Neural encoding of scene statistics for surface and object inference

Tai Sing Lee, Tom Stepleton, Brian Potetz, Jason Samonds

Computer Science Department, Robotics Institute,
and Center for the Neural Basis of Cognition
Carnegie Mellon University

In: *Object Categorization: Computer and Human Vision Perspective.* Ed. by Sven Dickinson, Ales Leonardis, Bernt Schiele, Michael Tarr, Cambridge University Press.

Correspondnece should be addressed to T.S. Lee (tai@cnbc.cmu.edu).

Dr. Tai Sing Lee
Mellon Institute, Rm 115
Center for Neural Basis of Cognition
Carnegie Mellon University
4400 Fifth Avenue, Pittsburgh, PA 15213.

Abstract:

Features associated with an object or its surfaces in natural scenes tend to vary coherently in space and time. In psychological literature, these coherent covariations have been described as important for neural systems to acquire models of objects and object categories. From a statistical inference perspective, such coherent covariation can provide a mechanism to learn statistical priors in natural scenes that are useful for probabilistic inference. In this article, we present some neurophysiological experimental observations in the early visual cortex that provide insights into how correlation structures in visual scenes are being encoded by neuronal tuning and connections among neurons. The key insight is that correlated structures in visual scenes result in correlated neuronal activities, which shapes the tuning properties of individual neurons and the connections between them, embedding Gestalt-related computational constraints or priors for surface inference. Extending these concepts to the inferotemporal cortex suggests a representational framework that is distinct from the traditional feed-forward hierarchy of invariant object representation and recognition. In this framework, lateral connections among view-based neurons, learned from the temporal association of the object views observed over time, can form a linked graph structure with local dependency, akin to a dense aspect graph in computer vision. This web-like graph allows view-invariant object representation to be created using sparse feed-forward connections, while maintaining the explicit representation of the different views. Thus, it can serve as an effective prior model for generating predictions of future incoming views to facilitate object inference.

Introduction

Visual scenes are often complex and ambiguous to interpret because of the myriad causes that generate them. To understand visual scenes, our visual systems have to rely on our prior experience and assumptions about the world. These priors are rooted in the statistical correlation structures of visual events in our experience. They can be learned and exploited for probabilistic inference in a Bayesian framework using graphical models. Thus, we believe that understanding the statistics of natural scenes and developing graphical models with these priors for inference are crucial for gaining theoretical and computational insights to guide neurophysiological experiments. In this paper, we will provide our perspective based on our works on scene statistics, graphical models and neurophysiological experiments.

An important source of statistical priors for inference is the statistical correlation of visual events in our natural experience. In fact, it has long been suggested in the psychology community that learning due to coherent covariation of visual events is crucial for the development of Gestalt rules (Koffka 1935) as well as models of objects and object categories in the brain (Gibson 1979, Roger and McClelland 2004). Nevertheless, there has been relatively little research on how correlation structures in natural scenes are encoded by neurons. Here, we will first describe experimental results obtained from multi-electrode neuronal recording in the primary visual cortex of awake-behaving monkeys. Each study was conducted at least on two animals. These results reveal mechanisms at the neuronal level for the encoding and the influence of scene priors in visual processing. Insofar as these mechanisms likely emerge from Hebbian learning (Hebb 1949) or its variant that is sensitive to the timing of the events (Markram et al. 1997), we conjecture that the same basic principles and mechanisms are universal, repeating themselves throughout the visual cortex. We argue that extending these principles to the inferotemporal cortex could provide a new perspective on object representation.

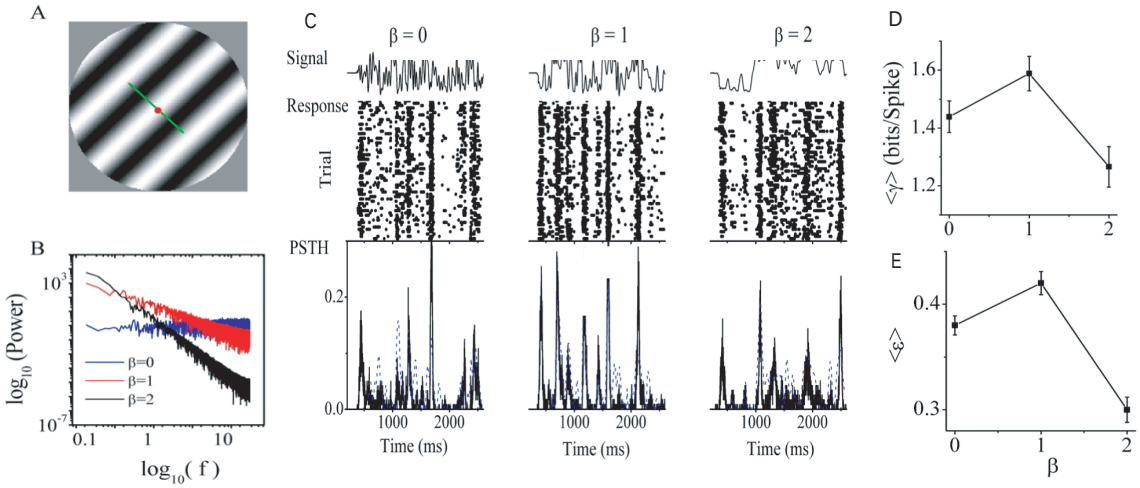

Figure 1: Neuronal preference for 1/f correlational structures in natural scenes. A: An example of the sine-wave grating stimulus input. The arrows indicate the directions of motion of the grating. B: The power spectra for the three classes of $1/f^{\beta}$ signals with $\beta = 0$ (white), $\beta = 1$ (pink) and $\beta = 2$ (brown) respectively. C: The top row depicts the phase of the motion of the input grating. The second row is the raster plot of a V1 neuron's response to the three sequences of input. The third row is the PSTH of the neuron's response. The $\beta = 1$ signal evokes the most robust response in the neuron, as indicated by the tall peaks, which reflect repeatability of the response when the same stimulus was presented. The solid lines represent the actual neural responses, and the dot lines represent the predicted responses based on the models recovered respectively from each class of signals. The reliability of the neuronal responses for 1/f signal also lead to better predictability of its recovered kernel. Coding efficiency (D) and information transmission rate (E) both exhibit a preference for the 1/f correlational structure. Adapted from Yu, Romero and Lee (2005).

Neural coding of statistical correlations in natural scenes

In natural scenes, there are a variety of correlation structures. First, at a single point in space, a visual signal is correlated over time. Second, different aspects of the visual signal, such as luminance and binocular disparity, can be correlated due to interaction of luminance and depth in three-dimensional (3D) scenes. Third, visual signals are correlated across space when they arise from a single surface. How are these correlations encoded in the nervous system? We found two potential mechanisms: (1) tuning properties of individual neurons, and (2) connectivity among neurons. Neurons develop tuning properties that can capture correlation structures in the feed-forward input at the earliest stages of processing, and correlation in the input signals will likely exhibit correlation in the tuning properties in the different feature dimensions. Spatially and temporally correlated visual events are likely encoded in recurrent (horizontal and feedback) connections between neurons. On a conceptual level, it might be meaningful to consider the former process as extracting unified information from earlier areas and the latter process as unifying associated representations in the same visual area.

Encoding correlation structures in tuning properties

Neurons are often characterized by tuning properties, i.e. whether they exhibit preferences for certain stimulus parameters along a certain feature dimension. To explore whether and how a correlation structure in visual features is encoded in the tuning properties of neurons, we have performed two neurophysiological experiments in primary visual cortex (V1). The first experiment concerns correlation structures with respect to time, and the second experiment concerns a correlation structure between depth and luminance cues at a single point in space. In these experiments, as well as all other physiological experiments presented in this article, the recordings were done on awake behaving monkeys performing a simple fixation task using multiple electrodes, each isolating single-unit activity of individual neurons.

Natural signals often exhibit similar statistical properties at all scales, and are thus described as having self-similar, or fractal, structure. One consequence of this scale-invariance property is that natural signals typically have a power spectrum that obeys a power-law, of the form $1/f^{\beta}$ (Ruderman and Bialek 1996, Potetz and Lee 2006). In the time domain, natural signals are characterized by a $1/f$ power spectrum, meaning that the total amount of power in each octave of frequency is the same for every octave. We evaluated how V1 neurons respond to noise signals with different power spectra, i.e. white noise ($\beta = 0$), pink or natural noise ($\beta = 1$), and brown noise ($\beta = 2$) as shown in Figure 1A and 1B. We found that signals with $\beta = 1$ were preferred over white and brown noise in the robustness and reliability of the response and in the amount of information about the stimulus transmitted in each spike (Yu et al. 2005). Figure 1C shows $\beta = 1$ signals generated more repeatable responses (higher peaks in the peri-stimulus time histograms PSTH). Figures 1D and 1E show that the coding efficiency and information transmission rate are highest for $\beta = 1$ signals. In a related experiment in the auditory system, Garcia-Lazaro et al. (2006) discovered that auditory neurons are also sensitive to this correlational structure. These findings suggest that neurons in early sensory areas are adapted for this important temporal correlation in natural signals, which might be a key factor underlying why neurons prefer natural stimuli over other stimuli as some studies have earlier observed.

