Models of visual cortex

Tomaso Poggio and Thomas Serre (2013), Scholarpedia, 8(4):3516. doi:10.4249/scholarpedia.3516 revision #149958 [link to/cite this article]

- **Tomaso Poggio**, MIT, Cambridge, MA
- **Thomas Serre**, Brown University, Providence, RI

Computational modeling has become a central approach in vision research. Recent developments in brain and cognitive sciences, machine learning, and computer vision have provided a wealth of new tools to study the computational mechanisms underlying primate vision. Models of the visual cortex provide the much-needed framework for summarizing and integrating existing data and for planning, coordinating, and interpreting new experiments. Models provide the glue for integrating knowledge across levels of analysis – from the level of neural circuits and networks to systems and, ultimately, behavior.

This article provides a brief overview of some representative computational models along with a description of the visual functions and operations that these models aim to explain. We first briefly discuss the computational approach to vision before reviewing computational models. In the process, we unfortunately had to omit many important models for brevity. Additional suggested readings are provided at the end of this article for the interested reader. *Disclaimer:* There also exist many computational models of vision that are not models of the visual cortex and will be excluded from the present review. These include, for instance, extensive work on the fly, the beetle, the rabbit, and others, as well as a large body of work from the past twenty years of research in computer vision.

Contents

The computational approach to vision

The computational approach, as stated by Marr & Poggio (1977), regards the visual system as an information processor, which performs computations on internal symbolic representations of visual information. As a consequence, one should distinguish between the meaning of symbols and operations that are carried out on them on the one hand, and the physical manifestation of these symbols on the other hand (Note 1). The understanding of a complex system such as the visual system needs to be considered at different levels of analysis:

1. Computations: What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?
2. Algorithm: How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the underlying algorithm for the transformation?
3. Hardware: How can the representation and algorithm be realized?

Marr (1982) notoriously emphasized that explanations at different levels are largely independent of each other; a software engineer does not need to know the hardware in any great detail. Today, this message has been somehow tempered, and computational neuroscience models typically span multiple levels of analysis from the computational to the hardware.

A classical distinction between modeling approaches is the bottom-up vs. top-down dichotomy (Note 2). The Blue Brain project (Markram 2006) is a noticeable example of the bottom-up approach (but several similar programs exist). Electrophysiology recordings are carried to try to reverse-engineer the visual system by systematically estimating the biophysical properties of neural circuits. The hope is to then be able to reconstruct the brain piece by piece and that, from the collective behavior of millions of artificial neurons, intelligent behavior will emerge. The top-down approach corresponds to forward engineering, which is the general goal of cognitive science, computer vision, machine learning, and artificial intelligence. Typically, this is done by building formal theories of visual perception in terms of computations and constraints that apply to an organism before building biologically-plausible implementations of these theories down to the level of neural circuits. The normative models mentioned below are examples of this top-down approach.

The notion of biological plausibility for models of the visual cortex has remained elusive. For instance, Ullman (1979) suggested that biologically-plausible computations should exploit the inherent parallelism of the cortex as well as its apparent uniformity (a somewhat more modern version of this is the idea of canonical circuits (Douglas & Martin 1991)). Models are by definition simplifications of reality, thus none of the models listed below are fully "realistic" (see Sturm & Konig 2001 for a discussion). At the level of explaining visual functions, only some of the models attempt to relate to actual cortical structures and specific areas and almost never go so far as neural circuits, like cortical layers or specific neurons or interneurons, for instance. We refer the reader to Koch's *Biophysics of computation* (Koch 1999) for a tentative list of mathematical operations that are theoretically implementable in neural hardware. While not a necessary condition, a model that uses operations from this list can, in principle, be considered biologically plausible.

Spiking vs. non-spiking is an important characteristic that is often used to distinguish between computational models. Non-spiking models (often abusively called rate-based) typically rely on rather simplified neural circuits and assume that neural activity or firing rates are encoded by analog values either from single cells over "long" time windows (i.e., >100 ms) or from small populations of cells (e.g., cortical columns) over shorter time windows (i.e., ~10 ms). McCulloch & Pitts' artificial neuron (McCulloch & Pitts 1943) and Rosenblatt's perceptron (Rosenblatt 1958) are typical examples of non-spiking models. Progress in hardware is now enabling the simulation of large-scale computational models with millions of units. Similarly, the biophysical realism of models has been improving whereby "point" neurons, which are highly simplified spiking models (e.g., SpikeNet (Thorpe 2002), are now being replaced with more detailed multi-compartment models that take into account the 3D layout of neurons.

At the other extreme, normative models do not try to mimic the detailed operation underlying visual processes. Such models typically assume that our visual cortex has evolved to become optimally adapted to the statistical structure of our visual world. The aim of these models is to provide an understanding the visual cortex on the basis of optimality; starting with a theoretical formulation of the problems and constraints of visual functions, normative models propose an answer for how visual tasks should be solved optimally given these constraints. Prominent examples of

the normative approach include sparse coding models of the primary visual cortex and visual statistics (see Simoncelli & Olshausen 2001), ideal observer models (Geisler 2011), and Bayesian inference models of visual perception (Yuille & Kersten 2006).

