeit relying on more complex kernels.

Here we suggest that 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 is transforming 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.

Such processing stages can be learned during development from temporal streams of natural images by exploiting the statistics of natural environments in two ways: Correlation over images provides *information-rich features* at various levels of complexity and sizes while correlations over time are 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 the efficient sharing of visual features between object categories and makes the learning of 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. Below we review evidence for this hierarchical architecture and the two mechanisms described above.

## 3. HIERARCHICAL ARCHITECTURE AND INVARIANT RECOGNITION

Several lines of evidence (both from human psychophysics and monkey electrophysiology studies) suggest that 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 about 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 in the past decades. One general class of computational models postulates that the hierarchical organization of the visual cortex is key to this process (see [4] for an alternative view-point). The processing of shape information in the visual cortex follows a series of stages, starting from the retina, 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 prefrontal cortex (PFC) involved in linking perception to memory and action (see [5] for references).

As one progresses along the ventral 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 combination of orientations) and complex objects and faces in higher visual areas such as IT. In parallel to this increase in the 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 of this increase in invariance properties, the receptive field size of neurons increases, from about one degree or less in V1 to several degrees in IT.

There is increasing evidence that IT, which has been critically linked with the monkey's ability to recognize objects, provides a representation of the image which facilitates recognition tolerant to image transformations. For instance, Logothetis and colleagues showed that monkeys could be trained to recognize paperclip-like wireframe objects at one specific location and scale [6]. After training, recordings in the IT cortex of these animals revealed some 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 for learning. More importantly, it was found that selective neurons also exhibited some range of invariance with respect to the exact position (between 2 and 4 degrees) and scale (around 2 octaves) of the stimulus – which was never presented before testing at these new positions and scales. More recently, work by Hung et al [7] showed that it was possible to train a (linear) classifier to robustly readout from a population of IT neurons, the category information of a briefly flashed stimulus. Furthermore it was shown that the classifier was able to generalize to a range of positions and scales (similar to Logothetis' data) that were never presented during the training of the classifier. This suggests that 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 need of object-specific learning (depending on the difficulty of the task).

[3] This is related to the point made by DiCarlo & Cox [1] about the main goal of the processing of information from the retina to higher visual areas to be "untangling object representations".