In a second experiment, we tested a prediction generated by the discovery of an inverse correlation between depth and luminance in 3D natural scenes. Using co-registered 2D color images and laser-acquired 3D range data, Potetz and Lee (2003) found that there is an inverse correlation (r = -0.18) between values of luminance and the depth of the camera from the point of observation. That is, brighter regions in an image tend to be nearer. Centuries ago, Leonardo da Vinci observed a perceptual phenomenon whereby brighter surfaces are perceived to be nearer, all other things being equal. This observation has been exploited by artists in paintings. This study therefore provides an ecological reason for such perception, demonstrating that correlational structures in natural scenes might be explicitly encoded as priors in our visual system. These correlational structures, we found, arise primarily from shadows in natural scenes (e.g., farther surfaces are more likely to lie within shadow) and turns out to be especially useful information for inferring depth from images (Potetz and Lee 2006) (Figure 2).

How are correlation structures between visual cues encoded in neurons? We know that neurons in V1 are tuned to binocular disparity (Cumming and DeAngelis 2001)–e.g., some neurons prefer near surfaces while other neurons prefer far surfaces relative to the fixation plane. Do neurons tuned to nearer surfaces also tend to prefer brighter surfaces? Indeed, we found this tendency to be the case at the population level in V1 (Potetz et al. 2006). Many neurons exhibit sensitivity (tunings) to both visual cues simultaneously. Among a

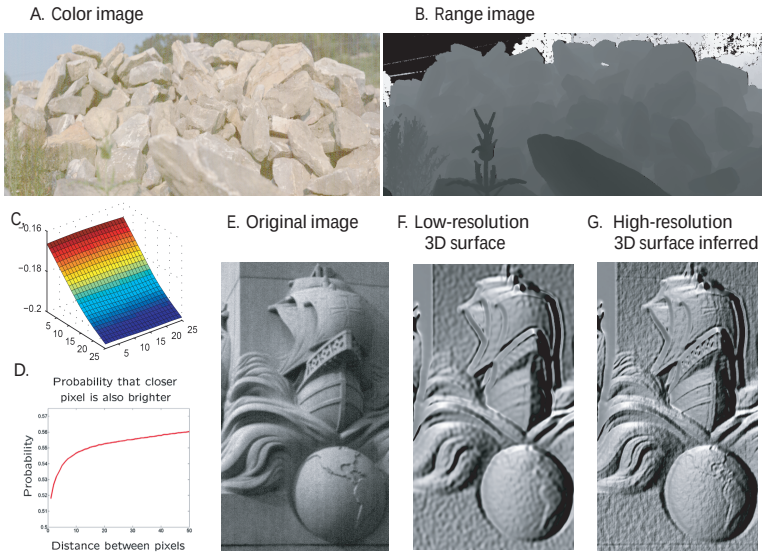

Figure 2: A: An example color image from our database. B: The corresponding range image. In this image, due to the shadowing in the rocky cliff face, the correlation between depth and pixel intensity is -0.37. C: Typical correlation between an intensity pixel and the surrounding range pixels across patches centered at intensity pixel's location. On average, the correlation between image intensity and range value at the same location is $r = -0.18$ – as shown by (13,13) in the graph. D: Given two pixels, the brighter pixel is usually closer to the observer. E: An example image from our database. F: The corresponding range image was subsampled to produce a low-resolution depth map, and then (for illustration purposes) rendered to create an artificial, computer-generated image. Next, a computer algorithm was used to learn the statistical relationship between the low-resolution 3D shape of (F) and the 2D image of (E). This includes both shading and shadow (nearness/brightness correlation) cues. In this example, shadow cues were stronger. This learned statistical relationship was then extrapolated into higher spatial frequencies to estimate the high-resolution 3D shape, shown in (G). Some high resolution depth features are 'hallucinated' by the algorithm correctly such as the cross on the sail, and the details on the globe. Adapted from Potetz and Lee (2003) and Potetz and Lee (2005).

population of 47 neurons, there is a strong trend for near-tuned cells to prefer a bright surface versus a dark surface with a statistically significant correlation between the disparity and brightness preference of $r = -0.39$ with $p = 0.01$ (Figure 3). Thus, correlation between the two cues in natural scenes is reflected in the joint tunings to the two cues at the population level. This is the first physiological finding relating the tuning curves of individual neurons across two different depth-defining cues, and might be the physiological underpinning of psychophysical studies that revealed the interaction of different visual cues on depth perception (Moreno-Bote et al. 2008).

The idea that the tuning properties of neurons are capable of capturing correlation structures in natural scenes is by no means new and is in fact the fundamental assumption for a number of seminal theoretical studies on the emergence of simple cell receptive fields (Olshausen and Field 1996), and of tuning properties of retinal ganglian neurons (Atick and

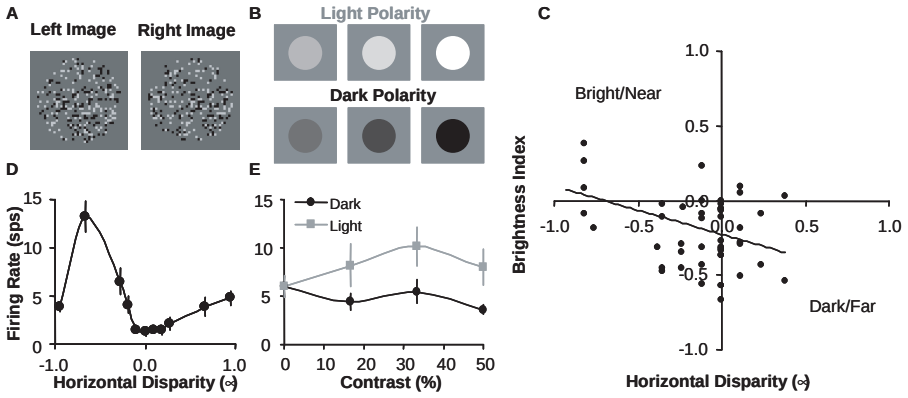

Figure 3: Testing for a correlation between disparity and luminance preference in V1. A: Random dot stereogram (RDS) stimulus. B: Light and dark spot stimuli. C: Scatter plot of brightness preference (light response - dark response)/(light response - dark response) versus disparity preference (disparity with peak response). D: Sample disparity tuning curve. E: Sample contrast response curve for light and dark polarity. From Potetz, Samonds and Lee (2006).

Redlich 1992). However, all these theoretical studies are ad-hoc, primarily providing explanations on well known existing physiological findings on neurons' tuning properties. Our neurophysiological experiments are motivated and predicted by scene statistics results, and, for the first time, yielded brand new neurophysiological evidence that is consistent with the theoretical assumption/prediction on neural encoding of correlational structures.

Encoding competitive constraints in neuronal interaction

While visual features or visual cues co-occurring at the same spatial regions can be encoded in the tuning properties of the neurons, as in the case of simple cells encoding the correlated activities of LGN neurons aligned in a particular orientation, some visual entities cannot occur simultaneously or are mutually exclusive. For example, given an observed surface, the hypothesis that it is at a particular depth is incompatible with the hypothesis it is at a different depth. This scenario requires neurons representing different hypotheses to compete with each other in explaining the observed image patch. The early computational model for stereopsis proposed by Marr and Poggio (1976) required a uniqueness constraint that stipulates that neurons coding for different disparities at the same location should inhibit one another. The independent component (sparse coding) explanation for the emergence of of simple cells' receptive fields (Olshausen and Field 1996) also requires similar competitive interaction. Later on, we will discuss how this uniqueness constraint is also relevant to object representations. Curiously, little is known about the competition between neurons in a cortical hypercolumn that are analyzing information within the same spatial window. To understand the neural implementation of mutual exclusion or the uniqueness constraint, we have carried out an experiment to study the interaction of neurons of different disparity tunings with spatially overlapping receptive fields.

In this experiment, while the monkeys fixate at a spot on a computer monitor, different depth planes rendered in dynamic random dot stereograms were presented in a 5 degree diameter

aperture for 1.2 seconds, one at a time. These are the stimuli typically used to measure the disparity tuning of a neuron. The novel component of our study was that we recorded from multiple neurons simultaneously, using multiple electrodes or a single electrode, and studied their interaction. The separated spikes from single electrodes or from two different electrodes recording from neurons with overlapping receptive fields were subject to cross-correlation analysis. Interaction strength between neurons is typically measured by cross-correlating spike trains, a measurement that can be positive or negative, reflecting the likelihood of a spike from one neuron coinciding with a spike from the other neuron. Correlation between two neurons' spike trains is first computed, and then the part of the the cross-correlation that can be attributed to the firing rate covariation of the two neurons is discounted, and finally the estimate is normalized by the firing rates or variation in firing rates. That is, if the interaction strength between a pair of neurons is fixed, this measurement will remain constant irrespective of the stimulus being presented and how the neurons respond to the stimulus. In the end, a strong positive or negative cross-correlation suggests that the neurons are connected in some manner within the cortical network.

