8

Graphical Models of the Visual Cortex

Thomas Dean

1 Pivotal Encounters with Judea

Post graduate school, three chance encounters reshaped my academic career, and all three involved Judea Pearl directly or otherwise. The first encounter was meeting Judea on a visit to the UCLA campus at a time when I was developing what I called temporal Bayesian networks and would later be called *dynamic belief networks* (an unfortunate choice of names for reasons I'll get to shortly). Judea was writing his book *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* [1988] and his enthusiasm for the subject matter was positively infectious. I determined from that meeting that I was clueless about all things probabilistic and proceeded to read each of Judea's latest papers on Bayesian networks multiple times, gaining an initial understanding of joint and marginal probabilities, conditional independence, etc. In those days, a thorough grounding in probability and statistics was rarely encouraged for graduate students working in artificial intelligence.

The second encounter was with Michael Jordan at a conference where he asked me a question that I was at a loss to answer and made it clear to me that I didn't really understand Bayesian probability theory at all, despite what I'd picked up from Judea's papers. My reaction to that encounter was to read Judea's book cover to cover and discover the work of I.J. Good. Despite being a math major and having met I.J. Good at Virginia Tech where I was an undergraduate and Good was a professor of statistics, I never took a course in probability or statistics. My embarrassment at being flummoxed by Mike's question forced me to initiate a crash course in probability theory based on the textbooks of Morris DeGroot [1970, 1986]. I didn't recognize it at the time, but Judea, Mike and like-minded researchers in central areas of artificial intelligence were in the vanguard of those changing the landscape of our discipline.

The third encounter was with David Mumford when our paths crossed in the midst of a tenure hearing at Brown University and David told me of his work on models of the visual cortex. I read David's paper with Tai Sing Lee [2003] as well as David's earlier related work [1991, 1992] and naïvely set out to implement their ideas as a probabilistic graphical model [Dean 2005]. Indeed, I wanted to extend their work since it did not address the representation of time passing, and I was interested in building a model that dealt with how a robot might make sense of its observations as it explores its environment.

Moreover, the theory makes no mention of how a robot might learn such a model, and, from years of working with robots, I was convinced that building a model by hand would turn out to be a lot of work and very likely prove to be unsuccessful. Here it was Judea's graphical-models perspective that, initially, made it easy for me to think about David's work, and, later, extend it. I also came to appreciate the relevance of Judea's work on causality and, in particular, the role of intervention in thinking about how biological systems engage the world to resolve perceptual ambiguity.

This chapter concerns how probabilistic graphical models might be used to model the visual cortex, and how the challenges faced in developing such models suggest areas where current theory falls short and might be extended. A graphical model is a useful formalism for compactly describing a joint probability distribution characterized by very large number of random variables. We are taking what is known about the anatomy and physiology of the primate visual cortex and attempting to apply that knowledge to construct probabilistic graphical models that we can ultimately use to simulate some functions of primate vision. It may be that the resulting probabilistic model also captures some important characteristics of individual neurons or their ensembles. For practical purposes, this need not be the case, though clearly we believe there are potential advantages to incorporating some lessons from biology into our models. Graphical models also suggest, but do not dictate, how one might use such a model along with various algorithms and computing hardware to perform inference and thereby carry out practical simulations. It is this latter use of graphical models that we refer to when we talk about implementing a model of the visual cortex.

2 Primate Visual Cortex

Visual information processing starts in the retina and is routed via the optic tract to the lateral geniculate nuclei (LGN) and then on to the striate cortex also known as visual area one (V1) located in the occipital lobe at the rear of the cortex. There are two primary visual pathways in the primate cortex: The ventral pathway leads from the occipital lobe into the temporal lobe where association areas in the inferotemporal cortex combine visual information with information originating from the auditory cortex. The dorsal pathway leads from the occipital to the parietal lobe which, among other functions, facilitates navigation and manipulation by integrating visual, tactile and proprioceptive signals to provide our spatial sense and perception of shape.

It is only in the earliest portion of these pathways that we have any reasonably accurate understanding of how visual information is processed, and even in the very earliest areas, the striate cortex, our understanding is spotty and subject to debate. It seems that cells in V1 are mapped to cells in the retina so as to preserve spatial relationships, and are tuned to respond to stimuli that appear roughly like oriented bars. Hubel and Wiesel's research on macaque monkeys provides evidence for and

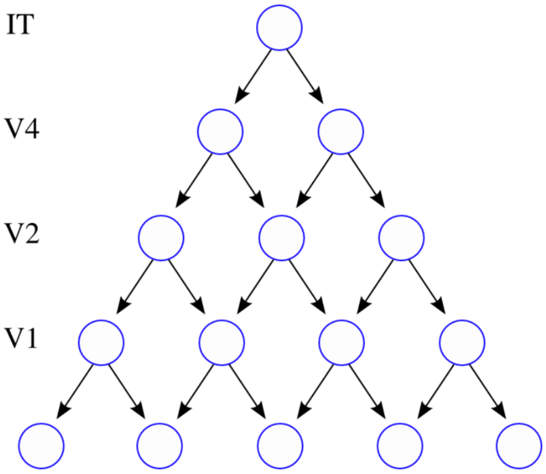


Figure 1. A simple hierarchical model of the ventral visual pathway.

subsequent studies confirm the latter characterization of function in primates [1962, 1968]. That said, there is still a good deal that we don't know about visual processing in V1 [Olshausen and Field 2005], and our understanding gets murkier as we progress along these pathways.

It is said that the ventral pathway is responsible for identifying "what" objects we see, and the dorsal for identifying "where" in our visual field these objects and their various parts are located and "how" we might interact with them by grasping, avoiding, etc. This is sufficiently vague that, given our current understanding of visual processing, it is probably a useful rule of thumb. However, there is ample evidence [Konen and Kastner 2008] to suggest that the "what", "where" and "how" are commingled via a myriad of connections and it is misleading to think of visual processing as a pipeline that leads in pure feed-forward fashion from simple features that subtend small portions of the visual field to more complex features that subtend greater and greater spatial (and temporal) extent. Indeed, there is no conclusive evidence that the brain is organized as a hierarchy of features, despite this being an elegant and comforting hypothesis to entertain.

