PROCEEDINGS OF THE IEEE

The visual system's internal model of the world

Journal:	*Proceedings of the IEEE*
Manuscript ID:	0209-SIP-2014-PIEEE.R1
Manuscript Categories:	Special Issue Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Lee, Tai Sing; Carnegie Mellon, Computer Science Department
Keyword:	Brain modeling, Visual system, Neural network architecture, Computation theory, Memory architecture

The visual system's internal model of the world

Tai Sing Lee, *Member, IEEE*

***Abstract*—The Bayesian paradigm has provided a useful conceptual theory for understanding perceptual computation in the brain. While the detailed neural mechanisms of Bayesian inference are not fully understood, recent computational and neurophysiological works have illuminated the underlying computational principles and representational architecture. The fundamental insights are that the visual system is organized as a modular hierarchy to encode an internal model of the world, and that perception is realized by statistical inference based on such internal model. In this paper, I will discuss and analyze the varieties of representational schemes of these internal models and how they might be used to perform learning and inference. I will argue for a unified theoretical framework for relating the internal models to the observed neural phenomena and mechanisms in the visual cortex.**

***Index Terms* – hierarchical model, visual cortex, Bayesian inference, neural circuits, computational theories, internal models.**

I. INTRODUCTION

The general theory of the perceptual computation is that the visual system performs Bayesian inference. This idea can be traced back to Helmholtz's [1] theory of unconscious inference, which states that the brain transforms the noisy and often impoverished 2D optical image impinged on the retina into a perceptual interpretation of the 3D world. This transformation involves a probabilistic computation that finds the most likely interpretation of the world, based in part on prior knowledge from experience, to explain the retinal input. This prior knowledge is a form of memory, or what we call an *internal model* of the world, for supporting Bayesian inference. What is the nature of this internal model? How does the brain build it and how it is used to make inferences of the world?

The complexity of the world and its images is daunting. The number of possible images that can be expressed in a gray-level image patch of 30 by 30 pixels is 900^256, practically infinite. Yet, moment by moment, we comfortably analyze a continuous stream of color visual images coming in through our retina as we parse the visual scene – recognizing, in a fraction of a second, objects, their spatial layouts and scene structures. It is almost impossible to encode prior knowledge of such a scale in the brain, even with its billions of neurons. Fortunately, natural images live in a restricted space, a much lower dimensional manifold inside this universe of infinite possibilities. Our visual system must have discovered and exploited the statistical structures of natural scenes in order to build an efficient internal model of the world.

Herb Simon [2] argued that the only way to model complexity is through hierarchical models, which should have the property of near-decomposability that allows modularization and compartmentalization of functions. A nearly decomposable modular hierarchical system separates high frequency dynamics and fast communication within a module, and low frequency computational dynamics with sparser and slower communication across modules. Simon argued that hierarchy and modularity are inevitable: Among evolving systems, only those that managed to obtain and then reuse stable subassemblies (modules) are likely to be able to search through the fitness landscape with reasonable speed. Thus, among possible complex forms, modular hierarchies are the ones that have time to evolve and survive.

The visual system is indeed such a modular hierarchical system, with its 30 or so visual areas arranged in a hierarchical organization (Figure 1). Each area specializes in certain functions [3], potentially concealing most aspects of its internal computations from others. These visual areas do interact with each other and perceptual experience emerges from the collective computation resulting from such interactions. Each visual area follows the design of a near-decomposable system, recursively organized in different modules and sub-modules. Thus, the visual cortex is in itself a form of a hierarchical memory system that encodes the brain's internal model of the visual world.

II. VARIETIES OF INTERNAL MODELS

Five major classes of computational models have been proposed over the last 40 years on how the hierarchical internal models in the visual cortex are constructed and function to support perceptual learning and inference. While they were all inspired by the hierarchical architecture of the biological visual system and share many fundamental characteristics, they represent different perspectives how the internal model can be learned and can be used for inference,

Tai Sing Lee is a professor in the Computer Science Department and the Center for the Neural Basis of Cognition, Carnegie Mellon University, Rm 115, Mellon Institute, 4400 Fifth Avenue, Pittsburgh, PA 15213, U.S.A. (e-mail: tai@ cs.cmu.edu).

This work was supported in part by the National Science Foundation (CISE) and National Institute of Health (NEI).

and in my opinion, each capturing or emphasizing certain elements of the reality of the brain.

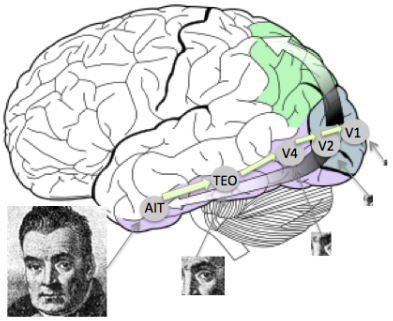

Figure 1: The visual cortex is arranged in a hierarchy of different visual areas (modules), starting from V1 that receives retinal-thalamic input, flowing to V2, V4, TEO, and AIT (anterior inferotemporal cortex). These areas form the ventral (WHAT) stream processing object forms, colored in violet. The dorsal stream, (WHERE) colored green, is associated with the parietal cortex, processing information about space and motion. The receptive fields of the neurons, indicated by the accompanying image patches, become progressively larger as the information propagates up the hierarchy, allowing the neurons to process and encode more global and abstract aspects of the input images.

A. Class I: Neocognitron, HMAX and CNN

The first class of hierarchical models of the visual cortex, starting with Fukushima's Neocognitron [4], is a feedforward multi-layer neural network. It primarily models the ventral stream of the visual hierarchy (V1, V2, V4, IT), i.e., the object recognition pathway. Along the model hierarchy, as in the visual system, neurons develop more complex and larger compound feature detectors from component detectors in the previous layer, with gradually increased tolerance to position, scale and rotation deformations of the feature detectors at each level. Orientation and position specific edge detectors in V1 are combined to articulate tunings to corners, junctions and curves in V2 and V4, culminating into "grandmother" neurons in the inferotemporal cortex (IT) that are selective to specific views of a particular class of objects. A central computational issue is how one can construct feature detectors that are highly specific on one hand and yet invariant to irrelevant variations on the other. For example, a neuron that responds only to an image of a cat (no matter how it is seen – in different views, lighting conditions, and spatial locations in the world) is called the specificity and invariance dilemma [5].

Inspired by Hubel and Wiesel's [6] discovery of simple and complex cells, Fukushima [4] suggested that simple cells compute conjunctions of features, and complex cells introduce local invariance to these feature detectors. In the case of V1, a simple cell is an oriented edge detector, and a complex cell integrates responses of simple cells of the same orientation in a small spatial neighborhood. This provides an oriented edge label that is insensitive to local shift, or the exact luminance appearance of the edge, representing a more abstract concept. Fukushima imagined that this simple/complex cell scheme could be repeated in each visual area along the hierarchy, as a strategy to gradually achieve specificity and invariance. A simple cell in one layer combines output from different complex cells in the previous layer to form compound feature detectors, as forming letters from strokes, then words from letters. A complex cell introduces tolerance to local deformation in position, scale and orientation for detectors in each layer. Such a cascade can develop neurons with high levels of selectivity to particular objects. At the same time, they can simultaneously achieve a high degree of invariance against variations in size, position and views of these objects.

LeNet [7] and HMAX [8] are recent refinements of similar ideas. They follow a similar architecture, but with richer representations empowered by more powerful learning and classification methods. When tuned well with supervised or unsupervised learning methods, these feedforward models can be highly effective in some object recognition tasks [9], [7], [10], [10]. Because these models conform to conventional wisdom that the visual cortex performs computation mostly in a series of feedforward modules, and can explain some behaviors of IT neurons, they are the most publicly recognized hierarchical models of the visual cortex. With the help of big data, convolution neural networks presently offer the best state-of-the-art performance in object recognition in computer vision and in speech recognition[11], [12].

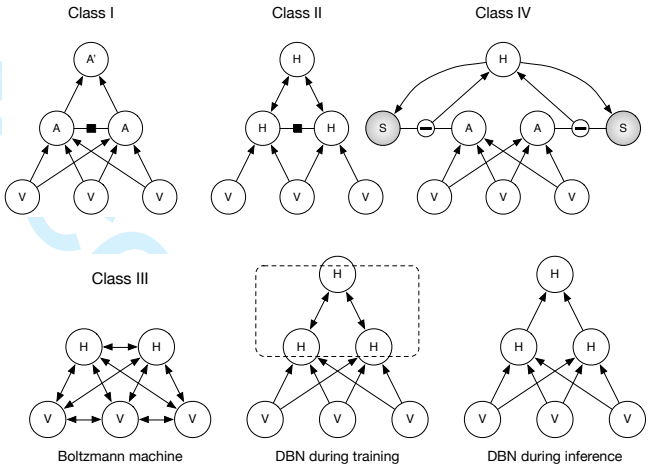

Figure 2: Schematics of the first four classes of models. Note that for class III, a DBN (deep belief net) uses a restricted Boltzmann machine with top-down inference and bottom-up fantasy to learn the internal model (dotted block). But once learned, inference is performed simply by feedforward computation.

B. Class II: Interactive Activation and Adaptive Resonance

The second class of models is called the interactive activation model [13] or the adaptive resonance model [14] in the connectionist literature. It is motivated by the psychological observation that visual perception is a constructive process and that global percept can influence low-level sensory processing. For example, visual perception of an individual letter can be influenced by the perception of the entire word. When the four letters WQRD are shown briefly, they can be perceived as WORD, with the Q

"mistakenly" perceived as an O. This is called word-superiority effect. The essential idea is that the visual system is interconnected in a recurrent fashion. Higher level interpretations can feed back to influence low-level processing. Within each area, there is a competition mechanism to suppress spurious noises or alternative interpretations. Thus, the activation of the WORD neuron in the higher area will feed back to suppress the representation of Q and enhance the representation of O at the earlier areas. The key distinction of this class of models with the first class is in its emphasis on perception being an interactive process, involving top-down feedback. The model is based on the association principle and can be tied to constraint satisfaction and probabilistic models of perception [15], [16]. O' Reilly et al.'s [17] Leabra cognitive architecture may be the most powerful instantiation of this class of models for recurrent processing in object recognition.

C. *Class III: Boltzmann machines and Deep Belief Nets*

Models in the first class compute by deterministic feedforward projection. Models in the second class are essentially dynamical systems that compute constraint satisfaction problems by gradient descent on an energy landscape. The energy functions or the objective functions for vision problems are usually not convex. Thus, these dynamical system models could be trapped into suboptimal solutions. To account for the flexibility in human inference, Hinton and Sejnowski [18] proposed the Boltzmann machine to model learning and inference in a statistical inference framework. The proposed Boltzmann machine is essentially a stochastic version of the dynamic Hopfield net, but now the energy function is interpreted as a joint distribution between the states of the different neurons. Sampling techniques such as Gibbs sampling and Markov chain Monte Carlo are used to estimate this distribution. The pairwise connections in the energy function of the Boltzmann machine are based on Hebbian learning – neurons firing together during a visual experience will be wired together to encode the statistical priors of stimulus correlations in the world. The Boltzmann machine learns by generating fantasy according to the encoded priors or internal models. Unsupervised learning is fueled by the need to match the statistical correlations between the states of neurons in the network during fantasy (or in dreams) to those that occur during natural experiences.

Despite its conceptual appeal, the Boltzmann machine was not very useful at the beginning because Gibbs sampling, used for both learning and inference in such a system, is very slow. When Boltzmann machines are restricted to have only the feedforward and feedback connections between visible and hidden units, without horizontal connections between units in each layer, the sampling process can become easy and fast, as one can fix one layer and make fast inference on the states of units in the second layer. Connections in these so-called restricted Boltzmann machines can be learned between two layers at a time, and then stacked up together to form deep belief nets [19], [20] to build a complex internal model of the world (see also the Helmholtz machine [21]). The dynamic interaction of these neurons in the Boltzmann machine corresponds to computing probabilistic inference, and tuning of the synaptic weights corresponds to learning the parameters of an internal probabilistic model of the visual input space. However, deep learning networks of this kind utilize feedback only during learning; inference is approximated by feedforward deterministic computation. Hinton believes that visual perception has to happen fast, yet Gibbs or MCMC sampling is slow. Thus, during perception, we don't fantasize.