We found that neurons with very different disparity tunings exhibited negative correlation (competitive interaction) in their spiking activity (Samonds et al. 2007, Samonds et al. 2008). Figures 4A,D show the receptive field locations and tuning relationships of a typical pair of neurons that exhibit competitive interaction. Figures 4B and 4C show the temporal evolution of neuronal interaction over time as a function of the depth as defined by the disparity of the presented random dot stereograms. These graph are population results, averaged across all competitive pairs in the population, aligned by the negative correlation peaks. Figure 4B shows a significant early negative correlation component (competitive interaction), superimposed on the baseline correlation, between the two neurons. This is most severe at where their disparity tunings diverge the most (Figure 4E, 4C). This interaction is accompanied by the emergence of the disparity tuning and the sharpening of disparity tuning over time(Figure 4F), i.e. an improved estimate of the image depth. These neurons that exhibit competitive interaction are different not only in their disparity tunings, but also in motion direction tuning as well. It remains to be resolved whether neurons common in some cue dimension but different in other cue dimensions would still engage in competitive interaction exclusively, or whether the interactions between these neurons are cue-dependent.

This is the first piece of evidence that neurons analyzing the same spatial location engage in competitive interaction that is consistent with the uniqueness constraint during stereopsis computation (Marr 1981, Marr and Poggio 1976). It is well known that inhibitory connections and suppressive interactions are restricted to be local (Lund et al. 2003), but earlier studies tend to suggest that these inhibitions are not specific to stimulus or the cells' tuning properties (Das & Gilbert 1999; Bosking et al. 1997; Shapley et al. 2003). Our results suggest the competitive interaction does depend on the tuning properties of the neurons. A uniqueness constraint however implies a winner-take-all scenario among the neurons, which might not be a desirable property in most cases. It is more desirable to encode a posterior probability distribution of the different hypotheses using the neuronal population at each location to enable a more robust inference and representation. This is analogous to beliefs at a node in a graphical model of a Bayes net (Rao 2004, Potetz and Lee 2008, see also Knill and Pouget 2004). The fact that neurons will continue to respond to suboptimal features also suggests the uniqueness constraint is probably a soft one. We will in later section discuss the role of such constraint in object representations.

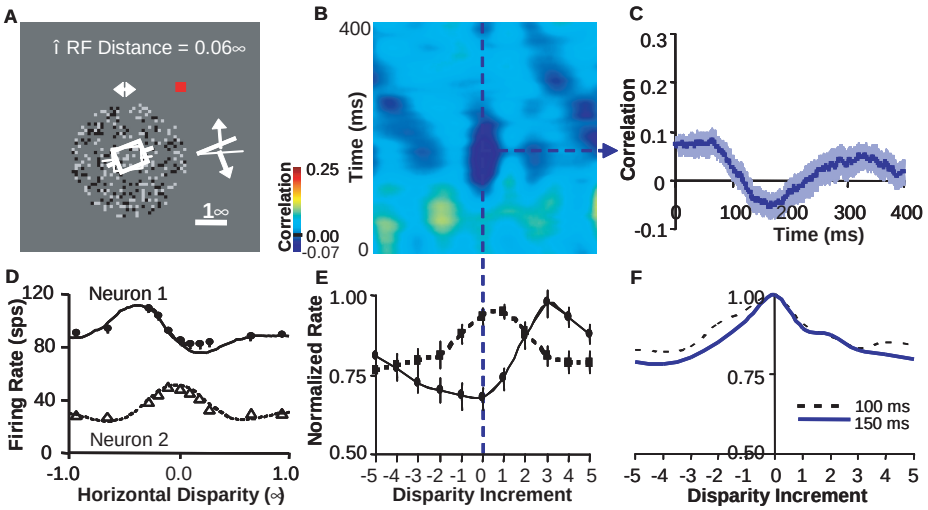

Figure 4: Local competitive interaction. A: Receptive fields, preferred orientation, and direction of motion (white) overlaying the RDS stimuli for an example competitive pair of neurons (red square is fixation point). B: Population summary of interaction strength (correlation) versus time and disparity for 17 neuronal pairs with antagonistic disparity tuning. C: A section of B noted by the dashed line. D: Disparity tuning for neuronal pair described in A. E: Population summary of disparity tuning for same pairs described in B. F: Population summary of sharpened disparity tuning after competitive interaction described in B. Adapted from Samonds et al. (2008).

Encoding spatial correlations in neuronal connectivity

Our next question concerns the neural encoding of correlated events across space, such as the co-occurrence of features belonging to a surface or parts belonging to an object at different spatial locations. The standard argument is that neurons coding for events that occur together simultaneously across space can lead to the formation of specific neurons downstream that encode this joint event. Recursively repeating this principle along the visual hierarchy can conceptually allow the formation of codes for features, subparts, parts, and objects in a compositional architecture. In order to avoid a combinatorial explosion in the number of codes required at the higher level, Geman (2006) proposed that higher order structures or representations can be dynamically constructed by composing reusable parts at each level along the visual hierarchy. The parts themselves are meaningful entities, learned from natural scene statistics, and are reusable in an enormous assortment of meaningful combinations. Such compositional hierarchies provide structured representations over which a probability distribution may be defined and used as prior models in scene interpretation. The key concept is that frequently co-occurring features are encoded explicitly by neurons, while occasionally co-occurring features are encoded transiently through the correlated activities or synchrony of the existing neurons.

There is indeed some evidence that the correlated activity or functional connectivity between

V1 neurons across space is dynamic and stimulus dependent. A number of multi-electrode neurophysiological studies have shown that the interaction between a pair of neurons is dynamic (Singer 1999, Samonds et al. 2006, Samonds et al. 2007, Kohn and Smith 2005). The main finding emerging from these studies is that the interaction strength (also termed effective or functional connectivity, spike correlation, and synchrony) between a pair of neurons is not fixed but varies as a function of the stimuli and stimulus context presented to the neurons in relation to the tuning properties of these neurons, and is typically greatest at the peak of the product of the two tuning curves. Furthermore, the vast majority of measurements of interaction between neurons has been positive or facilitatory in nature. Thus, these findings on stimulus-dependent correlated activities of V1 neurons can reflect such dynamic functional connectivity suggested by Geman.

Although Geman's composition machine is mediated primarily by bottom-up feedforward connections, in the visual cortex, besides these, there are vast numbers of horizontal and feedback connections. What are the functional roles of these recurrent connections, particularly the horizontal connections? Most of extra-classical surround effects neurophysiologists have observed are suppressive in nature. Thus, lateral inhibition or surround suppression are thought to be the dominant action of the horizontal connections, notwithstanding 80 percent of the synapses on the horizontal collaterals are excitatory in nature. We propose that the horizontal connections are implementing the constraints on how the different parts across space can vary relative to one another when the parts are being dynamically composed into a larger entity. In computer vision, this is modeled in terms of Markov random field models which allow information from a node's surrounding region to influence its interpretation of the stimulus in its analyzing window. A node here can be represented by a population of neurons analyzing the same spatial location. The connections define the statistical distribution of the relationship between different features or parts across space – how the distance and relative orientation between the different parts tend to vary in natural scenes, and what range of variation is permitted when higher order structures are composed and interpreted.

The simplest constraint used in computer vision for surface inference and segmentation is the ubiquitous *surface smoothness* constraint. That is, surfaces tend to be smooth locally. More precisely, the variation in the surface orientations might follow a statistical distribution such as a Gaussian or a Laplacian distribution. This constraint arises naturally from natural scene statistics. In our study of co-registered range and color images, we have examined the nature of this smoothness constraint (Potetz and Lee 2003). Figure 5A shows the correlation between pixels in range data as a function of distance, and Figure 5B shows the correlation between pixels in the image data as a function of distance. Both show some kind of exponential decay in correlation as a function of distance that can be fitted well with Laplacian distributions. The decay in correlation is significantly slower in the range data than in the luminance data, reflecting the fact that surfaces tend to be smooth, and that variation in surface depth was less than the variation in the luminance patterns or markings on the surface. This predicts that neurons with similar disparity tunings or other feature tunings should interact cooperatively, and their interaction strength should drop off exponentially as a function of distance according to scene statistics.