3 Static Graphical Models

Figure 1 depicts a simple graphical model of the ventral visual pathway, where the nodes in the bottom layer are meant to model retinal ganglion cells. Nodes in the second layer model cells in the striate cortex which is also known as Brodmann's area 17 or V1. The next layer corresponds to Brodmann's area 18 or V2 which is responsive to somewhat more complex patterns than V1. The penultimate layer encodes V4 which is tuned to object features of intermediate complexity, and the

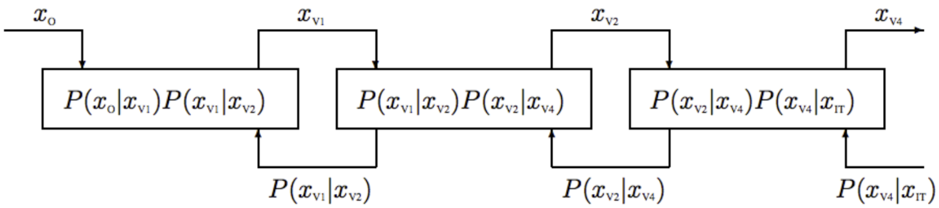

x_{O}
x_{V1}
x_{V2}
x_{V4}
$P(x_{\mathrm{O}}|x_{\mathrm{V1}})P(x_{\mathrm{V1}}|x_{\mathrm{V2}})$
$P(x_{\mathrm{V1}}|x_{\mathrm{V2}})P(x_{\mathrm{V2}}|x_{\mathrm{V4}})$
$P(x_{\mathrm{V2}}|x_{\mathrm{V4}})P(x_{\mathrm{V4}}|x_{\mathrm{IT}})$
$P(x_{\mathrm{V1}}|x_{\mathrm{V2}})$
$P(x_{\mathrm{V2}}|x_{\mathrm{V4}})$
$P(x_{\mathrm{V4}}|x_{\mathrm{IT}})$

Figure 2. A schematic of the hierarchical Bayesian framework proposed by Lee and Mumford [2003]. The regions of the visual cortex are linked together in a Markov chain. The activity in the ith region is influenced by bottom-up feed-forward data x_{i-1} and top-down probabilistic priors $P(x_i|x_{i+1})$ representing feedback from region $i+1$. The Markov property plays an important computational role by allowing units to depend only on their immediate neighbors in the Markov chain.

final layer represents inferotemporal cortex or IT which apparently responds to complex shapes. To get a more realistic picture, imagine that each layer represents a two-dimensional map with correspondence to the surface of the retina. Let's consider several ways in which this simple model falls short of the mark.

The graph in Figure 1 has no edges between nodes in the same layer. We know adjacent cells within each of V1, V2, V4 and IT communicate via many lateral connections. Given such connectivity, it would seem that two nodes representing adjacent cells in a layer are dependent in a statistical sense. However, the model as it stands has the property that any two nodes in a given layer are independent of one another when conditioned on the nodes in the layer immediately above. It turns out that it is very difficult to capture complex spatial relationships among adjacent regions of images using a graphical model having the intra-layer conditional independence implicit in model shown in Figure 1.

How might we capture the statistical properties of adjacent cells in the same cortical layer in a graphical model? First, note that our graphical model, while more complex than a tree, is simpler than an arbitrary directed graph being acyclic. Unfortunately, it is difficult to devise a plausible scheme to connect vertices within layers using directed edges and avoid cycles in the graph, and so we won't even attempt this. One possibility is that we model each layer as a Markov random field with the connectivity of a regular grid, but retain the directional edges between layers. The advantage of this approach is that the graph as a whole is a Markov chain and it is relatively simple to write down the joint distribution. Using the chain rule and the assumption made explicit in the graph shown in Figure 1 that each variable in the sequence $(x_{\mathrm{O}}, x_{\mathrm{V1}}, x_{\mathrm{V2}}, x_{\mathrm{V4}}, x_{\mathrm{IT}})$ is independent of the other variables given its immediate neighbors in the sequence, we write the equation relating the one retinal and three cortical regions as

$$P(x_{\mathrm{O}}, x_{\mathrm{V1}}, x_{\mathrm{V2}}, x_{\mathrm{V4}}, x_{\mathrm{IT}}) = P(x_{\mathrm{O}}, x_{\mathrm{V1}})P(x_{\mathrm{V1}}, x_{\mathrm{V2}})P(x_{\mathrm{V2}}, x_{\mathrm{V4}})P(x_{\mathrm{V4}}, x_{\mathrm{IT}})P(x_{\mathrm{IT}})$$

where x_{O} represents the retinal or *observation* layer. Moreover, we know that, although the edges all point in the same direction, information flows both ways in the hierarchy via Bayes rule (see Figure 2).

Despite the apparent simplicity when we collapse each layer of variables into a single, joint variable, exact inference in such a model is intractable. One might imagine, however, using a variant of the forward-backward algorithm to approximate the joint distribution over all variables. Such an algorithm might work one layer at a time, by isolating each layer in turn, performing an approximation on the isolated Markov network using Gibbs sampling or mean-field approximation, propagating the result either forward or backward and repeating until convergence. Simon Osindero and Geoff Hinton [2008] experimented with just such a model and demonstrated that it works reasonably well at capturing the statistics of patches of natural images.

One major problem with such a graphical model as a model of the visual cortex is that the Markov property of the collapsed-layer simplification fails to capture the inter-layer dependencies implied by the connections observed in the visual cortex. In the cortex as in the rest of the brain, connections correspond to the dendritic branches of one neuron connected at a synaptic cleft to the axonal trunk of a second neuron. We are reasonably comfortable modeling such a *cellular edge* as an edge in a probabilistic graphical model because for every cellular edge running forward along the visual pathways starting from V1 there is likely at least one and probably quite a few cellular edges leading backward along the visual pathways. Not only do these backward-pointing cellular edges far outnumber the forward-pointing ones, they also pay no heed to the Markov property, typically spanning several layers of our erstwhile simple hierarchy. Jin and Geman [2006] address this very problem in their hierarchical, compositional model, but at a considerable computational price. Advances in the development of adaptive Monte Carlo Markov chain (MCMC) algorithms may make inference in such graphical models more practical, but, for the time being, inference on graphical models of a size comparable to the number of neurons in the visual cortex remains out of reach.