In this article, we distinguish between *models of circuits and operations* and *models of visual functions*. Models of circuits and operations implement generic operations, such as gain control or normalization, and focus on the hardware and algorithmic levels of analysis. Such models constitute the building blocks for many models of the visual functions, such as motion computation or object recognition, which tend to focus on the computational and algorithmic levels of analysis. There is a certain tension between these two types of models, and very few models combine both. Both levels of modeling are in fact critically important, and they will eventually merge (as happened in physics, for example, with thermodynamics and statistical mechanics). A notable exception is the research program of Steve Grossberg (http://en.wikipedia.org/wiki/Stephen_Grossberg) , which has produced many models at all levels, from biophysics to system and beyond. In this article, we will review computational models of visual recognition in more depth because they provide a case in point to illustrate how computational models may help bridge the gap between multiple levels of analysis. Last, we discuss a key limitation from most existing models, which assume that core visual processing proceeds in a hierarchy of bottom-up feedforward processing and do not account either for the abundant back-projections found in the cortex nor human performance in visual tasks beyond fast visual processing.

Models of operations, circuits and learning

Basic operations and circuits

Many models developed today are at the level of circuits and microcircuits, some with quantitative details in terms of underlying biophysics of synapses and neurons. Typically, these models provide plausible hardware implementations for various computations that may be used and shared by several visual functions. Again, only a partial list is possible here. The first model of circuits in the visual cortex dates back to Hubel & Wiesel (1962), who first described a qualitative model of simple and complex cells in the primary visual cortex (V1) of cats (see Figure for details). The model suggests how selective pooling mechanisms could explain some of the key operations thought to take place in the primary visual cortex. For instance, the model suggests that the orientation selectivity of simple cells could be obtained via selective pooling mechanisms over afferent neurons with center-surround like receptive field organizations that are aligned along a preferred axis of orientation. Similarly, the tolerance to position of the complex cells could be obtained by considering afferent cells with the same preferred orientation but with receptive fields at slightly different positions. Fifty years after it was originally proposed, this model continues to be a key element in both machine and biological models of object recognition.

A refinement of Hubel & Wiesel's original model includes the energy model by Adelson & Bergen (1985) to account for the phase-tolerance of cortical complex cells. At the algorithmic level, the model is based on the summation of quadrature pairs of filters (i.e., simple cells) with the same orientation to build complex cells that are invariant to changes in the phase of the stimulus (and contrast reversal). At the same time, biophysical implementations have also been proposed for some of the model key components, such as the approximate squaring (Miller et al 2002). A different operation – gain control – is addressed by the divisive normalization model (Carandini and Heeger 1994), which suggests that divisive (so-called shunting) inhibition may play a key role in explaining the contrast dependent, sigmoid-like response profiles of many V1 neurons. Divisive normalization is also likely to constitute an important part in explaining the motion selectivity of cells in V1 and MT (Rust et al. 2005) as well as attentional mechanisms (Reynolds & Heeger 2009). It was introduced earlier as the basic operation to account for motion detection (Torre & Poggio 1978) and to describe quantitatively the detection of motion discontinuities in the fly (Poggio et al 1981).

Still, another operation common to many neurons in the visual cortex is the bell-shape tuning operation corresponding to a Gaussian-shaped response tuned to a specific, optimal pattern of activation from the presynaptic inputs (Poggio & Bizzi 2004), as induced, for instance, by a bar of a specific orientation for simple cells in V1 or by a face for neurons in specific patches of IT (Tsao et al. 2008). Other interesting models include 'winner-take-all' circuits (Yuille 1988; Rousselet et al. 2003). A close cousin of the 'winner-take-all' circuit is the softmax operation, which assumes the selection and transmission of the most active response among a set of neural inputs (Nowlan & Sejnowski 1995; Riesenhuber & Poggio 1999). It has been shown that such max-like operation successfully captures the response properties of a subset of the complex cells (Finn & Ferster 2007) in the primary visual cortex together with mechanisms based on the energy model by Adelson & Bergen (1985) described above.

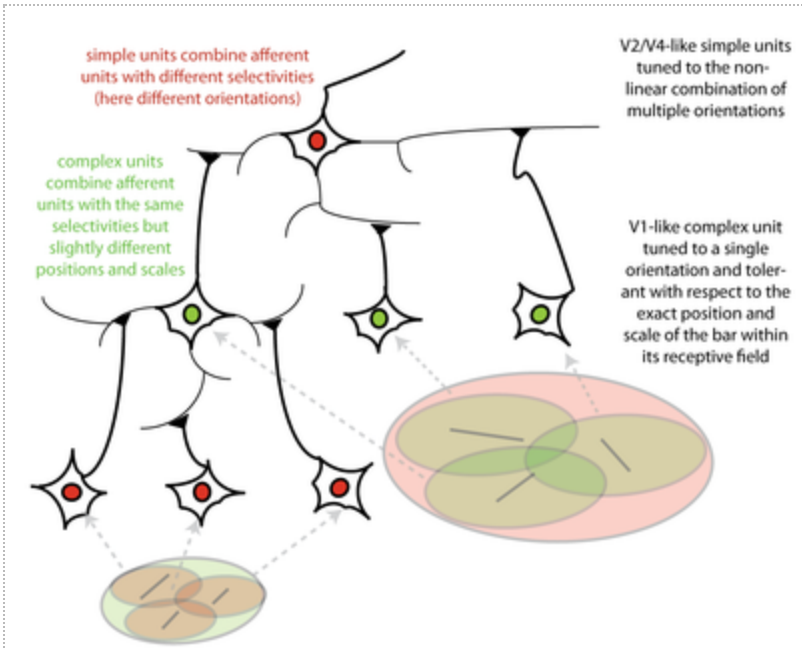

Figure 1: Hubel & Wiesel model of simple and complex cells. Hubel & Wiesel first described two classes of functional cells in the primary visual cortex: simple cells that respond best to bar-like (or edge-like) stimuli at a particular orientation, position, phase (i.e., bar-like vs. edge-like), and polarity (white bar on a black background or dark bar on a white background) within their relatively small receptive fields. Complex cells, while also selective for oriented bars and edges, tend to have larger receptive fields (about twice as large) and exhibit some tolerance with respect to the exact position (and phase of the bar) within their receptive fields. Hubel & Wiesel described a way where 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 afferent (input) 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 (Rust et al. 2005). 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 Serre et al. 2007) for a recent review and below). 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 .