To test this hypothesis, we carried out an experiment to measure how V1 disparity-tuned neurons interact across space when presented with different depth planes as rendered by dynamic random dot stereogram, one depth at a time. This is in fact the same experiment described in our earlier discussion on the uniqueness constraint, except now we are consid-

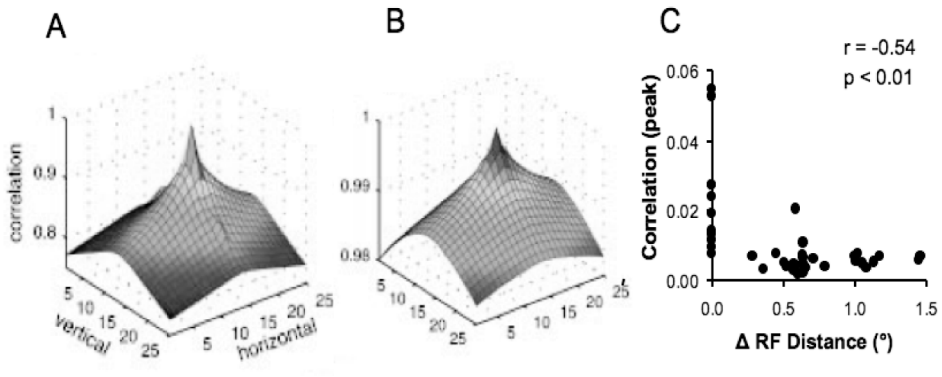

Figure 5: A: Correlation between intensity at a center pixel (13,13) and all the pixels in the same intensity patch, over all image patches (e.g. Figure 2A). B: Correlation between range at a center pixel (13,13) and all the pixels in the same range patch, over all image patches (e.g. Figure 2B). C: Peak in the correlogram of the activities of a pair of neurons that exhibit positive correlation as a function of distance between the receptive fields of the two neurons in visual space in degree visual angle.

ering the interaction between neurons with spatially distinct receptive fields. We observed significant positive cross-correlation of the spike trains for a variety of neuronal pairs, but most noticeably for neurons with similar disparity tunings. Excitatory interaction, as indicated by positive correlation in neural activities, extends a greater distance (a few mm in cortical distance) than the more local inhibitory interaction discussed earlier.

Figures 6A and 6D show the receptive field locations and disparity tuning curves of a typical pair of neurons exhibiting excitatory interaction. The neurons typically have spatially distinct receptive fields and very similar disparity tuning. The rest of the Figure 6 shows a population average of the interaction between 41 similar pairs of neurons, aligned by their peaks of strongest positive correlation, revealing that the strongest interaction occurred at the disparity where the two tuning curves intersect, i.e. shared the most in common. There appeared to be temporal dynamics in the neuronal interaction: the earlier phase (the first 100 msec) is non-stimulus specific, while the later phase (150-400 msec) is stimulus specific. The early phase in the correlated responses is likely due to the simultaneous burst in neuronal responses due to stimulus onset. That this early correlation tends to be strongest for stimuli that both neurons prefer the least also suggests that this correlation might arise from a common suppressive input shared by both neurons, presumably from neurons which prefer that disparity, as a manifestation of the uniqueness constraint (see also Figure 4). The second phase of strong and positive interaction likely reflects mutual facilitation, as it occurred only when the stimulus is precisely of their shared preferred disparity (Figures 6B, 6C), potentially reflecting the implementation of the continuity constraint. The initial competitive interaction is accompanied by a development of disparity tuning, while the later interaction is accompanied by further sharpening of the disparity tuning curves (Figure 6F). Figure 5C shows preliminary results that indicate the strength of positive correlation between neurons with similar disparity tunings dropped off with distance rather rapidly. More data however are required to allow a quantitative comparison with the prediction of the

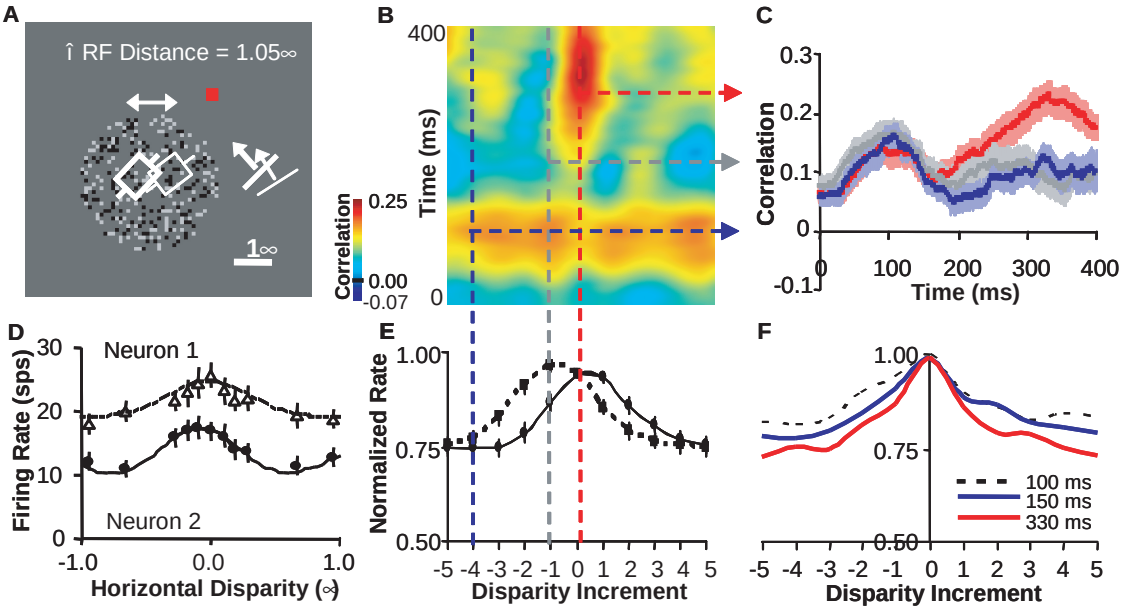

Figure 6: Cooperative interaction. A: Receptive fields, preferred orientation, and direction of motion (white) overlaying the RDS stimuli for an example cooperative pair of neurons. B: Population summary of interaction strength (correlation) versus time and disparity for 41 neuronal pairs with similar disparity tuning. C: Sections of B noted by the corresponding dashed lines. D: Disparity tuning for neuronal pair described in A. E: Population summary of disparity tuning for same pairs described in B. F: Population summary of sharpened disparity tuning after competitive interaction described in Fig. 4B (blue) and cooperative interaction described in B (red). Adapted from Samonds et al. (2008).

scene statistics.

The observation that correlation in neuronal activities is a function of disparity tunings of neurons (Samonds et al. 2007) extends the earlier observation based on orientation tunings (Ts'o et al. 1986, Singer 1999, Samonds et al. 2006) to the depth domain. It is worth noting that the correlation of neural activities discussed here is a measure computed within 25 to 50 msec window, distinct from the so-called fast-time synchrony (correlation within a 1-5 msec window) or slow-time spike count correlation (correlation within a temporal window ranging from 500 msec to seconds). Slow spike count correlation might arise from fluctuation of the states of the system such as attention, arousal or other mechanisms. Fast-time synchrony is important for understanding one-to-one neuronal connectivity, and might be important for von der Malsburgh's 'binding by synchrony' concept. It is however not certain whether fast-time correlation is necessary for Geman's compositional hierarchy. The intermediate time correlation we observe is informative of the dynamics of neuronal interaction during computation, and might be sufficient for generating the higher order codes. This hypothesis however needs to be confirmed by computational and neurophysiological experiments.

The surface smoothness constraint arises from the fact that visual cues (texture, disparity, color) tend to be smooth and continuous when they belong to the same surface. The co-occurrence and correlation of these visual cue events can lead to the formation of connections between neurons with similar tuning properties across space by classical Hebbian learning mechanisms. The contour version of this constraint is the contour smoothness constraint, or the association field, which has been demonstrated in both psychophysical experiments (Field et al. 1993) and scene statistics studies on the statistical distribution of luminance edge signals across space (Geisler et al. 2001; Sigman et al. 2001; Elder and Goldberg 2002). Such an association field has been shown to be useful for contour completion (Grossberg and Mingolla 1985, Williams and Jacobs 1997) and might be part of the underlying mechanisms for illusory contour representation in the early visual cortex (Lee and Nguyen 2001). Von der Malsburg and colleagues have shown that such facilitatory connectivity patterns can be learned by Hebbian learning based on the moving edges of objects in video in an unsupervised manner (Prodohl et al. 2003). Our evidence (Samonds et al. 2007, Samonds et al. 2008) on a neural substrate for a smoothness constraint in depth suggests that the association field concept might generalize beyond contours to statistical priors and Gestalt rules for organizing surfaces, and furthermore to statistical constraints for organizing configural parts of objects in object representation. In summary, our conjecture is that horizontal connectivity is not simply for mediating surround inhibition. Rather, it can enforce statistical constraints on the spatial and possibly temporal relationships between parts of surfaces and objects to facilitate the elimination of improbable solutions in generating the higher order representations, and to resolve ambiguity during perceptual inference.

Computational implications on cortical object representation

In the last section, we have discussed evidence for three major neural mechanisms for encoding statistical priors in the natural scenes: 1) feedforward convergent connections for encoding correlational or conjunctive structures between different visual cues/features occurring at the same spatial location in neuronal tuning properties; 2) competitive interaction among neurons at the same spatial location for encoding the uniqueness constraint or enforcing mutual exclusiveness of hypotheses; 3) cooperative recurrent (lateral and feedback) connections to encode spatial correlations of features or the distribution of the variations among their parameters. As anatomical architecture is fairly uniform across the different visual areas in the hierarchical visual cortex, these three fundamental mechanisms are likely repeated in each visual area to generate a hierarchy of priors to bring about hierarchical Bayesian learning and inference (Lee and Mumford 2003).