D. *Class IV: Predictive Coding Model*

Inspired by Grenander's [22] analysis by synthesis theory of inference, and by Burt and Adelson's [23] Laplacian pyramid for efficient coding, Mumford [24] proposed that the visual hierarchy might form an efficient image pyramid, and that vision is accomplished by an analysis through a synthesis loop. Here, the feedforward computation extracts features and proposals, and feedback computation synthesizes expectations based on the high level of interpretation of the proposals using the internal models. The mismatch between the top-down prediction and the input produces a prediction error signal or residue that can be used to update the internal models at the higher level to refine the synthesized top-down "prediction" until all the meaningful inputs are explained [24]. In his scheme, each layer of the hierarchy only needs to encode the residues as in the Laplacian pyramid to form an efficient *image pyramid* that can describe an image with an minimum description length (MDL) code (e.g., Figure 4) in accord with Barlow's [25] efficient coding principle. Rao and Ballard [26] developed a Kalman filter-based "predictive coding" model to implement this idea. They used it to explain some of contextual modulation effects and end-stopping effects in V1 neurons. The predictive coding idea has recently become quite popular in some circles of the cognitive neuroscience community. It has been generalized to non-visual systems [27], [28] and even to a "unified theory" of the brain [29]. In Rao and Ballard's [26] model, the representation of a static input image at each level was maintained over time. It is set up by the initial feedforward computation and subsequently modified by both top-down and bottom-up signals. While this does not exactly satisfy Mumford's MDL pyramid motivation, the system is efficient in that the lower level only needs to feed forward the residue error signals in subsequent feedforward communication after the initial volley. After all, why waste energy sending the bottom-up representation again once it has already been sent! It is important to note that deep belief nets (class III) can also be made to synthesize images during inference by unfolding the feedback path into a feedforward path in the form of auto-encoders [20]. Auto-encoders [20] or predictive coding models [26] can "code" current input with intermediate and higher level features, which are useful for classification. However, they cannot predict future events.

E. Class V: Hierarchical Bayes and Compositional System

The hierarchical Bayesian inference theory that I developed jointly with David Mumford [30] is an attempt to unify and reconcile the first four classes of models. It accepts the basic feed-forward computational architectural design of Neocognitron and convolutional neural networks, but emphasizes the critical functional roles of feedback shared by the interactive activation model and the predictive coding model, and argues for a statistical inference framework similar to that of the Boltzmann machine. The theory is motivated in part by the evidence of higher order feedback modulation observed in the primary visual cortex, and the enormous number of intrinsic horizontal connections in V1. Similar to the interactive activation model [13] and Ullman's [31] earlier proposal, signals, now in the form of beliefs, are propagating up and down the hierarchical network to update the representations at each level. However, every visual area is also endowed with its own intrinsic computational circuitry, modeling the visual concepts and their statistical relationship at a particular level, to perform probabilistic inference. In V1, its intrinsic horizontal connectivity encodes the geometric priors of contours and surfaces of the natural environment. Thus, V1 should be considered a geometric computing engine (a high-resolution buffer) engaging in all levels of visual computations that require high spatial resolution and fine details. This includes computations such as tracing contours, or "coloring" an object precisely with saliency signal, or constructing a mental representation of our perception.

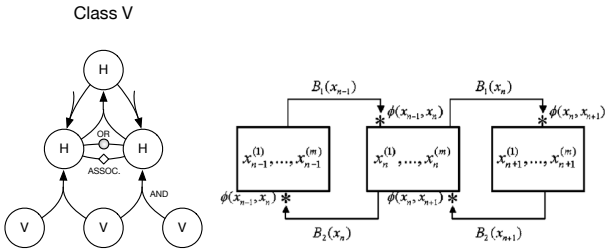

Figure 3: Schematic of the class V hierarchical Bayes model. Beliefs are passed bottom-up and top-down across modules (layers) via different paths. Within each module (layer), neurons are modulated by competitive mechanisms and facilitatory association field mechanisms.

In each visual area, computation is performed based on bottom-up, top-down and horizontal information. Horizontal intrinsic connections are important for encoding statistical relationships between the visual concepts encoded in a particular area. They go beyond local competition or the smoothness constraint in computer vision. For example, horizontal connections can encode the co-occurrence priors of edge detectors [32], [33] and higher order priors [34], [35] for contour completion, as well as co-occurrence in visual cues such as texture, shading, and binocular disparity in terms of association field for surface interpolation [36], [37], [38], [39], [35]. More generally, association fields and contextual relationships between different visual events across time and across space can be encoded in latent contextual variables to effect prediction and interpolation [40].

In our proposal, feedback itself does not have to explicitly synthesized an expected image to be compared with the input image or the representation in the previous level, as the predictive coding model [26] suggested. We reasoned that since V2 representation is more abstract and invariant, it does not have the precision to synthesize a high-resolution image to feed back to V1. Rather, feedback provides a set of global beliefs from a higher order extrastriate cortex as high level commands and instructions to condition and "work with" V1's own intrinsic circuitry to synthesize a precise internal representation of the prediction of the input to V1 (see also [41]. The prediction error of the synthesized representation and the bottom-up input representation is used to select, update and revise the cortical internal models for generating further predictions.

It is important to recognize that signals reaching V1 are 40 ms behind the occurrence of the actual events and it is 80 ms later in IT. At each moment in time, our visual cortex is basically processing signals of past events. Cortical computation for integrating evidence, reasoning about them and planning motor action will take additional time. Decisions based on past events will always be reactive. While a 100-150 ms delay might not be critical, if our internal models can "predict" what is actually happening in the world right now or in the future, our brain can then reason and act proactively. From this perspective, the brain could be operating entirely on a "grand illusion" of mental representations synthesized from internal models, and sensory signals (from the past), arriving later into the cortex and serving only to validate and revise past predictions of the internal models. The prediction error signals between the synthesized representation and the "observed" representation can be used to train the predictive aspects of the internal model. In this scheme, inference and learning happen continuously and simultaneously instead of happening in distinct phases, as in deep belief nets. The input signals or reality provide the teaching signals for training the internal model.

In order to allow multiple hypotheses to be simultaneously represented in terms of probability distribution, Mumford and I [30] suggested that probability distribution of hypotheses can be encoded non-parametrically as particles and that hierarchical inference can simultaneously proceed top-down, bottom-up and horizontally, using particle filtering or loopy Bayesian belief propagation.

A number of computational hierarchical models using belief propagation have been constructed to realize these ideas [42], [43], [44]. Hawkins' hierarchical temporal memory model [44], in particular, further emphasized the importance of the role of time and argued that visual scene processing is a serial dynamic process. Thus, it is important to encode visual events in terms of temporal memory in a modular hierarchy.

But the most concrete realization of these ideas in computer vision are the powerful hierarchical compositional models that S.C. Zhu and Mumford [45], L. Zhu and Yuille [46], [47] have developed. These are graphs of structured probabilistic distributions that allow explicit modeling of the relationship

between parts (horizontal edges), and the relationship between parts and whole (vertical edges). These allow a dynamic flexible recombination of reconfigurable parts to overcome the complexity issue. Figure 4 shows an example of an AND/OR graph with the highlighted edges indicating the parsing tree of a particular clock perceived (shown at the top of the graph) [45].

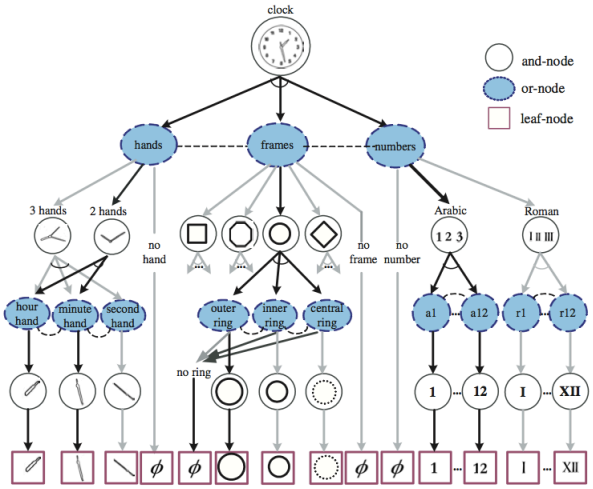

Figure 4: An example of an AND/OR graph that encodes the concept of a clock [45]. The graph encodes all varieties of clocks. The thick dark lines indicate a particular parsing tree related to the clock displayed at the top, which is also the input image tied to the leave nodes at the bottom. Reprinted from [45] with permission from authors.

These models have proved to be quite effective in unsupervised learning of object and scene representation, object recognition and scene parsing [46], [47], [45].

The top node of Figure 5 does not have to be an object; it could represent an entire scene. In fact, in a real visual scene, multiple objects are usually present simultaneously in a complex scene. The purpose of perception cannot simply be for identifying or classifying an object, scene or event, but instead is for understanding the relationship between the participating objects and the mutual consequence of their interactions. The hierarchical compositional models in computer vision mentioned above have begun to address the issues of both scene composition, and modeling the relationship of objects in a complex scene to solve the scene parsing problem. These models, sometimes modeled as an AND/OR graph, provide a stochastic grammar framework for understanding vision in a manner similar to that of language. From that perspective, scene parsing is analogous to parsing a paragraph, while object recognition is like parsing a sentence.

III. Elements of Representations

In this section, we will discuss some observations of neural mechanisms and architecture that provide insights into the construction and operations of the internal models in the visual system.

There are three major representational elements. First, at both the macro level and micro level, the visual system might be organized into modules in many different levels. This gives flexibility in representation to deal with the complexity and scalability issues by allowing flexible reconfiguration of parts and the flexible association of the parts and the whole concept. Second, at each module, there is a dictionary of visual concepts represented either in the tunings of individual neurons or in the population codes of neuronal assembly. In addition, the probability distributions and uncertainty of these visual concepts need to be represented. Third, the relationships between the dictionary elements can be encoded in the functional connectivity among the neurons as an integral part of the internal model for probabilistic inference.

A. Hierarchical and Modular Organization

Visual scene is a hierarchical composition of "stuff". A forest scene is composed of trees and rocks. A tree in turn is composed of its trunk, branches, leaves, and so on. That means it should be decomposable into its constituent parts. Our internal model of the world in the visual cortex might be best expressed in terms of a modular hierarchy, with a feedforward composition mechanism and feedback decomposition mechanism. The visual system appears to be factorized into a modular hierarchy recursively, allowing a division of labors at many levels, with each visual area representing specific information and performing specific functions. Within each visual area, it is divided into different modules or hypercolumns and mini-columns or neuronal pools. Neurons within a mini-column (e.g. orientation column) encodes similar features, whereas a hypercolumn denotes a unit containing a full set of values for any given set of receptive field parameters.

Mathematically, decomposition of a function into non-interacting components generally permits a more economical and compact representation of the function. For example, consider y is a function of 10 binary variables, $y = f(x_1, x_2, x_3, x_4, \dots, x_{10})$. The number of possible values it can assume is 2^10=1024. However, if f can be factored into $\{g_1, g_2, g_3\}$ where there are 2^3=8 values in $g_1(x_1, x_2, x_3)$ and 4 in $g_2(x_4, x_5,)$ and 32 values in $g_3(x_6, x_7, x_8, x_9, x_{10})$, then, with decomposition, the total number of possible values that need to be considered for f is only 44. Similarly, consider a 3D tensor, with each dimension = N, the tensor will be defined by N^3 possible numbers. If, however, it can be decomposed into and represented by a set of p principal components where $(p<<N)$, and each is defined by the outer product of three rank 1 vectors, then we only need $3N$ numbers to represent a component, and there are only $3pN$ numbers in the hypothesis space.

From a statistical modeling perspective, if the interdependency of variables is relatively local, the brain's internal model of the probability distribution of visual scenes can be factorized into a product of potential functions, each of which has much *lower dimensionality* than the original

distributions.

$$P(I,Z) \propto \prod_a \phi_a(\vec{x}_a)$$

where $\vec{x}_a$ is some subset of variables in set a in I and Z, and where I is the input image and Z the internal model. Such models, when certain constraints are satisfied, can be expressed in terms of a Bayes network, or for other constraints, as a Markov random field (MRF), organized in a hierarchy. MRFs are simply large, multivariate probability distributions that can be expressed as the product of factors, each of which is the function of a small subset of the variables in the original distribution. By isolating the variables within each potential function from the rest of the system, a modular hierarchical system allows more compact and economical representations, minimizing connections, and making learning and inference tractable.

The factorized modules allow intense and fast computation to happen within each module with dense connections. The modules need to interact, albeit with slower dynamics and sparser connections, in order to realize the global function collectively. Neurons with similar tunings often cluster together to form columns in a topological map to facilitate their interaction.

