Hierarchical Bayesian Inference in the Visual Cortex

Tai Sing Lee[1], **David Mumford**[2]

[1] *Computer Science Department*

Center for the Neural Basis of Cognition

Carnegie Mellon University

Pittsburgh, PA 15213

[2] *Division of Applied Mathematics,*

Brown University, Providence, Rhode Island 02912.

Corresponding author:

Dr. Tai Sing Lee, Rm 115, Mellon Institute, Ctr for the Neural Basis of Cognition, Computer Science Department, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213.

Email: tai@cs.cmu.edu. Phone: 412 268 1060. Fax: 412 268 5060

Keywords: visual computation, visual cortex, particle filtering, belief propagation, hierarchy, probabilistic reasoning.

Seven Figures.

Running title: Hierarchical Bayesian Inference in the Visual Cortex

Traditional views of visual processing suggest that early visual neurons in V1 and V2 are static spatiotemporal filters that extract local features from a visual scene. The extracted information is then channelled through a feedforward chain of modules in successively higher visual areas for further analysis. Recent electrophysiological recordings from early visual neurons in awake behaving monkeys reveal that there are many levels of complexity in the information processing of the early visual cortex, as seen in the long-latency responses of its neurons. These new findings suggest that activity in the early visual cortex is tightly coupled and highly interactive with the rest of the visual system. They lead us to propose a new theoretical setting based on the mathematical framework of hierarchical Bayesian inference for reasoning about the visual system. In this framework, the recurrent feedforward/feedback connections in the cortex serves to integrate top-down contextual priors and bottom-up observations to structure the concurrent probabilistic inference along the visual hierarchy. We suggest that the algorithms of particle filtering and Bayesian belief propagation are relevant for understanding these interactive cortical computations. We review some recent neurophysiological evidences that support the plausibility of these ideas.

OCIS codes: 330.4060

1. The traditional model of early visual cortex

Neurons in the primary visual cortex are known to be tuned to specific elementary local features in the visual scenes. These features include location, line orientation, stereo disparity, movement direction, color and spatial frequency.[1, 2] In traditional models,[3, 4] each of the higher visual areas is similarly tuned to specific features of a progressively more complex sort (such as shape) and the analysis of the current visual stimulus proceeds in a feedforward architecture, with each visual area using input from lower areas to compute its features.

It is also known that V1 neurons are influenced by the surrounding context of the stimuli.[5–7, 9] The contextual modulations in these studies have been modeled both biologically by the facilitation and inhibition carried by axon collaterals within V1 and psychophysically as being the neural correlate of pop-out or figure-ground saliency.[6–8, 10, 11] Functionally, these modulations are hypothesized to compute extended contours and region saliency.[12, 13] The effect on the model is that V1's role in visual computation is extended in time, allowing the computed features to integrate information from larger parts of the image.

Some of the observed contextual modulations are thought to result from feedback mediated by the massive recurrent connections from the extrastriate areas to V1. The traditional model incorporates these by assigning to feedback the role of attentional selection based on the mechanisms of biased competition.[14–16] The idea of biased competition is that when multiple stimuli are presented in a visual field, the different neuronal populations activated by these stimuli will engage in competitive interaction. Attending to a stimulus at a particular spatial location or to a particular object feature, however, could bias the competition in favor of the neurons representing the attended features or locations, enhancing their responses and suppressing the responses of the other neurons. However, all biased competition models (e.g.[15, 16]) work using only lateral inhibition.

While these models give an apparently complete explanation of almost all experimental data, they use the sophisticated machinery of feedback pathways in a rather impoverished way and they persist in viewing the computations in each visual area as predominantly independent processes. In this paper, we propose a Bayesian theory of hierarchical cortical computation based both on the mathematical ideas of pattern theory[17,18] and on recent experimental evidences. We believe this theory provides a much more tightly coupled model of the processing in visual areas and especially V1/V2. We will first sketch the general theoretical framework and then in subsequent sections discuss some experimental evidence to illuminate the plausibility of these ideas.

2. A Bayesian perspective

A. Hierarchical Bayesian inference

Bayesian inference and related theories have been proposed as a more appropriate theoretical framework for reasoning about top-down visual processing in the brain.[19–24] This idea can be traced back to the *unconscious inference* theory of perception by Helmholtz.[25]

Recall that Bayes's rule proposes that with observations x_0, hidden variables x_1 to be inferred and contextual variables x_h, then a probabilistic description of their effect on each other is given in the form:

$$P(x_0, x_1 | x_h) = P(x_0 | x_1, x_h) P(x_1 | x_h),$$

where $P(a|b)$ stands for the conditional probability of a, given b. The first term on the right, $P(x_0|x_1, x_h)$ is called the imaging model, and it describes the probability of the observations, given all the other variables. One often assumes that it does not depend on x_h, i.e. x_1 contains all the facts needed to predict the observations. The second term $P(x_1|x_h)$ is called the *prior* probability on x_1, i.e. its probability before the current observations. Then the second identity:

$$P(x_1|x_0, x_h)P(x_0|x_h) = P(x_0, x_1|x_h)$$

may be used to arrive at

$$P(x_1|x_0, x_h) = \frac{P(x_0|x_1, x_h)P(x_1|x_h)}{P(x_0|x_h)}.$$

The denominator $P(x_0|x_h)$ is the probability of the observations given x_h and is independent of x_1. Hence it can simply be viewed as the normalizing factor Z_1 needed so that the *posterior* probability $P(x_1|x_0, x_h)$ is a probability distribution, i.e. equals one when summed over all values of x_1.

In the example of early vision, we let x_0 stand for the current visual input, i.e. the output of the LGN; x_1 stands for the values of the features being computed by V1; and x_h stands for all higher level information – contextual information about the situation and more abstract scene reconstructions. Thus V1 arrives at the most probable values x_1 of its features by finding the *a posteriori* estimate x_1 that maximizes $P(x_1|x_0, x_h)$. If we make the simplifying Markov assumption that $P(x_0|x_1, x_h)$ does not depend on x_h, we can then interpret the formula above as saying that V1 computes its features by multiplying the probability of the sensory evidence $P(x_0|x_1)$ with the feedback biasing probabilities $P(x_1|x_h)$ and maximizing this by competition. Note that $P(x_1|x_h)$ is similar to the attentional bias factor used in the traditional model, but here it has a richer interpretation and carries much more information. This factor now includes *all* possible ways in which higher level information about the scene may affect the V1 features x_1. Two examples are high level illumination data making probable the low level fact that certain areas of the image are in shadow; or the high level knowledge of the identity of an individual suggesting that a face should have certain proportions, as measured from the low level data in V1.

This basic formulation can also capture the interaction between multiple cortical areas, such as V1, V2, V4 and IT. In this case, each area computes a set of features, called $x_{v1}, x_{v2}, x_{v4}, x_{it}$. We again make the simplifying assumption that if, in the sequence of variables $(x_0, x_{v1}, x_{v2}, x_{v4}, x_{it})$, any variable is fixed, then the variables before and after it are conditionally independent. This means we can factor the probability model for these variables and the evidence x_0 as

$$P(x_0, x_{v1}, x_{v2}, x_{v4}, x_{it}) = P(x_0|x_{v1})P(x_{v1}|x_{v2})P(x_{v2}|x_{v4})P(x_{v4}|x_{it})P(x_{it})$$

and make our model a *graphical model* or *Markov random field* based on the chain of variables:

$$x_0 \longleftrightarrow x_{v1} \longleftrightarrow x_{v2} \longleftrightarrow x_{v4} \longleftrightarrow x_{it}.$$

From this, it follows that:

$$\begin{aligned}
P(x_{v1}|x_0, x_{v2}, x_{v4}, x_{it}) =& P(x_0|x_{v1})P(x_{v1}|x_{v2})/Z_1, \\
P(x_{v2}|x_0, x_{v1}, x_{v4}, x_{it}) =& P(x_{v1}|x_{v2})P(x_{v2}|x_{v4})/Z_2, \\
P(x_{v4}|x_0, x_{v1}, x_{v2}, x_{it}) =& P(x_{v2}|x_{v4})P(x_{v4}|x_{it})/Z_4.
\end{aligned}$$

More generally, in a graphical model, one only needs *potentials* $\phi(x_i, x_j)$ indicating the preferred pairs of values of directly linked variables x_i and x_j, and we have:

$$\begin{aligned}
P(x_{v1}|x_0, x_{v2}, x_{v4}, x_{it}) =& \phi(x_0, x_{v1})\phi(x_{v1}, x_{v2})/Z(x_0, x_{v2}), \\
P(x_{v2}|x_0, x_{v1}, x_{v4}, x_{it}) =& \phi(x_{v1}, x_{v2})\phi(x_{v2}, x_{v4})/Z(x_{v1}, x_{v4}), \\
P(x_{v4}|x_0, x_{v1}, x_{v2}, x_{it}) =& \phi(x_{v2}, x_{v4})\phi(x_{v4}, x_{it})/Z(x_{v2}, x_{it}).
\end{aligned}$$

where $Z(x_i, x_j)$ is a constant needed to normalize the function to a probability distribution.