Even though our neurophysiological experiments, as discussed, have focused on the early visual cortex and on the issues of surface and depth inference, the lessons we learned should be relevant to understanding computational architecture in the higher visual areas such as IT (inferotemporal cortex) for object representation and analysis. The main question we asked is: what is the functional role of the horizontal connections in higher visual areas such as V4 and IT, particularly in the context of object representation and inference? Interestingly, almost all of the popular neural models on object representation construct a hierarchy primarily based on convergent feedforward connections (Fukushima 1980, Foldiak 1991, Riesenhuber and Poggio 1999, Wallis and Rolls 1997, Wiskott 2002, Geman 2006). Lateral connections, if considered at all in these models, are used to implement competition or inhibition within each level based on the prevalent neurophysiological reports on surround suppression. We have provided evidence in the last section that lateral connections in the early visual areas could be encoding spatial constraints such as the smoothness prior seen

in popular models in computer vision like Markov random fields. In higher visual areas, receptive fields become larger and larger as one traverses up the visual hierarchy, 4 times wider at V4 relative to V1 at the same eccentricity, and covering much of the visual field at the level of IT. IT neurons have been shown to be selective to specific views of objects but by and large invariant to their positions and scale. The cortical layout of IT is no longer retinotopic as in early visual areas, but exhibits some clustering among represented objects. Thus IT's horizontal connections must be encoding some relationships between objects that are no longer defined in terms of space. What could these relationships be?

IT horizontal connections for encoding temporal association of views

Our conjecture is that horizontal connections between IT neurons can be used to encode temporal association of different views of objects in our visual experience. These connections can serve two purposes: 1) achieving invariance with flexibility, 2) generating predictions during imagination and inference.

Learning features that are invariant within each class while maintaining enough specificity to allow discrimination between different classes is the central issue in object representation. An object's appearance can change dramatically in different poses, perspectives and lighting conditions. How does the visual system learn to recognize an object being the same despite its multitude of possible views?

In our visual experience, the world is dynamic either due to changing illumination, object shape deformation, or relative motions of objects in the scene. Under these conditions, the temporal contiguity of visual events offers a powerful cue for our visual systems to link the different appearances of an individual object together as its appearance changes. That is, most objects present in one instant will likely be present in the near future. Any visual pattern that can be measured for specific object will likely exhibit relatively smooth changes within small-to-medium time intervals. A dining couple in a restaurant scene might tilt their heads, smile, and speak, but neither is likely to disappear or spontaneously transform into another object. This principle of persistence or smooth variation is either implicit or explicit in a wide variety of computer vision tasks, especially tracking. Thus, by observing the dynamic behavior of objects in a visual scene over time, one can develop a set of dynamically linked observations of the objects themselves, which in turn can inform the construction of equivalence classes of visual patterns representing the same object.

The idea that temporal correlation of visual events can promote invariance learning has been explored by a number of earlier neural models (Foldiak 1991, Wallis and Rolls 1997, Wiskott 2002), even though learning from video is still at its infancy in computer vision. The earlier neural models however used a feed-forward network to learn the invariance based on gradual convergence of inputs from neurons coding different views onto the downstream neuron. Typically, a downstream neuron learns to associate two visual stimuli appearing in rapid succession as same based on the trace-learning rule, which stipulates that lingering activity of a downstream neuron responding to a first stimulus will potentiate its synapses to respond to a second stimulus as well, provided the two stimuli appear within a short time window (300-500 msec). A Hebbian-like mechanism will eventually cause the cell to respond equally well to both stimuli after repeated viewings. The disadvantages of such feedforward networks are threefold: first, specificity of a particular view (pose) of an object is lost when multiple views are converged into a single entity, so that a neuron coding an object in an invariant manner will necessarily have no idea of the pose being seen; second, a hierarchy

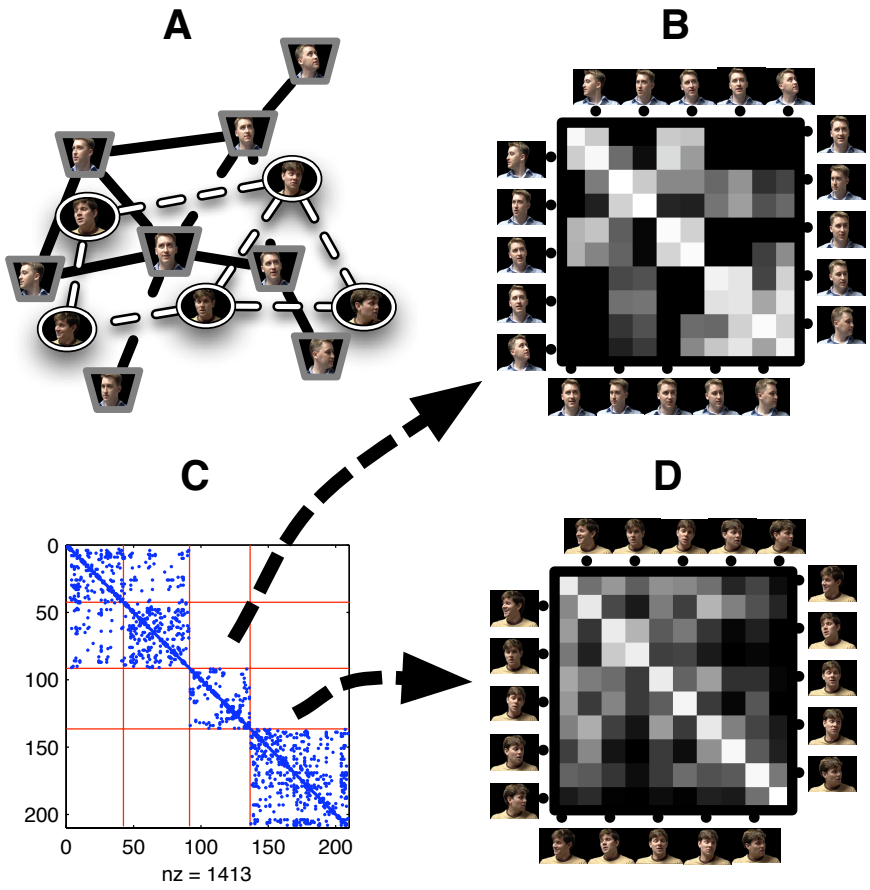

Figure 7: A: Diagrammatic depiction of linking together object views into view based object models with probabilistic transitions between views (see text for details). C: Sparsity structure of a Markov transition matrix over a collection of object views: blocks along the diagonal correspond to different objects. B, D: Connectivity between some views of two learned objects. Transitions in matrices are created by marginalizing out other object views; the intensity map is nonlinearly compressed to show detail.

with many layers is required, as invariance has to be achieved gradually by combining a few views at a time at each level; third, these feedforward networks cannot be used to predict what views will be seen next for the purpose of furnishing prediction to facilitate inference.

While the trace-learning rule might be useful for associating events based on convergent input, we argue here that the spike-timing dependent plasticity learning rule (Markram et al. 1997) might be useful for learning a lateral association network in each visual area based on temporal contiguity of visual events: as the synapses between a pre-synaptic neuron representing one view, and a post-synaptic neuron representing a subsequent view will be reinforced during visual experience. The lateral connections in such a network link the different views of an object in a graph, which is disconnected from members or nodes of the graphs representing other objects.

Figure 7A illustrates the essence of these ideas. Here, views of two distinct objects (persons) are arranged in a simulated perceptual space, where distances between views reflect a measure of low-level perceptual similarity. There is considerable overlap between these two objects within the space: frontal views of the two different persons' faces are more similar to each other than the frontal view and the profile view of the same person. The horizontal connections, which can be considered as transition edges in a hidden Markov model, offer new ways to measure similarity based on the temporal correlations of the views observed in our experience. For example, we might propose that the similarity between two different views is the marginal probability of making transition from one view to the other in a fixed number of steps. The similarity between views (e.g. frontal views) that are weakly connected or not connected by any path is thereby zero or nearly zero, even though they resemble each other in low-level image space.

We have implemented an unsupervised learning system that takes many short clips of videos with objects exhibiting a variety of pose changes and learns the horizontal connections based on the temporal contiguity and the perceptual similarity of the visual events. Figure 7 presents the horizontal connection matrix among 209 views learned from four different objects the system observed. This is essentially a transition matrix of a hidden Markov model with each state representing a particular view. The sparsity structure of the transition matrix for the learned graph connecting these views appears in Figure 7C; it exhibits a mostly block-diagonal structure corresponding to the four objects, which indicates that transitions between two different objects are unlikely (top left two blocks) or impossible (bottom right). For two of the objects, we have selected 10-member subsets of the learned views (indicated by the frames from which the view models were trained) and created probabilistic transition matrices for these subsets by marginalizing out other views from the Figure 7C matrix: the results appear in Figures 7B and 7D and exhibit a sensible connectivity structure for the chosen views (elements further from the diagonal, corresponding to transitions between non-adjacent views, are darker).