Interestingly, all of these distinct neural operations – and in particular the energy model, the tuning, and the softmax operation described above – may be computed by a similar circuitry, involving divisive normalization and biophysically plausible polynomial nonlinearities, for different parameter values within the circuit (Kouh & Poggio 2008). Thorpe and colleagues have described a related circuit for template matching based on the timing of the arrival of spikes from a group of neurons (Thorpe, 2002). In fact, there is now limited evidence for such type of encoding in somatosensory areas (see (VanRullen 2005) for a review). Recent advances in available technologies for dissecting the detailed circuitry of the cortex will soon allow neuroscientists to study neural circuits at an unprecedented level of detail and test detailed models of columnar and laminar organization (i.e., the arrangement of cortical neurons into six distinct layers) as proposed by the LAMINART model (see (Grossberg 2007) for a review).

A cortical circuit that has also emerged from a combination of anatomical and physiological studies was proposed earlier as a canonical circuit for the cortex by (Douglas & Martin 1991). The circuit explains the intracellular responses to pulse stimulation in V1 in terms of the interactions between basic populations of neurons. There also exist models of oscillations and synchronization (see Scholarpedia articles on Binding by synchrony and Synfire chains). There is a very large literature on the latter subject – often related to computational models of attention (e.g., Niebur & Koch 1994; Borgers and Kopell 2003; Tiesinga and Sejnowski 2004; see corresponding Scholarpedia article on oscillatory-models of attention). Several groups have recently developed detailed biophysical models with spiking neurons and conductance-based synapses of the attentional effects in V4 neurons, which go beyond previous phenomenological models (Reynolds et al, 1999). For instance, (Buia & Tiesinga 2008) described a model that includes populations of spiking excitatory cells and two types of interneurons that can predict different types of oscillatory behaviors for spatial vs. feature-based attention.

Models of learning and development

Learning is arguably the key to understanding intelligence (Poggio & Smale 2003). One of the most striking features of the cortex is its ability to wire itself. Understanding how the visual cortex wires up through development and how plasticity refines connections into adulthood is likely to give necessary constraints to computational models of visual processing. An active area of research concerns the learning of invariances at the level of complex cells (see Figure) in the visual cortex such as invariance to 2D transformations (e.g., translation and scale) as well as invariance to view-point (see also invariance learning and slow feature analysis). In the context of the simple and complex cells described above, Foldiak (1991) proposed how both simple and complex cells could learn to pool over the right kinds of afferent units; simple cells 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 cells 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).

Further, several models of self-organization of brain function based on Kohonen networks have been described as plausible mechanisms for the development of cortical columns, ocular dominance, direction selectivity, spatial frequency selectivity, disparity and even color. These models are based on two key operations; a competitive neural circuit that implements a winner-take-all function as described above and a simple Hebbian-like learning mechanism similar to the one described above for simple cell plasticity.

Models of visual functions

Early vision

Many of our perceptual abilities in vision have been extensively studied in psychophysics, and corresponding algorithms usually exist in computer vision. Many of the models that have been developed do not have a direct correspondence with the visual cortex and hence should be considered at the computational level. Such models have played a key role in further motivating the development of models at the algorithmic level that take into account the anatomy and the physiology of cortex. A representative example of this tension is dispartity, which correspond to our

ability to recover 3D information by combining two (or more) 2D views from each one of the two eyes. Some of the first computational models of visual cortex were about stereopsis (Marr & Poggio 1976). For instance, the first model of stereo disparity by Marr & Poggio (1976) is not per se a model of cortex since it mainly spells out some of the constraints to be used to solve the matching problem of stereo and an algorithm. It can, however, be mapped out to circuits of neurons that correspond to cortical disparity tuned cells. The second stereo model by (Marr & Poggio 1979) was a model of the visual cortex in the sense that it took into account physiology data about cortical neurons and made specific predictions about them. Still it lacked much of the details that one would expect from a model of cortex today. In the meantime, the disparity-specific responses of simple disparity tuned neurons in V1 have been described by an energy model (Qian 1994) based on local, feedforward interactions (Read et al 2002). Little progress has been made on global properties of stereopsis and their representation at the neural level. This is an area where computational efforts have been lacking.

Since then, a tremendous amount of data has been collected about the visual cortex and a large number of models of the visual cortex have been developed. A number of models for early vision have been described (mostly in the eighties, following the work of Marr, Poggio, Ullman, Horn, Grimson, Richards, Winston, Ballard, Koch, Hildreth and others). These include models of edge detection (Marr 1979), spatio-temporal interpolation and approximation, computation of optical flow and direction selectivity (Ullman 1979; Marr 1981), computation of lightness and albedo, shape form contours, shape form texture, shape from shading, binocular stereo matching (Marr & Poggio 1976), structure from motion, structure from stereo, surface reconstruction (Grimson 1982) and filling-in (Ullman 1976), and computation of surface color (Barrow & Tenebaum 1981; Marr 1982; Hurlbert 1989).

Several models exist for explaining the orientation tuning of simple cells in the primary visual cortex (V1). Various forms of feedforward models and recurrent models are the two major examples. Though it is still unclear which model best fits the data, it is likely that both feedforward input geometry from the LGN -– as originally proposed by Hubel and Wiesel -– and recurrent intracortical circuits are involved in shaping the selectivity of simple cells. It is also possible that a version of the feedforward model is well suited for explaining orientation tuning in simple cells, whereas recurrent models involving recurrent normalization are suited to account for complex cells (Teich & Qian 2006).