In this framework, each cortical area is an expert for inferring certain aspects of the visual scene, but its inference is constrained by both the bottom-up data coming in on the feedforward pathway (the first factor in the LHS of each equation) and the top-down data feeding back (the second factor) (see Figure 1a). Each cortical area seeks to maximize the probability of its computed features x_i by combining the top-down and bottom-up data using the above formulae (the Z's can be ignored). The system as a whole moves, game-theoretically, toward an equilibrium in which each x_i has an optimum value given all the other x's. In particular, feedback from all higher areas can ripple back to V1 and cause a shift in the preferred features computed by V1. Thus long latency responses in V1 will tend to reflect increasingly more global feedback from abstract higher-level features, such as illumination and the segmentation of the image into major objects. For instance, a faint edge could turn out to be an important object boundary after the whole image is interpreted, although the edge was suppressed as a bit of texture during the first bottom-up pass.

The feedforward input drives the generation of the hypotheses, and the feedback from higher inference areas provides the priors to shape the inference at the earlier levels. Hierarchical Bayesian inference is concurrent across multiple areas, so that each piece of information does not need to flow forward to IT, return V1 and then back to IT, etc. Such a large loop would take too much time per iteration and is infeasible for real time inference. Rather, successive cortical areas in the visual hierarchy can constrain each other's inference in small loops instantaneously and continuously. It is plausible that such a system, as a whole, might converge rapidly to a consistent interpretation of the visual scene incorporating low level and high level sources of information.

B. Particle filtering

A major complication in this approach is that, unless the image is simple and clear, each area cannot be completely sure of its inference until the whole image is understood. More precisely, if the computation proceeds in a 'greedy' fashion with each cortical area settling on one seemingly best value for its features x_i in terms of the other areas signals, it may settle into an incorrect local maximum of the joint probability. Even allowing an iteration in which each x_i is updated when one of its neighbors updates its features, one may well find a situation in which changing one x_i decreases the joint probability but still a radical change of *all* x_i might find a still more probable interpretation. In computer vision experiments, this occurs frequently.

The remedy is not to 'jump to conclusions', but to allow multiple high probability values for the features or hypotheses to stay alive until longer feedback loops have had a chance to exert an influence. This approach is called *particle filtering*, and its use has been developing explosively in the computer vision and artificial intelligence communities.[26] The essential idea is to compute not with one guess for the true value of each set of features x_i, but with a moderate number of guesses (e.g. n in visual area i) $\{x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)}\}$ which are assigned weights $w_{i,1}, w_{i,2}, \cdots, w_{i,n}$ so that the weighted sum of these guesses is a discrete approximation to the full posterior probability distribution on x_i. In the broadest terms, particle filtering is simply replacing a full probability table by a weighted set of samples. When the number of values of a random variable becomes astronomical (as in perception), this is quite possibly the best way to deal with distributions on it, known to probabilists as using a 'weak approximation'.

This technique has produced the most successful computer vision programs to date for tracking moving objects in the presence of clutter and irregular motion (situations where all other techniques have failed).[27,28] It has also found wide spread application in solving mapping and localization in mobile robots.[29] Let us illustrate how particle filtering works with an example on robot localization. Suppose a robot has a map of the environment and has a knowledge of the data measurements that are associated with a particular location. Initially, it has no clue where it is, so it scatters an even distribution of the particles over the map, with each particle indicating a hypothesis of the robot's location. When the robot makes an observation of its surrounding, it can narrow down the space of hypotheses but nevertheless is not certain of its exact location, because such an observation is often corrupted by noises and is usually ambiguous as similar observations could be made in multiple locations. To account for this gain in knowledge and the narrowing down of the hypothesis, it updates the probability weight of particle (hypothesis) i by $w_i = P(obs|x_i)$, i.e. the likelihood of making this observation obs assuming the robot is at location x_i. The probability of all the particles are then renormalized so that they sum to one. A new set of samples are then drawn from the pool of existing particles. The probability of a particular particle being drawn is equal to its weight or $P(obs|x_i)$. The drawn particle is then thrown back to the map with some small random jitters from the original location. The variance of the jitter is proportional to the level of confidence, which can progressively be reduced as time goes on. A more probable (or heavier) particle could be drawn a number of times, giving a denser set of particles in a neighborhood around its original location. This is called the condensation process. Very light particles could vanish altogether, corresponding the extinction of those hypotheses. When the robot moves, the particles will move on the map as predicted by the motion equation of robot. When a new observation is made, because of the condensation due to resampling, the distribution of hypotheses are sampled according to the priors on how well they could predict the observation. The density of the resampled particles thus summarizes all the prior knowledge the robot has based on the earlier observations. As time progresses, the condensation process produces a convergence of the particles toward the most probable location(s). The great virtue of particle filtering is that it does not need to assume the underlying distribution is Gaussian, as required for the Kalman filtering machinery, hence it can approximate an arbitrary probability distribution, and has been considered as a generalized version of Kalman filtering.

In the low level/high level vision context, the algorithm is similar but not identical. In tracking or robot localization, the algorithm proceeds forward in time and information from many observations is integrated. One can also go 'backwards' in time and reconstruct from later data the most likely place where the robot was located at some point in the past, using future data to clear up ambiguities. This is exactly the way the forward/backward algorithm works in speech recognition. But in the vision situation, the time axis is replaced by the axis from low to high level, starting with local elementary image features and progressing to more global and abstract features. The algorithm should then work at all levels simultaneously, communicating by what is called *message passing* or *belief propagation* in each cycle of computation.[30] More formally, one has a set of particles $\{x_i^{(1)}, \cdots, x_i^{(n)}\}$ at each level i, bottom-up messages $B_1(x_n^{(i)})$ and top-down messages $B_2(x_n^{(i)})$ and one alternates between propagating the messages up and down via:

$$B_1(x_n^{(i)}) = \max_j \left(B_1(x_{n-1}^{(j)}) \phi(x_{n-1}^{(j)}, x_n^{(i)}) \right)$$

$$B_2(x_n^{(i)}) = \max_j \left(B_2(x_{n+1}^{(j)}) \phi(x_n^{(i)}, x_{n+1}^{(j)}) \right)$$

and updating the particles by resampling and perturbing using the weights:

$$w_{n,i} = B_1(x_n^{(i)}) B_2(x_n^{(i)}) / (\text{normalizing factor } Z_n).$$

A schematic of this forward/backward algorithm is shown in Figure 1b. Note that B_1 and B_2 are beliefs. They are not particles, but are sets of numbers that represent the conditional probabilities of the particles conditional on whatever part of the data/context which has been incorporated by the belief propagation so far. Algorithm of this type, although with separate sets of particles for bottom-up and top-down messages, are under active investigation in the computer vision community.[31,32]

Is such an algorithm neurally plausible? For the belief propagation algorithm to work, the bottom-up and top-down terms need to be represented separately, allowing their strengths to be conveyed to further areas. We can imagine that the bottom-up and top-down messages are represented by the activity of superficial (layers 2 and 3) and deep (layer 5) pyramidal cells respectively, as they project to higher and lower areas. More specifically, the variables $B_1(x_n)$ would correspond to the activity of superficial pyramidal cells and $B_2(x_n)$ to the activity of deep pyramidal cells. If the factors ϕ were equal to the weights of synapses of these pyramidal cells on their targets in remote areas, then the displayed updating rule for B_1 and B_2 (or a soft version of it) could be given by integration of inputs in the remote neurons. The particle itself needs to be represented by the activity of an ensemble of neurons, which could be bound by timing (e.g. synchrony)[33] or by synaptic weights after short-term facilitation.[34] Note that the top-down messages can utilize the same recurrent excitatory mechanism that has been proposed for implementing biased competition for attentional selection.[14–16] In fact, visual attention itself could be considered as 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. Because 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'.

INSERT Figure 1: Illustration of Hierarchical Bayes/Particle filtering

C. V1 as the high-resolution buffer

What is the distinctive role of V1 in such a hierarchical model? In terms of the probability model on which the theory rests, x_{v1} are the only variables *directly* connected to the observations x_0. Neurally, this is reflected in the fact that V1 is the recipient of the vast majority of the projections of the retina (via the LGN). Thus V1's activity should reflect firstly the greedy computation of the best values x_{v1} depending only on the visual stimulus and secondly, the progressive modification of these values as higher level aspects of the stimulus are recognized or guessed at, and updated posteriors on x_{v1} are computed using feedback. If any visual computation affects the local interpretation of the image, it will change the posterior on x_{v1} and hence be reflected in the firing of V1 neurons. This led us to propose that, instead of being the first stage in a feedforward pipeline, V1 is better described as the unique 'high-resolution buffer' in the visual system.[10,35]

In computer vision, algorithms for perceptual organization often operate on an image array of numbers, and the results are represented in arrays with resolution as high and sometimes higher than that of the input arrays. While the representations in the early visual areas (LGN, V1 and V2) are precise in both space and feature domains because of their small receptive fields arranged in retinotopic coordinates,[1] the size of the receptive fields of neurons increases dramatically as one traverses successive visual areas along the two visual streams (dorsal "where" and ventral "what" streams). For example, the receptive fields in V4 or MT are at least four times larger in diameter than those in V1 at the corresponding eccentricities,[36] and the receptive fields in the inferotemporal cortex (IT) tend to cover a large portion of the visual field.[37] This dramatic increase in receptive field size corresponds to a successive convergence of visual information that is necessary for extracting invariance and abstraction (e.g. translation, size) but it also results in the loss of spatial resolution and fine details in the higher visual areas.