Information flow in cortical modules is characterized by vertical connections in columns and horizontal connections in layers. Each visual area (module) is composed of 6 main layers, with similar topological maps across layers. In the early visual area, neurons in each column have overlapping receptive fields, i.e., they analyze the same window in the visual space. The vertical flow of information is across layers within the same column. The basic information is as follows: bottom-up input primarily arrives at layer 4 (though some also reach layer 2 and layer 5), projects to layers 2 and 3, then to layers 5 and 6, and then back to layer 4, forming a closed loop. Layers 2+3 neurons project up to layer 4 of the higher visual areas. Layers 5+6 project back to layers 2+3 and 5+6 of the lower visual areas. Neurons within layers 2+3 emanate dense local horizontal axonal collaterals to form networks [48]. Neurons within layers 5+6 form their own horizontal networks. Layers 2+3 project to and receive feedback from higher areas, while layers 5+6 project to lower visual areas or subcortical areas. The recurrent horizontal and feedback connections introduce contextual effects on the responses of the neurons. Neurons do not fire simply based on the stimulus presented to their receptive fields, but depend on local and global context [49], [50], [30].

Neurons in one module provide both convergent and divergent projections to neurons in the other modules: V1 neurons from multiple hypercolumns will project to a single column in V2, and a single V1 neuron will also provide a divergent projection to multiple columns in V2. This results in the expansion of the receptive fields of neurons across different visual areas. A V2 neuron's receptive field is twice the diameter of a V1 neuron's receptive field, while a V4 neuron's receptive field is twice the diameter of that of a V2 neuron at the same eccentricity [51]. As one moves up the hierarchy, the neurons become more selective to more complex features, yet with greater and greater tolerance to translation, scaling and, to a certain extent, rotation. The modular architecture at different scales or levels of the hierarchy allows flexible composition of the visual concepts represented in the modules to create more global new visual concepts.

B. *Dictionary of Visual Concepts*

Let us consider the visual cortex as a graph, then the nodes of the graph represent "visual concepts". How could these concepts be learned? How could these concepts be represented by neurons and neuronal populations?

A visual concept can be considered as a conjunction of a set of more elementary concepts or features. They are produced by feedforward connections through an AND operation, and are encoded in the form of receptive field tuning of individual neurons, or potentially in the tuning of neuronal populations.

The first principle of concept learning is that of suspicious coincidence, proposed by Barlow [25]. If multiple features occur frequently together, it is likely they originate from the same visual event, e.g., when parts of an object always occur together, they will be grouped together, and a neuron will be created to encode the conjunction of these features as a "higher order concept". This is the basic point of the "simple cell" operation in each layer of the Neocognitron or the AND operation in the AND/OR graph.

The second principle of concept learning is called redundancy reduction, also from Barlow [25]. If a neuron in the next layer is encoding a particular conjunction of features, i.e., expressing a specific visual concept, there should be no need for the other neurons in that layer to encode the same event. This can be implemented by local competition or a winner-take-all mechanism among the neurons [52].

These two principles have been successfully applied to explain the emergence of simple cell receptive fields in the primary visual cortex, in the framework of sparse coding [53], [54], [90], [55], [9], [10]. The local competition for removing redundancy between similarly tuned neurons can be learned by the anti-Hebbian rule [52], [42]. By demanding that the responses of the neuronal population be sparse, the receptive fields of the hidden units in the first layer of many classes of generative models produced by training with whitened natural images all resemble the simple cell receptive fields in V1.

Many of these sparse coding models are based on the "linear receptive fields" assumption, minimizing reconstruction cost function with L1 norm. Recent neurophysiological studies suggest that feedforward computations, even those in the retina, might not be so linear [56]. It was also found that synchronized spikes from distinct LGN neurons provide a much stronger drive to V1 neurons [57], [58] than when they are not synchronized, and that synchronized spikes coming into the same dendritic branch of a neuron produce super-linear responses [59]. Thus, feedforward computation could work as a soft version of an AND operator that integrates evidence to create a conjunctive feature detector. A soft-AND requires the coexistence of several pieces of evidence, but tolerates some missing parts, which could arise due to occlusion or noises, and hence is less

rigid than a simple AND operator. From this perspective, simple cells could be tuned to a great variety of specific features, similar to the rich dictionary of feature detectors learned using K-mean related methods in computer vision [54], [60], [61].

In fact, the ratio between the number of V1 neurons and the number of retinal ganglion cells covering the same visual space is about 2000:1 near the fovea and 200:1 in the peripheral. Such an overcomplete representation can enhance resolution of the analysis in the feature domain and provide flexibility in representation for building models. For any computational problem, finding the best way to represent the problem state space is crucial; some problems can be solved much more easily given the right representation. A single complete representation forces us to decide on only one representation, whereas over-complete representations allow us to retain the benefits of multiple complete representations.

From a feature detector perspective, an over-complete representation allows V1 neurons to tune to a great variety of specific complex local features. A recent stunning Ca imaging experiment of macaque V1 by Tang Shi-Ming and colleagues at Beijing University, monitoring thousands of neurons simultaneously in response to a large array of visual stimuli, suggests that macaque V1 might have an extremely rich and highly selective set of nonlinear feature detectors that are very sparse in their activities at both the individual and population levels.

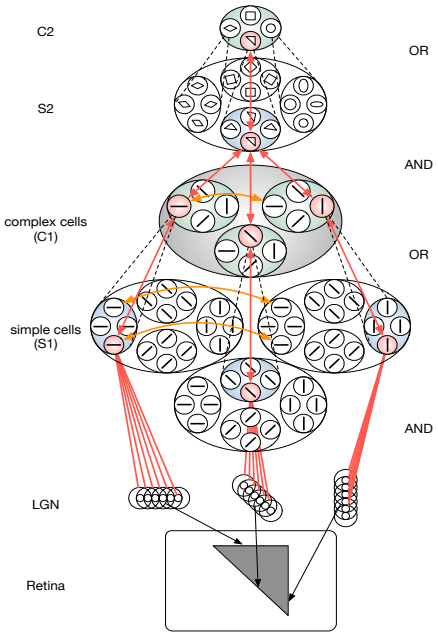

Figure 5: Relating the neural circuit to the theoretical AND/OR graph (Figure 4). Local competition between simple cells coding similar features (tinted blue) at the same location implements redundancy reduction, similar to the OR operation. Local competition between complex cells coding distinct concepts at the same receptive field location (tinted green) normalizes the probability of the different hypotheses. Surround suppression (tinted gray) computes Bayesian surprise based on the predictive coding principle. The association field is learned based on stimulus correlation, linking co-occurring visual events to model relationships between parts. The C1 complex cells converge to form a higher order S2 simple cell (coding for a triangle in this example) using an AND operator, and the competitions start at this level again. The up-down red lines dynamically "bind" together feature detectors across the different levels of the hierarchy during perception, relating parts to the whole shape. Neurons linked together by the red lines will show enhanced firing activities [62], [63], [64] and/or enhanced synchronous activity [65]. This linking implements the parsing tree in an AND/OR graph shown in Figure 4.

Hubel and Wiesel [6], [66] suggested that simple cells feed their output to complex cells to form more abstract and invariant feature detectors, insensitive to the precise location and appearance of the stimuli. Riesenhuber and Poggio [8] suggested this operation can be characterized as max-pooling. A complex cell of a particular orientation will fire when any one of the simple cells of the same orientation in a spatial neighborhood gets excited. More generally, a complex cell can be considered to indicate a more abstract visual concept. This operation allows any instance that belongs to the complex cell's visual concept to turn it on, allowing it to select from many possible options of the same label. For example, a vertically tuned complex cell indicating the concept of a vertical edge can be activated by many different types of vertical edges represented by simple cells, or simple cells of the same kind within a spatial neighborhood. This can be considered as an OR operation. This local competitive OR circuit, however, retains "memory" of which a simple cell excited that complex cell. This allows the backward tracking along the hierarchy when a parsing tree is instantiated in the AND/OR graph (see Figure 4 and Figure 5).

Both the OR operation and the max-pooling operation are nonlinear operation that can be implemented by a *local competitive circuit* that forces *similarly tuned* neurons at a spatial location (within a mini-column) to compete. This can happen between simple cells in an orientation column in V1 (blue tinted circle in Figure 5) for *instance selection*. Thus, the OR operation can be implemented by simply summing the output of a set of related instance feature detectors (or hypothesis particles) after they are subjected to strong local competition for redundancy reduction, which can also be interpreted as max-pooling.

The complex cells themselves will engage in a second type of local competitive mechanism. This concerns the competition among neurons tuned to very *different* visual concepts at the *same spatial location* (within a hypercolumn) to mediate concept selection. This is an OR operation on distinct hypotheses to implement the uniqueness constraint [67]. For example, in V1, competition between different disparity-tuned neurons in the same hypercolumn of V1 falls into this category [38], [39]. It might also be related to the classical phenomena of cross-orientation inhibition, as well as the ubiquitous normalization mechanism [68], [69]. Divisive normalization can be implemented using a shunting inhibition mechanism: The sum of the inhibitory input coming into a neuron is gated by that neuron's activity. The stronger its activity, the stronger the effect of the inhibition will be. Shunting inhibition could be imposed on the axonal initial segments, or possibly the soma, making parvalbumin (PV) expressing inhibitory basket cells or chandelier cells prime

suspects for mediating such action.

C. *Encoding the Probability of Visual Concepts*

Natural scenes are complex and inherently ambiguous. Thus, the visual concepts inferred also need to be associated with certain probabilities or uncertainties. Probability distributions and uncertainty of the visual concepts can be encoded implicitly or explicitly. Barlow and Levick [70], among others [71], [72], [30], [73], advocated an explicit representation of probability by neurons' activities. That is, a neuron is supposed to be encoding a specific stimulus feature, and its activity is monotonically related to the probability density or the log probability of that feature. Uncertainty is represented implicitly by a wider and lower activation pattern, i.e., greater entropy, across the population. Typical data structures in computer vision are similar to this, where the probability distributions and marginalized beliefs are represented over a histogram of discretized values of a variable [32], [74] to make the computation more tractable. This is called the *explicit probability distribution* code.

We [30] have proposed that probability distributions of hypotheses can be represented non-parametrically by samples, called particles. This *particle sample* code is inspired by the particle filtering computational framework used effectively in visual tracking [75] and robot spatial localization at the time. The essential idea is that each visual area has to compute and represent not just one guess for the true value of its set of features, but a moderate number of guesses. There could be n sets of values for the features in a visual area. In addition, each of these guesses is assigned a weight in such a way that the weighted sum of these guesses is a discrete approximation of the full posterior probability distribution. The neurons are the particles, representing specific hypotheses. The weight or importance is represented by the firing rates of the neurons. In the broadest terms, particle filtering is simply replacing a full probability table with a weighted set of samples. When the number of values of a random variable becomes astronomical (as happens in perception), this is quite possibly the best way to deal with distributions on it, which is known to probabilists as using a ''weak approximation''. In this representation scheme, a probability distribution is represented or approximated by particle samples. Shi and Griffiths (2009) [76] provided a rigorous formulation of similar ideas in terms of hierarchical importance sampling. The particle filter framework dictates that the density of the particles encoding the prior distribution of visual events should be high in regions where events happen often, and sparse in region where events are rarely observed. Thus, the distribution of samples should reflect the first order statistics or the frequency of occurrence of the visual events. In early visual areas, neurons tuned to different stimulus parameters can be considered as particles. As an example, Figure 6 shows that the distribution of preferred tunings of disparity detectors in V1 match the first order statistics or frequency of occurrence of the disparity signals of 3D natural scenes [77], [78].

In this framework, a neuron can be considered as a particle sample, labeling a particular hypothesis at a particular level. Its firing rate reflects its confidence in the hypothesis it represents.

A problem with this representation is that the tuning curve is measured in terms of firing rates obtained by averaging spikes of a neuron over many trials. Neurons tend to exhibit Poisson variability in their spiking activities from trial to trial. One can obtain such averaged spike counts by summing the spikes over a longer period of time, or pooling spikes from over a population of neurons with similar tunings. That is, a neuronal pool, rather than an individual neuron, is representing a visual concept.