The analysis of motion is also fairly well studied. This includes problems such as the analysis of motion information in early visual areas, such as MT as well as the recognition of biological motion (e.g., object/person moving left or right, etc.). A number of models (Nowlan and Sejnowski 1995; Simoncelli & Heeger 1996) have quantitatively described the properties of motion sensitive neurons in MT, a visual area involved in motion processing. In one of the models (Rust, Mante et al. 2006), which is itself the evolution of previous models of MT, the computation is performed in two stages, corresponding to neurons in cortical areas V1 and MT. Each stage computes a weighted linear sum of inputs followed by rectification and divisive normalization. The output of the model corresponds to the steady-state firing rates of a population of MT neurons, which form a distributed representation (population encoding) of image velocity for each local spatial region of the visual stimulus. The model accounts for a wide range of physiological data.

Invariant recognition

The most studied visual function is probably object recognition, which reflects our ability to assign a label or meaning to an image of an object irrespective of the precise size, position, illumination or context and clutter. The main computational problem in object recognition is achieving invariance while preserving selectivity (Riesenhuber & Poggio 1999); cells found in IT are typically tuned to views of complex objects such as a faces (Tsao et al 2008)-- they discharge strongly to a face but very little or not at all to other objects. A hallmark of these cells is the robustness of their responses to stimulus transformations such as scale and position changes. This finding presents

an interesting question: how could these cells respond differently to similar stimuli (for instance, two different faces) that activate the retinal photoreceptors in similar ways, but respond consistently to scaled and translated versions of the preferred stimulus, which produce very different activation patterns on the retina?

It has been postulated that the goal of the ventral stream of the visual cortex is to achieve an optimal tradeoff between selectivity and invariance via a hierarchy of processing stages whereby neurons at higher and higher levels exhibit an increasing degree of invariance to image transformations such as translations and scale changes. Many of the models are feedforward (they allow for local recurrent circuits to implement key operations such as the ones described above within areas but not between areas) and are restricted to describe the first ~100 ms or so of information processing after an image is flashed on the retina. Of course, a model of vertebrate vision must take into account multiple fixations, image sequences, as well as top-down signals, attentional effects and the structures mediating them (e.g., the extensive back-projections present throughout cortex), but feedforward hierarchical models assume that the effect of cortical feedback will be seen after some time (>100 ms).

Most of the models of visual functions involve a single visual area or a small number of them. There are very few examples of system-level computational models that consider most visual areas, and all of these models are models of visual recognition. It is natural, then, to consider object recognition as the best example for hierarchical models that reflect one of the most obvious features of cortex – the hierarchy of areas from V1 to IT. Feedforward hierarchical models have a long history, beginning in the 1970s with Marko & Giebel (1970)'s homogeneous multilayered architecture and later Fukushima's Neocognitron. One of their key computational mechanisms originates from the pioneering physiological studies and models of Hubel & Wiesel (see Figure 1). 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. Since then, many models have been proposed (see (Serre & Poggio, 2010) for a recent review) which extend the classical simple-to-complex cells model by Hubel & Wiesel to extrastriate areas and have been shown to account for a host of experimental data. Such models assume two functional classes of simple and complex cells with specific predictions about their respective wiring and resulting functionalities. Figure 2 shows the HMAX model originally proposed by Riesenhuber & Poggio (1999) and later extended by Serre et al (2007). Closely related hierarchical architectures include VisNet, hierarchical versions of normative approaches such as slow feature analysis, convolutional networks and other hierarchical architectures (Wersing & Konig, 2003, Ullman 2007, Masquelier & Thorpe 2007, Pinto et al. 2008).

Now, why hierarchies? The answer - for models in the Hubel & Wiesel spirit - is that the hierarchy may provide a solution to the invariance-selectivity trade-off problem by decomposing a complex task such as invariant object recognition in a hierarchy of simpler ones (at each stage of processing). Hierarchical organization in cortex is not limited to the visual pathways, and thus a more general explanation may be needed. Interestingly, from the point of view of classical learning theory (Poggio & Smale 2003), there is no need for architectures with more than three layers. So, why hierarchies? There may be reasons of efficiency, such as the efficient use of computational resources. For instance, the lowest levels of the hierarchy may represent a dictionary of features that can be shared across multiple classification tasks.

There may also be the more fundamental issue of sample complexity, the number of training examples required for good generalization. An obvious difference between the best classifiers derived from learning theory and human learning is in fact the number of examples required in tasks such as object detection. The theory shows that the complexity of the hypothesis space sets the speed limit and the sample complexity for learning. If a task - like a visual recognition task – can be decomposed into low-complexity learning tasks for each layer of a hierarchical

learning machine, then each layer may require only a small number of training examples. Neuroscience suggests that what humans can learn can be represented by hierarchies that are locally simple. Thus, our ability to learn from just a few examples, and its limitations, may be related to the hierarchical architecture of cortex.

Beyond invariant object recognition, several models exist that try to account for the neural mechanisms underlying the processing of dynamic body stimuli (see Giese & Poggio 2003) for a review. These models are based on hierarchical neural architectures, including detectors that extract form or motion features from image sequences. Position and scale invariance has been accounted for by pooling neural responses along the hierarchy. It has been shown that such models reproduce several properties of neurons that are selective for body movements and behavioral and brain imaging data (Giese & Poggio 2003). Recent work proves the high computational performance of biologically inspired architectures for the recognition of body movement, which lies in the range of the best non-biological algorithms in computer vision (Jhuang et al. 2007).