In the hierarchical inference framework, the recurrent feedback connections between the areas would allow the areas to constrain each other's computation. This perspective dictates that the early visual areas do not merely perform simple filtering[2] or feature extraction operations.[1] Rather, they continue to participate in all levels of perceptual computations, if such computations require the support of their intrinsic machinery. In this framework, image segmentation, surface inference, figure-ground segregation and object recognition do not progress in a bottom-up serial fashion, but most occur concurrently and interactively in constant feedforward and feedback loops that involve the entire hierarchical circuit in the visual system at the same time. The idea that various levels in cognitive and sensory systems have to work together interactively and concurrently has been proposed in the neural modeling community[19, 23, 38–41] based primarily on psychological literature. But it is not until recently that solid neurophysiological evidence started to emerge to champion for this idea.

INSERT Figure 2: Illustration of High-resolution buffer hypothesis

3. Experimental Evidence

When the high-resolution buffer hypothesis was first proposed,[10, 35] it was primarily conjectural and based on data that are open to multiple interpretations. There is, however, increasing evidence from various laboratories[11, 42–47] that seems to support the high-resolution buffer hypothesis and, more generally, the hierarchical inference framework.

A. Timing

First of all, the timing studies of Thorpe's lab[48] show clearly that high level visual judgements (e.g. whether an image contains an animal or not) could be computed within 150 ms.

His work involves EEG recordings on humans, and he finds significant changes in frontal lobe activity between two conditions, in which the subject responds by pressing a button or not, starting at 150 ms post-stimulus. Thus a rather complete analysis including very high level abstract judgements seems to be formed in 150 ms. On the other hand, trans-cranial magnetic stimulation (TMS) studies from Shimojo's lab[42] show that TMS *over V1 alone* can produce visual scotomas in the subjective experience of human subjects at up to 170 ms latency. Thus V1 activity, over a period overlapping with activity expressing high level knowledge of scene properties, is essential for conscious visual perception. Taken together, these two pieces of evidence suggest that concurrent activation of V1 and the prefrontal cortex might be necessary for computing and representing a global coherent percept. Intact activities in V1 might be critical for the integrity of perception.

While data from Thorpe's lab[48] and Schall's lab[49] clearly showed that perceptual decision signals appear in the prefrontal cortex at about 150 ms post-stimulus onset, this does not necessarily mean object recognition can be done on a feedforward one-pass basis. In hierarchical Bayesian inference, the coupling is continuous between adjacent cortical areas. There is therefore plenty of time within the 150 ms period for the different cortical areas to interact concurrently. Several recent neurophysiological experiments suggest that relevant perceptual and decision signals emerge in the early visual cortex and the prefrontal cortex almost simultaneously. Schall and colleagues[49] showed that when a monkey has to choose a target among multiple distractors in a conjunctive search task, the neural signal at target location starts to differentiate from the signals at distractor locations at about 120 ms to 150 ms post stimulus onset. In a similar experiment in which we[11] monitored the responses of V1 and V2 neurons to target and distractors, the target response was found to become distinguished from the distractor response at about 100-120 ms post stimulus onset – 60 ms after the initial response onset of the V1 neurons, and only 10-20 ms before the similar activity emerges in the prefrontal area. This suggests computation in the cortex is rather concurrent – the distribution of the hypotheses converges rapidly through the continuous dynamics of the cortical interaction. It is conceivable that the whole hierarchy could converge to a single hypothesis within 60-80 ms of cortical interaction.

B. Scale of Analysis

Lamme[7] found that a V1 neuron (RF size < 0.8 degree) fires more strongly when its receptive field is inside a 4 degree diameter figure than when it is in the background, as if the neuron is sensitive to abstract construct of figure-ground. We[10] also found the initial response of the neuron is sensitive only to local features, and that it takes another 40 ms post-response onset for the signals start to become sensitive to the figure-ground context. Thus the early visual neurons's computation seems to progress in a local-to-global (fine-to-coarse) manner. On the other hand, recordings in IT have showed that higher level neurons behave in the opposite way.[50] In response to images of human faces, the initial responses of the neurons contain more information at coarse scale (such as gender of the face), and the later responses contain information of finer details, such as the individual specific information, suggesting IT's computations progress in a coarse to fine manner. These observations are consistent with the picture that the higher level area and the lower level area interact continuously to constrain each other's computation: the early areas first process local information, while the higher level areas first become sensitive to the global context. As the computation evolves under recurrent interaction, the early areas become sensitive to global context, while the higher areas become sensitive to the relevant precise and detailed information over time. Figure 2 illustrates why a high-resolution buffer is essential for visual reasoning. One may imagine that the higher areas in this case can instantly 'recognize' the face image based on the bottom-up cues (B_1 path) present in the illuminated subparts of the face, but feedback (B_2 path) from the face recognition area is critical for us to detect the faint edge and conclude that this is indeed the boundary of the face. This conclusion is mandatory, for if that boundary of the face cannot be detected under the same illumination condition, we will be alarmed and may form a different interpretation about what we actually saw.

Not every computation has to work all its way back to V1. Kosslyn[51] showed that in fMRI studies that a subject's V1 will light up differentially only when he is asked to imagine things or perform tasks that involved information of fine details. Scale of analysis is therefore a key factor. Given that feedback does consume energy, V1 would be consulted only when a scene is ambiguous without some high-resolution details. For computations that involve only detecting large objects, discriminating coarse features, or recognizing the gist of the scene, V2 and V4's involvement might be sufficient. All the experiments that managed to demonstrate a high-level or attentional effect in V1 seem to require the monkeys utilize information of fine details in their tasks. In Roelfsema and colleagues's experiment,[46] for example, the monkey was asked to trace one of the two curves displayed on the screen. He found that a neuron responds more strongly when its receptive field lies on the curve being traced than when its receptive field lies on the curve not being traced, as if there is a top-down attentional beam that traces and highlights the curve. This finding is significant in that many earlier studies from Desimone's lab[14] suggested that V1 activity is not modulated by attention particularly when the stimuli used are relatively large. In that context, V1 might not be needed in the computation.

In another study, Motter[52] also found that it is very difficult to demonstrate 'attentional modulation' (i.e. top-down effect) when there is only one single object on the screen, but that attentional modulation could be revealed when multiple objects are present. Apparently when multiple objects are presented on the screen, they engage in competition. Asking the monkey to pay attention to a particular location often results in the removal of the inhibition imposed by the surround on that location. Gilbert and colleagues[47] demonstrated an attentional effect in V1 only after the the monkeys were trained to perform a vernier acuity test – aligning two small vertical lines. These findings suggest that when the monkeys are performing tasks that require the discrimination of fine features, feedback can penetrate back to V1. Interestingly, Shimojo[42] found that when different spatial frequency gratings were used as stimulus in their TMS experiment, the optimal range of TMS delay is systematically increased as the spatial frequency increases, indicating that a finer resolution analysis might require an earlier visual area.

C. Interactive Hierarchy

While the above experiments demonstrate the emergence of attention effect in early visual areas during high-resolution analysis, it is unclear to what degree feedback is involved in normal and complex perceptual processing. In a hierarchical inference framework, feedback could be more or less automatic. We[11,44] have conducted two experiments to investigate V1 and V2 involvement in more complex perceptual processes that likely involves interaction among multiple cortical areas: the first is contour completion, and the second is shape from shading.

It has been hypothesized since the time of Hubel and Wiesel[1] that V1 is involved in edge detection. Some findings suggest it is involved in signaling more abstract and cue-invariant boundary such as texture boundary.[10] The findings from Gilbert's lab[9] that an additional bar along the longitudinal direction outside the receptive field could exert a facilitatory effect on a V1 neuron suggest a plausible computational mechanism for contour continuation.[12, 13, 54–56] In addition, a number of experiments (e.g.[53]) found that additional bars on the two sides of the neuron's longitudinal axis tend to suppress the response of a neuron, suggesting non-maximum noise suppression.[57] Curiously, there is no direct evidence for contour completion in V1. In fact, neural correlates of illusory contour as in Kanizsa triangle have only been observed in V2 but not in V1.[58]

On the other hand, the high-resolution buffer hypothesis suggests that V1 is the ideal machinery for computing geometrical curvilinear structures, as illustrated by the curve tracing experiment of Roelfsema.[46] In light of these considerations, we[44] decided to re-examine the issue of neural responses to illusory contours in area V1 and V2, using a static display which allowed tracking the temporal evolution of responses. We found that neurons in area V1 do indeed respond to illusory contours, e.g. completing the contour induced by the corner discs shown in Figure 3, although at a latency greater than that in V2.