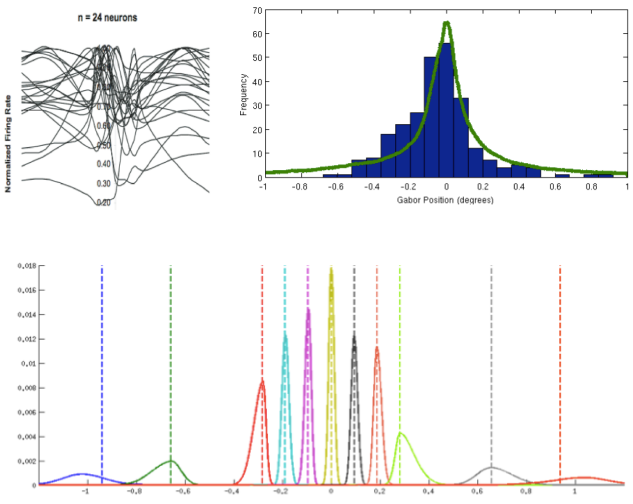

Figure 6. Top left panel: samples of disparity-tuning curves recorded simultaneously using a Utah array in primate V1. Top right panel: the distribution of the preferred disparity tunings (peak of the tuning curves) match the distribution of the occurrence frequency of disparity signals in 3D natural scenes [77], as in a particle filter code. Bottom panel: The probabilistic population code was based on the spike counts of a population of 248 neurons, obtained for each of the 11 disparities. Each was indicated by the dotted lines, and tested in a stereo experiment. Data from [79], [39]. Each color curve indicates the geometric mean of posterior $p(s|\mathbf{r})$ computed from the spike counts in 1 s (with 30 repeats) of the neuronal population [77]. Note the posteriors near 0 are sharper, while more distant disparities have a broader posterior, corresponding to higher uncertainty, as in human perception.

Hoyer and Hyvärinen [80] suggested that the temporal variability of neuronal spiking activity could potentially be interpreted as Monte Carlo sampling of the posterior. The idea is that at each moment in time, the activity of a neuronal population represents a "sample" of the posterior probability distribution with each neuron representing a variable (feature) in a high-dimensional multivariate distribution, and each population activity pattern representing a point in this high dimensional space. Variability in a neuron's spiking activity, and co-variability in the response among neurons, reflect the network dynamics for sampling from the posterior distribution. Each sample could be a binary vector, indicating whether each cell in the population has fired or not fired within a 20 ms time bin. *Over time*, the samples could also provide a non-parametric approximation of the posterior distribution. Fiser et al. [81] suggested that the spontaneous activity of neurons might be such "samples" from the prior distribution. There is some evidence showing that spike

patterns of the spontaneous activity in the V1 neuronal populations of newborn ferrets initially was very different from the spike patterns when they were watching movies, but the distribution of the spike patterns during spontaneous activity and that during natural movie watching grew much closer over the course of two weeks [82]. This observation supported the idea that spontaneous activity of the neurons can serve to encode priors. Bayesian inference then involves the interaction of the "spontaneous activity" and the activity evoked by visual input. This *sampling code* is an intriguing hypothesis that merits further investigation.

The particle codes can also be related to the so-called probabilistic population code. The *probabilistic population code* uses the Poisson model of spike count variability of neurons to model the likelihood function $p(r_i|s)$ for each neuron i. With Bayes rule, the posterior distribution of the stimuli s given the spike counts of a neuronal population, $\vec{r} = (r_1, r_2, r_3, \dots, r_N)$, is given by

$$p(s \mid \vec{r}) \propto \prod_i \frac{e^{-f_i(s)} f_i(s)^{r_i}}{r_i!} p(s)$$

where $f_i(s)$ is the tuning curve of neuron i, the expected or mean number of spikes for each stimulus or behavior s parameter in a time window of a particular duration (e.g. 200 ms, or 1 second), r_i is the actual observed spike count within a particular time window, and $p(s)$ is the prior on s. The tuning curves can be quite general, and can be coarse or fine, but the neurons are assumed to be independent. Such a model has been extensively used in Bayesian decoding of movement and location based on the neural signals in the motor system and in the hippocampus. A number of researchers [83], [84], [85], [86] have suggested that such a representational scheme might in fact be used by the brain to represent probability distributions using neuronal population activity. The bottom panel of Figure 6, for example, shows the averaged posterior distribution of the "visual concepts", in this case, 11 disparity of the input stimulus across. Whether the trial-by-trial posterior distribution reflects the uncertainty associated with behavior remains to be shown. Computational algorithms exploiting such code to solve real vision problems remain to be developed.

D. Relationships between Visual Concepts

The vertical and horizontal edges in the probabilistic graph model of the visual cortex (e.g., Figure 4) specify relationships between the visual concepts. The vertical edges specify the relationship between a visual concept (parent) and its parts (children). They are implemented by feedforward and feedback connections. The horizontal edges specify the relationship between one visual concept and the other visual concepts (siblings), which are implemented by horizontal connections. These ideas are illustrated in Figures 3 and 5.

Over the years, two computational principles concerning the vertical and horizontal interaction in the cortex have emerged. The first is the predictive coding principle, which is a generalization of the efficient coding principle from tunings in the individual neuron level to the hierarchical system. As discussed earlier in the context of predictive coding models (Class IV internal models), this principle allows efficient representation of an entire visual concept in the system by minimizing the redundancy between the higher order representation and the lower level representation. Predictive coding proposes that when a global concept is inferred at a higher level, a prediction is synthesized to suppress the redundant lower level representations in the early level via feedback (Figure 3). These different levels could mean different visual areas in the visual cortex, but could also mean different layers within a visual area, i.e., the predictive feedback can be intra-areal or inter-areal.

It has been proposed [24], [26] that surround suppression, which is ubiquitous in the visual cortex [87], [88], [89], [90], can be considered a manifestation of predictive coding.

The basic phenomenon is that while neurons would not fire unless their receptive fields are directly stimulated, their responses to receptive field stimuli are often modulated by the stimuli in the surrounding context. Most of the time, particularly when the stimuli are sinusoidal grating stimuli, the surround modulation observed is suppressive. The strongest suppression often occurs when the stimulus parameters of the receptive field (orientation, direction of motion, disparity, color) are the same as those in the surround. Testing with orientation stimuli, Knierim and Van Essen [91] found that 2/3 of the V1 neurons experienced stimulus-unspecific suppression (independent of the orientation of the stimuli), and 1/3 of the cells experienced the greatest suppression when the orientation of the surround and the receptive field stimuli were the same. Zipser et al. [92] found that neurons experienced surround suppression in response to many different cues (color, movement, orientation and disparity), though most neurons could experience surround suppression responding to one cue but not to others. Lee et al. [64] showed that such surround suppression in V1 generalized to higher order visual constructs such as 3D shape from shading and as such, the relative enhancement of the signals inside a small region can be considered a neural correlate of perceptual saliency. The spatial extent of the surround suppression can be quite large, about 3-6 times larger than the receptive fields. The effect was observed typically at 40-60 ms after the initial response onset, and was significantly attenuated when V2 was deactivated by cooling [93], [113], supporting the idea that a significant fraction of the surround modulation is mediated by recurrent feedback interaction [24], [26], [30]. In real life, the visual system has to process a continuous temporal sequence of events. The concept of predictive coding can be extended to the time domain to generate predictions of future events. Recently neurophysiological results found that neurons in IT experienced more significant suppression when an incoming event is predicted than when such event is not predicted [94], [95]. Prediction, apart from facilitating visual processing, can be used to obtain an error signal to drive both the learning process and the inference process [40].

When the global context of a stimulus predicts that the features appear in the receptive field of a neuron, the neuron's response to these features is suppressed. Thus, such a neuron's

response can be considered to carry a prediction error signal. We have suggested the observation that the later part of early cortical visual neurons' responses are often lower than the initial response which is, in fact, a reflection of the recurrent feedback's predictive suppression. A recent experiment provides clear evidence that this might at least be part of the story. When pictures of a set of large objects were presented repeatedly to the receptive fields of earlier cortical neurons across days, we [96] found that the neurons' later responses became more suppressed as the system became familiar with the object. The acquisition of familiar global object representation appears to introduce an additional suppressive effect on the later part of the neurons' responses of these early cortical neurons with localized receptive fields. The effect was likely to change in global representation and not in the transformation of the receptive field tunings of the neurons. This is because the receptive field stimuli are no different from those in the other images that the animal was constantly exposed to.

However, the neural responses of early cortical neurons usually are completely suppressed even when the global context and/or the global objects in the visual stimuli are very clear, i.e., when there should not be any prediction error signals, indicating that the neural responses of most early visual neurons are not coding simply the prediction error generated by subtraction [26] between the top-down prediction and the bottom-up input. This, we reasoned, is because V1 and the extra-striate areas represent different and complementary information; we still want V1 to maintain representation of the high-resolution fine details of the image. The emergence dynamic or long-term linking of a neuron representing a particular visual concept and neurons representing the higher order parent concept and other sibling concepts could support a *normalization of the responses to the parts and to the whole*, resulting in attenuation of this concept signal as it becomes increasingly associated with a larger context. Error signals might still be calculated, possibly by another group of neurons, which can be used to update and refine the internal models during both learning and inference. But such signals remain to be identified in early cortical neurons' responses by appropriate experiments.

The second principle is the associative coding principle, derived from Hebbian learning, which is the cornerstone of the interactive activation class of models. Association coding associates low-level concepts to high-level concepts through the feedforward and feedback vertical connections between different layers in the hierarchical system to encode the parent-child relationships in a graphical model. It can also be used to encode sibling relationships between concepts at the same level using horizontal connections. It allows networks to learn spatial relationships between parts of objects, or between objects in a visual scene, as a key part of the hierarchical compositional system.

Deep belief nets [97] and convolution neural nets [7] focus primarily on vertical connections. Without loopy horizontal connections, inference can progress very fast in a feedforward manner. Here, I want to focus on the discussion of horizontal connections, which are well known in neurophysiology but often willfully ignored or neglected in deep learning models.

Tso et al. [98] first demonstrated that V1 neurons of similar orientation tunings in the superficial layers tend to be connected together horizontally. Such connections can provide facilitation to presumably enable boundary completion and surface interpolation in V1. Anatomical evidence [99] and physiological evidence [100] suggested that these connections are not isotropic, but stronger along the longitudinal direction of the preferred orientation of the neurons, consistent with the contour association field discovered in psychophysical studies [101] and scene statistical studies [32], [33]. Steve Zucker's [35] works suggested that individual V1 neurons might be associated with a curvature transport field. The classical association field [101], [32], [33] could simply result from the superposition of many of these curvature fields. There is indeed evidence for curvature sensitive cells even in V1 [102], [103]. Moreover, the extensive and dense local horizontal connections in V1 should not just serve contour completion, but likely serve a number of other early vision computations such as surface interpolation and image reconstruction. Indeed, Zucker [35] suggested that there might also be transport fields for interpolating surface cues such as color, texture, shading and disparity along with the contour curvature transport fields. Our laboratory has found converging evidence from statistical analysis of 3D scenes [78], [104] as well as neurophysiological recordings [38], [39] that implicated the existence of a disparity association field in V1 that encodes the pairwise correlation of disparity signals experienced in 3D natural scenes, which can facilitate stereo surface interpolation computation.

Association fields are natural outcomes of Hebbian learning due to stimulus correlations in natural scenes. Ko et al. [84] demonstrated the existence of direct monosynaptic and bidirectional connectivity between pairs of neurons that exhibited correlated responses to natural movies. It has been proposed that Markov random fields can be used to conceptualize cortical computation in the primary visual cortex [105], [106]. Markov random fields can encode the relationship between features and concepts in each visual area. A node in the MRF is a continuous variable which has to be represented by a population of neurons. Each is a binary unit tuned to a particular range of the stimulus space. The Boltzmann machine is a form of MRF with binary variables, which are closest to a neural implementation. When we trained a Boltzmann machine using disparity signals derived from 3D natural scenes, we found that the Boltzmann machine learns a neural circuit that is very similar to the neural circuit inferred from our neurophysiological experiments [78], with cooperation between similarly disparity-tuned neurons across space and competition between different disparity-tuned neurons at the same location [39], [38]. In addition, we found in scene statistical analysis that there is a strong spatial correlation in disparity signals along the cardinal direction than along the oblique, which can be traced to the prevalence of vertical and horizontal structures in natural scenes and a cardinal effect in surface tilt. We found that this characteristic signature in natural scene statistics manifests in a stronger

observed functional connectivity among similarly disparity tuned V1 neurons along the cardinal direction than along the oblique [104]. This observation further supports the idea that neural circuitry is shaped by the statistical structures in the environment [104]. These findings suggest that either the Markov random field or the Boltzmann machine is a viable model for conceptualizing the computation and predicting the circuitry in the visual cortex's internal models.

