The "Standard Model" of object recognition in cortex

Object recognition in cortex is thought to be mediated by the ventral visual pathway [Ungerleider 94] running from primary visual cortex, V1, over extrastriate visual areas V2 and V4 to inferotemporal cortex, IT. Based on physiological experiments in monkeys, IT has been postulated to play a central role in object recognition. IT in turn is a major source of input to PFC, "the center of cognitive control" [Miller 00] involved in linking perception to memory.

Over the last decades, several physiological studies in non-human primates have established a core of basic facts about cortical mechanisms of recognition that seem to be widely accepted and that confirm and refine older data from neuropsychology. A brief summary of this consensus knowledge begins with the groundbreaking work of Hubel and Wiesel first in the cat [Hubel 62,65] and then in the macaque [Hubel 68]. Starting from *simple cells* in primary visual cortex, V1, with small receptive fields that respond preferably to oriented bars, neurons along the ventral stream [Perrett 93,Tanaka 96,Logothetis 96] show an increase in receptive field size as well as in the complexity of their preferred stimuli [Kobatake 94]. At the top of the ventral stream, in anterior inferotemporal cortex (AIT), cells are tuned to complex stimuli such as faces [Gross 72,Desimone 84,Desimone 91,Perrett 92]. A hallmark of these IT cells is the robustness of their firing to stimulus transformations such as scale and position changes [Tanaka 96,Logothetis 96,Logothetis 95,Perrett 93]. In addition, as other studies have shown [Perrett 93,Booth 98,Logothetis 95,Hietanen 92], most neurons show specificity for a certain object view or lighting condition.

A comment about the architecture is important: In its basic, initial operation — akin to "immediate recognition" — the hierarchy is likely to be mainly feedforward (though local feedback loops almost certainly have key roles) [Perrett 93]. ERP data [Thorpe 96] have shown that the process of object recognition appears to take remarkably little time, on the order of the latency of the ventral visual stream [Perrett 92], adding to earlier psychophysical studies using a rapid serial visual presentation (RSVP) paradigm [Potter 75,Intraub 81] that have found that subjects were still able to process images when they were presented as rapidly as 8/s.

In summary, the accumulated evidence points to six mostly accepted properties of the ventral stream architecture:

- A hierarchical build-up of invariances first to position and scale and then to viewpoint and more complex transformations requiring the interpolation between several different object views;
- in parallel, an increasing size of the receptive fields;
- an increasing complexity of the optimal stimuli for the neurons;
- a basic feedforward processing of information (for "immediate" recognition tasks);
- plasticity and learning probably at all stages and certainly at the level of IT;
- learning specific to an individual object is not required for scale and position invariance (over a restricted range).

These basic facts lead to a *Standard Model*, likely to represent the simplest class of models reflecting the known anatomical and biological constraints. It represents in its basic architecture the average belief — often implicit — of many visual physiologists. In this sense it is definitely not "our" model. The broad form of the model is suggested by the basic facts; we have made it quantitative, and thereby predictive (through computer simulations).