Beyond the processing of biological motion and actions, face processing is another key function of high ecological significance. It is becoming well accepted that a network of visual cortical modules, mostly in IT, may constitute a system to process various aspects of face recognition – from the detection of a face in an image to its identification to a classification of its expression (see Tsao & Livingstone 2008). It is still unclear whether computational strategies similar to the hierarchical models proposed for general object recognition (see above), but different

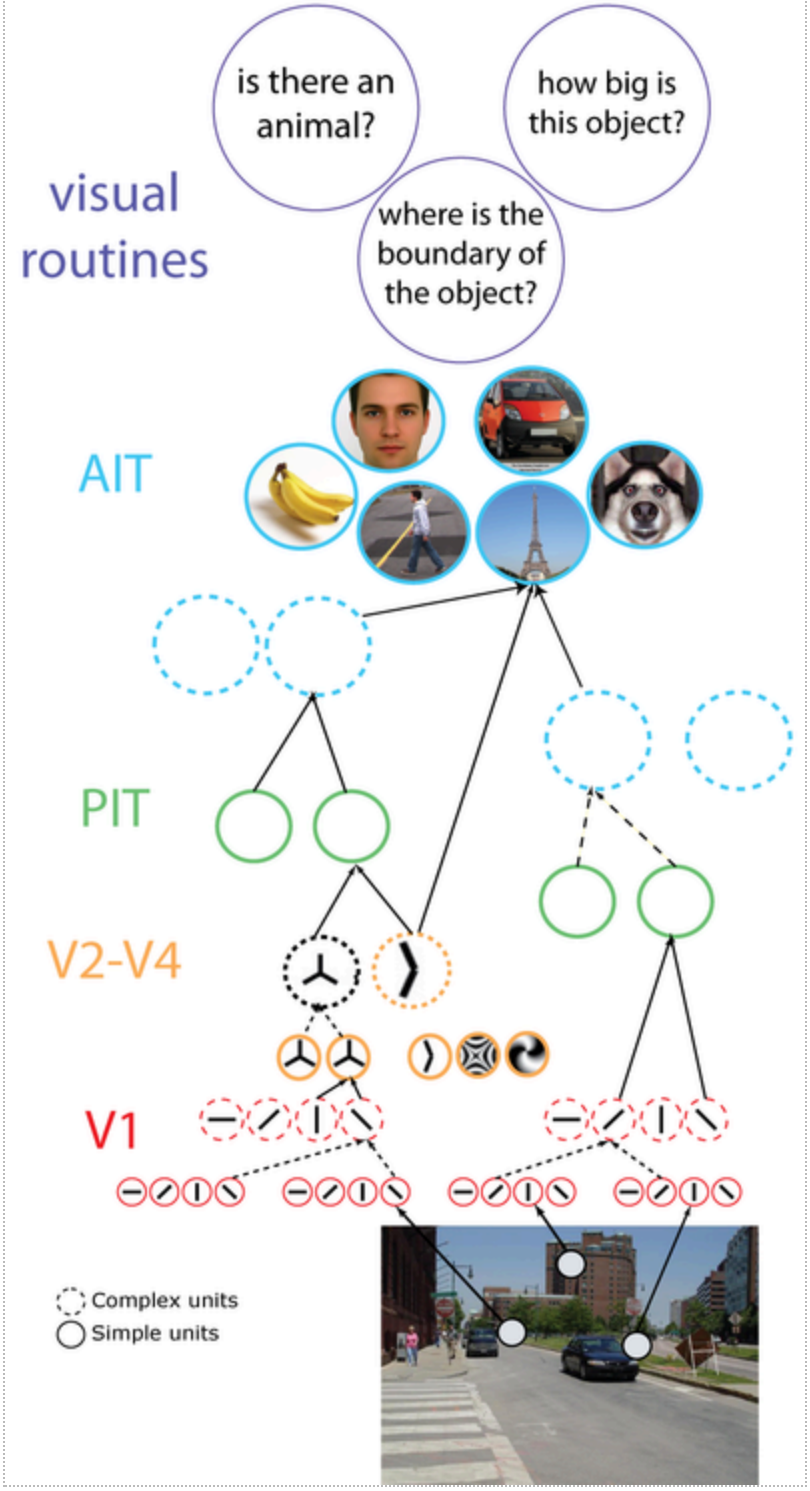

Figure 2: Sketch of the HMAX hierarchical computational model of the visual cortex (tentative mapping with areas of the visual cortex shown in color, some areas of the parietal cortex and dorsal streams not shown). Acronyms: V1, V2 and V4 correspond to primary, secondary and quaternary visual areas, PIT and AIT to posterior and anterior inferotemporal areas, respectively.

parameters (e.g., number and selectivity of tuned cells and the spatial extent of the features coded by neurons at intermediate levels), may be able to account for the properties of the face neurons and for the psychophysical signatures of face perception.

Backprojections, image inference and how visual cortex really works

Feedforward models of recognition and of other visual abilities have been useful to explore the power of fixed hierarchical organization as originally suggested by Hubel & Wiesel. These models assume that our core object recognition capability proceeds through a cascade of hierarchically organized areas along the ventral stream of the visual cortex with computations at each successive stage being largely feedforward (Riesenhuber & Poggio 1999; DiCarlo et al 2012). They have led, for instance, to algorithms competitive with the best computer vision systems (Serre & Poggio, 2010). Their limitations, however, are becoming increasingly obvious. Not only top-down effects are key to normal, everyday vision, but backprojections are also likely to be a key part of what cortex is computing and how. Thus, a major question for modeling visual cortex revolves around the role of backprojections and the related fact that vision is more than object recognition and requires interpreting and parsing visual scenes (as opposed to simply finding out whether a specific object is present in the visual scene or not). A human observer can essentially answer an infinite number of questions about an image (one could in fact imagine a Turing test of vision).