INSERT Figure 3: Illustration of Illusory Contour Experiment

In this experiment, the monkey was asked to fixate at a spot on the screen, while the Kanizsa square was presented at different locations on the computer monitor in different trials. Over successive trials, the responses of the neurons to different locations relative to the illusory contour was studied (Figure 3). At the beginning of the experiment, consistent with Von der Heydt's earlier report, we found that V1 neurons in fact do not respond to the illusory contours. We then realized that because the corner discs (pacmen) were shown in the periphery, all the monkey might be seeing was just the flashing on and off of corner discs on the screen without perceiving the illusory square. We took several measures to enhance the awareness of the monkeys to the illusory square. First, we placed the fixation spot inside the illusory square, so that the monkey was looking at or inside the illusory square. Second, we presented the stimuli in a sequence: four black circular discs were first presented for 400 ms; then they were turned into the corner discs, creating an illusion that a white square had abruptly appeared in front of the circular disks, occluding them. The sudden onset of the illusory square also serves to capture the attention of the monkey to the square. Third, we introduced in our presentation a series of 'teaching' stimuli, i.e. real squares that are defined by line or contrast to help the monkey 'see' the illusion. Remarkably, in the third sessions after this change in paradigm, we started to find V1 neurons responding to the illusory contour in the stimulus (Figures 4).

INSERT Figure 4: Neural Correlates of Illusory Contour

The neural correlate of the illusory contour signal emerged in a V1 neuron at precisely the same location where a line or luminance contrast elicited the maximum response from the cell (Figure 4a). The response to the illusory contour was delayed relative to the response to the real contours by 55 ms (Figure 4b), emerging about 100 ms after stimulus onset. The response to the illusory contour was significantly greater than the response to the controls, including the amodal contour or when the corner discs were rotated. At the population level, we found that sensitivity to illusory contours emerged at 65 ms in V2 and 100 ms in the superficial layer of V1 (Figures 4c,d). A possible interpretation of these data is that V2 detects the existence of an illusory contour by integrating information from a more global spatial context, and then generates a prior $P(x_{v1}|x_{v2})$ to constrain the contour inference in V1. The resulting contour is the hypothesis particle that maximizes $P(x_o, x_{v1}, x_{v2}, x_{v4}, x_{it})$ which is the product of a cascade of feedback priors and bottom-up hypotheses. The particle filtering implementation of the contour completion process in V1 might be similar to Williams and Jacob's[56] stochastic random-walk model for contour continuation, except that it contains, in addition, many hierarchical layers of computations involving greater and greater chunks of information.

A second experiment that provides more conclusive evidence in support of feedback and hierarchical inference in V1 and V2 is related to the cortical processing of 3D shape from shading.[11] When viewing the display shown in Figure 5a, we perceive a set of convex shapes automatically segregating from a set of concave shapes. Perceptually, the convex shapes can be seen as concave and the concave shapes as convex if somehow we assume the scene is lit from below. These two interpretations can alternate in perceptual rivalry as in the Necker cube illusion. These interpretations of 3D shapes emerge purely from the shading information and hence are called shape from shading (SFS). Ramachandran[59] points out that this fast segregation suggests that 3D shape interpretation can influence the parallel process of perceptual organization. A case in point is that a similar image with contrast elements, but without a 3D interpretation, does not readily segregate into groups (Figure 5b). These pairs of stimuli are therefore ideal for probing the interaction between high-level interpretation (3D inference) and low-level parallel processes.

INSERT Figure 5: Shape from Shading versus 2D contrast

In this experiment, we[11] studied how V1 and V2 neurons respond to shape from shading stimulus, particularly in their sensitivity to perceptual pop-out saliency due to 3D interpretation. We tested the responses of V1 and V2 neurons when their receptive fields were placed at the center of a stimulus element. Typically, the receptive field is less than 0.7 degree while the diameter of the stimulus element is 1 degree visual angle. When comparing the neuronal responses to the black-and-white (BW) stimulus in comparision to the shape from shading (SFS) stimulus (Figure 6), we found that V1 neurons are sensitive mostly to contrast, and invariably respond better to the BW stimulus than the SFS stimulus, which has a weaker contrast. A significant number of V2 neurons, however, responded better to the SFS elements than to the BW elements particularly in their later responses (Figures 7a,b). This shows the V2 neurons might be more interested in a more abstract representation of 3D surface than the bottom-up luminance contrast. Furthermore, we found that while both V1 and V2 neurons do not exhibit pop-out response for the BW stimulus, V2 but not V1 neurons do exhibit the pop-out response for the SFS stimulus in a passive fixation task at the beginning. The pop-out response is defined by the ratio of the response of the neuron to the odd-ball condition over its response to the uniform condition. In both conditions, the stimulus on the receptive field is the same, but the surrounding elements are different for the odd-ball condition and the same for the uniform condition (Figures 7a,b). The finding that V2 exhibits preattentive pop-out response to shape from shading further reinforces the conjecture that V2 neurons might be representing 3D shape primitives, and provide an infrastructure for computing parallel pop-out through lateral inhibition. Recently, Von der Heydt's lab[60] found that V2 neurons are indeed sensitive to convex shape defined by both shape from shading as well as random-dot stereogram, providing a more direct evidence supporting the idea that 3D representational elements exist in V2.

INSERT Figure 6: Shape from Shading Stimulus

We found that while V1 neurons were not sensitive to the pop-out signals defined by the SFS stimulus at the beginning, they become sensitive to the pop-out signals after the monkeys were trained on a task that requires them to detect the location of the pop-out target. Interestingly, even though the monkeys can detect the oddball in the BW stimulus as well as the SFS stimulus, their V1 and V2 neurons exhibited the pop-out effect only for the SFS stimulus but not for the BW stimulus. As a population, the SFS pop-out emerged in V2 at around 95 ms post-stimulus onset, while it emerged in V1 5 ms later, at 100 ms

(Figure 7). The strength of these pop-out signals were found to be inversely correlated with the reaction time and positively correlated with the accuracy of the monkeys' performance in detecting the oddball.

INSERT Figure 7: Neural Responses to SFS

What is the purpose for these pop-out signals to appear in V1 after the monkeys have utilized the stimulus? Could this be simply a passive reflection of the stronger response in V2 enhanced by attention and awareness? We found that the pop-out signal is spatially precise in V1 – that it can be observed only on the target, but not on the distractor elements right next to it. This suggests that when the monkeys have to detect the location of a small target, a more spatially precise signal needs to be established in the high-resolution buffer. In addition, the fine interpretation of the 3D shape might also involve constant interaction with V1 as well. Note that the pop-out signal can be observed in a passive task even when the monkeys have not performed the detection task for several months. This suggests that once the monkey becomes 'aware' of the shape from shading pop-out elements, with practice, its visual system has enhanced the automatic coupling between V2 and V1 in this particular computation.

The 3D information from V2 provides the priors for facilitating V1's parallel pop-out computation and the precise localization of the pop-out target. Our findings suggest that these priors not only contain 3D information, but also the information about behavioral relevance or saliency.[11] We could change the top-down priors, for example, by manipulating the presentation frequency of the different odd-ball stimuli. When a particular oddball stimulus was presented more frequently than others, the monkey detected and reacted to this oddball faster and more accurately. The change in stimulus statistics often produce a change in the behavioral performance of the monkeys, which was accompanied by a parallel change in the relative pop-out strength in the neural signals. This interactive coupling between V1 and V2 are consistent with the hierarchical Bayesian inference framework.

Hierarchical inference is most evident in cases where the brain needs to resolve ambiguous stimuli in which the correct percept requires integration of multiple factors. In the illusory contour experiment, there are multiple hypotheses to explain the bottom-up data. The brain somehow chooses the simplest explanation – that a white square is occluding four black circular discs, even at the extra expense of hallucinating a subjective contour where there is really no visual evidence for it. It is only in this ambiguous situation that one can see a feedback effect in V1. In the shape from shading experiment, it is the need to finely localize the pop-out stimulus that finally drives the processing back to V1. The experiment of Bullier's lab[43] is also consistent with this idea. They found the effect of feedback is most evident in V1 only when the stimuli are of low visibility, low saliency and high ambiguity. In this context, the psychophysical stimuli used by Adelson[61] to study the disambiguation of edges and shapes could provide useful probes for investigating this hypothesis further.

D. Multiple Hypotheses

Is there any neurophysiological evidence that is suggestive of particle filtering in the cortex? A hall-mark of particle filtering is that multiple hypotheses are kept alive during the computation, so that the system does not need to jump into conclusions and can change its mind to entertain other possibilities.