Association field, in the form of a pairwise horizontal connection, is capable of encoding primarily pairwise or second order correlation in the feature responses in each particular visual area. Thus, first order correlations are encoded in neuronal tunings, and second order correlations in the signals are captured by pairwise connections. MRF with only pairwise connections, however, cannot capture the higher order correlations in natural images. Higher order correlations can potentially be encoded by higher order cliques, i.e., potentials that constrain the simultaneous co-activation of multiple neurons [74], [107], or encoded in the feedforward/feedback recurrent connections to a downstream unit along the hierarchy. It is important for the horizontal connections to capture the pairwise correlation. Factoring out the second order statistics allows the more expensive vertical connections to focus their learning on the higher order correlational structures in the signals [41].

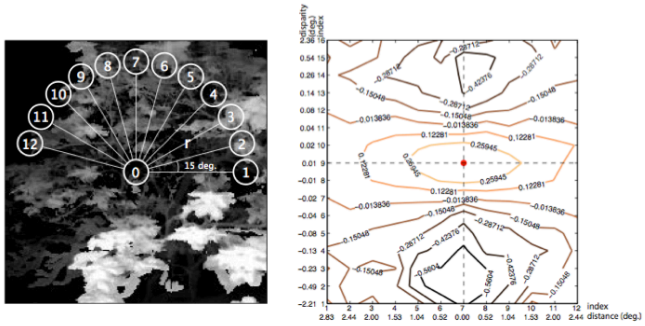

Figure 7: Left panel: A Boltzmann machine with 13 hypercolumns, each with 16 neurons tuned to different disparities, is trained with 3D natural scene data. The spiking activities of the neurons are determined by the disparity signals at each location during training. The Boltzmann machine learns connections that can spontaneously generate the same statistical distribution of population spike patterns during spontaneous activities as those observed during training by 3D natural scenes. This implements the encoding of stimulus correlations in natural scenes in neuronal connectivity by Hebbian learning mechanisms. The right panel shows the connectivity matrix of one neuron, tuned to 0 disparity at location 7, to all the other neurons in the other hypercolumns. The resulting connections show cooperation between similarly disparity-tuned neurons across spatial locations (rows 8, 9, 10) and competition between neurons with distinct disparity tunings at the same location (center column) and across columns [78], as predicted in [67] based on computational considerations.

Traditionally, the horizontal pairwise connections encode the pairwise "smoothness constraints" in Markov random field for contour and surface interpolation. However, given the sophisticated neural circuits and the great variety of interneurons within each columnar modules in each visual area, it is very likely more sophisticated contextual information can be encoded in the neurons within each visual area to implement higher order constraints or higher order priors [74].

A case in point is a recent module model we developed [40], called the contextual predictive encoding model, which attempts to integrate sparse coding, predictive coding and associative coding strategies into a unified framework. This model proposes a generalization of the predictive coding principle, and is also related to Memisevic and Hinton's gated Boltzmann machine [34]. The idea is to introduce a set of contextual latent variables into an autoencoder framework to allow visual events across space and across time to predict each other based on contextual priors. The proposed mutual predictability principle allows the system to fill in missing details in space and time (interpolation) as well as to extrapolate and to predict the future. The model learns a set of contextual latent variables as internal models of the spatial temporal feature transforms observed in the training sequences. These latent variables (see Figure 8) serve to modulate the synthesis process by scaling the importance (weights) of the learned basis functions. The inference of these latent variables is nonlinear, based on an iterative expectation/maximization-like procedure and is driven by prediction error minimization. The distinction between this "predictive encoding" model and other predictive coding models [26] is that the prediction is generated by local contextual and transformational priors, rather than higher order dictionaries in a downstream area, and that the residue signals update the context but are not necessarily the only signals being fed forward to downstream neurons. Thus, the model should be considered as a module in a hierarchy, and can be stacked up together to form a new kind of hierarchical deep network.

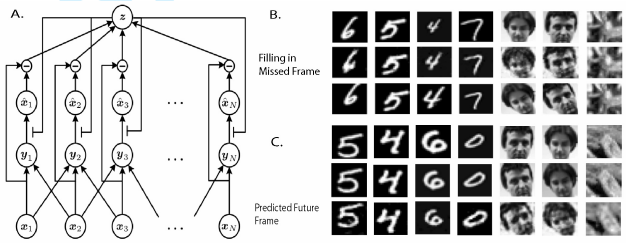

Figure 8: (A) Latent contextual variables **z** learn the spatiotemporal context. The activation of a set of contextual variables **z** by the input frames feeds back to scale the basis function during the synthesis process to maximize the predictability of the frames from one another in that spatial temporal context. Prediction error signals update **z** in a EM like process. (B) Interpolation results: second row shows the interpolated images for the missing frames, given the first and the third frames. (C) Prediction results: third row shows the predicted images for the future frames, given the first and the second frames.

We [40] found this model capable of interesting computation such as interpolating missing frames, and predicting future frames in a movie (see Figure 8). This model is an example of how sparse (efficient) coding, predictive coding and associative coding can be integrated in a

unified framework to perform nontrivial computation.

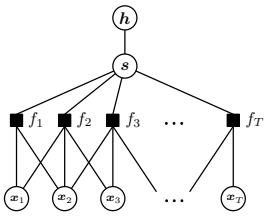

Figure 9: The neural model in Figure 8 is in fact the realization of a graphical model with factors (**f**'s) and a set of latent variables (**h**) depicted here. The latent variables **h** are connected to the weighted sum of the output of **f. f** are related to **y**, and **h** to **z** in the neural circuit model in Figure 8. The learning of the weights in the network is driven by the prediction error signals.

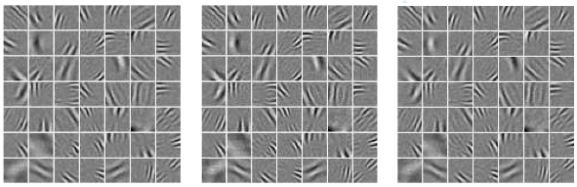

Figure 10: Spatiotemporal receptive fields of neurons trained by natural movies. Gabor-like spatiotemporal filters are learned as efficient codes ("sparse codes" because L1 norm is used).

IV. Models, Mechanisms and Algorithms

A. Varieties of Inference Algorithms

We know little about the computational algorithms the brain uses to exploit the internal models to make perceptual inference. But decades of algorithmic developments in computer vision could potentially delineate the scope of the possible neural algorithms. In computer vision, there are four major classes of optimization algorithms that have been used with a certain degree of success for performing statistical visual inference in probabilistic graphical models: mean field approximation, Markov chain Monte Carlo (MCMC) sampling, graph cut and belief propagation. These are approximation methods because exact inference is possible only for a small set of problems. All these methods have certain elements that are biologically plausible, and have implications on how internal models are encoded and used.

Neurons and neuronal networks have typically been modeled in terms of dynamic systems that follow a gradient descent dynamic to minimize energy functions that specify the constraints for solving a problem. As energy functions for vision problems tend to be highly non-convex, gradient descent on such an energy landscape tends to be trapped in local minima. Computation in dynamical systems can be considered as performing a mean-field approximation of the computation in a stochastic system. There are a number of strategies that might allow a neural system to escape the curse of local minima. One potential strategy is to introduce stochastic noise to the system, which in some balanced networks could allow different clusters of neurons to become active at different times [108]. Such a model can provide a good account of the recently observed persistent slow-dynamics of cortical state fluctuation over time in the cortex, and can potentially enable stochastic computation in a dynamical system.

Sampling approaches such as MCMC are widely used in machine learning and computer vision. If time is not an issue, sampling can produce very robust results for both learning and inference. In some cases, when the energy function is too non-convex, sampling is often the only method that works. Gibbs sampling is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of samples from a probability distribution that can be used to approximate the joint distribution in graphical models such as Markov random field or deep belief nets. It generates a sample from the distribution of each variable in turn, conditional on the current values of the other variables. The sequence of the samples constitutes a Markov chain, and the stationary distribution of that Markov chain is the joint distribution. When a graph can be factorized as in a modular hierarchy network such as a Bayesian network or a Markov random field, Gibbs sampling can be particularly effective.

Hoyer and Hyvärinen [80] had suggested earlier that the variability in neural response could reflect a Monte Carlo sampling of the posterior distribution. It has also been recently suggested [81], [82] that the spontaneous population spike patterns might represent the priors, as "samples" drawn from a distribution defined by network connections that encode the statistical priors of the natural scenes. During inference, the spontaneous activity or priors' samples interact with "likelihood" samples elicited by the input to produce samples for the posterior distributions. Thus, the sampling approach potentially can provide a natural explanation for spiking variability of individual neurons, as well as co-variability between neurons, and might be related to the ongoing activities that reflect the large fluctuations in cortical states. The problem with the sampling approach, particularly in a hierarchy with many loops, is that it might be too slow to be practical for inference. Drawing samples in the brain is limited by the duration of an action potential, which is in the order of a few milliseconds if the refractory period is included.

Dynamical system is fast but not accurate. Sampling is robust but too slow. Mumford and I [30] have proposed that belief propagation potentially offers a good compromise, particularly when the internal models can be factorized into a product of potential functions. BP computes by passing and weighting messages and beliefs, and can naturally keep multiple hypotheses alive, yet it can still be fast and parallelizable. When the probability distribution is represented non-parametrically using particles, a variant of belief propagation is called particle filtering.

Belief propagation has been successfully applied to a wide variety of computer vision problems with impressive results [109], [75], [74]. Figure 11 shows how an efficient belief propagation [110], [107] my student Brian Potetz developed

can solve the shape from shading problem as defined by Horn. The analysis path inferred the underlying 3D shape; the synthesis path is simply computer graphics that render the 3D shape based on lighting direction.

The algorithm proceeds forward in time, and information from many observations is integrated into the current set of particles and their weights. One can also go backward in time and reinterpret earlier evidence i.e., using current data to clear up ambiguities in past interpretations. This is exactly the way the forward/backward algorithm works in speech recognition, except that using particles allows one to overcome explosions in the number of states of the system. In the vision situation, information flow progresses both along the time axis and along the visual hierarchy, starting with local elementary image features and progressing to more global and more abstract features in the higher areas. The recurrent interaction across the hierarchy helps to collapse the hypothesis space over time [30].

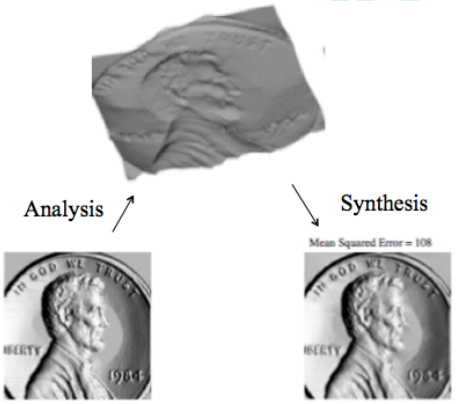

Figure 11. The results of inferring the 3D shape of a penny given its 2D optical image (shown on the left) using an efficient Belief Propgation algorithm. Top: the inferred 3D shape, a rendering based on the inferred latent surface orientation *(p,q)* variables at each location. Right: The image synthesized from the inferred latent *(p,q)* variables match almost perfectly the input image. This illustrates analysis by synthesis in perceptual inference [110], [107].

The efficient belief propagation algorithms [111], [74] typically exploit a coarse-to-fine strategy to speed up computation. For example, the hypothesis space is initially coarsely partitioned to represent the beliefs in a histogram or a finite set of states. Over time, the belief space is more finely partitioned to allow successive convergence of the beliefs. This is similar to particle filtering. What exact strategy the brain uses remains to be elucidated. A number of researchers have seriously explored the plausible neural implementation of BBP algorithms [42], [43], [72], [44], [73]. Neural implementation of exact belief propagation usually requires more elaborate dendritic computations [73], [107].

Figure 12 shows another example from neural decoding to help illustrate the general concept. Here, the task is to decode the movement of a sinewave grating seen by a neuron with local receptive field [112]. At a given moment in time, when an input signal is measured or sampled, the system has a set of hypotheses over the possible trajectory in the last 200-300 ms. The hypothesis particles are indicated by the different trajectories radiating from the right (most recent past) to the left (more distant past) in the "pre-resampling stage" block, with the more probable hypothesis represented by thicker trajectories. There are many equally likely hypotheses on the right because inference is made based on the most recent, temporally local information, which is inherently ambiguous and uncertain. On the left, when a more global view or historical view is available, the hypothesis space has collapsed into 2-3 hypotheses.