Computational models of top-down processes fall into two broad categories; one class of models sees the ventral stream as performing a continuous, task-independent inference about the world between eye movements. In this scenario, the ventral stream continuously develops a complex data structure representing the components of the scene and their relationships (see for instance, hierarchical generative and related models Hinton 2007; Lee & Mumford 2003). This class of models is related to prediction-verification recursions – an old approach known in AI as “hypothesis-verification” (Mumford 1996; Rao & Ballard 1999; Hawkins & Blakeslee 2002). Hierarchical generative models assume that the main function of the backprojections is to carry top-down hypotheses in order to compare them with bottom-up sensory information. One of the most attractive versions of this class of models is due to Lee & Mumford (2003). In their Bayesian framework, the recurrent feedforward/feedback loops in the cortex serve to integrate top-down contextual priors and bottom-up observations so as to implement concurrent probabilistic inference along the visual hierarchy. Here, recognition proceeds by iteration, with higher levels in the ventral stream generating a guess about the meaning of the incoming signal, and feeding this back to lower levels to generate what is in essence a synthetic image that can be compared with the image delivered from the retina. This “generative” point of view in its strongest form implies that neurons in the ventral pathways represent – after a series of bottom-up and top-down iterations – mutually and globally consistent conditional probabilities of certain features given the sensory input and the high level hypothesis or priors.

The second class of models follows the more conventional intuition that seeing amounts to computing a high resolution buffer in areas higher than the visual cortex with the goal to perform inference in a task dependent way only when needed, through a search involving attention and eye movements (see corresponding Scholarpedia article on computational models of attention). Within this framework, one key role for back-projections is to select and modulate specific connections in early areas in a top-down fashion (in addition to managing and controlling learning processes). In this extension, the backprojections mediate attentional modulations in lower areas; they also route information from specific lower areas to specialized task dependent categorization routines running in higher areas such as Pre-Frontal Cortex (PFC) . This class of models correspond to the belief that our subjective feeling of the richness of vision is based on the ability of *looking again* - possibly by shifting attention or gaze – whenever needed. Thus, back-projections may not only control the gain of specific neurons (this is the simplest model of a spotlight of attention), but they may also effectively run *routines* for reading out specific task-dependent information from IT

(for instance, one program may correspond to the question “is the object in the scene an animal?”, another may read out information about the size of the object in the image from activity in IT.) They may also select *programs* in areas lower than IT (possibly by modulating connection weights) to carry image inference tasks (is the animal to the right or to the left of the tree?). During normal vision, back-projections are likely to control, in a dynamic way, routines running at all levels of the visual system throughout attentional shifts (and fixations). In particular, small areas of the visual fields may be *routed* from the appropriate early visual area (as early as V1) by covert attentional shift controlled from the top to circuits specialized for any number of specific tasks (Poggio 1984). This highly speculative framework fits best with the point of view described by Hochstein & Ahissar (2002). They suggested that explicit vision advances in reverse hierarchical direction, starting with fast visual processing (corresponding to our “immediate recognition”) at the top of the cortical hierarchy and returning downward as needed in a “vision with scrutiny” mode in which reverse hierarchy routines focus attention to specific, active, low-level units.

Concluding remarks

Computational modeling is a burgeoning field, and in recent years, there has been an explosion in the number of models of visual processing that have been proposed. The models described in this article are computational in nature. Computational models, however, are limited in their explanatory power; ideally, they should eventually lead to a deeper and more general mathematical theory. As noted by Chuck Stevens, *“Models are common; good theories are scarce”*. Progress in the field of computational vision will thus require synergies between multiple fields from mathematics and statistics to computer and brain sciences.

Notes

Note 1: An alternative approach which denies the notion of symbol manipulation is the embodied approach (http://en.wikipedia.org/wiki/Embodied_cognitive_science) .

Note 2: This should not be confused with the notion of bottom-up (i.e., feedforward) vs. top-down (i.e., feedback, recurrent) processes.

Additional resources

The visual cortex (http://webvision.med.utah.edu/VisualCortex.html) , Matthew Schmolesky at the University of Utah.

Eye, Brain and Vision (http://hubel.med.harvard.edu/book/bcontex.htm) , David Hubel at Harvard University.

Computational Vision: Information Processing in Perception and Visual Behavior, H.A. Mallot. MIT Press, Cambridge (MA).

Computational models of visual processing, M.S. Landy & J.A. Movshon, MIT Press, Cambridge (MA).

Seeing: The Computational Approach to Biological Vision, J.P. Frisby & J.V. Stone, MIT Press, Cambridge (MA).

Theoretical Neuroscience . Computational and Mathematical Modeling of Neural Systems, P. Dayan & L.F. Abbott, MIT Press, Cambridge (MA).

References

Adelson, E.H. & J.R. Bergen (1985). Spatiotemporal energy models for the perception of motion. *J Opt Soc Am A*, 2(2):284-99.

Barrow, H.G. & Tenenbaum, J.M. (1981). Computational vision. *Proc IEEE*, 69(5):572-595.

Borgers, C. & N. Kopell (2003). Synchronization in networks of excitatory and inhibitory neurons with sparse, random connectivity. *Neural Comp* 15(3): 509-38.

Buia, C.I. & P.H. Tiesinga (2008). Role of interneuron diversity in the cortical microcircuit for attention. *J Neurophysiol* 99(5): 2158-82.

Carandini, M. & D.J. Heeger (1994). Summation and division by neurons in primate visual cortex. *Science* 264(5163): 1333-6.

Dicarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3), 415–34. doi:10.1016/j.neuron.2012.01.010

Douglas, R.J. & K.A. Martin (1991). A functional microcircuit for cat visual cortex. *J Physiol* (Lond) 440:735–69.

Felleman, D.J. & D.C. van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex* 1:1–47, 1991.

Finn, I.M. & D. Ferster (2007). Computational diversity in complex cells of cat primary visual cortex. *J Neurosci* 27(36):9638-48.

Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural Comp* 3:194–200.

Giese, M. & T. Poggio (2003). Neural mechanisms for the recognition of biological movements and action. *Nat Rev Neurosci* 4:179–192.

Grimson, W.E.L. (1982) A computational theory of visual surface interpolation. *Philos Trans R Soc London Ser B* 298:395-427.

Geisler, W.S. (2008). Visual perception and the statistical properties of natural scenes. *Annu Rev Psychol*, 59, 10.1-10.26.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Res*, 51, 771–781.

Grossberg, S. (2007). *Towards a unified theory of neocortex: Laminar cortical circuits for vision and cognition*. In Computational Neuroscience: From Neurons to Theory and Back Again, Eds: Paul Cisek, Trevor Drew, John Kalaska; Elsevier, Amsterdam, pp. 79-104.

Hahnloser, R.H., R. Sarpeshkar, et al. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* 405(6789): 947-51.

Hawkins, J. & S. Blakeslee (2002). *On Intelligence*. New York, Times Books, Holt.

Hinton, G. E. (2007). To recognize shapes, first learn to generate images. *Progress in Brain Research*, 165.

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

Hubel, D.H. & T.N. Wiesel (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol* 160:106–154.

Hurlbert, A. (1989). *The computation of color*. PhD Thesis. Harvard Medical School and Massachusetts Institute of Technology, Department of Brain & Cognitive Sciences.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans Pattern Anal Machine Intell*, 20(11).

Jhuang H., T. Serre, L. Wolf & T. Poggio (2007). A biologically-inspired system for action recognition, Proc IEEE International Conference on Computer Vision.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annu Rev Psychol*, 55, 271-304.

Koch, C. (1999). *Biophysics of Computation: Information processing in single neurons*. Oxford University Press. ISBN 0-19-510491-9.

Koch, C., & S. Ullman (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, 4(4):219-27.

Kouh, M. & T. Poggio (2008). A canonical neural circuit for cortical nonlinear operations. *Neural Comput*, 20(6): 1427-51.

Kveraga, K., J. Boshyan, et al. (2007). Magnocellular projections as the trigger of top-down facilitation in recognition. J Neurosci 27(48): 13232-40.

Lee, D.K., L. Itti, et al. (1999). Attention activates winner-take-all competition among visual filters. *Nat Neurosci*, 2(4): 375-81.

Lee, T.S. & D. Mumford (2003). Hierarchical Bayesian inference in the visual cortex. *J Opt Soc Am A Opt Image Sci Vis* 20(7): 1434-48.

Marko, H. & Giebel, H. (1970). *Recognition of handwritten characters with a system of homogeneous Layers*. Nachrichtentechnische Zeitschrift*, 23(455-459)*.

Markram, H. (2006). The blue brain project. *Nat Rev. Neurosci.*, 7, 153–160.

Marr, D. & T. Poggio (1976). Cooperative computation of stereo disparity. *Science* 194(4262): 283-7.

Marr, D. & T. Poggio (1977). From understanding computation to understanding neural circuitry. In: Neuronal Mechanisms in Visual Perception, E. Poppel, R. Held and J.E. Dowling (eds.), *Neurosciences Res. Prog. Bull.*, 15, 470-488.

Marr, D. & T. Poggio (1979). A computational theory of human stereo vision. *Proc R Soc Lond B Biol Sci* 204(1156): 301-28.

Marr, D. & S. Ullman (1981). Directional selectivity and its use in early visual processing. *Proc R Soc Lond B* 211(1183):151-180.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

Masquelier, T., & Thorpe, S. J. (2007). Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Comput Biol*, 3(2), e31. doi:06-PLCB-RA-0472R2 [pii] 10.1371/journal.pcbi.0030031

McCulloch, W. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull Math Biol* 7:115 - 133.

Miller, K.D., & T.W. Troyer (2002). Neural noise can explain expansive, power-law nonlinearities in neural response functions. *J Neurophysiol* 87(2): 653-659.

Mumford, D. (1996). *Pattern theory: a unifying perspective*. New York, NY USA, Cambridge University Press.

Najemnik, J. & Geisler, W.S. (2005). Optimal eye movement strategies in visual search. *Nature*, 434, 387-391.

Niebur, E. & C. Koch (1994). A model for the neuronal implementation of selective visual attention based on temporal correlation among neurons. *J Comput Neurosci* 1(1-2): 141-58.

Nowlan, S. J. & T. J. Sejnowski (1995). A selection model for motion processing in area MT of primates. *J Neurosci* 15(2): 1195-214.

Pinto, N., Doukhan, D., DiCarlo, J. J., & Cox, D. D. (2009). A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS computational biology*, 5(11), e1000579. doi:10.1371/journal.pcbi.1000579

Poggio, T., W. Reichardt & W. Hausen (1981). A Neuronal circuitry for relative movement discrimination by the visual system of the fly, *Naturwissenschaften*, 68,9, 43-466.

Poggio, T. (1984). Routing thoughts. Massachusetts Institute of Technology. Artificial Intelligence Laboratory, AI Working Paper (#258).

Poggio, T. & E. Bizzi (2004). Generalization in vision and motor control. *Nature* 431: 768-774.