One line of evidence that might support such an idea is the binocular rivalry experiment from Logothetis' lab.[62] When two different images are presented to the two eyes, we can only see one image at a time. This is known as binocular rivalry. It turns out that this is a rivalry between two perceptual hypotheses represented in the brain rather than the rivalry between information from the two eyes.[62] A curious fact is that almost all the *relevant* IT neurons respond consistently with perception, while only 10 percent of the V1 neurons, and 20 percent of the V2 neurons, responded in accordance with the percept. The rest of the V1 and V2 neurons just seem to mind their own business, representing 'bottom-up information' independent of the current perception.

This gradual increase in the percentage of neurons whose responses are correlated with perception along the visual hierarchy have also been observed in both the illusory contour and the shape from shading experiments.[11,44] This gradual increase in the neural correlate of perception along the visual hierarchy is quite a long-standing puzzle and has been taken as to mean that V1 is less 'conscious' than IT. Particle filtering might furnish a different perspective on this phenomenon. Maybe the early visual cortex, by keeping a significant number of neurons independent of the current perception is the area for keeping different evidence and hypotheses alive. As one moves up the hierarchy, the hypothesis' space becomes smaller and smaller, as the hypothesis distribution is successively resampled by the top-down priors (see also Geman and Bienenstock's compositional machine[63]). In binocular rivalry, when IT or prefrontal cortex is 'tired' of one hypothesis, the remaining hypotheses that are kept alive lower in the visual hierarchy are able to rebel and push the alternative hypothesis up to dominance. By perserving a variety of low-level sensory information in intact forms, independent of cognitive and perceptual decision, the early visual cortex is in a position to furnish alternative evidence to change the opinions of the high level neurons. Alternatively, when the high level area changes its idea, the early visual cortex can provide the necessary data to support the alternative hypothesis.

E. Resonance and Predictve Coding

Although we have been thinking primarily of top-down influences as *enhancing* activity in lower areas by reinforcing belief with high level context, there are recent very striking experiments[64] showing relative suppression of low level activity when an integrated simple high level percept can 'explain' the low level data. This is work by Murray and Kersten using fMRI on human subjects and showed that when similar sets of stimuli were presented – one a relatively complex two-dimensional pattern and one with a simple three-dimensional interpretation – V1 activity was less for the 3D pattern. This was even the case for a bistable stimulus, which alternates between a simple 3D percept with occlusion and a more complex 2D percept. Here they find a correlation between the times in which the subject reported seeing the 3D percept and the times in which V1 activity decreased. Roe and her colleagues[65] also found that in their optical imaging and single-unit experiments, while V2 neurons' activities were enhanced by illusory contours defined by abutted sinewave gratings, V1 neurons' end-stopping responses were suppressed! These experiments support our earlier work[19] and related ideas[38,40,66] in which it has been proposed that certain bottom-up pathways carried 'error' signals indicating the mismatch between data and their reconstruction or prediction with contextual priors and that when there was no error, the lower area would be relatively inactive. However, quite a different interpretation is possible using the theory of multiple hypotheses or particles. In a situation in which complex data is present for which no coherent or simple high level interpretation has been found, one would expect that many particles are needed to approximate the relatively spread out and multimodel posterior on the low level features. In psychophysical terms, many bits and pieces of the stimulus are trying to assemble into larger groupings, but none are very successful. However, when one high level interpretation emerges, this set of particles collapses and only one set of confirmed groupings remains. Thus the experiments of Murray and Kersten,[64] and Roe and colleagues[65] are quite consistent with the present framework.

4. Conclusion

Recent neurophysiological experiments have provided a variety of evidence suggesting that feedback from higher order areas can modulate the processing of the early visual cortex. The popular theory in the biological community to account for feedback is attention modulation. From that perspective, visual processing is still primarily a series of feedforward computation, only the computation and information flow is regulated by selective attention.[14] On the other hand, within the neural modeling community, there has been a number of models or theories,[19, 23, 38–40] with increasing sophistication, that emphasize directly or indirectly the feedback from higher order areas which might serve as contextual priors for influencing lower-level inference. Here, we suggest that these ideas could be formulated in the form of hierarchical Bayesian system, and that ideas from Bayesian belief propagation[30] and particle filtering[26, 27, 29] are relevant to understanding these interactive computations in the visual cortex. From this perspective, attention should not be conceptualized in terms of biased competition, but may be more appropriately viewed in terms of *biased inference*, or providing top-down priors in a hierarchical Bayesian inference framework. These priors reshaped the probabilistic distribution of the hypotheses by interacting with the data, so that a maximum likelihood conclusion can be arrived.

We reviewed a number of recent neurophysiological findings that are highly suggestive of such a hierarchical inference system, and, in particular, of the unique role of the primary visual cortex as a high-resolution buffer in this hierarchy. The impact of feedback is often subtle and becomes evident only when high-resolution details are required in certain computations or when the visual stimuli are ambiguous. In order to keep multiple hypotheses alive, the early visual areas have to continue to maintain data (and particles) that are not consistent with the current dominant hypothesis. This might be one of the reasons why a smaller percentage of early visual neurons are correlated with the perception than the neurons in the higher areas.

Central to our framework is the forward/backward mechanism which is conceptually already embodied in many existing neural models[19,23,38–40] and is critical to the success of many powerful vision and robotic algorithms. Our current attempts to reconcile an important difference between two competing schools of thoughts by emphasizing both the beliefs and the errors need to be propagated in these recurrent interactions. In the adaptive resonance[38] or interactive activation models,[39] an active global concept will feed back to enhance the neural activities in the early areas, disambiguating and enhancing some of the lower level informations that are vague but nevertheless are consistent with the global percept. These ideas are supported by neurophysiological experiments that show higher order information can enhance early visual responses.[7,11,44] On the other hand, pattern theory[19] and the predictive coding model[40] emphasize that feedback serves to suppress the activities in the early areas as a way of 'explaining away' the evidence. This idea is supported particularly by some recent imaging experiments.[40,64] In the latter class of models, only error residues are projected forward to the higher areas. In our current proposal on hierarchical bayesian inference, beliefs are embodied in both the bottom-up signals and the top-down signals. The activity of the deep layer cells reflects the top-down belief or probability, while the activity of the superficial layer cells reflects bottom-up belief or probabilities. A particle is then an ensemble of deep AND superficial cells whose strength as an ensemble (the binding strength via synchrony or rapid synatpic weight changes) is something like the weight of the particle. While other models keep particles for the forward and backward streams separate and non-interacting until the last step for mathematical reasons,[31] it might be beneficial to combine top-down and bottom-up information as soon as possible so as to form particles that reflect both bottom-up and top-down information as we suggest.

While the precise computational and neural implementation of many aspects of Bayesian belief propagagtion and particle filtering is not entirely clear, and remains to work out in both computational and neurophysiological experiments, we think the parallel and resonance between recent AI work on BBP/ particle-filtering and recent neurophysilogical findings in the visual cortex are striking and should not be ignored. This article summarizes our thoughts on their plausible connections and aims at stimulating more precise experimental research along this line. We expect these ideas will grow explosively in the next few years in the computatinal vision and biological vision community and will revolutionarize how we think about neural and computational processses underlying vision.

Acknowledgements

Tai Sing Lee was supported by NSF CAREER 9984706, McDonnell-Pew Foundation, NIH P50-MH64445 and NIH EY08098. David Mumford was supported by NSF DMS-0074276 and Burrough's Wellcomb Foundation grant 2302.

References

1. D.H. Hubel, T.N. Wiesel, "Functional architecture of macaque monkey visual cortex". Proc. Royal Soc. B (London) **198,** 1-59 (1978).

2. R.L., De Valois, K.K. De Valois, *Spatial vision*, Oxford University Press, New York, 1988.

3. D. Marr, *Vision*, Freeman, 1983.

4. D.J. Felleman, and D.C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex," Cereb. Cortex **1,** 1-47 (1991)

5. L. Maffei, A. Fiorentini, "The unresponsive regions of visual cortical receptive fields", Vision Research **16,** 1131-1139 (1976).

6. J.J. Knierim, D.C. Van Essen, "Neuronal responses to static texture patterns in area V1 of the alert macaque monkey," J. Neurophysiology **67,** 961-980 (1992).

7. V.A.F. Lamme, "The neurophysiology of figure-ground segregation in primary visual cortex". J. Neuroscience **15(2),** 1605-1615 (1995).

8. K. Zipser, V.A.F. Lamme, P.H. Schiller, "Contextual modulation in primary visual cortex," Journal of Neuroscience 16, 7376-7389 (1996).

9. M.K. Kapadia, G. Westheimer, C.D. Gilbert, "Spatial distribution of contextual interactions in primary visual cortex and in visual perception", J. Neurophysiol. **84(4),** 2048-2062 (2000).

10. T.S. Lee, D. Mumford, R. Romero, V.A.F. Lamme, "The role of the primary visual cortex in higher level vision", Vision Research **38(15-16)**, 2429-54 (1998).

11. T.S. Lee, C. Yang, R. Romero, D. Mumford, "Neural activity in early visual cortex reflects behavioral experience and higher order perceptual saliency", Nature Neuroscience **5(6),** 589-597 (2002).