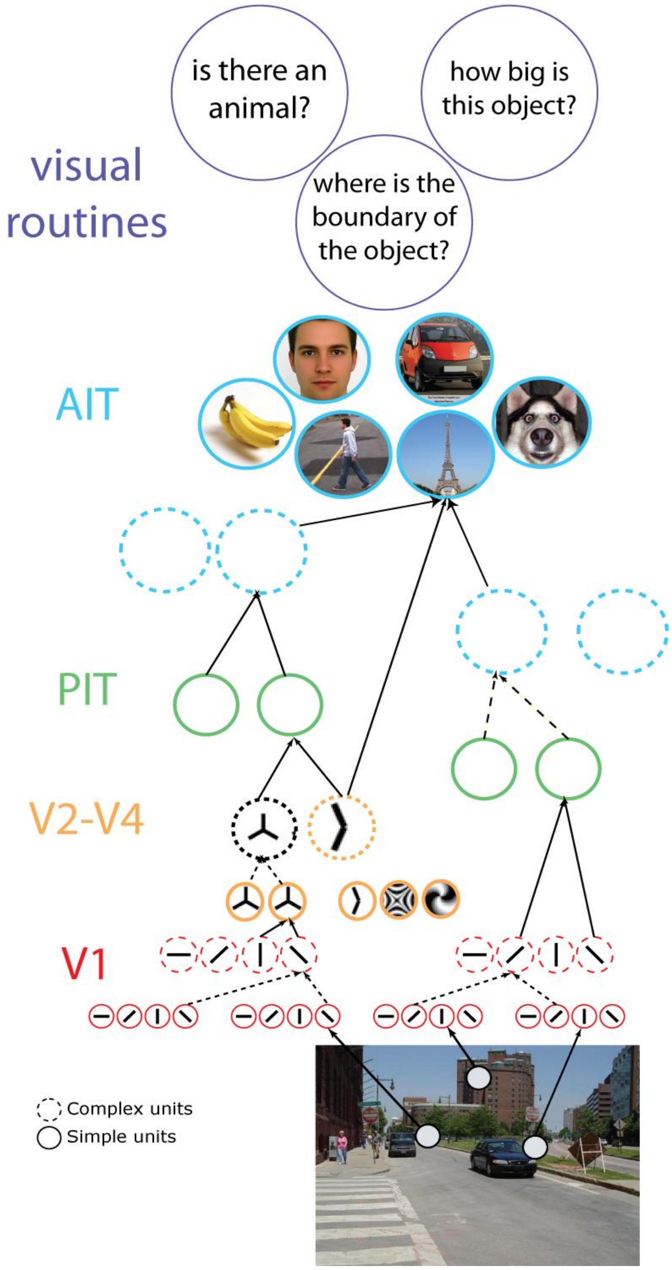


**Figure 2: Hierarchical feedforward models of the visual cortex (see text for details).**

# 4. COMPUTATIONAL MODELS OF OBJECT RECOGNITION IN CORTEX

We have developed [5, 8] – 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) that characterizes the properties of neurons along the object recognition processing hierarchy. In addition, the model is sufficient to mimic human performance in difficult visual recognition tasks [9] (while performing at least as well as most current computer vision systems [10]).

Feedforward hierarchical models have a long history starting with Marko & Giebel's homogeneous multi-layered architecture [11] in the 70's and later Fukushima's Neocognitron [12]. One of the key computational mechanisms in these, and other hierarchical models of visual processing, originates from the pioneering physiological studies and models of Hubel and Wiesel (see Box 1). The basic idea in these models is to build an increasingly complex and invariant object representation in a hierarchy of stages by progressively integrating (i.e., pooling) convergent inputs from lower levels. Building upon several existing neurobiological models [13-19], conceptual proposals [20-23] and computer vision systems [12, 24], we have been developing [5, 15] (see also [25, 26]) a similar computational theory (see Fig. 1) that attempts to quantitatively account for a host of recent anatomical and physiological data.

The feedforward hierarchical model of Figure 2 assumes two classes of functional units: ***simple*** and ***complex*** units. Simple units act as local template matching operators: They increase the complexity of the image representation by pooling over local afferent units with selectivities for different image-features (for instance edges at different orientations). Complex units on the other hand 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.

## 4.1 Learning and plasticity

How much of the organization of the visual cortex is influenced by development vs. genetics remains a matter of debate. A recent fMRI study [27] showed that the patterns of neural activity elicited by certain ecologically important classes of objects such as faces and places in monozygotic twins were significantly more similar than in dizygotic twins. These results thus suggest that genes may play a significant role in the way the visual cortex is wired to process certain object classes. At the same time, several electrophysiological studies have demonstrated learning and plasticity in the adult monkey (see for instance [28]). Learning is likely to be both faster and easier to elicit in higher visually responsive areas such as PFC or IT [28] than in lower areas.

This makes intuitive sense: For the visual system to remain stable, the time scale for learning should increase ascending the ventral stream[4]. In the model of Fig. 2, we assumed that unsupervised learning from V1 to IT happens during development in a sequence that starts with the lower areas. In reality, learning may continue throughout adulthood (certainly at the level of IT and perhaps in intermediate and lower areas).

[4] In the hierarchical model described in Figure 1, this process is done layer-by-layer starting from the bottom. This is similar to recent work by Hinton and colleagues [29] and quite different from the original neural networks that used back-propagation and learned simultaneously all layers at the same time. Our implementation (described in Box 1) includes the unsupervised learning of features from natural images but assumes the learning of position and scale tolerance which are thus hardwired in the model (but see [26] for an initial attempt).