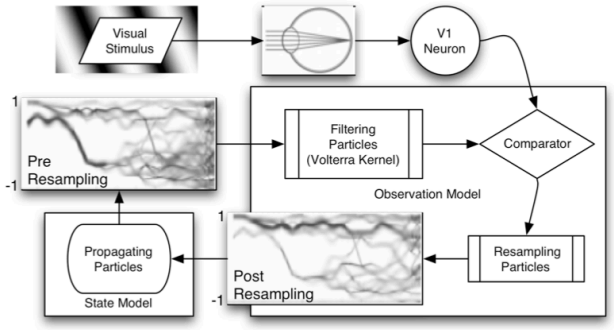

Figure 12. Dynamics of a particle filtering algorithm. Data input and the different predictions of the internal model are compared at the comparator at each point in time. The hypothesis that generates the prediction that best matches the observation is given stronger weight (importance), as indicated by the thickness of the line in the stimulus trajectory hypothesis. The beliefs at the post-resampling stage could become substantially different from that the pre-resampling stage. Observe the change of the favored hypothesis.

In this example, as a new measurement comes in, the new evidence "resamples" the particles, reevaluating the importance of each of the probable hypotheses. The "post-sampling stage" block shows that the most important hypothesis in the pre-sampling stage has become less important, while the runner-up hypothesis in the pre-sampling stage now becomes the most probable or important hypothesis. Thus, new information that comes in over time or across space can drastically change the system's interpretation of the visual scene.

As discussed earlier, the information flow along the time axis here is very similar to the information flow up and down the hierarchy over time. When the inference reaches the top of the hierarchy, the recurrent interaction across the hierarchy helps to collapse the hypothesis space over time.

In the context of the AND/OR graph of a clock, each of this particle will link the concept of the clock at the top of the hierarchy to its parts, down to the image represented in V1. The images of two different clocks trigger two different parsing graphs, selecting different parts in the AND/OR graph as illustrated in Figure 13.

B. Binding and Grouping Mechanism

The concept of a particle in the form of a parsing graph does involve the concept of binding of the neuronal ensemble representing different visual concepts at different levels along the visual hierarchy. When lower order concepts are exciting a higher order concept, and the higher order concepts feed back

to reinforce the lower order concepts, this interaction, as in the interactive activation, will increase the functional connectivity of the concerned neurons, leading the neurons to fire synchronously or oscillate together. This is related to the controversial *binding by synchrony* hypothesis [113], [114], [115], [63], [116]. Neurons that fire together synchronously or in oscillation together are said to be "bound" together as a group. This can be seen as a mass-spring model, with neurons as masses, and their connections as springs. The strength of the neuronal connectivity is dynamically set based on the priors and the visual input. Each parsing graph will fire synchronously and exhibit oscillation, segmented from other parsing graphs or hypotheses. This is very similar to the computer vision grouping and segmentation algorithms based on spectral graph partitioning [117], [118], [119]. It is important to note that from our perspective, binding simply means the different parts of a hypothesis particle are hand-shaking and dynamically coupled together. Synchronous activity is more of a reflection of interaction, rather than a "code" to be read. This allows the visual system to couple the concepts from the most abstract level down to V1 or even the LGN level, producing a coherent holistic perception of the object or the scene. Perception of a face is not simply the activation of a "face neuron" in IT, but the holistic activation of the entire chain of neurons involved in the parsing graph, coding a nose with its wrinkles, and the eyes with their sparkles. The co-activation or synchrony of relevant neurons along the hierarchy corresponds to the coupling of nodes in a parsing tree in a AND/OR graph (thicker line in Figure 4, and red lines in Figure 5).

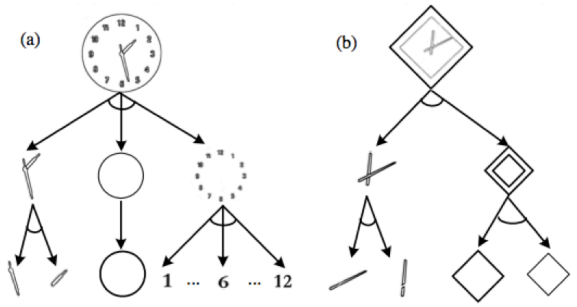

Figure 13. Two parsing graphs or selection paths in the AND/OR graph in response to the images of two different clocks [45]. Neurons coding the different visual concepts along the parsing graph will be functionally coupled together through their interaction. Reprinted from [45] with permission from authors.

Mumford and I [30] have argued that the particle itself might need to be represented by the concerted activity of an ensemble of neurons, which could be bound by timing (e.g., synchrony or by synaptic weights after short-term facilitation). Note that the top-down messages can utilize the same recurrent excitatory mechanism that has been proposed for implementing biased competition for attentional selection. In fact, visual attention itself could be considered a special case in this framework. The recurrent excitatory connections across the multiple modules in the visual hierarchy allow the neurons in different areas to link together to form a larger hypothesis particle by firing concurrently and/or synchronously. Since its implementation requires that groupings of mutually reinforcing alternative values of features in different areas be formed, this algorithm might be linked to the solution of the binding problem.

Sometimes, it is possible that a particular concept in the middle of the hierarchy can be shared by a few particles. In this event, those particles need to be entertained simultaneously; the visual concept will need to participate in multiple hypotheses at the same time. This might not be a serious issue, as each visual concept might be redundantly represented by multiple identical or similar neurons, and these neurons can be coupled with different hypothesis particles, each participating in only one particle. Alternatively, a neuron representing a visual concept can hand shake with multiple higher order neurons by firing synchronously.

Lamme's figure enhancement effect or the enhancement of neuronal activities inside a figure relative to the background observed in V1 [63], [120], [64], [129] can be considered a consequence of a higher order concept feeding back to the early visual areas in the hierarchy. This enhances the activities of those neurons that have contributed to this higher order concept, resulting in an increase in firing rates of these neurons. Recently, it has been observed that this enhancement effect appears first in higher areas such as V4 before appearing in V1 [62]. This phenomenon is consistent with the general picture that neurons in the same parsing graph will interactively activate one another through recurrent feedback connections, establishing stronger functional connectivity or coupling. Such interaction could cause neurons along the parsing tree to be more synchronous and exhibit oscillation phenomena. This might be more appropriately called *synchrony by binding*, rather than binding by synchrony.

Pascal Fries [121] proposed that the various rhythms well known in the brain might not simply be an epiphenomenon of interaction, but a way to shape the communication between the different brain areas by changing the excitability of a neuronal population. This Communication through Coherence (CTC) theory argues that the rhythmic changes in neuronal populations' excitability can open and close the communication channels between populations of neurons. This might provide a flexible communication structure and mechanism to couple visual concepts together as a particle along the hierarchy, creating and selecting the appropriate parsing graph in visual scene analysis. There is some evidence, particularly based on the simultaneous analysis of spikes and local field potentials, in support of such a hypothesis. Salazar et al. (2012) found that neuronal synchronization across the fronto-parietal network carries content-specific information, i.e., what objects are being held in the visual working memory of monkeys.

V. Conclusions

In this paper I have discussed some observations on the computational principles governing the construction and the representation of the internal models of the world in the visual

cortex. These observations can be summarized into the following five major principles, which are by no means exhaustive.

*1) Principle of **statistical inference***

The stream of visual sensory signals of the world coming into the visual cortex through the retina are inherently ambiguous because of the stochastic nature of spike messages, noises, the constant jitters of the eyes, and the locality of the sensors in the early visual system. To resolve this ambiguity, the brain has to make statistical inferences using internal models that encode the prior knowledge of the world that it constructs from experience [1]. From our current conceptualization, these internal models are best expressed in terms of probabilistic graphical models that are structured and compositional. The nodes in these graphical models represent visual concepts. The probability and importance of these concepts are represented by neural activities as particle samples non-parametrically [30]. An ensemble of neurons representing concepts at different levels of scale and abstraction form a macro-particle as a coherent hypothesis of the world.

*2) Principle of **compositional hierarchy***

The visual world is extremely complex. The only way to scale up an internal model to deal with this complexity is to use a compositional model to exploit the fact that images themselves are generated by composing objects [2], [122]. A compositional internal model allows information to be decomposed and represented in parts, allowing whole concepts to be constructed and represented using flexible recombination and reconfigurations of these reusable part concepts [5]. A compositional system might also offer a nearly decomposable system to allow parallel and distributed processing of information [2]. We conjecture that this hierarchy can be organized conceptually with the AND/OR graph architecture [45], [123], which is very much in the spirit of the alternating simple cell/complex cell architecture of the Neocognitron [4]. A node in such a graph represents a concept as a flexible template, which is activated by "pooling" [8] or "routing" [124] during feedforward inference to achieve invariance. It can also be steered by "memorized switches or routing signals" to reconstruct by the input precisely, as in the deconvolutional network [125] to instantiate the parsing graph of an object, and more generally of a whole scene (Figure 13).

*3) Principle of **efficient and associative coding***

There are two important aspects to the internal model: visual concepts that are encoded in the tunings of the neurons, and the relationship among the concepts encoded in the functional connectivity of the neurons. Visual concepts are learned based on the detection of suspicious coincidence of conjunctive simpler visual concepts through Hebbian learning. The representation is made efficient by a variety of local competitive interaction: (1) local competition between neurons with similar tunings at the same location to remove redundancy (e.g. sparse coding) [55], [53]; (2) local competition across space of similar tuned neurons to achieve invariance, as a form of max-pooling [8]; (2) local competition between neurons of different tunings, coding distinct visual concepts at the same locations during inference to implement the uniqueness constraint [67], [38] or to normalize the probability distribution of the visual concept hypotheses for explaining the images [30], [83]. The second order relationship between the visual concepts can be modeled by pairwise horizontal connections [126], [78]. Higher order relationships among visual concepts can be encoded by the feedforward/feedback vertical connections as well as interneurons [19], [127], [40]. The learning of the relationships among visual concepts is governed by Hebbian rule-based associative coding principle [18].

*4) Principle of **generative inference and learning***

Predictive coding is an extension of the efficient coding principle to the hierarchical system to minimize the redundancy of the visual concepts at different levels of the hierarchy [24][26]. However, apart from efficiency in representation, predictive coding mechanism might serve a more important functional role: validating the correctness of the internal model. The validity of an internal model can be confirmed by the system's performance. Organisms with false internal models will have poor performance and will likely be destroyed by natural selection. The difficulty with such validation approach is the lack of immediate teaching signals. It might be more advantageous to use the reality or the visual input as the teaching signal. The hierarchical visual system can uses its internal model to generate an explanation or expectation of what we are seeing and will see. The residue errors between the model's prediction and the input signals at each level can be used to validate, update and refine the internal model state during inference and the model itself at a long time scale as learning [40]. In essence, comparing predictions of our internal models with the stream of incoming data allows us to continuously update and validate our internal model of the world.

*5) Principle of **exploration and robustness***

The brain needs to constantly adapt its internal models to the changing environment to ensure their predictions can produce effective behaviors and an accurate representation of reality. This means that neurons and neural circuits have to constantly explore the space of internal models by modifying their tunings to come up with better visual concepts, and modifying existing connections to encode new relationships, in a manner similar to particle filtering [30], [128]. The external environment is constantly changing and full of "noises" and there are always events that we don't fully understand. Thus, the neural system might have to explicitly generate stochastic noise and random fluctuations [129], [108] to encourage exploration of the hypothesis space during learning and inference to increase the adaptability and robustness of the system.

These principles have been useful in our conceptualization of the computational organization of the primate's hierarchical visual system. The last principle is particularly important. Indeed, vision, through our incessant eye movements, is an active process of constant exploration and search. The brain is

always looking for useful clues to build a better internal model of the world. Imagination, exploration and experimentation are critical processes that allow the brain to search through the fitness landscape rapidly to ensure the survival of the species. From this perspective, the other principles related to attention selection, motor action and planning, and sensory-motor integration, which we have not discussed, are also crucial for understanding the visual system. Ultimately, the usefulness of the internal models in the visual system must rest on the usefulness and effectiveness of their predictions for generating appropriate and rewarding behaviors.

ACKNOWLEDGEMENT

I thank many colleagues, including David Mumford, Alan Yuille, Carl Olson, Song-Chun Zhu, Jay McClelland, Jason Samonds, Brian Potetz, Herb Simon, Steve Zucker and many other colleagues and students for their insightful discussion and collaboration over the years. For this paper, I express my special thank to Yimeng Zhang and Yuke Li for the preparation of the manuscript and Maureen Kelly for proof-reading. This work is supported by NSF CISE IIS 1320651 and NIH R01 EY022247.