12. J. Braun, "On the detection of salient contours", Spatial Vision **12(2)**, 211-225 (1999).

13. Z. Li, "A neural model of contour integration". Neural Computation **10**, 903-940 (2001).

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

15. M. Usher, E. Niebur "Modeling the temporal dynamics of IT neurons in visual search: A mechanism for top-down selective attention". Journal of Cognitive Neuroscience **8**,

311-327 (1996).

16. G. Deco, T.S. Lee, "A unified model of spatial and object attention based on inter-cortical biased competition", Neurocomputing **44-46**: 769-774 (2001).

17. U. Grenander, *General Pattern Theory*, Oxford Univ Press, 1993.

18. D. Mumford, "Pattern Theory: the Mathematics of Perception", in Li, T. (Ed), *Proceedings of the International Congress of Mathematicians* (2002, Beijing), Vol.1, Higher Education Press (Beijing), 2002.

19. D. Mumford, "On the computational architecture of the neocortex II", Biological Cybernetics **66**, 241-251 (1992).

20. D. Mumford, "Pattery theory: a unifying perspective", in *Perception as Bayesian inference,* Knill, D.C., Richards, W. (Eds), (Cambridge University Press, 1996), pp. 25-62.

21. D.C. Knill, W. Richards, (Eds) *Perception as Bayesian inference*, Cambridge University Press, Cambridge, U.K., 1996.

22. T.S. Lee, "A Bayesian framework for understanding texture segmentation in the primary visual cortex". Vision Research **35(18)**, 2643-2657 (1995).

23. G. Hinton, P. Dayan, B. Frey, R. Neal, "The wake-sleep algorithm for unsupervised neural networks", Science **268**, 1158-1161 (1995).

24. M.S. Lewicki, T.J. Sejnowski, "Bayesian unsupervised learning of higher order structure," Advances in Neural Information Processing Systems 9. 1998.

25. H.V. Helmholtz, Handbuch der physiologischen Optik, Leipzig: Voss; 1867.

26. Doucet, A., de Freitas, N., Gordon, N., (Eds) *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, 2001.

27. M. Isard, and A. Blake, "ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework" in *Lecture Notes in Computer Science 1406*, Burkhardt, H.

and Neumann, B. (Eds). (Springer-Verlag, Berlin, 1998), pp. 893-908.

28. M. Isard and A. Blake, "A Smoothing Filter for Condensation", Proc. 5th European Conf on Comp Vision, pp.767-781, 1998.

29. S. Thrun, D. Fox, W. Burgard, F. Dellaert, "Robust Monte Carlo localization for mobile robots" Artificial intelligence, **101**, 99-141 (2001).

30. J. Yedidia, W.T. Freeman, Y. Weiss, "Understanding belief propagation and its generalization," International Joint conference on Artificial Intelligence (IJCAI 2001).

31. E. Sudderth, A. Ihler, W. Freeman and A. Willsky, "Nonparametric Belief Propagation", MIT Lab for Information and Decision Systems Tech Report P-2551, 2002.

32. M. Isard, "Belief Propagation in Particle Networks", Tech Report, Microsoft Research Silicon Valley, 2002.

33. C.M. Gray, "The temporal correlation hypothesis of visual feature integration: still alive and well," Neuron **24**, 31-47.

34. P.J. Sjostrom and S. B. Nelson, "Spike timing, calcium signals and synatpic plasticity", Current Opinion in Neurobiology **12**, 305-314, 2002.

35. D. Mumford, "Commentary on article of H.Barlow", In: *Perception as Bayesian inference*, Knill, D.C., Richards, W. (Eds), (Cambridge University Press, 1996), pp. 25-62.

36. R. Gattass, A.P. Sousa, C.G. Gross, "Visuotopic organization and extent of V3 and V4 of the macaque", J. Neurosci. **8(6)**, 1831-1845, (1988).

37. C.G. Gross, "Visual function of inferotemporal cortex," In *Handbook of Sensory Physiology,* 7/3B:L, R. Jung, (Eds), (Springer-Verlag, Berlin, 1973), pp. 451-482.

38. G. Carpenter, S. Grossberg, "A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition", Machine. Comp. Vision, Graphics and Image Proc. **37**, 54-115 (1987).

39. J.L. McClelland, D.E. Rumelhart, “An interactive activation model of context effects in letter perception. Part I: an account of basic findings”, Psychol. Rev. **88**, 375-407 (1981).

40. R. Rao, D. Ballard, “Dynamic model of visual recognition predicts neural response properties in the visual cortex,” Neural Computation, **9** 721-763 (1997).

41. J. Bullier, “Integrated model of visual processing,” Brain Res. Review, **36(2-3)**, 96-107 (2001).

42. Y. Kamitani, S. Shimojo, “Manifestation of scotomas by transcranial magnetic stimulation of human visual cortex”, Nature Neuroscience, **2**, 767-771 (1999).

43. J.M. Hupe, A.C. James, B.R. Payne, S.G. Lomber, P. Girard, J. Bullier, “Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons”, Nature **394(6695)**, 784-787 (1998).

44. T.S. Lee, M. Nguyen, “Dynamics of subjective contour formation in the early visual cortex,” Proc. Natl Acad. Sci. **98(4)**, 1907-1911 (2001).

45. H. Super, H. Spekreijse, V.A.F. Lamme, “Two distinct modes of sensory processing observed in monkey primary visual cortex (V1)”, Nat. Neurosci. **4(3)**, 304-310 (2001).

46. P.R. Roelfsema, V.A.F. Lamme, H. Spekreijse, “Object-based attention in the primary visual cortex of the macaque monkey”, Nature, **395(6700)**, 376-381 (1998)

47. M. Ito, C.D. Gilbert, “Attention modulates contextual influences in the primary visual cortex of alert monkeys”, Neuron **22**: 593-604 (1999).

48. R. VanRullen, S. Thorpe, “Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects”, Perception, **30**, 655-668 (2001).

49. N.P. Bichot, J.D. Schall, “Effects of similarity and history on neural mechanisms of visual selection,” Nature Neuroscience, **2(6)**, 549-554 (1999).

50. Y. Sugase, S. Yamane, S. Ueno, K. Kawano, "Glboal and fine information coded by single neurons in the temporal visual cortex". Nature, **400(6747)**, 869-873 (1999).

51. S. Kosslyn, W.L. Thompson, I.J. Kim, N.M. Alpert, (1995). "Topographical representations of mental images in primary visual cortex," Nature, **378**, 496-498 (1995).

52. B.C. Motter, "Focal attention produces spatially selective processing in visual cortical areas V1, V2, V4 in the presence of competing stimuli," J. Neurophysiology **70(3)**, 909-919 (1993).

53. R.T. Born, R.B.H. Tootell, "Single-unit and 2-deoxyglucose studies of side inhibition in macaque striate cortex," Proc. Natl. Acad. Sci **88**, 7071-7075 (1991).

54. W.S. Geisler, J.S. Perry, J. Super, D.P. Gallogly, "Edge co-occurrence in natural images predicts contour grouping performance," Vision Research, **41**, 711-724 (2001).

55. J. August, S.W. Zucker, "The curve indicator random field: curve organization via edge correlation," in *Perceptual organization for artificial vision systems*, K. Boyer, S. Sarka, (Eds.) (Kluwer Academic, Boston, 2000), pp. 265-288.

56. L. Williams, D. Jacobs, "Stochastic completion fields: a neural model of illusory contour shape and saliency," Neural computation, **9(4)**, 837-858 (1997).

57. J.F. Canny, "A computational approach to edge detection," IEEE Trans. Pattern Analysis and Machine Intelligence, 679-698 (1986).

58. R. von der Heydt, E. Peterhans, G. Baumgarthner, "Illusory contours and cortical neuron responses," Science, **224(4654)**, 1260-1262 (1984).

59. V.S. Ramachandran, "Perception of shape from shading," Nature **331**, 163-166 (1988).

60. F.T. Qiu, R. Endo, R. von der Heydt, "Selectivity for structural depth in neurons of monkey area V2," Abstracts of Society of Neuroscience's 30th Annual meeting, New Orleans, L.A. 1582, (2000).

61. E. Adelson, A. Pentland, "The Perception of Shading and Reflectance," in *Perception as Bayesian Inference*, D. Knill and W. Richards (Eds.), (Cambridge University Press, 1996), pp. 409-423.

62. N.K. Logothetis, "Object vision and visual awareness," *Curr. Opin. Neurobiol*, **8(4)**, 536-44 (1998).

63. E. Bienenstock, S. Geman, D. Potter, "Compositionality, MDL priors, and object recognition", in: *Advances in Neural Information Processing Systems*, **9**, Mozer, M.C. M.I. Jordan, M.I and T. Petsche, T. (Eds), (MIT Press, 1997), pp. 838-844.

64. S. Murray, D. Kersten, B. Olshausen, P. Schrater and D. Woods, "Shape perception reduces activity in human primary visual cortex", to appear in Proc. Nat. Acad. Sci. 2002.

65. B.M. Ramsden, C.P. Hung and A.W. Roe, "Real and illusory contour processing in area V1 of the primate: a cortical balancing act," Cerebral Cortex **11**, 648-665 (2001).