### *4.1.1 Unsupervised learning in the ventral stream of the visual cortex*

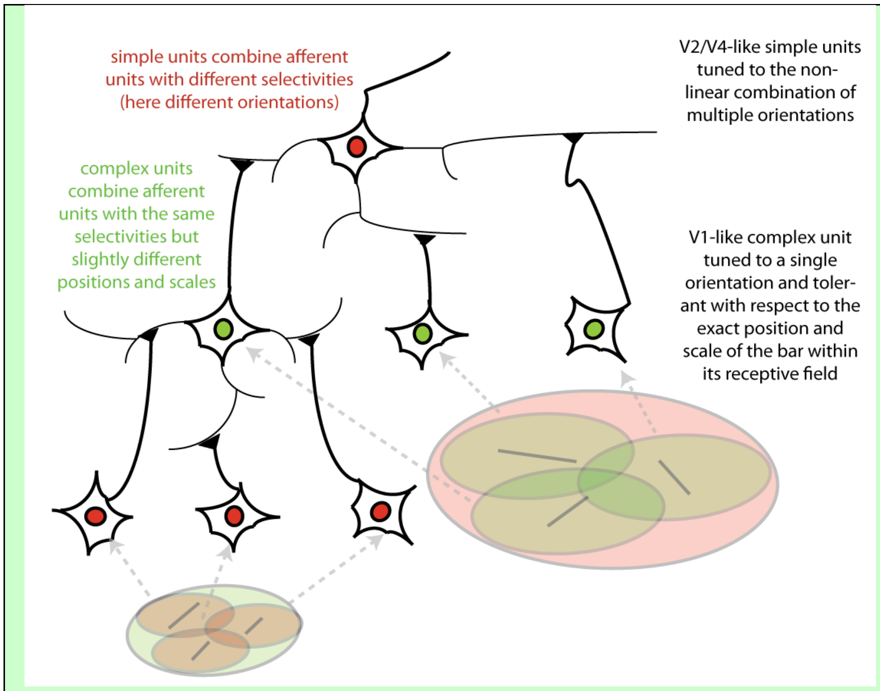
With the exception of the task-specific units at the top of the hierarchy (denoted ‘visual routines’), learning in the model described in Figure 2 remains unsupervised thus closely mimicking a developmental learning stage.

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. One of the key goals of the visual system may be to adapt to the statistics of its natural environment through visual experience and perhaps evolution. In the model of Figure 2, the selectivity of simple and complex units can be learned from natural video sequences (see Box 1 for details).

### *4.1.2 Supervised learning in higher areas*

After this initial developmental learning stage, learning of a new object category only requires training of task-specific circuits at the top of the ventral stream hierarchy. The ventral stream hierarchy thus provides 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 have to be learned anew for every 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).

In a default state (when no specific visual task is set) there may be a default routine running (perhaps the routine: *What is there?*). As an example of a simple routine consider a classifier, which 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 (using a perceptron-like learning rule).


simple units combine afferent units with different selectivities (here different orientations)
V2/V4-like simple units tuned to the non-linear combination of multiple orientations
complex units combine afferent units with the same selectivities but slightly different positions and scales
V1-like complex unit tuned to a single orientation and toler-ant with respect to the exact position and scale of the bar within its receptive field

**Box 1: Functional classes of cells and learning.**

***Simple and complex cells.*** Following their work on striate cortex [20], Hubel & Wiesel first described two classes of functional cells. Simple cells that respond best to bar-like (or edge-like) stimuli at a particular orientation, position and phase (i.e., white bar on a black background or dark bar on a white background) within their relatively small receptive fields. Complex cells, on the other hand, while also selective for bars, tend to have larger receptive fields (about twice as big) and exhibit some tolerance with respect to the exact position (and phase of the bar) within their receptive fields. Hubel & Wiesel described a way by which specific pooling mechanisms could explain the response properties of these cells. Simple-cell-like receptive fields could be obtained by pooling the activity of a small set of cells tuned to spots of lights (as observed in ganglion cells in the retina and the Lateral Geniculate Nucleus) aligned around a preferred axis of orientation (not shown on the figure). Similarly, position tolerance at the complex cell level (green color on the figure), could be obtained by pooling over afferent simple cells (at the level below) with the same preferred orientation but slightly different positions. Recent work has provided evidence for such selective pooling mechanisms in V1 [30]. Extending these ideas from primary visual cortex to higher areas of the visual cortex led to a class of models of object recognition, the feedforward hierarchical models (see [5] for a recent review). Illustrated at the top of the figure on the left is a V2-like simple cell obtained by combining several V1 complex cells tuned to bars at different orientations. Iterating these selective pooling mechanisms leads to a hierarchical architecture like the one described in Figure 2. Along the hierarchy, units become selective for increasingly complex stimuli and at the same time exhibit more and more invariance properties with respect to position (and scale).