4 Temporal Relationships

Each neuron in the visual cortex indirectly receives input from some, typically contiguous, region of retinal ganglion cells. This region is called the neuron's *receptive field*. By introducing lags and thereby retaining traces of earlier stimuli, a neuron can be said to have a receptive field that spans both space and time — it has a *spatiotemporal* receptive field. A large fraction of the cells in visual cortex and V1 in particular have spatiotemporal receptive fields. Humans, like most animals, are very attentive to motion and routinely exploit motion to resolve visual ambiguity,

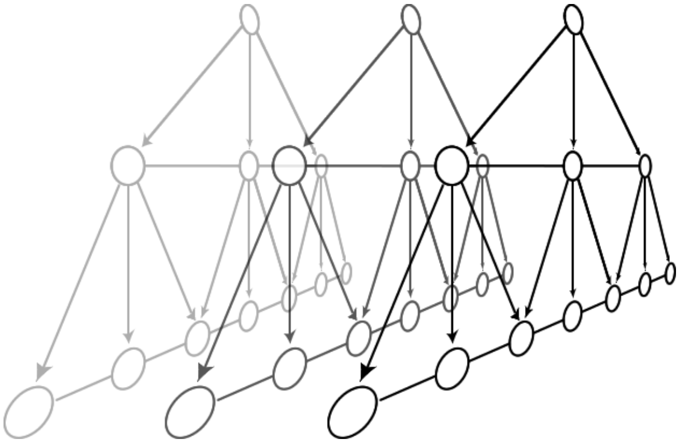


Figure 3. A cartoon of the hierarchical hidden Markov model described in [Dean 2006] showing the same basic structure as in Figure 1, but with additional undirected, intra-layer edges, and replicated for some number of time steps to form a hierarchy of hidden Markov models, so that nodes in a each layer represent different spatial and temporal extent. Not shown, but present in the full model, are edges that span the temporal slices thus modeling neural circuits that have spatiotemporal receptive fields.

and, generally, deal with the four-dimensional, space-time continuum in which we live.

Figure 3 depicts the obvious temporal analog of Figure 1. Dean [2006] presents a model of visual cortex based on the Hierarchical Hidden Markov Model of Fine *et al* [1998]. In the model described in [2006], nodes have edges that span both nodes within the layers of individual time slices and nodes that reside within the same layer of adjacent time slices.

The ability to represent the passage of time is clearly important in enabling us to make predictions and plan our actions, but it also allows us to expedite learning. The cortex evolved to take advantage of *temporal coherence*, the property that, for the most part, the appearance of the objects present in our visual field, animate or otherwise, do not change a great deal from one instant to the next. We can exploit this property to learn about the stable features of our environment. Földiák [Földiák 1991] describes a biologically plausible theory of how neural circuits might exploit temporal coherence to learn useful features by looking for signals that change slowly over time. Wiskott and Sejnowski [2002], Hyvärinen*et al* [2003], George and Hawkins [2005], Dean [2006] and others have proposed various algorithms for improving on this same basic idea.

Using proximity in space and time to group similar visual features was recognized early on by the Gestalt psychologists, and it is one of many characteristics of our visual experience that biology has evolved to exploit to its advantage. However, in this chapter, I want to explore a different facet of how we make sense of and, in some cases, take advantage of spatial and temporal structure to survive and thrive, and how these aspects of our environment offer new challenges for applying graphical models.

5 Dynamic Graphical Models

Whether called temporal Bayesian networks [Dean and Wellman 1991] or dynamic Bayesian networks [Russell and Norvig 2003], these graphical models are designed to model properties of our environment that change over time and the events that precipitate those changes. The networks themselves are not dynamic: the numbers of nodes and edges, and the distributions that quantify the dependencies among the random variables that correspond to the nodes are fixed. At first blush, graphical models may seem a poor choice to model the neural substrate of the visual cortex which is anything but static. However, while the graph that comprises a graphical model is fixed, a graphical model can be used to represent processes that are highly dynamic, and contingent on the assignments to observed variables in the model. In the remainder of this section, we describe characteristics of the visual system that challenge our efforts to model the underlying processes required to simulate primate vision well enough to perform such tasks such as object recognition and robot navigation.

The retina and the muscles that control the shape of the lens and the position of the eyes relative to one another and the head comprise a complex system for acquiring and processing visual information. A mosaic of photoreceptors activate several layers of cells, the final layer of which consists of retinal ganglion cells whose axons comprise the optic nerve. This multi-layer extension of the brain performs a range of complex computations ranging from light-dark adaptation to local contrast normalization [Brady and Field 2000]. The information transmitted along the optic tract is already the product of significant computational processing.

Visual information is *retinotopically* mapped from the retinal surface to area V1 so as to preserve the spatial relationships among patches on the retina that comprise the receptive fields of V1 cells. These retinotopic mappings are primarily sorted out *in utero*, but the organization of the visual cortex continues to evolve significantly throughout development — this is particularly apparent when children are learning to read [Dehaene 2009]. Retinotopic maps in areas beyond V1 are more complicated and appear to serve purposes that relate to visual tasks, *e.g.*, the map in V2 anatomically divides the tissue responsible for processing the upper and lower parts of the visual fields. These retinotopic maps, particularly those in area V1, have led some computer-vision researchers to imagine that early visual processing proceeds via transformations on regular grid-like structures with cells analogous to pixels.

The fact is that our eyes, head, and the objects that we perceive are constantly

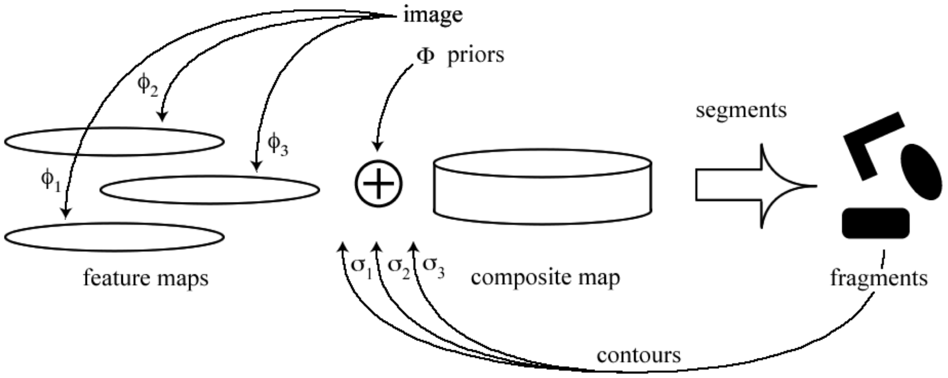
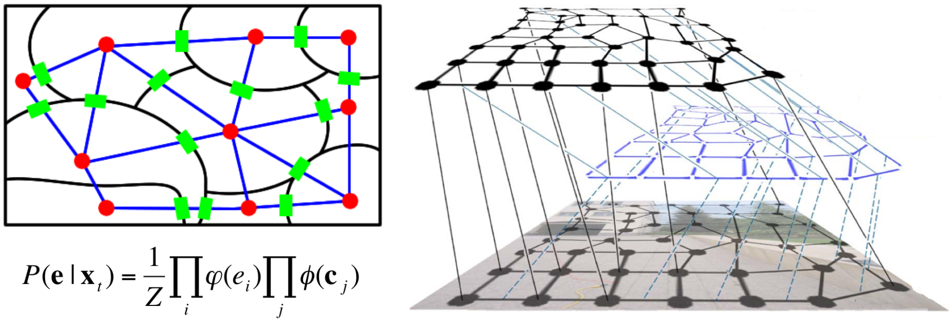