66. , C. Koch, T. Poggio, "Predicting the visual world: silence is golden", Nature Neuroscience, **2(1)**, 9-10 (1999).

Figure Legends:

Figure 1:

(a) A schematic of the proposed Hierarchical Bayesian inference framework in the cortex. The different visual areas (indicated by the boxes) are linked together as a Markov chain. The activity in V1, x_1, is influenced by the bottom-up feedforward data x_0 and the probabilistic priors $P(x_1|x_2)$ fed back from V2. The concept of Markov chain is important computationally because each area can be mainly influenced by its direct neighbors. (b) An alternative way of implementing Hierarchical Bayesian inference using particle-filtering and belief propagation. B_1 and B_2 are bottom-up and top-down beliefs respectively. They are sets of numbers that reflects the conditional probabilities of the particles conditioned on the context that has been incorporated by the belief propagation so far. The top-down beliefs are the responses of the deep layer pyramidal cells that project backward, and the bottom-up beliefs are the activities of the responses of the superficial layer pyramidal cells that projected to the higher areas. The potentials ϕ are the synaptic weights at the terminals of the projecting axons. A hypothesis particle may link a set of particles spanning several cortical areas, and the probability of this hypothesis particle could be signified by its binding strength either via synchrony or rapid synaptic weight changes.

Figure 2:

V1 is reciprocally connected to all the expert visual modules either directly or indirectly. It therefore can serve as a high-resolution buffer to integrate various information together into a coherent percept. In this example of the high-resolution buffer, the bottom-up cues from the illuminated part of the face caused a face hypothesis particle to respond, this particle provides the contextual priors of the face to re-examine the data at the high-resolution buffer, locating the faint edge in the shadow as a part of the face.

Figure 3:

Selected stimuli in the subjective contour experiment. (a) An example sequence of stimulus presentation in a single trial. (b) Receptive field of the tested neuron was 'placed' at 10 different positions across the illusory contour, one per trial. (c) amodal contour – the subjective contour was interrupted by intersecting lines. (d) One of the several rotated pac-men controls. The surround stimulus was roughly the same, but there was no illusory contour. (e) One of the several types of real squares defined by luminance contrast. (f). Square defined by lines. These controls were used to assess the the neuron's positional sensitivity to real contour as well as for comparing the temporal responses between real and illusory contours. Adapted from Ref 44 with permssion from authors.

Figure 4:

(a) The spatial profile of a V1 neuron's response to the contours of both real and illusory squares, in a temporal window 100-150 ms after stimulus onset. The real or illusory square was placed at different spatial locations relative to the receptive field of the cell. This cell responded to the illusory contour when it was at precisely the same location where a real contour evoked the maximal response from the neuron. This cell also responded significantly better to the illusory contour than to the amodal contour (T-test, $p < 0.003$) and did not respond much when the pac-men were rotated. (b) Temporal evolution of the cell's response to the illusory contour compared to its response to the real contours of a line square, or a white square, as well as to the amodal contour. The onset of the response to the real contours was at 45 ms, about 55 ms ahead the illusory contour response. (c) Population averaged temporal response of 49 V1 neurons in the superficial layer to the illusory contours and controls. (d) Population averaged temporal response of 39 V2 neurons in the superficial layer to the illusory contours and controls. Adapted from Ref 44 with permssion from authors.

Figure 5:

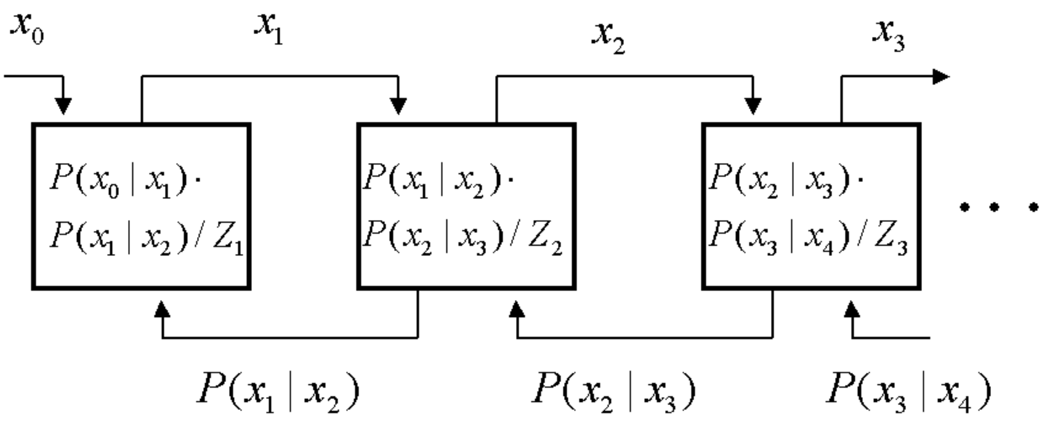
Ramachandran[59] showed that shape from shading stimuli produce instantaneous segregation, whereas black-and-white contrast stimuli did not. Given the main distinction between the two types of stimuli is that the stimulus elements in (a) but not those in (b) afford 3D interpretation, 3D information must have directly influenced the early parallel processes of perceptual grouping.

Figure 6:

Higher order perceptual pop-out. (a) A typical stimulus display was composed of 10 x 10 stimulus elements. Each element was 1^o visual angle in diameter. The diameter of the classical receptive field (RF) of a typical cell at the eccentricities tested ranged from 0.4^o to 0.8^o visual angle. Displayed is the LA (Lighting from Above) oddball condition, with the LA oddball placed on top of the cell's receptive field, indicated by the open circle. The solid dot indicates the fixation spot. (b) Each stimulus set had four conditions: singleton, oddball, uniform, and hole. Displayed are the iconic diagrams of all the conditions for the LA set and the LB set, as well as the oddball conditions for the other four sets. The center element in the iconic diagram covered the receptive field of the neuron in the experiment. The surround stimulus elements were placed outside the RF of the neuron. The comparison was between the oddball condition and the uniform condition, while the singleton and the hole conditions were controls. Adapted from Ref 11 with permission from authors.

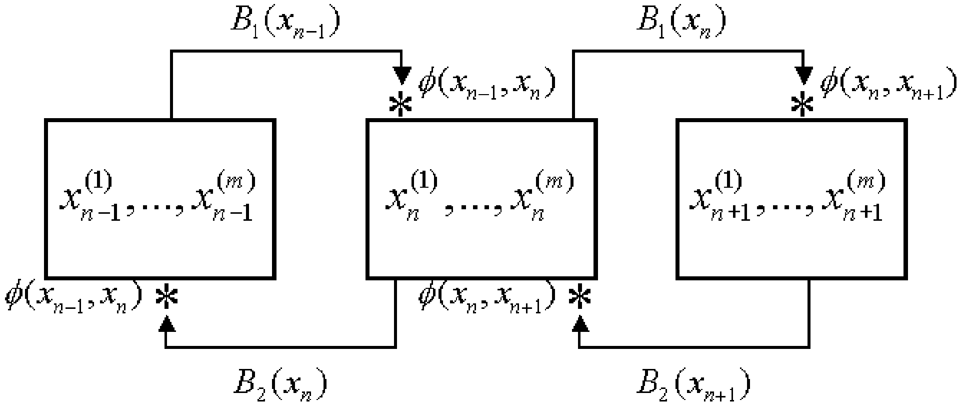
Figure 7:

Temporal evolution of the average population average response of 22 V2 units and 30 V1 units from a monkey to the LA set and the WA set in a stage after the monkey had utilized the stimuli in its behavior. Each unit's response was first smoothed by a running average within a 15 ms window, then averaged across the population. A significant difference (pop-out response) was observed between the population average response to the oddball



x_0
x_1
x_2
x_3
$P(x_0 | x_1)\cdot$ $P(x_1 | x_2)/Z_1$
$P(x_1 | x_2)\cdot$ $P(x_2 | x_3)/Z_2$
$P(x_2 | x_3)\cdot$ $P(x_3 | x_4)/Z_3$
$\cdots$
$P(x_1 | x_2)$
$P(x_2 | x_3)$
$P(x_3 | x_4)$

(a)



(b)

Figure 2:

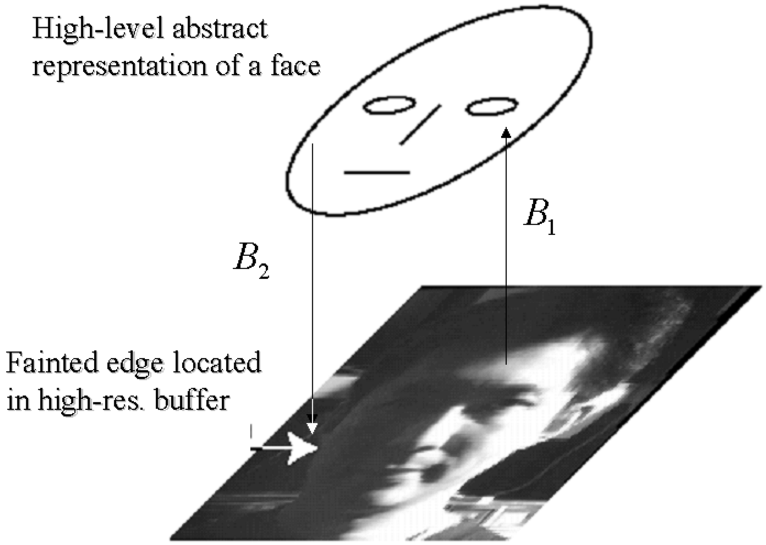

High-level abstract
representation of a face
B_1
B_2
Fainted edge located
in high-res. buffer