***Learning of selectivity and invariance.*** In the model of Figure 1, simple units are selective for specific conjunctions of inputs (i.e., similar to an and-like operation). Their wiring thus corresponds to learning correlations between inputs at the same time-points (i.e., for simple cells in V1, the bar-like arrangements of LGN inputs, and beyond V1, more elaborate arrangements of bar-like subunits, etc). This corresponds to learning what combinations of features appear most frequently in images (i.e., which sets of inputs are consistently co-active) and to become selective to these patterns. Conversely the wiring of complex units may correspond to learning how to associate frequent transformations in time – such as translation and scale – of specific image features coded by afferent (simple) cells. The wiring of the complex units reflects learning of correlations across time (because of the object motion), e.g., for V1-like complex units, learning which afferent units with the same orientation and neighboring locations should be wired together because, often, such a pattern changes smoothly in time (under translation) [31].

## 5. IMMEDIATE RECOGNITION

An important aspect of the visual object recognition hierarchy (see Figure 2), i.e., the role of the anatomical back-projections abundantly present between almost all of the areas in visual cortex, remains a matter of debate. A commonly accepted hypothesis is that the basic processing of information is feedforward [32]. This is 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 monkey [7] show that the activity of small neuronal populations, over very short time intervals (as small as 12.5 ms) and only about 100 ms after stimulus onset, contains surprisingly accurate and robust information supporting a variety of recognition tasks. While this 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 visual cortex, aiming to explain ***immediate recognition***, the initial phase of recognition before eye movements and high-level processes take place.

### 5.1 Agreement with experimental data

Since it was originally developed [5, 15], the model of Fig. 2 has been able to explain a number of new experimental data. This includes data that were not used to derive or fit model parameters. The model seems to be qualitatively and quantitatively consistent with (and in some cases actually predicts, see [5]) several properties of subpopulations of cells in V1, V4, IT, and PFC as well as fMRI and psychophysical data (see Box 2 for a complete list of findings).

**Box 2: Summary of quantitative data that are compatible with the model described above.** Black corresponds to data that were used to derive the parameters of the model, red to data that are 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), IT (= inferotemporal cortex). Data from these areas correspond to monkey electrophysiology studies. LOC (=Lateral Occipital Complex) involves fMRI with humans; the Psych. 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. T. Serre, A. Oliva, T. Poggio, *Proc Natl Acad Sci U S A* **104**, 6424 (2007).
2. M. Riesenhuber *et al.*, *Proc Biol Sci* **271**, S448 (2004).
3. X. Jiang *et al.*, *Neuron* **50**, 159 (2006).
4. D. J. Freedman, M. Riesenhuber, T. Poggio, E. K. Miller, *J Neurosci* **23**, 5235 (2003).
5. M. Riesenhuber, T. Poggio, *Nature Neuroscience* **2**, 1019 (1999).
6. N. K. Logothetis, J. Pauls, T. Poggio, *Curr Biol* **5**, 552 (1995).
7. C. P. Hung, G. Kreiman, T. Poggio, J. J. DiCarlo, *Science* **310**, 863 (2005).
8. T. Serre *et al.*, MIT AI Memo 2005-036 / CBCL Memo 259 (2005).
9. T. Serre *et al.*, *Prog Brain Res* **165**, 33 (2007).
10. D. Zoccolan, M. Kouh, T. Poggio, J. J. DiCarlo, *J Neurosci* **27**, 12292 (2007).
11. T. J. Gawne, J. M. Martin, *J Neurophysiol* **88**, 1128 (2002).
12. J. H. Reynolds, L. Chelazzi, R. Desimone, *J Neurosci* **19**, 1736 (1999).
13. K. Taylor, S. Mandon, W. A. Freiwald, A. K. Kreiter, *Cereb Cortex* **15**, 1424 (2005).
14. A. Pasupathy, C. Connor, *Journal of Neurophysiology* **82**, 2490 (1999).
15. C. Cadieu *et al.*, *Journal of Neurophysiology* **98**, 1733 (2007).
16. J. L. Gallant *et al.*, *J Neurophysiol* **76**, 2718 (1996).
17. P. H. Schiller, B. L. Finlay, S. F. Volman, *J Neurophysiol* **39**, 1288 (1976).
18. D. H. Hubel, T. N. Wiesel, *J Physiol* **160**, 106 (1962).
19. R. L. De Valois, D. G. Albrecht, L. G. Thorell, *Vision Res* **22**, 545 (1982).
20. I. Lampl, D. Ferster, T. Poggio, M. Riesenhuber, *J Neurophysiol* **92**, 2704 (2004).