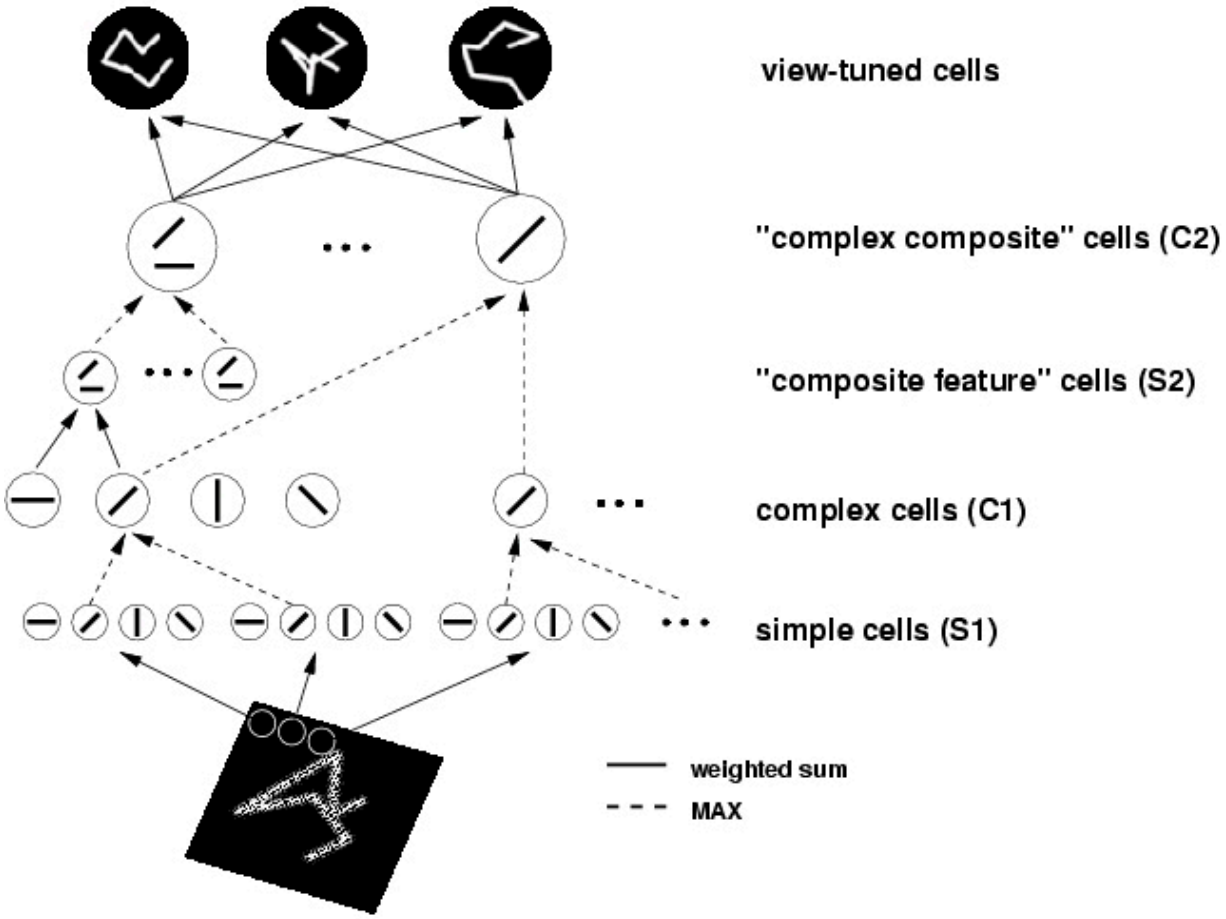
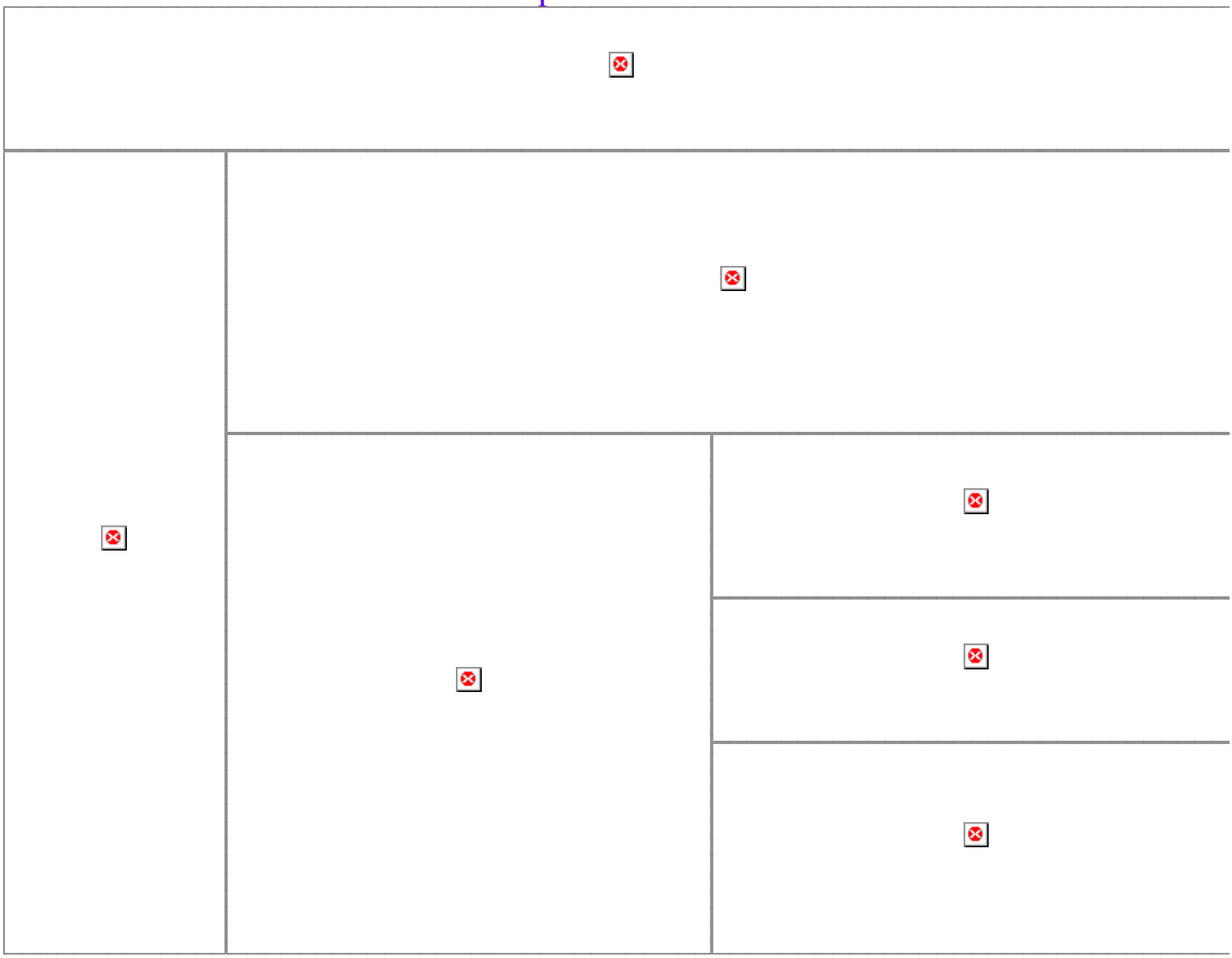