These networks might offer some advantages over the earlier models based purely on feedforward convergence in invariance learning. First, when the node coding for one view is activated by bottom-up input, the activation will spread across the network along the horizontal connections. The activation of the entire network in the next moment represents a probability distribution over what the input might become in the future. This activation spreading with its probabilistic interpretation can facilitate inference by potentiating the sensitivity of neuronal detectors to particular incoming stimuli. IT neurons related by such a graph preserve the specificity of the 'view' they are coding, but can also spread predictions about the incoming visual stimulus. These predictions specify potential appearances of whole objects and salvaging the interpretation of ambiguous and obscured stimuli. Second, these networks of facilitative, lateral connections offer a more efficient, one-layer mechanism for invariance learning: rather than gradually building invariance within a hierarchy, strongly connected components give rise to patterns of propagated co-activation within ensembles of views that together, collectively represent an object. A view-invariant neuron need only "tap" a few locations in this network of view-selective neurons with sparse, long-range connections to detect this co-activation. In addition to the improved efficiency of this method when compared to models with gradually converging layers, this account may also draw support from the fact that few intermediate view-invariant neurons have been found.

These lateral association networks are similar to an aspect graph in computer vision, with some differences. In a traditional aspect graph, nodes reflect topologically identical con-

figurations of image components; here, a node represents a range of appearance for which the model associated with the node, which characterizes an object view as configurations of component parts, is sufficiently accurate. Here, each node is a view of a face, and other nodes are different views of the face of a same person that this view is likely evolves to. The lateral connectivity describes edges in the graph whose strengths reflect the probability of the transitions between the different views of a person. Interestingly, Tanifuji and colleagues (Wang et al. 1996) found that neurons coding different views of a face are arranged in spatially adjacent cortical locations in the IT cortex.

From a statistical learning perspective, the hidden Markov model (HMM) characterization of object representation just described presents several interesting challenges. Like the mammalian visual system, we would like a computational mechanism for learning such dynamically-linked view-based object models to flexibly infer both how many objects exist in its visual world and how many views are necessary to model the objects. The latter problem, which is equivalent to inferring how many states an HMM needs to model data, has been addressed by recent "Infinite HMM" techniques based on the hierarchical Dirichlet process from non-parametric Bayesian statistics (Beal et al. 2002, Teh et al. 2006). To tackle the first issue, we are developing an extension of these models embodying the notion that views of the same object are temporally clustered in our visual experience. The key to this model is a prior that favors a nearly block-diagonal structure in the transition matrix describing the HMM's dynamic behavior. Each object's ensemble of views thus corresponds to a block of states within the model, and transitions between views in the same block are generally much more likely than transitions between views in different blocks. A particular view is assigned to one and only one block, and as such visually similar views of separate objects will be modeled with two distinct, object-specific states—a joint representation of appearance and identity that permits finer predictions of future appearance through conditioning on the knowledge of what the viewed object actually is. Finally, as with the views, the model flexibly accommodates varying numbers of objects using similar non-parametric Bayesian machinery.

V4 horizontal connections encoding spatial relationships between parts

In the above discussion, we have assumed IT neurons encode specific views of objects for simplicity in exposition. Such a view-based scheme might require explicitly storing a huge number of views of an almost infinite number of objects and their parts. Geman (2006) suggested that one can have a hierarchy of composable and reusable parts to construct object representation dynamically to avoid this combinatorial explosion problem. Each view neuron, rather than encoding an image, should really be encoding an ensemble of parts, with specific spatial configural relationship constrained by the horizontal connections one visual area below. Each of these parts in turn represent a cluster or distribution of appearances of that part, computed by some invariance transforms. The parts themselves are meaningful entities, learned from natural scene statistics and are reusable in different combinations for representing the multitude of objects.

Again, as in IT, the temporal association of the parts of objects can be learned from our dynamic visual experience and represented explicitly in intra-areal (within-area) connections in the intermediate visual areas such as V4 in the form of a Markov transition matrix. This matrix will produce predictions on how a given appearance of a part will evolve over time to produce a more invariant object representation using 'fuzzier parts'.

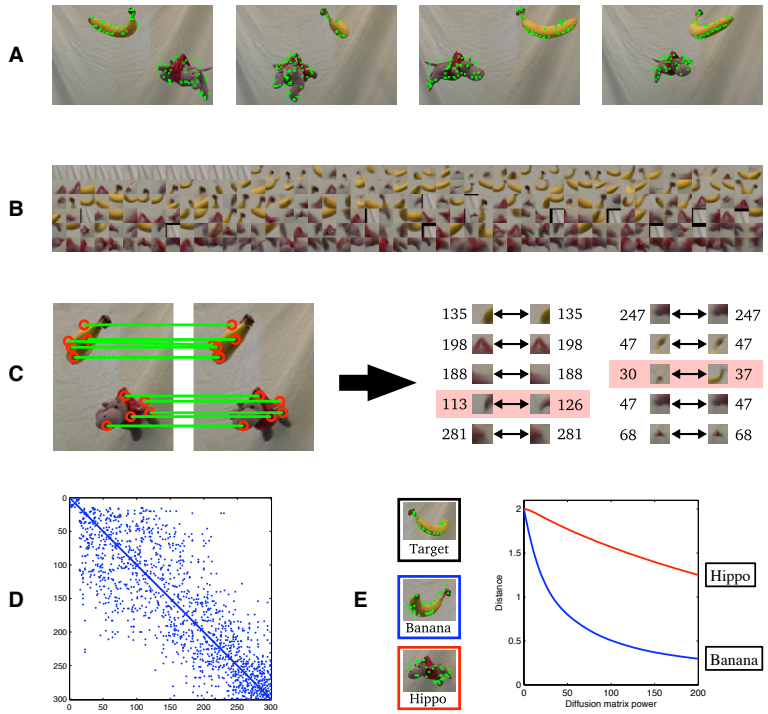

Figure 8: A demonstration example of the method for learning the appearance dynamics of low-level image features. A: The input video contains a banana and a toy hippo rotating in space. Sparse, local image features of various scales are detected in each frame. Descriptors for those features are simply the 15x15 pixel, 3 color image patches themselves. Next, K-means is run on a random assortment of extracted descriptor patches to yield a 300-bin discretization of the feature descriptor space; cluster centers appear in B. C: Features are tracked between video frames (left). Discretized descriptors of tracked features yield the appearance transitions at right, which then inform the Markov transition matrix whose sparsity structure appears in D. E: L1 diffusion distance comparison of one image feature bag-of-words histogram (Target) with two others. The banana histogram becomes similar after a few diffusion steps; the hippo histogram remains distinct.

Figure 8 illustrates these ideas. Let us assume V4 neurons encode fragments and corners with some positional slack within relatively large receptive fields. The Markov transition matrix is learned by first tracking the parts over time based on feature similarity and spatiotemporal proximity, then quantizing the stimulus space into clusters of distinct part appearances, and finally building links between corresponding parts that are sufficiently different in appearance. In this demonstration example, our system analyzed a video of a banana and a hippo moving around in a baby mobile (Figure 8A). The part-features of the moving objects are automatically learned and partitioned into 300 discrete part feature clusters using K-means (Figure 8B). These part-features are tracked over time based on spatial and temporal contiguity of the parts, as shown in Figure 8C. This allows a Markov transition matrix between the features to be built (as shown in Figure 8D). Note that this matrix is much less block diagonal than the transition matrix for the object-view representation above, since parts are more local and less object-specific.

Multiplying the transition matrix with the input observation (a delta function in the 300-tuple vector, or more generally a data likelihood function) gives you the predicted distribution of hypotheses of part appearances in the next time step. This prediction serves as a prior distribution for visual interpretation of the incoming image. Successive multiplication of the observation by the matrix simulates an experience-based diffusion process that predicts distribution of the possible appearance at different time points in the future. The diffused representation is more robust when matching the input representation to the stored representation for the following reason. For simplicity, let us consider an object's view is represented by a histogram of the occurrence frequency of a certain set of features. Without diffusion, only identical views will have identical feature count histograms. A slight change of view on an object will produce a drastically different histogram. Blurring the histogram using this temporal association matrix will create a histogram that is more tolerant against variations in the appearance of the object, so that the incoming view does not have to be exactly the same as any of the stored views in order to be recognized. Figure 8E shows the L1 distance (sum of the absolute difference between each bin for two histograms) comparison of one image feature histogram (Target) with two others, after some blurring with the Markov transition matrix. The banana histograms become more similar to each other after a few diffusion steps, while remaining distinct from the hippo histogram. This illustrates how such association of parts derived from observed dynamics can increase invariance for object recognition by integrating information about the temporal association of part appearance through the relatively local facilitatory connections between the parts.

The mathematics behind such experience-based metric for data similarity have recently been studied by Lafon and Lee (2006), among others. Conceptually, the idea is also related to Ullman's features of intermediate complexity (Ullman et al. 2002) in which fuzzy intermediate representation is shown to promote a certain degree of invariance and slack that can promote object recognition. However, our proposed blurring with HMM is more general and might be more sensible, as it reflects the invariance that is learned based on temporal association of visual events in natural scenes.