We recently compared the performance of this model and the performance of human observers in a rapid animal vs. non-animal recognition task [9] for which recognition is fast and cortical back-projections are possibly less relevant. Results indicate that the model predicts human performance quite well during such task suggesting that the model may therefore 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 amount of clutter. Interestingly the performance of both the model and human observers was highest (~90% correct for both human participants and the model) on images for which the amount of information is maximal and the amount of clutter minimal and decreases monotically as the amount of clutter in the image increases. This decrease in performance with increasing amount of clutter is likely to reflect a key limitation of this type of feedforward architectures. 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 [5] with neurophysiology data [33].

### 5.2 Application to Computer Vision

How does the model [5]) perform in real-world recognition tasks and how does it compare to state-of-the-art AI systems? Given the many specific biological constraints that the theory had to satisfy (e.g., using only biophysically plausible operations, receptive field sizes, range of invariances, etc) it was not clear how well the model implementation described above would perform in comparison to systems that have been heuristically engineered for these complex tasks.

At the time – about 5 years ago – we were surprised to find that the model is capable of recognizing well complex images (see [10]). The model performed 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 [10] and also the extension by Mutch & Lowe [25]). A related system with fewer layers, less invariance and more units has an even better recognition rate on the CalTech data set [34].

In parallel we also developed an automated system for the parsing of street scene images [10] based in part on the class of models described above. The system is able to recognize well seven different object categories (cars, bikes, skies, roads, buildings, trees) from natural images of street scenes despite very large variations in shape (e.g., trees in summer and winter, SUVs as well as compact cars under any view point).

An emerging application of computer vision is content-based recognition and search in videos. Again, neuroscience may suggest an avenue for approaching this problem. We have developed an initial model for the recognition of biological motion and actions from video sequences. The system is based on the organization of the dorsal stream of the visual cortex [35], which has been 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 that are very similar to those used in the model of the ventral stream described above but starts with spatio-temporal filters modeled after motion-sensitive cells in the primary visual cortex.

Recently we evaluated the performance of the system for the recognition of actions (both humans and animals) in real-world video sequences [35]. We found that the model of the dorsal stream competed with a state-of-the-art system (which itself outperforms many other systems) on all three datasets (see [35] for details). In addition we found that the learning in this model produces a large dictionary of optic-flow patterns, which seems to be consistent with the response properties of cells in the Medial Temporal (MT) area in response to both isolated gratings and plaids (i.e., 2 gratings superimposed).

## 6. CONCLUSION AND FUTURE DIRECTIONS

The demonstration that a model designed to mimic known anatomy and physiology of the visual system led to good performance with respect to computer vision benchmarks may suggest that neuroscience is on the verge of providing novel and useful paradigms to computer vision and perhaps to other areas of computer science. The model we described can obviously be modified and improved by taking into account new experimental data (for instance more detailed properties of specific visual areas such as V1 [36]), implementing several of its implicit assumptions such as the learning of invariances from sequences of natural images, taking into account additional sources of visual information such as binocular disparity and color and extending it to describe the dynamics of neural responses. The recognition performance of models of this general type can be improved by exploring the space of parameters (e.g., receptive field sizes, connectivity, etc.), for instance by using computer intensive iterations of a mutation-and-test cycle (Cox *et al.*, abstract #164 presented at Cosyne, 2008).