Figure 4. This graphic depicts a variant of the process of segmentation as proposed by Tenenbaum and Barrow [1977] in which the low-level features ϕ_i are used to construct feature maps that are combined with priors Φ derived from context to form composite maps which produce candidate fragments that suggest the boundaries or contours σ_i of objects which in turn guide subsequent interpretation and assignment of fragments to objects.

in flux. Even with our head fixed while viewing a still image, our eyes quickly jump or *saccade* up to 90° of visual angle several times a second, and perform much smaller adjustments or *microsaccades* at around 30–50 Hertz. A two-week-old baby can already track objects by *smooth pursuit*, thereby keeping an object centered in the *fovea* which is the central, high-acuity and high-color-sensitivity portion of the retina. Both smooth pursuit and (macro) saccades are driven by attentional mechanisms which combine a bottom-up, small-image-patch, data-driven component with a top-down, whole-image-gist, prior-knowledge component. The important point here is that how we perceive the visual world has little resemblance — beyond the earliest processing stages — to a sequence of orderly transformations performed on a regular grid, and has everything to do with assembling a puzzle out of fragmentary glimpses snatched as our gaze quickly shifts relative to the frames of reference of our head, our body and the ground on which we stand.

Even if we concentrate on the hundred milliseconds or so during which the fovea remains focused on a small patch of a still image between saccades, the representation of this process as a graphical model becomes complicated as we abstract from the "pixel" level. Figure 4 depicts a process whereby the responses of low-level feature detectors are used to construct feature maps. Typically, the feature detectors report information about intensity, color, texture, etc. These maps are combined so that every location in an image is summarized by a vector of features. In bottom-up segmentation, such summaries alone are used to aggregate locations into segments that correspond to object surfaces or at least respect object boundaries. Tenenbaum



$$P(\mathbf{e} \mid \mathbf{x}_t) = \frac{1}{Z}\prod_i \varphi(e_i)\prod_j \phi(\mathbf{c}_j)$$

Figure 5. The graphic in the upper left depicts an over-segmentation of an image superimposed with a graph whose nodes (depicted as circles) correspond to segments or *superpixels* and whose edges (marked with rectangular boxes) correspond to boundaries between segments. The graphic on the right (adapted from Saxena *et al* [2007]) shows a graphical model in the form of multi-layer factor graph where nodes in the lowest layer correspond to superpixels, those in the second layer to boundary classes and those in the top to object surfaces.

and Barrow [1977] suggested that this low-level information has to be supplemented by *prior* knowledge, which provides a more global context in which to interpret the low-level information, and combined in an iterative of process of agglomeration.

This process has been realized in a graphical model format using variants of factor graphs [Saxena, Chung, and Ng 2007] and conditional random fields [Hoiem, Efros, and Hebert 2007] and hierarchical Dirichlet-process hidden-Markov trees [Kivinen, Sudderth, and Jordan 2007]. Figure 5 characterizes the basic structure of the algorithm adopted by Saxena *et al* [2007] and by Hoiem *et al* [2007]. First, the raw image is divided into *superpixels* — regions of pixels homogeneous in their intensity, color or texture — which correspond to the segments obtained from an over-segmentation of the image that is assumed to respect image boundaries. Next, the superpixels are used to construct a graphical model consisting of nodes corresponding superpixels, information that pertains to their status as object (occlusion) boundaries, and their relationships *vis a vis* their contribution to the same objects. Inference serves to identify boundaries or the absence thereof, some superpixels can be merged, and the process repeated until no further merging is possible.

In the implementation of this process, the topology of the graphical model is generated on an image-by-image basis, and the iterative refinement process requires adjustments to the size and connectivity of the graph that are particular to the boundaries of object surfaces in the image being observed. Such a process can be implemented within a fixed-graph structure but the model of a dynamically changing graph is simple to implement and apply recursively. One advantage, however, of a graph in which the nodes in a fixed grid of nodes are dynamically assigned to segments as part of inference is that this model is potentially more elegant, and even biologically plausible, in that the recursive process might be represented as a single hierarchical graphical model allowing inference over the entire graph, rather than over sequences of ever more refined graphs.

The above discussion of segmentation is but one example in which nodes in a graphical model might serve as *generic* variables that are bound as required by circumstances. But perhaps this view is short sighted; why not just assume that there are enough nodes that every possible (visual) concept corresponds to a unique combination of existing nodes. In this view, visual interpretation is just mapping visual stimuli to the closest visual "memory". Given the combinatorics, the only way this could be accomplished is to use a hierarchy of features whose base layer consists of small image fragments at many different spatial scales, and all subsequent layers consist of compositions of features at layers lower in the hierarchy [Bienenstock and Geman 1995; Ullman and Soloviev 1999; Ullman, Vidal-Naquet, and Sali 2002]. This view accords well with the idea that most visual stimuli are not determined to be novel and, hence, we construct our reality from bits and pieces of existing memories [Hoffman 1998]. Our visual memories are so extensive that we can almost always create a plausible interpretation by recycling old memories. It may be that in some aspects of cognition we have to employ generic neural structures to perform the analog of binding variables, but for much of visual intelligence this may not be necessary given a large enough memory of reusable fragments. Which raises the question of how we might implement a graphical model that has anywhere near the capacity of the visual cortex.

6 Distributed Processing at Cortex Scale

The cortex consists of a layered sheet with a more-or-less uniform cellular structure. Neuroanatomists have identified what are called *columns* corresponding to groups of local cells running perpendicular to the cortical surface. Vernon Mountcastle [2003] writes "The basic unit of cortical operation is the *minicolumn* [...] [containing] on the order of 80–100 neurons [...] The minicolumn measures of the order of 40-50μ in transverse diameter, separated from adjacent minicolumns by vertical cell-sparse zones which vary in size in different cortical areas." These minicolumns are then grouped into cortical columns which "are formed by the binding together of many minicolumns by common input and short-range horizontal connections."

If we take the cortical column — not the minicolumn — as our basic computational module as in [Anderson and Sutton 1997], then the gross structure of the neocortex consists of a dense mat of inter-columnar connections in the outer-most layer of the cortex and another web of connections at the base of the columns. The inter-columnar connectivity is relatively sparse (something on the order of 10^{15} connections spanning approximately 10^{11} neurons) and there is evidence [Sporns and Zwi 2004] to suggest that the induced inter-columnar connection graph exhibits the properties of a *small-world graph* [Newman, Watts, and Strogatz 2002]. In particular, evidence suggests the inter-columnar connection graph has low diameter (the length of the longest shortest path separating a pair of vertices in the graph) thereby enabling relatively low-latency communication between any two cortical columns.