While we envision that the horizontal connections in IT encode temporally associated views, we expect the horizontal connections in the earlier retinotopic visual areas such as V4 and V1 to encode constraints about spatial relationships between the configural parts for representing objects, as in the compositional AND/OR graphs in computer vision (Zhu and Mumford 2006, Zhu et al. 2008). The temporal association matrix of the HMM model described above will likely be represented by the connections within a local cortical neighborhood such as hypercolumn. Interestingly, that would mean that within a hypercolumn in the early visual areas, there will be inhibitory connections to enforce the uniqueness constraint as well as facilitatory connections to enforce the temporal association prior.

In a recent hierarchical composition model proposed by Yuille and colleagues (Zhu et al. 2008), features and the spatial configural relationship between the features in a hierarchy can be learned in an unsupervised manner from a set of unlabeled images that contain the object to be learned based on the principle of *suspicious coincidence* and the principle of *competitive exclusion*. The first principle dictates that proposals based on frequent co-occurrence of a set of features, subject to some invariant transformation, across all the images in the training set, will be learned as higher order feature while proposals based on spurious co-occurrence of features that do not recur often enough will be considered suspicious and eliminated. When co-occurring features come together to generate a higher order proposal, these conjunctive features will undergo an invariant transform to map a

class of conjunctive features to one higher order proposal according to some rules. The invariant transformation is important as it is rather unlikely that identical image fragments will be seen across a significant number of the images in the training set. The Markov transition network discussed above can be one way to implement this 'invariant transform' by blurring each of the observed image fragments in an experience-dependent manner, but other clustering mechanisms such as K-means clustering in our first example (Figure 7) or as Zhu and colleagues' (2008) clustering method are also reasonable. The competitive exclusion principle suggests that multiple proposed higher order concepts will compete to explain an image fragment represented to that level, and that only the one that provides the best explanation across the entire training set will be chosen and remembered over the others. Thus proposed, higher order features engage in the same competition as the disparity neurons in V1 engaged in during depth inference under the 'uniqueness constraint'. This is also similar to the competition among the representations of the different objects (the graphs or webs of views) for explaining each observed view during learning and inference, as enforced by the block-diagonal prior in our extended Infinite HMM machine, discussed above. It is important to understand that the competition interaction takes place during both learning and inference, as learning requires inference at each level.

Summary and Future Directions

In this chapter, we present some of our neurophysiological evidence for how spatial and temporal correlational structures in natural scenes could be encoded in neurons in terms of their tuning properties and in terms of their connectivity. Some of these structures, such as correlation between luminance and depth, and the spatial correlation of visual cues within a surface, can serve as surface priors useful for robust probabilistic 3D surface inference. The evidence on neuronal tuning to temporal correlation reflects neurons' sensitivity to temporal events, and is partly the inspiration for our conjecture on temporal association networks in the visual cortex. While theories on the importance of correlation structures in shaping the nervous system are long standing, there has been almost no direct physiological evidence demonstrating neural encoding of correlations of natural scenes, particularly those resulting from theoretical predictions, except for the work of Dan et al. (1996) on LGN. The findings discussed here serve to strengthen those theoretical claims, as well as to reveal the diversity of strategies for encoding correlation structures in natural scenes.

We found neural evidence in support of three basic mechanisms for learning and encoding priors: (1) individual neurons' basic tuning properties, likely based on feedforward connections, are sensitive to correlation structures in natural scenes – the principle of coincidence conjunction, (2) neurons representing different hypotheses compete with one another to explain the input from the same visual window of analysis – the principle of competitive exclusion, (3) neurons with similar tunings tend to exhibit excitatory interaction with neurons with similar tunings at the same spatial location or across spatial location, possibly encoding temporal association and spatial co-occurrence of features respectively – the principle of spatiotemporal association. These mechanisms or principles, to a first approximation, have direct correspondence with the necessary and maybe sufficient computational mechanisms deployed in models that perform unsupervised learning of object representation in hierarchical composition system (Zhu et al. 2008).

The general principle underlying all these mechanisms and principles is redundancy reduction or minimum-length description codes (MDL). Mumford (1992) had argued that visual cortex encodes a hierarchy of efficient codes that minimize the redundancy in image descrip-

tion (necessary for behavior) as a whole. During learning and inference, the representations in a higher order area produce 'hallucinations' or 'image hypotheses' to earlier visual areas to explain away the bottom-up proposals they provide. Given the relative uniformity in anatomical structures and computational architecture across the different visual areas, we expect the mechanisms we observed in V1 for encoding spatiotemporal correlations are relevant to understanding the strategies of the higher visual areas such as IT and beyond for object and category representations.

However, while the three mechanisms discussed above are general, the functional roles of horizontal connections in different visual areas might be different because of the difference between neurons in these areas in terms of spatial and temporal tuning properties. We conjecture that the horizontal connections in early visual areas are enforcing spatial constraints among features and parts, while those in IT are enforcing temporal constraints among views. We envision the competitive interactions mediated by the vertical interactions within each hypercolumn within each area to be enforcing the uniqueness or competitive exclusion constraint, but the facilitatory interactions mediated by the vertical connections within each area are enforcing the temporal association constraint to generate prediction and invariance. It is worth noting that both the vertical connections and horizontal connections in IT might be shaped by temporal association, with vertical connections linking local clusters of views that are similar to one another – views separated by short time span during learning, and horizontal connections linking views that are more distinct. Diffusion among neurons in the vertical column makes representation more fuzzy and robust, but competition among these neurons will make view or pose interpretation more precise. This suggests that the interaction between groups of neurons might be dynamic in nature, exhibiting facilitatory or inhibitory interactions at different points in time. In our experiments, we did find the interaction or the effective connectivity between neurons evolve and change over time with an intermediate time scale of 30-50 msec. This dynamics reflects possibly the evolution of neuronal interaction associated with the progression of perceptual computation. It also provides some possible constraints on the time scale of interaction as well as the permissible time frame of spike integration or coincidence for learning the higher order structures by downstream neurons.

There are two new elements in our proposal for object representation that worth emphasizing. First, that lateral connections in IT can serve to encode the temporal association of visual events (e.g. views). This organization allows the generation of predictions about how an object would look over time to facilitate object recognition. We have argued that this organization might be more efficient in terms of implementation in IT for achieving invariance without losing specificity. Second, the implication of similar temporal association networks in intermediate visual areas is that these can provide *invariance transforms* on the parts or a experience-dependent measure for evaluating data similarity, which also provide a more robust object representation for learning and inference. While many of these ideas are still in the realm of speculation, they are nevertheless precise enough to be tested experimentally.

The research described here represents only baby steps in our understanding of how scene statistical priors might be encoded in the brain. Many questions and challenges along this line of research remain unanswered. First, after demonstrating that neurons are sensitive to these correlational structures, a logical next step is to understand whether and how these sensitivities can be used effectively for learning and inference. In our opinion, developing computational models that actually work is critically important for guiding neurophysio-

logical experimental research for understanding the neural representations and mechanisms underlying the solution of a problem. To this end, we have developed a computational framework based on graphical models with efficient belief propagation algorithms that can flexibly learn and incorporate a variety of priors, including higher order cliques in markov random field, for depth inference (Potetz 2007, Lee and Potetz 2008). With this framework, we have already produced state-of-the-art techniques for inference of 3D shape from shading in Lambertian surfaces. It would be important to explore how the correlational structures between depth and luminance images in natural scenes can be harnessed to improve shape inference on non-Lambertian surfaces. However, incorporating too many priors or loops in the model will cause efficiency, stability and convergence issues. Neurophysiological and psychophysical investigation can help us to select the appropriate priors and representations to use, which could be useful for overcoming these obstacles.

The AND/OR graph or hierarchical composition model of Geman (2006), Zhu and Mumford (2006) and Zhu et al. (2008) provides an elegant computational framework for conceptualizing feedforward and feedback connections in the visual cortex. Coming up with experiments to test this class of models is an important challenge in neuroscience because this framework offers an interesting perspective on how the visual system might operate that is fundamentally different from the current feedforward hierarchical model (Fukushima 1980, Riesenhuber and Poggio 1999) or the popular conceptualization of feedback in terms of attention mediated by biased competition (Desimone and Duncan 1995). In this chapter, we have advocated the important role of encoding temporal association of visual events at different levels of the cortical hierarchy for constructing invariant object representation. Although neural modelers have long recognized the importance of coherent covariation and temporal association in concept learning and invariant object representation learning, this is a relatively unexplored territory in computer vision. As discussed earlier, we are developing a computational framework based on hierarchical dirichelet processes that can organize video data into coherent view-based object classes based on temporal correlation of these visual events (Stepleton et al. 2008). This effort might provide new insights to the amazing cortical organization of visual information during learning and development.