It is important however to realize the intrinsic limitations of the specific computational framework we have described here and why it is at best a first step towards understanding the visual cortex. First, from the anatomical and physiological point of view the class of feedforward models described here is incomplete, as it does not take into account 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 have to be taken into account in order to describe visual perception beyond the first 100-200 msec.

Given enough time, humans make eye movements to scan an image and performance in many object recognition tasks can increase significantly over that obtained during fast presentations. Extensions of the model to incorporate feedback are possible and under way [37]. We think that feedforward models may well turn out to be approximate descriptions of the first 100-200 msec of the processing required by more complex theories of vision, which are based on back-projections [38-44]. The computations involved in the initial phase are however non trivial and are essential for any scheme involving feedback to work. A second, related point is that normal visual perception is much more than classification as it involves interpreting and parsing visual scenes. In this sense again, the class of models we described is limited, since it deals with classification tasks only. Thus, more complex architectures are needed (see [8] 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 several aspects of monkey physiology and of human recognition, does not yield a good understanding of the computational principles of cortex and of their power. What is needed is a mathematical *theory* – to explain the hierarchical organization of the cortex.

## 7. ACKNOWLEDGMENTS

We would like to thank Jake Bouvrie as well as the referees for valuable comments on this manuscript.

## 8. REFERENCES

[1] DiCarlo, J. J. and Cox, D. D. Untangling invariant object recognition. Trends Cogn Sci, 11, 8 (Aug 2007), 333-341.

[2] Bengio, J. and Le Cun, Y. Scaling learning algorithms towards AI, 2007.

[3] Heisele, B., Serre, T. and Poggio, T. A component-based framework for face detection and identification. Int J Comput Vis, 74, 2 (Jan 1 2007), 167-181.

[4] Hegdé, H. and Felleman, D. J. Reappraising the functional implications of the primate visual anatomical hierarchy. The Neuroscientist, 13, 5 (2007), 416-421.

[5] 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 (2005).

[6] Logothetis, N. K., Pauls, J. and Poggio, T. Shape representation in the inferior temporal cortex of monkeys. Curr Biol, 5 (May 1 1995), 552-563.

[7] Hung, C. P., Kreiman, G., Poggio, T. and DiCarlo, J. J. Fast read-out of object identity from macaque inferior temporal cortex. Science, 310, (2005), 863-866.

[8] Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U. and Poggio, T. A quantitative theory of immediate visual recognition. Prog Brain Res, 165, (2007), 33-56.

[9] Serre, T., Oliva, A. and Poggio, T. A feedforward architecture accounts for rapid categorization. Proc Natl Acad Sci, 104, 15 (Apr 10 2007), 6424-6429.

[10] Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M. and Poggio, T. Object recognition with cortex-like mechanisms. IEEE TPAMI, 29, 3 (2007), 411-426.

[11] Marko, H. and Giebel, H. Recognition of handwritten characters with a system of homogeneous Layers. Nachrichtentechnische Zeitschrift, 23 (1970), 455-459.

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

[13] Wallis, G. and Rolls, E. T. A model of invariant recognition in the visual system. Prog Neurobiol, 51 (1997), 167-194.

[14] Mel, B. W. SEEMORE: Combining color, shape and texture histogramming in a neurally-inspired approach to visual object recognition. Neural Comp, 9, 4 (1997), 777--804.

[15] Riesenhuber, M. and Poggio, T. Hierarchical models of object recognition in cortex. Nature Neurosci, 2, 11 (1999), 1019-1025.

[16] Ullman, S., Vidal-Naquet, M. and Sali, E. Visual features of intermediate complexity and their use in classification. Nat Neurosci, 5, 7 (Jul 2002), 682-687.

[17] Thorpe, S. Ultra-Rapid Scene Categorization with a Wave of Spikes. In Proc of BMCV (2002).