REFERENCES

[1] H. von Helmholtz, *Handbuch der physiologischen Optik*. Leipzig Voss, 1896.
[2] H. A. Simon, *The architecture of complexity*. Springer, 1991.
[3] D. J. Felleman and D. C. Van Essen, "Distributed Hierarchical Processing in the Primate Cerebral Cortex," *Cereb. Cortex*, vol. 1, no. 1, pp. 1–47, 1991.
[4] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, 1980.
[5] S. Geman, "Invariance and selectivity in the ventral visual pathway," *J. Physiol.*, 2006.
[6] D. H. Hubel and T. N. Wiesel, "Ferrier Lecture: Functional Architecture of Macaque Monkey Visual Cortex," *Proc. R. Soc. London B Biol. Sci.*, vol. 198, no. 1130, pp. 1–59, 1977.
[7] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *Handb. brain theory neural networks*, vol. 3361, p. 310, 1995.
[8] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nat Neurosci*, vol. 2, no. 11, pp. 1019–1025, 1999.
[9] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How Does the Brain Solve Visual Object Recognition?," *Neuron*, vol. 73, no. 3, pp. 415–434, 2012.
[10] D. Cox and N. Pinto, "Beyond simple features: A large-scale feature search approach to unconstrained face recognition," in *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011, pp. 8–15.
[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," *CoRR*, vol. abs/1409.4, 2014.
[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.
[13] J. L. McClelland and D. E. Rumelhart, "An interactive activation model of context effects in letter perception: I. An account of basic findings.," *Psychol. Rev.*, vol. 88, no. 5, pp. 375–407, 1981.
[14] S. Grossberg, "Competitive learning: From interactive activation to adaptive resonance," *Cogn. Sci.*, vol. 11, no. 1, pp. 23–63, 1987.
[15] J. L. McClelland, D. Mirman, D. J. Bolger, and P. Khaitan, "Interactive Activation and Mutual Constraint Satisfaction in Perception and Cognition," *Cogn. Sci.*, vol. 38, no. 6, pp. 1139–1189, 2014.
[16] J. L. McClelland, "Integrating probabilistic models of perception and interactive neural networks: a historical and tutorial review," *Front. Psychol.*, vol. 4, p. 503, 2013.
[17] R. C. O'Reilly, D. Wyatte, S. Herd, B. Mingus, and D. J. Jilk, "Recurrent Processing during Object Recognition," *Front. Psychol.*, vol. 4, p. 124, 2013.
[18] G. E. Hinton and T. J. Sejnowski, "Learning and Relearning in Boltzmann Machines," in *Parallel Distributed Processing: Foundations*, MIT Press, 1986.
[19] G. E. Hinton, "Learning multiple layers of representation," *Trends Cogn. Sci.*, vol. 11, no. 10, pp. 428–434, 2007.
[20] G. E. Hinton, "Reducing the Dimensionality of Data with Neural Networks," *Science (80-.).*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.
[21] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel, "The Helmholtz Machine," *Neural Comput.*, vol. 7, no. 5, pp. 889–904, 1995.
[22] U. Grenander, *Lectures in pattern theory. 2. , Pattern analysis*. New York, Heidelberg, Berlin: Springer, 1978.
[23] P. J. Burt and E. H. Adelson, "The Laplacian Pyramid as a Compact Image Code," *Commun. IEEE Trans.*, vol. 31, no. 4, pp. 532–540, Apr. 1983.
[24] D. Mumford, "On the computational architecture of the neocortex," *Biol. Cybern.*, vol. 66, no. 3, pp. 241–251, 1992.
[25] H. B. Barlow and W. A. Rosenblith, "Possible principles underlying the transformations of sensory messages," in *Sensory Communication*, Cambridge, MA: MIT Press, 1961, pp. 217–234.
[26] R. P. N. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nat Neurosci*, vol. 2, no. 1, pp. 79–87, 1999.
[27] M. Bar, "The proactive brain: using analogies and associations to generate predictions," *Trends Cogn. Sci.*, vol. 11, no. 7, pp. 280–289, 2007.
[28] A. Todorovic, F. van Ede, E. Maris, and F. P. de Lange, "Prior Expectation Mediates Neural Adaptation to Repeated Sounds in the Auditory Cortex: An MEG Study," *J. Neurosci.*, vol. 31, no. 25, pp. 9118–9123, 2011.
[29] K. Friston, "The free-energy principle: a unified brain theory?," *Nat Rev Neurosci*, vol. 11, no. 2, pp. 127–138, 2010.
[30] T. S. Lee and D. Mumford, "Hierarchical Bayesian inference in the visual cortex," *J. Opt. Soc. Am. A*, vol. 20, no. 7, pp. 1434–1448, 2003.
[31] S. Ullman, "Sequence Seeking and Counter Streams: A Computational Model for Bidirectional Information Flow in the Visual Cortex," *Cereb. Cortex*, vol. 5, no. 1, pp. 1–11, 1995.
[32] J. H. Elder and R. M. Goldberg, "Ecological statistics of Gestalt laws for the perceptual organization of contours," *J. Vis.*, vol. 2, no. 4, pp. 324–353, Jun. 2002.
[33] W. S. Geisler, J. S. Perry, B. J. Super, and D. P. Gallogly, "Edge co-occurrence in natural images predicts contour grouping performance," *Vision Res.*, vol. 41, no. 6, pp. 711–724, Mar. 2001.
[34] R. Memisevic and G. E. Hinton, "Learning to Represent Spatial Transformations with Factored Higher-Order Boltzmann Machines," *Neural Comput.*, vol. 22, no. 6, pp. 1473–1492, 2010.
[35] S. W. Zucker, "Stereo, Shading, and Surfaces: Curvature Constraints Couple Neural Computations," *Proc. IEEE*, vol. 102, no. 5, pp. 812–829, May 2014.
[36] O. Ben-Shahar, P. S. Huggins, T. Izo, and S. W. Zucker, "Cortical connections and early visual function: intra- and inter-columnar processing," *J. Physiol.*, vol. 97, no. 2–3, pp. 191–208, 2003.
[37] G. Li and S. W. Zucker, "Differential Geometric Inference in Surface Stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 72–86, Jan. 2010.
[38] J. M. Samonds, B. R. Potetz, and T. S. Lee, "Cooperative and Competitive Interactions Facilitate Stereo Computations in Macaque Primary Visual Cortex," *J. Neurosci.*, vol. 29, no. 50, pp. 15780–15795, Dec. 2009.
[39] J. M. Samonds, B. R. Potetz, C. W. Tyler, and T. S. Lee, "Recurrent Connectivity Can Account for the Dynamics of Disparity Processing in V1," *J. Neurosci.*, vol. 33, no. 7, pp. 2934–2946, Feb. 2013.
[40] M. Zhao, C. Zhuang, Y. Wang, and T. S. Lee, "Predictive Encoding of Contextual Relationships for Perceptual Inference, Interpolation and Prediction," *Int. Conf. Learn. Represent. (workshop Pap. Arch. (http//arxiv.org/abs/1411.3815)*., Nov. 2015.
[41] M. Ranzato, V. Mnih, J. M. Susskind, and G. E. Hinton, "Modeling Natural Images Using Gated MRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2206–2222, 2013.

[42]S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and where: A Bayesian inference theory of attention," *Vision Res.*, vol. 50, no. 22, pp. 2233–2247, 2010.
[43]T. Dean, "A Computational Model of the Cerebral Cortex," in *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2*, 2005, pp. 938–943.
[44]D. George and J. Hawkins, "Towards a Mathematical Theory of Cortical Micro-circuits," *PLoS Comput Biol*, vol. 5, no. 10, p. e1000532 EP –, 2009.
[45]S.-C. Zhu and D. Mumford, *A stochastic grammar of images*. Now Publishers Inc, 2007.
[46]L. Zhu, Y. Chen, and A. L. Yuille, "Recursive Compositional Models for Vision: Description and Review of Recent Work," *J. Math. Imaging Vis.*, vol. 41, no. 1–2, pp. 122–146, 2011.
[47]L. Zhu, C. Lin, H. Huang, Y. Chen, and A. L. Yuille, "Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence and competitive exclusion," in *European Conference on Computer Vision, 2008*, 2008, pp. 1–14.
[48]C. D. Gilbert and T. N. Wiesel, "Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex," *J. Neurosci.*, vol. 9, no. 7, pp. 2432–2442, 1989.
[49]C. D. Gilbert and T. N. Wiesel, "Morphology and intracortical projections of functionally characterised neurones in the cat visual cortex," *Nature*, vol. 280, no. 5718, pp. 120–125, 1979.
[50]E. R. Kandel, J. H. Schwartz, T. M. Jessell, and S. Mack, Eds., *Principles of neural science*. New York, Chicago, San Francisco: McGraw-Hill Medical, 2013.
[51]R. Gattass, A. P. Sousa, and C. G. Gross, "Visuotopic organization and extent of V3 and V4 of the macaque," *J. Neurosci.*, vol. 8, no. 6, pp. 1831–1845, 1988.
[52]P. Földiák, "Forming sparse representations by local anti-Hebbian learning," *Biol. Cybern.*, vol. 64, no. 2, pp. 165–170, 1990.
[53]B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, Jun. 1996.
[54]A. Coates, A. Karpathy, and A. Y. Ng, "Emergence of Object-Selective Features in Unsupervised Feature Learning," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 2681–2689.
[55]C. J. Rozell, D. H. Johnson, R. G. Baraniuk, and B. A. Olshausen, "Sparse Coding via Thresholding and Local Competition in Neural Circuits," *Neural Comput.*, vol. 20, no. 10, pp. 2526–2563, 2008.
[56]T. Gollisch and M. Meister, "Eye Smarter than Scientists Believed: Neural Computations in Circuits of the Retina," *Neuron*, vol. 65, no. 2, pp. 150–164, 2010.
[57]P. Kara and R. C. Reid, "Efficacy of Retinal Spikes in Driving Cortical Responses," *J. Neurosci.*, vol. 23, no. 24, pp. 8547–8557, 2003.
[58]J.-M. Alonso, W. M. Usrey, and R. C. Reid, "Precisely correlated firing in cells of the lateral geniculate nucleus," *Nature*, vol. 383, no. 6603, pp. 815–819, 1996.
[59]A. Polsky, B. W. Mel, and J. Schiller, "Computational subunits in thin dendrites of pyramidal cells," *Nat Neurosci*, vol. 7, no. 6, pp. 621–627, 2004.
[60]X. Hu, J. Zhang, P. Qi, and B. Zhang, "Modeling response properties of V2 neurons using a hierarchical K-means model," *Neurocomputing*, vol. 134, no. 0, pp. 198–205, 2014.
[61]G. Papandreou, L.-C. Chen, and A. L. Yuille, "Modeling Image Patches with a Generic Dictionary of Mini-epitomes," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 2059–2066.
[62]M. Chen, Y. Yan, X. Gong, C. D. Gilbert, H. Liang, and W. Li, "Incremental Integration of Global Contours through Interplay between Visual Cortical Areas," *Neuron*, vol. 82, no. 3, pp. 682–694, 2014.
[63]V. A. F. Lamme and H. Spekreijse, "Neuronal synchrony does not represent texture segregation," *Nature*, vol. 396, no. 6709, pp. 362–366, 1998.
[64]T. S. Lee, C. F. Yang, R. D. Romero, and D. Mumford, "Neural activity in early visual cortex reflects behavioral experience and higher-order perceptual saliency," *Nat Neurosci*, vol. 5, no. 6, pp. 589–597, 2002.
[65]C. M. Gray, "The temporal correlation hypothesis of visual feature integration: still alive and well," *Neuron*, vol. 24, no. 1, pp. 31–47, 1999.
[66]D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, no. 1, pp. 106–154, 1962.
[67]D. Marr and T. Poggio, "Cooperative Computation of Stereo Disparity," *Science (80-.).*, vol. 194, no. 4262, pp. pp. 283–287, 1976.
[68]S. Ellias and S. Grossberg, "Pattern formation, contrast control, and oscillations in the short term memory of shunting on-center off-surround networks," *Biol. Cybern.*, vol. 20, no. 2, pp. 69–98, 1975.
[69]D. J. Heeger, "Normalization of cell responses in cat striate cortex," *Vis. Neurosci.*, vol. 9, no. 02, pp. 181–197, 1992.
[70]H. B. Barlow and W. R. Levick, "Three factors limiting the reliable detection of light by retinal ganglion cells of the cat," *J. Physiol.*, vol. 200, no. 1, pp. 1–24, 1969.
[71]M. J. Barber, J. W. Clark, and C. H. Anderson, "Neural Representation of Probabilistic Information," *Neural Comput.*, vol. 15, no. 8, pp. 1843–1864, 2003.
[72]S. Deneve, "Bayesian Spiking Neurons I: Inference," *Neural Comput.*, vol. 20, no. 1, pp. 91–117, 2007.
[73]R. P. N. Rao, "Bayesian computation in recurrent neural circuits," *Neural Comput.*, vol. 16, no. 1, pp. 1–38, 2004.
[74]B. R. Potetz and T. S. Lee, "Efficient belief propagation for higher-order cliques using linear constraint nodes," *Comput. Vis. Image Underst.*, vol. 112, no. 1, pp. 39–54, 2008.
[75]A. Blake, B. Bascle, M. Isard, and J. MacCormick, "Statistical models of visual shape and motion," *Philos. Trans. R. Soc. London. Ser. A Math. Phys. Eng. Sci.*, vol. 356, no. 1740, pp. 1283–1302, 1998.
[76]L. Shi and T. L. Griffiths, "Neural Implementation of Hierarchical Bayesian Inference by Importance Sampling," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Curran Associates, Inc., 2009, pp. 1669–1677.
[77]B. Poole, I. Lenz, G. Lindsay, J. M. Samonds, and T. S. Lee, "Connecting scene statistics to probabilistic population codes and tuning properties of V1 neurons," in *Neuroscience 2010*, 2010.
[78]Y. Zhang, X. Li, J. M. Samonds, B. Poole, and T. S. Lee, "Relating functional connectivity in V1 neural circuits and 3D natural scenes using Boltzmann machines," in *Computational and Systems Neuroscience 2015*, 2015.
[79]J. M. Samonds, B. R. Potetz, and T. S. Lee, "Relative luminance and binocular disparity preferences are correlated in macaque primary visual cortex, matching natural scene statistics," *Proc. Natl. Acad. Sci.*, vol. 109, no. 16, pp. 6313–6318, Apr. 2012.
[80]P. O. Hoyer and A. Hyvärinen, "Interpreting Neural Response Variability as Monte Carlo Sampling of the Posterior," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. MIT Press, 2003, pp. 293–300.
[81]J. Fiser, P. Berkes, G. Orbán, and M. Lengyel, "Statistically optimal perception and learning: from behavior to neural representations," *Trends Cogn. Sci.*, vol. 14, no. 3, pp. 119–130, 2010.
[82]P. Berkes, G. Orbán, M. Lengyel, and J. Fiser, "Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment," *Science (80-.).*, vol. 331, no. 6013, pp. 83–87, 2011.
[83]W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget, "Bayesian inference with probabilistic population codes," *Nat. Neurosci.*, vol. 9, no. 11, pp. 1432–1438, Oct. 2006.
[84]T. D. Sanger, "Probability density estimation for the interpretation of neural population codes," *J. Neurophysiol.*, vol. 76, no. 4, pp. 2790–2793, 1996.
[85]H. S. Seung and H. Sompolinsky, "Simple models for reading neuronal population codes," *Proc. Natl. Acad. Sci.*, vol. 90, no. 22, pp. 10749–10753, 1993.
[86]R. S. Zemel, P. Dayan, and A. Pouget, "Probabilistic interpretation of population codes," *Neural Comput.*, vol. 10, no. 2, pp. 403–430, 1998.
[87]J. Allman, F. Miezin, and E. McGuinness, "Stimulus Specific Responses from Beyond the Classical Receptive Field: Neurophysiological Mechanisms for Local-Global Comparisons in Visual Neurons," *Annu. Rev. Neurosci.*, vol. 8, no. 1, pp. 407–430, 1985.
[88]R. T. Born and R. B. Tootell, "Single-unit and 2-deoxyglucose studies of side inhibition in macaque striate cortex.," *Proc. Natl. Acad. Sci.*, vol. 88, no. 16, pp. 7071–7075, 1991.
[89]J. R. Cavanaugh, W. Bair, and J. A. Movshon, "Nature and Interaction of Signals From the Receptive Field Center and Surround in Macaque V1 Neurons," *J. Neurophysiol.*, vol. 88, no. 5, pp. 2530–2546, 2002.
[90]J. J. Knierim and D. C. van Essen, "Neuronal responses to static texture patterns in area V1 of the alert macaque monkey," *J. Neurophysiol.*, vol. 67, no. 4, pp. 961–980, Apr. 1992.