It is estimated that there are about a quarter of a billion neurons in the primary visual cortex — think V1 through V4 — counting both hemispheres, but probably only around a million or so cortical columns. If we could roughly model each cortical column with a handful of random variables, then it is at least conceivable that we could implement a graphical model of early vision.

To actually implement a graphical model of visual cortex using current technology, the computations would have to be distributed over many machines. Training such a model might not take as long as raising a child, but it could take many days — if not years — using the current computer technology, and, once trained, we presumably would like to apply the learned model for much longer. Given such extended intervals of training and application, since the mean-time-til-failure for the commodity-hardware-plus-software that comprise most distributed processing clusters is relatively short, we would have to allow for some means of periodically saving local state in the form of the parameters quantifying the model.

The data centers that power the search engines of Google, Yahoo! and Microsoft are the best bet that we currently have for such massive and long-lived computations. Software developed to run applications on such large server farms already have tools that could opportunistically allocate resources to modify the structure of graphical model in an analog of neurogenesis. These systems are also resistant to both software and equipment failures and capable of reallocating resources in the aftermath of catastrophic failure to mimic neural plasticity in the face of cell death.

In their current configuration, industrial data centers may not be well suited to the full range of human visual processing. Portions of the network that handle very early visual processing will undoubtedly require shorter latencies than is typical in such server farms, even among machines on the same rack connected with high-speed Ethernet. Riesenhuber and Poggio [1999] use the term *immediate recognition* to refer to object recognition and scene categorization that occur in the first 100-200ms or so from the onset of the stimuli. In that short span of time — less than the time it takes for a typical saccade, we do an incredibly accurate job of recognizing objects and inferring the gist of a scene. The timing suggests that only a few steps of neural processing are involved in this form of recognition, assuming 10–20ms per synaptic transmission, though given the small diameter of the inter-columnar connection graph, many millions of neurons are likely involved in the processing. It would seem that at least the earliest stages of visual processing will have to be carried out in architectures capable of performing an enormous number of computations involving a large amount of state — corresponding to existing pattern memory — with very low latencies among the processing units. Hybrid architectures that combine conventional processors with co-processors that provide fast matrix-matrix and matrix-vector operations will likely be necessary to handle even a single video stream in real-time.

Geoff Hinton [2005, 2006] has suggested that a single learning rule and a relatively simple layer-by-layer method of training suffices for learning invariant features in text, images, sound and even video. Yoshua Bengio, Yann LeCun and others have also had success with such models [LeCun and Bengio 1995; Bengio, Lamblin, Popovici, and Larochelle 2007; Ranzato, Boureau, and LeCun 2007]. Hyvärinen *et al* [2003], Bruno Olshausen and Charles Cadieu [2007, 2008], Dean *et al* [2009] and others have developed hierarchical generative models to learn sparse codes resembling the responses of neurons in the medial temporal cortex of the dorsal pathway. In each case, the relevant computations can be most easily characterized in terms of linear algebra and implemented using fast vector-matrix operations best carried out on a single machine with lots of memory and many cores (graphics processors are particularly well suited to this sort of computation).

A more vexing problem concerns how we might efficiently implement any of the current models of Hebbian learning in an architecture that spans tens of thousands of machines and incurs latencies measured in terms of milliseconds. Using super computers at the national labs, Eugene Izhikevich and Gerald Edelman [2008] have performed spike-level simulations of millions of so-called *leaky integrate and fire* neurons with fixed, static connections to study the dynamics of learning in such ensembles. Paul Rhodes and his team of researchers at Evolved Machines have taken things a step further in implementing a model that allows for the dynamic creation of edges by simulating dendritic tree growth and the chemical gradients that serve to implement Hebbian learning. In each case, the basic model for a neuron is incredibly simple when compared to the real biology. It is not at all surprising that Henry Markram and his colleagues at EPFL (Ecole Polytechnique Fédérale de Lausanne) require a powerful supercomputer to simulate even a single cortical column at the molecular level. In all three of these examples, the researchers use high-performance computing alternatives to the cluster-of-commodity-computers distributed architectures that characterize most industrial data warehouses. While the best computing architecture for simulating cortical models may not be clear, it is commonly believed that we either how have or soon will have the computing power to simulate significant portions of cortex at some level of abstraction. This assumes, of course, that we can figure out what the cortex is actually computing.

7 Beyond Early Visual Processing

The grid of columnar processing units which constitutes the primate cortex and the retinotopic maps that characterize the areas participating in early vision, might suggest more familiar engineered vision systems consisting of frame buffers and graphics processors. But this analogy doesn't even apply to the simplest case in which the human subject is staring at a static image. As pointed out earlier, our eyes make large — up to 90° of visual angle — movements several times a second and tiny adjustments much more often.

A typical saccade of, say, 18° of visual angle takes 60–80ms to complete [Harwood, Mezey, and Harris 1999], a period during which we are essentially blind. During the subsequent 200–500ms interval until the next saccade, the image on the fovea is relatively stable, accounting for small adjustments due to micro saccades. So even a rough model for the simplest sort of human visual processing has to be set against the background of two or three fixations per second, each spanning less than half a second, and separated by short — less than 1/10 of a second — periods of blindness.

During each fixation we have 200–500ms in which to make sense of the events projected on the fovea; simplifying enormously, that's time enough to view around 10–15 frames of a video shown at 30 frames per second. In most of our experience, during such a period there is a lot going on in our visual field; our eyes, head and body are often moving and the many objects in our field of view are also in movement, more often than not, moving independent of one another. Either by focusing on a small patch of an object that is motionless relative to our frame of reference or by performing smooth pursuit, we have a brief period in which to analyze what amounts to a very short movie as seen through a tiny aperture. Most individual neurons have receptive fields that span an even smaller spatial and temporal extent.

If we try to interpret movement with too restrictive a spatial extent, we can mistake the direction of travel of a small patch of texture. If we try to work on too restrictive a temporal extent, then we are inundated with small movements many of which are due to noise or uninteresting as they arise from the analog of smooth camera motion. During that half second or so we need to identify stable artifacts, consisting of the orientation, direction, velocity, etc., of small patches of texture and color, and then combine these artifacts to capture features of the somewhat larger region of the fovea we are fixating on. Such a combination need not entail recognizing shape; it could, for example, consist of identifying a set of candidate patches, that may or may not belong to the same object, and summarizing the processing performed during the fixation interval as a collection of statistics pertaining to such patches, including their relative — but not absolute — positions, velocities, etc.

In parallel with processing foveal stimuli, attentional machinery in several neural circuits and, in particular, the lateral intraparietal cortex — which is retinotopically mapped when the eyes are fixated — estimates the saliency of spatial locations throughout the retina, including its periphery where acuity and color sensitivity are poor. These estimates of "interestingness" are used to decide what location to saccade to next. The oculomotor system keeps track of the dislocations associated with each saccade, and this locational information can be fused together using statistics collected over a series of saccades. How such information is combined and the exact nature of the resulting internal representations is largely a mystery.