[18] Amit, Y. and Mascaro, M. An integrated network for invariant visual detection and recognition. Vision Research, 43, 19 (2003), 2073-2088.

[19] Wersing, H. and Koerner, E. Learning optimized features for hierarchical models of invariant recognition. Neural Comp, 15, 7 (2003), 1559-1588.

[20] Hubel, D. H. and Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol, 160, (Jan 1962), 106-154.

[21] Perrett, D. and Oram, M. Neurophysiology of shape processing. Image Vision Comput, 11 (1993), 317-333.

[22] Hochstein, S. and Ahissar, M. View from the top: hierarchies and reverse hierarchies in the visual system. Neuron, 36, 5 (Dec 5 2002), 791-804.

[23] Biederman, I. Recognition-by-Components: A Theory of Human Image Understanding. Psych. Rev., 94 (1987), 115--147.

[24] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. Gradient-Based Learning Applied to Document Recognition. Proc. of the IEEE, 86, 11 (1998), 2278--2324.

[25] Mutch, J. and Lowe, D. Multiclass Object Recognition Using Sparse, Localized Features. In Proc of IEEE CVPR (2006).

[26] Masquelier, T., Serre, T., Thorpe, S. and Poggio, T. Learning complex cell invariance from natural videos: a plausibility proof. MIT-CSAIL-TR #2007-060 (2007).

[27] 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. J Neurosci, 27, 51 (2007), 13921-13925.

[28] Li, N. and DiCarlo, J. J. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science, 321, 5895 (Sep 12 2008), 1502-1507.

[29] Hinton, G. E. Learning multiple layers of representation. Trends Cogn Sci, 11, 10 (Oct 2007), 428-434.

[30] Rust, N., Schwartz, O., Simoncelli, E. P. and Movshon, J. A. Spatiotemporal elements of macaque V1 receptive fields. Neuron, 46, 6 (Jun 16 2005), 945-956.

[31] Foldiak, P. Learning invariance from transformation sequences. Neural Comp, 3, (1991), 194-200.

[32] Thorpe, S., Fize, D. and Marlot, C. Speed of processing in the human visual system. Nature, 381, 6582 (1996), 520-522.

[33] Reynolds, J. H., Chelazzi, L. and Desimone, R. Competitive mechanisms subserve attention in macaque areas V2 and V4. J Neurosci, 19, 5 (Mar 1 1999), 1736-1753.

[34] Pinto, N., Cox, D. D. and DiCarlo, J. J. Why is real-world visual object recognition hard? PLoS Comp Biol, 4, 1 (2008).

[35] Jhuang, H., Serre, T., Wolf, L. and Poggio, T. A Biologically Inspired System for Action Recognition. In Proc of IEEE ICCV (2007).

[36] Rolls, E. T. and Deco, G. Computational Neuroscience of Vision. Oxford University Press, Oxford, 2002.

[37] Chikkerur, S., Tan, C., Serre, T. and Poggio, T. An integrated model of visual attention using shape-based features. MIT-CSAIL-TR-2009-029 (2009).

[38] Rao, R. P. and Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neurosc., 2, 1 (1999), 79-87.

[39] Lee, T. S. and Mumford, D. Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am A, 20, 7 (Jul 2003), 1434-1448.

[40] Dean, T. A Computational Model of the Cerebral Cortex. In Proc of AAAI (2005).

[41] George, D. and Hawkins, J. A hierarchical Bayesian model of invariant pattern recognition in the visual cortex. In Proceedings of IJCNN (2005).

[42] Yuille, A. and Kersten, D. Vision as Bayesian inference: analysis by synthesis? Trends Cogn Sci, 10, 7 (Jul 2006), 301-308.

[43] Epshtein, B., Lifshitz, I. and Ullman, S. Image interpretation by a single bottom-up top-down cycle. Proc Natl Acad Sci (2008).

[44] Grossberg, S. Towards a unified theory of neocortex: Laminar cortical circuits for vision and cognition. Prog Brain Res, 165 (2007), 79-104.