[91]J. J. Knierim and D. C. van Essen, "Neuronal responses to static texture patterns in area V1 of the alert macaque monkey," *J. Neurophysiol.*, vol. 67, no. 4, pp. 961–980, Apr. 1992.
[92]K. Zipser, V. A. F. Lamme, and P. H. Schiller, "Contextual Modulation in Primary Visual Cortex," *J. Neurosci.*, vol. 16, no. 22, pp. 7376–7389, 1996.
[93]J. M. Hupe, A. C. James, B. R. Payne, S. G. Lomber, P. Girard, and J. Bullier, "Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons," *Nature*, vol. 394, no. 6695, pp. 784–787, 1998.
[94]T. Meyer, S. Ramachandran, and C. R. Olson, "Statistical Learning of Serial Visual Transitions by Neurons in Monkey Inferotemporal Cortex," *J. Neurosci.*, vol. 34, no. 28, pp. 9332–9337, 2014.
[95]T. Meyer and C. R. Olson, "Statistical learning of visual transitions in monkey inferotemporal cortex," *Proc. Natl. Acad. Sci.*, vol. 108, no. 48, pp. 19401–19406, 2011.
[96]S. Ramachandran, T. S. Lee, and C. R. Olson, "Effect of image familiarity on neuronal responses in areas V2 and V4 of monkey visual cortex," in *Neuroscience 2014*, 2014.
[97]G. E. Hinton, "Learning multiple layers of representation," *Trends Cogn. Sci.*, vol. 11, no. 10, pp. 428–434, 2007.
[98]D. Y. Ts'o, C. D. Gilbert, and T. N. Wiesel, "Relationships between horizontal interactions and functional architecture in cat striate cortex as revealed by cross-correlation analysis," *J. Neurosci.*, vol. 6, no. 4, pp. 1160–1170, 1986.
[99]W. H. Bosking, Y. Zhang, B. Schofield, and D. Fitzpatrick, "Orientation Selectivity and the Arrangement of Horizontal Connections in Tree Shrew Striate Cortex," *J. Neurosci.*, vol. 17, no. 6, pp. 2112–2127, 1997.
[100] M. K. Kapadia, G. Westheimer, and C. D. Gilbert, "Spatial Distribution of Contextual Interactions in Primary Visual Cortex and in Visual Perception," *J. Neurophysiol.*, vol. 84, no. 4, pp. 2048–2062, 2000.
[101] D. J. Field, A. Hayes, and R. F. Hess, "Contour integration by the human visual system: Evidence for a local ``association field'," *Vision Res.*, vol. 33, no. 2, pp. 173–193, Jan. 1993.
[102] A. Dobbins, S. W. Zucker, and M. S. Cynader, "Endstopped neurons in the visual cortex as a substrate for calculating curvature," *Nature*, vol. 329, no. 6138, pp. 438–441, 1987.
[103] J. Hegdé and D. C. Van Essen, "A Comparative Study of Shape Representation in Macaque Visual Areas V2 and V4," *Cereb. Cortex*, vol. 17, no. 5, pp. 1100–1116, 2007.
[104] X. Li, J. M. Samonds, Y. Liu, and T. S. Lee, "Pairwise interaction of V1 disparity neurons depends on spatial configural relationship between receptive fields as predicted by 3D scene statistics," in *Society of Neuroscience Conference Abstract*, 2012.
[105] C. Koch, J. Marroquin, and A. Yuille, "Analog 'neuronal' networks in early vision," *Proc. Natl. Acad. Sci.*, vol. 83, no. 12, pp. 4263–4267, 1986.
[106] T. S. Lee, "A Bayesian framework for understanding texture segmentation in the primary visual cortex," *Vision Res.*, vol. 35, no. 18, pp. 2643–2657, 1995.
[107] B. Potetz and T. S. Lee, "Scene statistics and 3D surface perception," *Comput. Vis. From Surfaces to Objects. Boca Raton, FL Chapman Hall*, pp. 1–25, 2010.
[108] A. Litwin-Kumar and B. Doiron, "Slow dynamics and high variability in balanced cortical networks with clustered connections," *Nat Neurosci*, vol. 15, no. 11, pp. 1498–1505, 2012.
[109] P. Felzenszwalb and D. Huttenlocher, "Efficient Belief Propagation for Early Vision," *Int. J. Comput. Vis.*, vol. 70, no. 1, pp. 41–54, 2006.
[110] B. Potetz, "Efficient Belief Propagation for Vision Using Linear Constraint Nodes," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007, pp. 1–8.
[111] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 2004, vol. 1, pp. I–261–I–268 Vol.1.
[112] R. C. Kelly and T. S. Lee, "Decoding V1 Neuronal Activity using Particle Filtering with Volterra Kernels," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. K. Saul, and B. Schölkopf, Eds. MIT Press, 2004, pp. 1359–1366.
[113] C. von der Malsburg, "The What and Why of Binding: The Modeler's Perspective," *Neuron*, vol. 24, no. 1, pp. 95–104, 1999.
[114] P. M. Milner, "A model for visual shape recognition.," *Psychol. Rev.*, vol. 81, no. 6, pp. 521–535, 1974.
[115] W. Singer and C. M. Gray, "Visual feature integration and the temporal correlation hypothesis," *Annu. Rev. Neurosci.*, vol. 18, no. 1, pp. 555–586, 1995.
[116] M. N. Shadlen and J. A. Movshon, "Synchrony Unbound: A Critical Evaluation of the Temporal Binding Hypothesis," *Neuron*, vol. 24, no. 1, pp. 67–77, 1999.
[117] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[118] D. A. Tolliver and G. L. Miller, "Graph Partitioning by Spectral Rounding: Applications in Image Segmentation and Clustering," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, vol. 1, pp. 1053–1060.
[119] S. X. Yu and J. Shi, "Segmentation given partial grouping constraints," *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. 26, no. 2, pp. 173–183, Feb. 2004.
[120] T. S. Lee, D. Mumford, R. D. Romero, and V. A. F. Lamme, "The role of the primary visual cortex in higher level vision," *Vision Res.*, vol. 38, no. 15–16, pp. 2429–2454, 1998.
[121] P. Fries, "A mechanism for cognitive dynamics: neuronal communication through neuronal coherence.," *Trends Cogn. Sci.*, vol. 9, no. 10, pp. 474–80, Oct. 2005.
[122] A. L. Yuille and R. Mottaghi, "Complexity of Representation and Inference in Compositional Models with Part Sharing," *arXiv Prepr. arXiv1301.3560*, 2013.
[123] Y. Chen, L. Zhu, C. Lin, H. Zhang, and A. L. Yuille, "Rapid Inference on a Novel AND/OR graph for Object Detection, Segmentation and Parsing," in *Advances in Neural Information Processing Systems 20*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2008, pp. 289–296.
[124] B. A. Olshausen, C. H. Anderson, and D. C. Van Essen, "A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information.," *J. Neurosci.*, vol. 13, no. 11, pp. 4700–4719, Nov. 1993.
[125] M. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *Computer Vision ,{Â}{i} ECCV 2014*, vol. 8689, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer International Publishing, 2014, pp. 818–833.
[126] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.
[127] S. Osindero and G. E. Hinton, "Modeling image patches with a directed hierarchy of Markov random fields," in *Advances in Neural Information Processing Systems 20*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2008, pp. 1121–1128.
[128] D. Ganguli and E. P. Simoncelli, "Implicit encoding of prior probabilities in optimal neural populations," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 658–666.
[129] R. Kelly, M. A. Smith, R. Kass, and T. S. Lee, "Local field potentials indicate network state and account for neuronal response variability," *J. Comput. Neurosci.*, vol. 29, no. 3, pp. 567–579, 2010.

Tai Sing Lee (S'1985, M'1996), received his S.B. in Engineering Physics in 1986 from Harvard College, and Ph.D. in 1993 in Engineering Sciences from Harvard University, and in Medical Physics Medical Engineering from the Harvard-MIT Division of Health Sciences and Technology. He completed his postdoctoral training in primate neurophysiology in the Department of Brain and Cognitive Science at MIT in 1996, and became a faculty at Carnegie Mellon University, where he conducted interdisciplinary research, using a combination of computational and experimental techniques to study the neural mechanism and circuitry underlying visual perception. He is a full professor of computer science in the Computer Science Department and in the Center for the Neural Basis of Cognition at Carnegie Mellon University, and a recipient of the NSF Career Award, and IEEE's Helmholtz award.