The main point of the above discussion is that, while human visual processing may begin early in the dorsal and ventral pathways with something vaguely related to computer image processing using a fixed, spatially-mapped grid of processing and memory units, it very quickly evolves into a process that requires us to combine disjoint intervals of relatively stable imagery into a pastiche from which we can infer properties critical to our survival. Imagine starting with a collection of snapshots taken through a telephoto lens rather than a single high-resolution image taken with a wide-angle lens. This is similar to what several popular web sites do with millions of random, uncalibrated tourist photos.

The neural substrate responsible for performing these combinations must be able to handle a wide range of temporal and spatial scales, numbers and arrangements of inferred parts and surfaces, and a myriad of possible distractions and clutter irrelevant to the task at hand. We know that this processing can be carried out on a more-or-less regular grid of processors — the arrangement of cortical columns is highly suggestive of such a grid. We are even starting to learn the major pathways — bundles of axons sheathed with myelin insulation to speed transmission — connecting these biological processors using diffusion-tensor-imaging techniques. What we don't know is how the cortex allocates its computational resources beyond those areas most directly tied to the peripheral nervous system and that are registered spatially with the locations of the sensors arrayed on the periphery.

From a purely theoretical standpoint, we can simulate any Turing machine with a large enough Boolean circuit, and we can approximate any first-order predicate logic representation that has a finite domain using a propositional representation. Even so, it seems unlikely that even the cortex, with its 10^{11} neurons and 10^{15} connections, has enough capacity to cover the combinatorially many possible arrangements of primitive features that are likely inferred in early vision. This implies that different portions of the cortex must be allocated dynamically to perform processing on very different arrangements of such features.

Bruno Olshausen [1993] theorized that neural circuits could be used to *route* information so that stimuli corresponding to objects and their parts could be transformed to a standard scale and pose, thereby simplifying pattern recognition. Such transformations could, in principle, be carried out by a graphical model. The neural circuitry that serves as the target of such transformations — think of it as a specialized frame buffer of sorts — could be allocated so that different regions are assigned to different parts — this allocation being an instance of the so-called symbol binding problem in connectionist models [Rumelhart and McClelland 1986] of distributed processing.

8 Escaping Retinotopic Tyranny

While much of the computational neuroscience of primate vision seems mired in the first 200 milliseconds or so of early vision when the stimulus is reasonably stable and the image registered on the fovea is mapped retinotopically to areas in V1 through V4, other research on the brain is revealing how we keep track of spatial relationships involving the frames of reference of our head, body, nearby objects, and the larger world in which we operate. The brain maintains detailed maps of the body and its surrounding physical space in the hippocampus and somatosensory, motor, and parietal cortex [Rizzolatti, Sinigaglia, and Anderson 2007; Blakeslee and Blakeslee 2007]. Recall that the dorsal — "where" and "how" — visual pathway leads to the parietal cortex, which plays an important role in visual attention and our perception of shape. These maps are dynamic, constantly adapting to changes in the body as well as reflecting both short- and long-term knowledge of our surroundings and related spatial relationships.

When attempting to gain insight from biology in building engineered vision systems, it is worth keeping in mind the basic *tasks* of evolved biological vision systems. Much of primate vision serves three broad and overlapping categories of tasks: recognition, navigation and manipulation. Recognition for foraging, mating, and a host of related social and survival tasks; navigation for exploration, localization and controlling territory; manipulation for grasping, climbing, throwing, tool making, etc.

The view [Lengyel 1998] that computer vision is really just inverse graphics ignores the fact that most of these tasks don't require you to be able to construct an accurate 3-D representation of your visual experience. For many recognition tasks it suffices to identify objects, faces, and landmarks you've seen before and associate with these items task-related knowledge gained from prior experience. Navigation to avoid obstacles requires the ability to determine some depth information but not necessarily to recover full 3-D structure. Manipulation is probably the most demanding task in terms of the richness of shape information apparently required, but even so it may be that we are over-emphasizing the role of static shape memory and under-emphasizing the role of dynamic visual servoing — see the discussion in [Rizzolatti, Sinigaglia, and Anderson 2007] for an excellent introduction to what is known about how we understand shape in terms of affordances for manipulation.

But when it comes right down to it, we don't know a great deal about how the visual system handles shape [Tarr and Bülthoff 1998] despite some tantalizing glimpses into what might be going on the inferotemporal cortex [Tsunoda, Yamane, Nishizaki, and Tanifuji 2001; Yamane, Tsunoda, Matsumoto, Phillips, and Tanifuji 2006]. Let's suppose for the sake of discussion that we can build a graphical model of the cortex that handles much of the low-level feature extraction managed by the early visual pathways (V1 through V4) using existing algorithms for performing inference on Markov and conditional random fields and related graphical models. How might we construct a graphical model that captures the part of visual memory that pools together all these low-level features to provide us with such a rich visual experience? Lacking any clear direction from computational neuroscience, we'll take a somewhat unorthodox path from here on out.

As mentioned earlier, several popular web sites offer rich visual experiences that are constructed by combining large image corpora. Photo-sharing web sites like Flickr, Google Picasa and Microsoft Live Labs PhotoSynth are able to combine

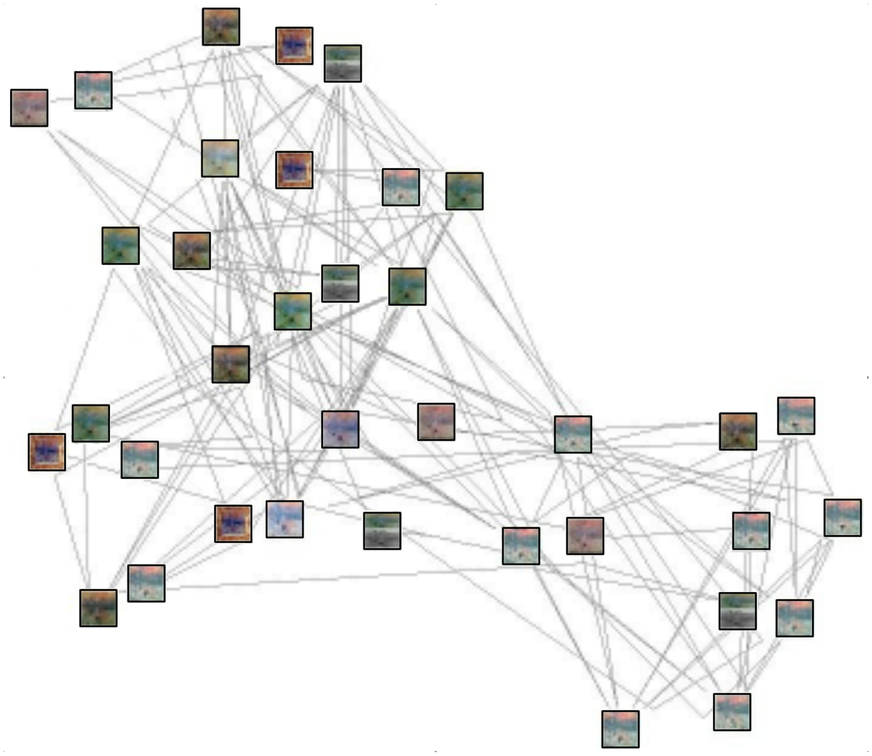


Figure 6. Graphical model with vertices corresponding to image patches and edges representing various relationships among the image patches (after Jing and Baluja [2008]).