Figure 1

The model reflects the general organization of visual cortex in a series of layers from V1 to IT to PFC. From the point of view of invariance properties, it consists of a sequence of two main modules based on two key ideas. The first module, shown schematically in Figure 1, leads to model units showing the same scale and position invariance properties as the view-tuned IT neurons of [Logothetis 95], using the same stimuli. This is not an independent prediction since the model parameters were chosen to fit Logothetis' data. It is, however, not obvious that an hierarchical architecture using plausible neural mechanisms could account for the measured invariance *and* selectivity. Computationally, this is accomplished by a scheme that can be best explained by taking striate complex cells as an example: invariance to changes in the position of an optimal stimulus (within a range) is obtained in the model by means of a *maximum operation (max)* performed on the simple cell inputs to the complex cells, where the strongest input determines the cell's output. Simple cell afferents to a complex cell are assumed to have the same preferred orientation with their receptive fields located at different positions. Taking the maximum over the simple cell afferent inputs provides position invariance while preserving feature specificity. The key idea is that the step of filtering followed by a max operation is equivalent to a powerful signal processing technique: select the peak of the correlation between the signal and a given matched filter, where the correlation is either over position or scale. The model alternates layers of units combining simple filters into more complex ones — to increase pattern selectivity — with layers based on the max operation — to build invariance to position and scale while preserving pattern selectivity.

Figure 2: Click on the model to see what different projects are being done with that part of HMAX!



In the second part of the architecture, shown in Figure 2, learning from multiple examples, i.e. different view-tuned neurons, leads to view-invariant units as well as to neural circuits performing specific tasks. The key idea here is that interpolation and generalization can be obtained by simple networks, similar to Gaussian Radial Basis Function networks [Poggio 90] — that learn from a set of examples, that is input-output pairs. In this case, inputs are views and the outputs are the parameters of interest such as the label of the object or its pose or expression (for a face). The Gaussian Radial Basis Function (GRBF) network has a hidden unit for each example view, broadly tuned to the features of an example image (see also [deBeeck 01]). The weights from the hidden units to the output are learned from the set of examples, that is input-output pairs. In principle two networks sharing the same hidden units but with different weights (from the hidden units to the output unit), could be trained to perform different tasks such as pose estimation or view-invariant recognition. Depending just on the set of training examples, learning networks of this type can learn to categorize across exemplars of a class [Riesenhuber AI Memo 00] as well as to identify an object across different illuminations and different viewpoints. The demonstration [Poggio 90] that a view-based GRBF model could achieve view-invariant object recognition in fact motivated psychophysical experiments [Buelthoff 92,Gauthier 97]. In turn the psychophysics provided strong support for the view-based hypothesis against alternative theories (for a review see [Tarr 98])

and, together with the model, triggered the physiological work of [Logothetis 95].

Thus the two key ideas in the model are 1) the max operation to provide invariance at several steps of the hierarchy and 2) the RBF-like learning network to learn a specific task, based on a set of cells tuned to example views.

HMAX home

Last modified: Dec 2002 by uk
Last modified: Dec 2003 by er