Poggio, T. & S. Smale (2003). The mathematics of learning: Dealing with data. Notices of the American Mathematical Society 50(05): 537-544.

Qian, N. (1994). Computing stereo disparity and motion with known binocular cell properties. *Neural Computation*, 6(3), 390–404.

Rao, R.P.N. & D.H. Ballard (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci* 2:79-87.

Read, J.C.A., A J. Parker & B.G. Cumming (2002). A simple model accounts for the response of disparity-tuned V1 neurons to anticorrelated images. *Visual Neurosci* 19(6): 735-753.

Reynolds, J.H & D. J. Heeger (2009). The normalization model of attention. *Neuron*, 61(2), 168-185.

Reynolds, J.H., L. Chelazzi, et al. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *J Neurosci*, 19(5): 1736-53.

Riesenhuber, M. & T. Poggio (1999). Hierarchical models of object recognition in cortex. *Nat Neurosci* 2:1019–1025.

Rolls,E.T. (2012). Invariant visual object and face recognition: neural and computational bases, and a model, VisNet. *Frontiers in Computational Neuroscience* 6: (35) 1-70.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Cornell Aeronautical Laboratory, Psychol Rev, v65, No. 6, pp. 386–408.

Rousselet G.A., S.J. Thorpe & M. Fabre-Thorpe (2003). Taking the MAX from neuronal responses. *Trends Cogn Sci* 7: 99-102.

Rust, N. C., O. Schwartz, et al. (2005). Spatiotemporal elements of macaque V1 receptive fields. *Neuron* 46(6): 945-56.

Rust, N. C., V. Mante, et al. (2006). How MT cells analyze the motion of visual patterns. *Nat Neurosci* 9(11): 1421-31.

Serre, T., G. Kreiman, M. Kouh, C. Cadieu, U. Knoblich & T. Poggio. (2007). *A quantitative theory of immediate visual recognition*. In: Progress in Brain Research, Computational Neuroscience: Theoretical Insights into Brain Function, 165, pp. 33-56.

Serre, T., A. Oliva & T. Poggio (2007). A feedforward architecture accounts for rapid categorization. *Proc Natl Acad Sci U S A* 104(15): 6424-9.

Serre, T., & Poggio, T. (2010). A neuromorphic approach to computer vision. *Communications of the ACM*, 53(10), 54. doi:10.1145/1831407.1831425

Simoncelli, E.P. & B.A. Olshausen (2001). Natural image statistics and neural representation. *Annu Rev Neurosci* 24:1193-216.

Simoncelli, E.P. & D.J. Heeger (1998). A model of neuronal responses in visual area MT. *Vision Res* 38(5):743-761.

Sturm, A.K. & P. Koenig (2001). Mechanisms to synchronize neuronal activity. *Biol Cybern* 84:153-172.

Teich, A.F. & N. Qian (2006). Comparison among some models of orientation selectivity. *J Neurophysiol* 96(1): 404-19.

Tiesinga, P.H. & T.J. Sejnowski (2004). Rapid temporal modulation of synchrony by competition in cortical interneuron networks. *Neural Comput* 16(2): 251-75.

Thorpe, S.J. (2002). Ultra-rapid scene categorization with a wave of spikes. Proc. of Biologically Motivated Computer Vision: 2nd International Workshop, Tübingen, Germany.

Torre, V. & T. Poggio. A synaptic mechanism possibly underlying directional selectivity motion. *Proc R Soc Lond B Biol Sci* 202, 409-416, 1978.

Tsao, D.Y., N. Schweers, et al. (2008). Patches of face-selective cortex in the macaque frontal lobe. *Nat Neurosci* 11(8): 877-9.

Tsao, D.Y., Linvingston, M.S. (2008). Mechanisms of Face Perception. Annu Rev Neurosci 31: 411-437.

Tsotsos, J. (1997). Limited capacity of any realizable perceptual system is a sufficient reason for attentive behavior. *Consciousness and cognition*, 6(2–3), 429–436.

Ullman, S. (1976). "Filling in the gaps: the shape of subjective contours and a model for their generation. *Biol Cybern*, 25:1--6.

Ullman, S. (1979). The interpretation of visual motion. MIT Press, Cambridge, MA.

Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Sciences*, 11(2), 58–64. doi:10.1016/j.tics.2006.11.009

VanRullen, R., R. Guyonneau, & S.J. Thorpe (2005). "Spike times make sense." *Trends in Neurosci*, 28(1).

Wersing, H., & Koerner, E. (2003). Learning optimized features for hierarchical models of invariant recognition. *Neural Comp.*, 15(7), 1559–1588.

Yuille, A., & Kersten, D. (2006). "Vision as Bayesian inference: analysis by synthesis?" *Trends Cogn Sci*, 10(7), 301-308.

Yuille, A. & N.M Grzywacz (1988). "A computational theory for the perception of coherent visual motion." *Nature* 333: 71-74.

Sponsored by: Eugene M. Izhikevich, Editor-in-Chief of Scholarpedia, the peer-reviewed open-access encyclopedia
Reviewed by (http://www.scholarpedia.org/w/index.php?title=Models_of_visual_cortex&oldid=130934) : Mr. Tim Kietzmann, University of Osnabrück, Germany
Reviewed by (http://www.scholarpedia.org/w/index.php?title=Models_of_visual_cortex&oldid=130950) : Peter König, Inst. Cognitive Science, Universität Osnabrück, Germany

Accepted on: 2013-04-01 11:02:10 GMT (http://www.scholarpedia.org/w/index.php?title=Models_of_visual_cortex&oldid=131433)

Categories: Computational Neuroscience | Brain Models | Vision