multiple snapshots to construct views of popular landmarks from viewpoints not represented by any one snapshot. Google StreetView stitches together video and high-resolution wide-angle images to provide a seamless experience of virtually driving down a street in your home town. Google Earth combines images from satellite, aircraft, ground-based vehicles and, now, sonar-equipped ships and submersibles to allow us explore extensive regions of the planet. These applications are possible due to fast structure-from-motion algorithms that use reliably recoverable and locally distinctive features called *keypoints* extracted from pairs of images to align the images and stitch them together, blending the shared portions and adjusting the color and contrast of the composite to create the illusion of a single image. It is worth noting that these popular web sites facilitate the basic tasks of biological vision that were listed earlier: find me images of a particular famous face, show me a photo of Mount Rainier taken from a site in Tacoma, Washington, tell me what my hotel in New York would look like if I approached it from the direction of Penn Station.

What if the cortex simply memorizes every *novel* fixated foveal patch that spans some fixed-width receptive field and relates them by using low-level features extracted in V1 through V4 as keypoints to estimate geometric and other meaningful relationships among patches? The use of the word “novel” in this context is meant to convey that some method for statistical pooling of similar patches is required to avoid literally storing every possible patch. This is essentially what Jing and Baluja [2008] do by taking a large corpus of images, extracting low-level features from each image, and then quantifying the similarity between pairs of images by analyzing the features that they have in common. The result is a large graph whose vertices are images and whose edges quantify pair-wise similarity (see Figure 6). By using the low-level features as indices, Jing and Baluja only have to search a small subset of the possible pairs of images, and of those only the ones that pass a specified threshold for similarity are connected by edges. Jing and Baluja further enhance the graph by using a form of spectral graph analysis to rank images in much the same way as Google ranks web pages. Torralba *et al* [2007] have demonstrated that even small image patches contain a great deal of useful information, and furthermore that very large collections of images can be quickly and efficiently searched to retrieve semantically similar images given a target image as a query [Torralba, Fergus, and Weiss 2008].

In principle, such a graph could be represented as a probabilistic graphical model and the spectral analysis reformulated in terms of inference on graphical models. The process whereby the graph is grown over time, incorporating new images and new relationships, currently cannot be formulated as inference on a graphical model, but it is interesting to speculate about very large, yet finite graphs that could evolve over time in response to new evidence. Learning the densities used to quantify the edges in graphical models *can* can be formulated in terms of hyper-parameters directly incorporated into the model and carried out by traditional inference algorithms [Buntine 1994; Heckerman 1995]. Learning graphs whose size and topology change over time is somewhat more challenging to cast in terms of traditional methods for learning graphical models. Graph size is probably not the determining technical barrier however. Very large graphical models consisting of documents, queries, genes, and other entities are now quite common, and, while exact inference in such graphs is typically infeasible, approximate inference is often good enough to provide the foundation for industrial-strength tools.

Unfortunately, there is no way to tie up the many loose ends which have been left dangling in this short survey. Progress depends in part on our better understanding the brain and in particular the parts of the brain that are further from the periphery of the body where our senses are directly exposed to external stimuli. Neuroscience has made significant progress in understanding the brain at the cellular and molecular level, even to the point that we are now able to run large-scale simulations with some confidence that our models reflect important properties of the biology. Computational neuroscientists have also made considerable progress developing models — and graphical models in particular — that account for features that appear to play an important role in early visual processing. The barrier to further progress seems to be the same impediment that we run into in so many other areas of computer vision, machine learning and artificial intelligence more generally, namely the problem of representation. How and what does the brain represent about the blooming, buzzing world in which we are embedded? The answer to that question will take some time to figure out, but no doubt probabilistic graphical models will continue to provide a powerful tool in this inquiry, thanks in no small measure to the work of Judea Pearl, his students and his many collaborators.

References

Anderson, J. and J. Sutton (1997). If we compute faster, do we understand better? *Behavior Ressearch Methods, Instruments and Computers 29*, 67–77.

Bengio, Y., P. Lamblin, D. Popovici, and H. Larochelle (2007). Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems 19*, pp. 153–160. Cambridge, MA: MIT Press.

Bienenstock, E. and S. Geman (1995). Compositionality in neural systems. In M. Arbib (Ed.), *The Handbook of Brain Theory and Neural Networks*, pp. 223–226. Bradford Books/MIT Press.

Blakeslee, S. and M. Blakeslee (2007). *The Body Has a Mind of Its Own*. Random House.

Brady, N. and D. J. Field (2000). Local contrast in natural images: normalisation and coding efficiency. *Perception 29*(9), 1041–1055.

Buntine, W. L. (1994). Operations for learning with graphical models. *Journal of Artificial Intelligence Research 2*, 159–225.

Cadieu, C. and B. Olshausen (2008). Learning transformational invariants from time-varying natural images. In D. Schuurmans and Y. Bengio (Eds.), *Advances in Neural Information Processing Systems 21*. Cambridge, MA: MIT Press.

Dean, T. (2005). A computational model of the cerebral cortex. In *Proceedings of AAAI-05*, Cambridge, Massachusetts, pp. 938–943. MIT Press.

Dean, T. (2006, August). Learning invariant features using inertial priors. *Annals of Mathematics and Artificial Intelligence 47*(3-4), 223–250.

Dean, T., G. Corrado, and R. Washington (2009, December). Recursive sparse, spatiotemporal coding. In *Proceedings of the Fifth IEEE International Workshop on Multimedia Information Processing and Retrieval*.

Dean, T. and M. Wellman (1991). *Planning and Control*. San Francisco, California: Morgan Kaufmann Publishers.

DeGroot, M. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

DeGroot, M. H. (1986). *Probability and Statistics.* Reading, MA: Second edition, Addison-Wesley.