References:

Angelucci, A., and Bressloff, P.C. (2006). The contribution of feedforward, lateral and feedback connections to the classical receptive field and extra-classical receptive field surround of primate V1 neurons. Prog. Brain Res. 154:93-121.

Atick, J.J., and Redlich, A.N. (1992). What does the retina know about natural scenes? Neural Computation 4: 196-210.

Beal, M., Ghahramani, Z., and Rasmussen, C.E. (2002). The infinite hidden Markov model. In Dietterich, T.G., Becker, S. and Ghahramani, Z. (eds) Neural Information Processing Systems 14: 577-585. Cambridge, MA, MIT Press.

Blake, A., and Zisserman, A. (1987). Visual Reconstruction. Cambridge, MA: MIT Press.

Bosking, W.H., Zhang, Y., Schofield, B., and Fitzpatrick, D. (1997). Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. J Neurosci 17: 2112-2127.

Cumming, B. G., and DeAngelis, G. C. (2001). The physiology of stereopsis. Annu Rev Neurosci. 24: 203-238.

Dan, Y., Atick, J.J., Reid, R.C. (1996). Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. J Neurosci. 1996 16(10):3351-3362.

Das, A., and Gilbert, C.D. (1999). Topography of contextual modulations mediated by short-range interactions in primary visual cortex. Nature 399: 655661.

Desimone, R., and Duncan J. (1995). Neural mechanisms of selective visual attention. Annu Rev Neurosci. 18:193-222.

Elder, J.H., and Goldberg, R.M. (2002). Ecological statistics of Gestalt laws for the perceptual organization of contours. J Vision 2: 324-353.

Field, D.J., Hayes A., and Hess R.F. (1993). Contour integration by the human visual system: evidence for a local association field Vision Res 33: 173-193.

Foldiak, P. (1991). Learning invariance from transformation sequences. Neural Computation, 3, 194-200.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4): 93-202.

Garcia-Lazaro, J., Ahmed, B., and Schnupp J. (2006). Tuning to Natural Stimulus Dynamics in Primary Auditory Cortex. Current Biology, 16(3) 264-271.

Geisler, W.S., Perry, J.S., Super, B.J., and Gallogly, D.P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. Vision Res 41: 711-724.

Geman, S. (2006). Invariance and selectivity in the ventral visual pathway. Journal of Physiology - Paris, 100, 212-224.

Gibson, J. (1979). The Ecological Approach to Visual Perception. New Jersey, USA: Lawrence Erlbaum Associates.

Grossberg, S., and Mingolla, E. (1985). Neural dynamics of form perception: Boundary completion, illusory figures, and neon color spreading. Psychological Review , 92 (2), 173-211.

Hebb, D.O. (1949). The Organization of Behavior: A Neuropsychological Theory. New York: Wiley.

Kapadia, M.K., Ito, M., Gilbert, C.D., Westheimer, G. (1995). Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in V1 of alert monkeys. Neuron 15, 843-856.

Koffka, K. (1935). Principles of Gestalt Psychology. London: Lund Humphries.

Kohn, A., and Smith, M.A. (2005). Stimulus dependence of neuronal correlation in primary visual cortex of the macaque. Journal of Neuroscience, 25: 3661-3673.

Knill, D.C., and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. Trends in Neurosciences 27(12): 712-719.

Lafon, S., and Lee, A.B. (2006). Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization. IEEE Trans. Pattern Anal. Mach. Intell. 28(9): 1393-1403.

Lee, T.S., Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. Journal of Optical Society of America, A. . 20(7): 1434-1448.

Lee, T.S., Nguyen, M. (2001). Dynamics of subjective contour formation in early visual cortex. Proceedings of the National Academy of Sciences, U.S.A. , 98(4) 1907-1911.

Lund, J.S., Angelucci, A., and Bressloff P.C. (2003). Anatomical substrates for functional columns in macaque monkey primary visual cortex. Cereb Cortex 13: 15-24.

Markram, H., Lubke, J., Frotscher, M. and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science 275: 213 - 215.

Marr, D., Poggio, T. (1976) Cooperative computation of stereo disparity. Science 194(4262):283-287.

Marr, D. (1982). Vision. San Francisco: Freeman.

Moreno-Bote, R., Shpiro, A., Rinzel, J., Rubin, N. (2008). Bi-stable depth ordering of superimposed moving gratings. Journal of Vision 8(7):20, 1-13.

Olshausen, B.A, and Field, D.J. (1996). Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. Nature, 381: 607-609.

Potetz, B. (2007) Efficient belief propagation for vision using linear constraint nodes CVPR (IEEE Conference on Computer Vision and Pattern Recognition).

Potetz, B., and Lee, T.S. (2003). Statistical correlations between 2D images and 3D structures in natural scenes. Journal of Optical Society of America, A. 20(7): 1292-1303.

Potetz, B., and Lee, T.S. (2006). Scaling Laws in Natural Scenes and the Inference of 3D Shape. NIPS – Advances in Neural Information Processing Systems 18, 1089-1096, MIT Press.

Potetz, B.R., Samonds, J.M., Lee, T.S. (2006). Disparity and luminance preference are correlated in macaque V1, matching natural scene statistics. Soc Neurosci abstract.

Potetz, B., and Lee, T.S. (2008). Belief propagation for higher order cliques using linear constraint nodes. Computer Vision and Image Understanding. In Press.

Prodhl, C, Wurtz, R. P., and von der Malsburg, C. (2003). Learning the Gestalt rule of

collinearity from object motion. Neural COmputation 15:1865-1896.

Rao, R. (2004.) Bayesian computation in recurrent neural circuits. Neural Computation, 16(1), 1-38.

Riesenhuber, M., and Poggio, T. (1999). Hierarchical Models of Object Recognition in Cortex. Nature Neuroscience 2: 1019-1025.

Rogers, T.T., McClelland, J.L. (2004). Semantic Cognition: A Parallel Distributed Processing Approach. Cambridge, MA: MIT Press.

Ruderman, D.L. and Bialek, W. (1994). Statistics of natural images: scaling in the woods. Phys. Rev. Lett. 73, 814-817.

Samonds, J., Potetz, B., Lee, T.S., (2007). Neurophysiological evidence of cooperative mechanisms for stereo computation. Advances in Neural Information Processing Systems 19, Ed. by McCallum A. MIT Press.

Samonds, J.M., Potetz, B.R., Lee, T.S. (2008). Cooperative and competitive interactions facilitate stereo computations in macaque primary visual cortex. Submitted.

Samonds, J.M., Zhou, Z., Bernard, M.R., Bonds, A.B. (2006). Synchronous activity in cat visual cortex encodes collinear and cocircular contours. J Neurophysiol 95:2602-2616.

Sigman, M., Cecchi, G.A., Gilbert, C.D., and Magnasco, M.O. (2001). On a common circle: natural scenes and Gestalt rules. Proc Nat Acad Sci USA 98: 1935-1940.

Singer, W. (1999). Neuronal synchrony: a versatile code for the definition of relations? Neuron 24: 11-125.

Stepleton, T., Gordon, G., Lee, T.S. (2008). The Block Diagonal Infinite HMM for learning object representation. In preparation.

Tappen, M.F., Freeman, W.T. (2003). Comparison of graph cuts with belief propagation for stereo using identical MRF parameters, IEEE Intl. Conference on Computer Vision (ICCV).

Teh, Y.W., Jordan, M.I., Beal, M.J., and Blei, D.M. (2006). Hierarchical dirichlet processes. J. Amer. Stat. Assoc., 101(476):1566-1581.

Ts'o, D.Y., Gilbert, C.D., and Wiesel, T.N. (1986). Relationships between horizontal interactions and functional architecture in cat striate cortex as revealed by cross-correlation analysis. J Neurosci 6: 1160-1170.

von der Malsburg, C. (1995). Binding in models of perception and brain function. Current Opinion in Neurobiology, 5:520–526.

Wallis, G., Rolls, E. (1997). Invariant face and object recognition in the visual system. Progress in Neurobiology, 51: 167-194.

Wang, G., Tanaka, K., Tanifuji, M. (1996). Optical Imaging of Functional Organization in the Monkey Inferotemporal Cortex Science, 272, No. 5268. 1665-1668.

Williams, L.R., Jacobs, D.W. (1997). Stochastic Completion Fields: A Neural Model of Illusory Contour Shape and Salience, Neural Computation, 9(4) 837-858.

Wiskott, L. (2002). Slow Feature Analysis: Unsupervised Learning of Invariances Neural Computation. 14:715-770.

Yu, Y., Romero, R., Lee, T.S. (2005). Preference of sensory neural coding for 1/f signals. Physics Review Letters, 94, 108103, 1-4.

Zhu, S.C., and Mumford, D. (2006). A Stochastic Grammar of Images. Foundations and Trends in Computer Graphics and Vision, Vol.2, No.4, pp 259-362.

Zhu, L., Lin, C., Huang, H., Chen, Y., Yuille, A. (2008). Unsupervised Structure Learning: Hierarchical Recursive Composition, Suspicious Coincidence and Competitive Exclusion. Proceedings of ECCV.