Dehaene, S. (2009). *Reading in the Brain: The Science and Evolution of a Human Invention.* Viking Press.

Fine, S., Y. Singer, and N. Tishby (1998). The hierarchical hidden Markov model: Analysis and applications. *Machine Learning 32*(1), 41–62.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation 3*, 194–200.

George, D. and J. Hawkins (2005). A hierarchical Bayesian model of invariant pattern recognition in the visual cortex. In *Proceedings of the International Joint Conference on Neural Networks*, Volume 3, pp. 1812–1817. IEEE.

Harwood, M. R., L. E. Mezey, and C. M. Harris (1999). The spectral main sequence of human saccades. *The Journal of Neuroscience 19*, 9098–9106.

Heckerman, D. (1995). A tutorial on learning Bayesian networks. Technical Report MSR-95-06, Microsoft Research.

Hinton, G. and R. Salakhutdinov (2006, July). Reducing the dimensionality of data with neural networks. *Science 313*(5786), 504–507.

Hinton, G. E. (2005). What kind of a graphical model is the brain? In *Proceedings of the 19th International Joint Conference on Artificial Intelligence.*

Hoffman, D. (1998). *Visual Intelligence: How We Create What we See.* New York, NY: W. W. Norton.

Hoiem, D., A. Efros, and M. Hebert (2007). Recovering surface layout from an image. *International Journal of Computer Vision 75*(1), 151–172.

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

Hubel, D. H. and T. N. Wiesel (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology 195*, 215–243.

Hyvärinen, A., J. Hurri, and J. Väyrynen (2003). Bubbles: a unifying framework for low-level statistical properties of natural image sequences. *Journal of the Optical Society of America 20*(7), 1237–1252.

Izhikevich, E. M. and G. M. Edelman (2008). Large-scale model of mammalian thalamocortical systems. *Proceedings of the National Academy of Science 105*(9), 3593–3598.

Jin, Y. and S. Geman (2006). Context and hierarchy in a probabilistic image model. In *Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition*, Volume 2, pp. 2145–2152. IEEE Computer Society.

Jing, Y. and S. Baluja (2008). Pagerank for product image search. In *Proceedings of the 17th World Wide Web Conference*.

Kivinen, J. J., E. B. Sudderth, and M. I. Jordan (2007). Learning multiscale representations of natural scenes using dirichlet processes. In *Proceedings of the 11th IEEE International Conference on Computer Vision*.

Konen, C. S. and S. Kastner (2008). Two hierarchically organized neural systems for object information in human visual cortex. *Nature Neuroscience 11*(2), 224–231.

LeCun, Y. and Y. Bengio (1995). Convolutional networks for images, speech, and time-series. In M. Arbib (Ed.), *The Handbook of Brain Theory and Neural Networks*. Bradford Books/MIT Press.

Lee, T. S. and D. Mumford (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America 2*(7), 1434–1448.

Lengyel, J. (1998). The convergence of graphics and vision. *Computer 31*(7), 46–53.

Mountcastle, V. B. (2003, January). Introduction to the special issue on computation in cortical columns. *Cerebral Cortex 13*(1), 2–4.

Mumford, D. (1991). On the computational architecture of the neocortex I: The role of the thalamo-cortical loop. *Biological Cybernetics 65*, 135–145.

Mumford, D. (1992). On the computational architecture of the neocortex II: The role of cortico-cortical loops. *Biological Cybernetics 66*, 241–251.

Newman, M., D. Watts, and S. Strogatz (2002). Random graph models of social networks. *Proceedings of the National Academy of Science 99*, 2566–2572.

Olshausen, B. and C. Cadieu (2007). Learning invariant and variant components of time-varying natural images. *Journal of Vision 7*(9), 964–964.

Olshausen, B. A., A. Anderson, and D. C. Van Essen (1993). A neurobiological model of visual attention and pattern recognition based on dynamic routing of information. *Journal of Neuroscience 13*(11), 4700–4719.

Olshausen, B. A. and D. J. Field (2005). How close are we to understanding V1? *Neural Computation 17*, 1665–1699.

Osindero, S. and G. Hinton (2008). Modeling image patches with a directed hierarchy of markov random fields. In J. Platt, D. Koller, Y. Singer, and S. Roweis (Eds.), *Advances in Neural Information Processing Systems 20*, pp. 1121–1128. Cambridge, MA: MIT Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, California: Morgan Kaufmann.

Ranzato, M., Y. Boureau, and Y. LeCun (2007). Sparse feature learning for deep belief networks. In *Advances in Neural Information Processing Systems 20*. Cambridge, MA: MIT Press.

Riesenhuber, M. and T. Poggio (1999, November). Hierarchical models of object recognition in cortex. *Nature Neuroscience 2*(11), 1019–1025.

Rizzolatti, G., C. Sinigaglia, and F. Anderson (2007). *Mirrors in the Brain How Our Minds Share Actions, Emotions, and Experience*. Oxford, UK: Oxford University Press.

Rumelhart, D. E. and J. L. McClelland (Eds.) (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume I: Foundations*. Cambridge, Massachusetts: MIT Press.

Russell, S. and P. Norvig (2003). *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ: Second edition, Prentice Hall.

Saxena, A., S. Chung, and A. Ng (2007). 3-D depth reconstruction from a single still image. *International Journal of Computer Vision 76*(1), 53–69.

Sporns, O. and J. D. Zwi (2004). The small world of the cerebral cortex. *Neuroinformatics 2*(2), 145–162.

Tarr, M. and H. Bülthoff (1998). Image-based object recognition in man, monkey and machine. *Cognition 67*, 1–20.

Tenenbaum, J. and H. Barrow (1977). Experiments in interpretation-guided segmentation. *Artificial Intelligence 8*, 241–277.

Torralba, A., R. Fergus, and W. Freeman (2007). Object and scene recognition in tiny images. *Journal of Vision 7*(9), 193–193.

Torralba, A., R. Fergus, and Y. Weiss (2008). Small codes and large image databases for recognition. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, pp. 1–8. IEEE Computer Society.

Tsunoda, K., Y. Yamane, M. Nishizaki, and M. Tanifuji (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience 4*, 832–838.

Ullman, S. and S. Soloviev (1999). Computation of pattern invariance in brain-like structures. *Neural Networks 12*, 1021–1036.

Ullman, S., M. Vidal-Naquet, and E. Sali (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience 5*(7), 682–687.

Wiskott, L. and T. Sejnowski (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation 14*(4), 715–770.

Yamane, Y., K. Tsunoda, M. Matsumoto, N. A. Phillips, and M. Tanifuji (2006). Representation of the spatial relationship among object parts by neurons in macaque inferotemporal cortex. *Journal Neurophysiology 96*, 3147–3156.