Figure 3:

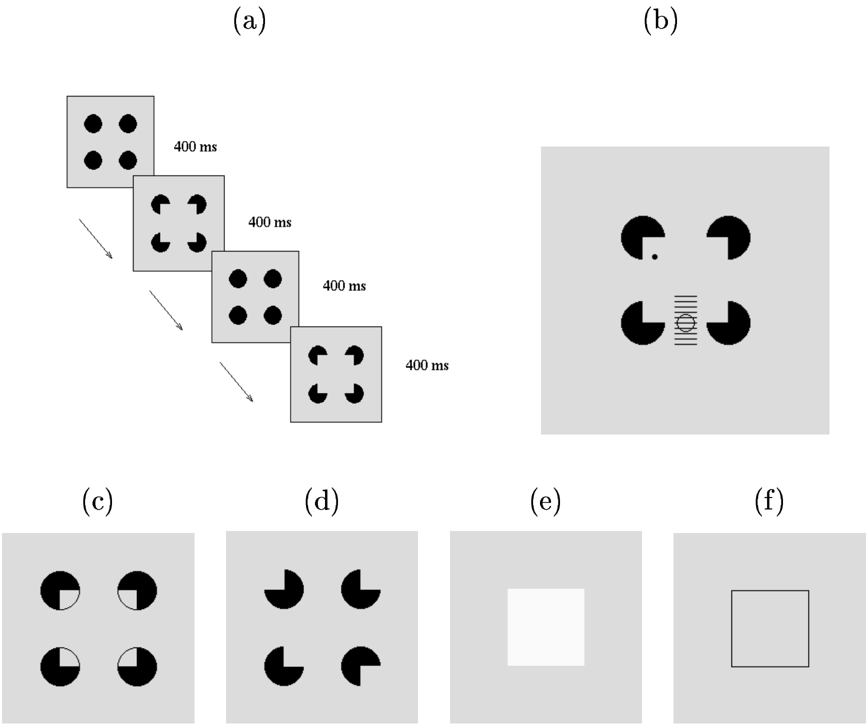


Figure 4:

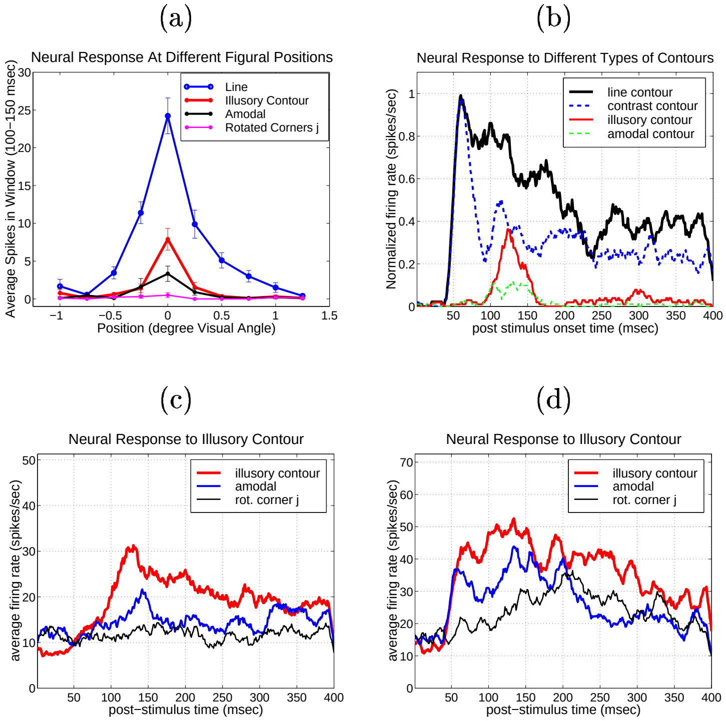

(a)
Neural Response At Different Figural Positions
Average Spikes in Window (100−150 msec)
0
5
10
15
20
25
30
−1
−0.5
0
0.5
1
1.5
Position (degree Visual Angle)
Line
Illusory Contour
Amodal
Rotated Corners j
(b)
Neural Response to Different Types of Contours
Normalized firing rate (spikes/sec)
0
0.2
0.4
0.6
0.8
1
50
100
150
200
250
300
350
400
post stimulus onset time (msec)
line contour
contrast contour
illusory contour
amodal contour
(c)
Neural Response to Illusory Contour
average firing rate (spikes/sec)
0
10
20
30
40
50
50
100
150
200
250
300
350
400
post−stimulus time (msec)
illusory contour
amodal
rot. corner j
(d)
Neural Response to Illusory Contour
average firing rate (spikes/sec)
0
10
20
30
40
50
60
70
50
100
150
200
250
300
350
400
post−stimulus time (msec)
illusory contour
amodal
rot. corner j

Figure 5:

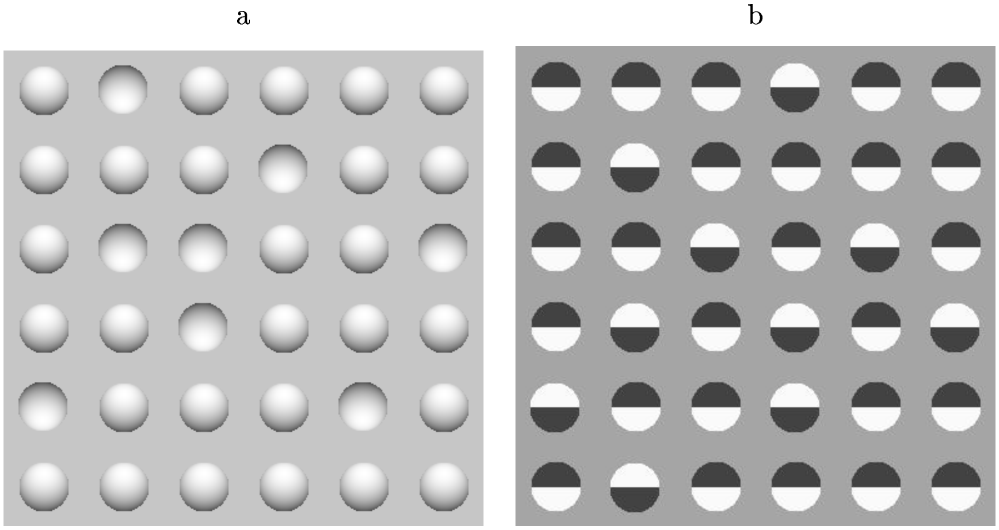


Figure 6:

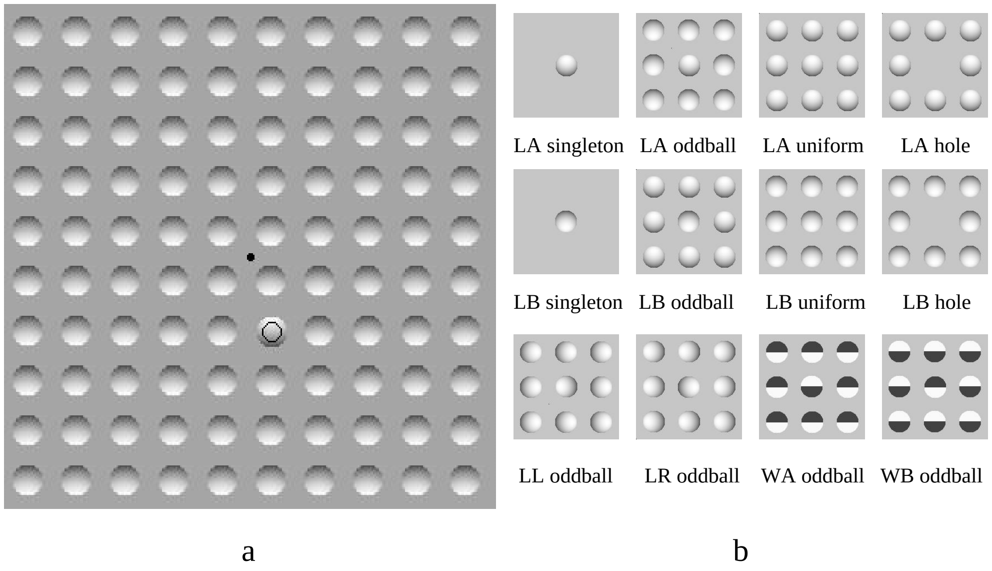
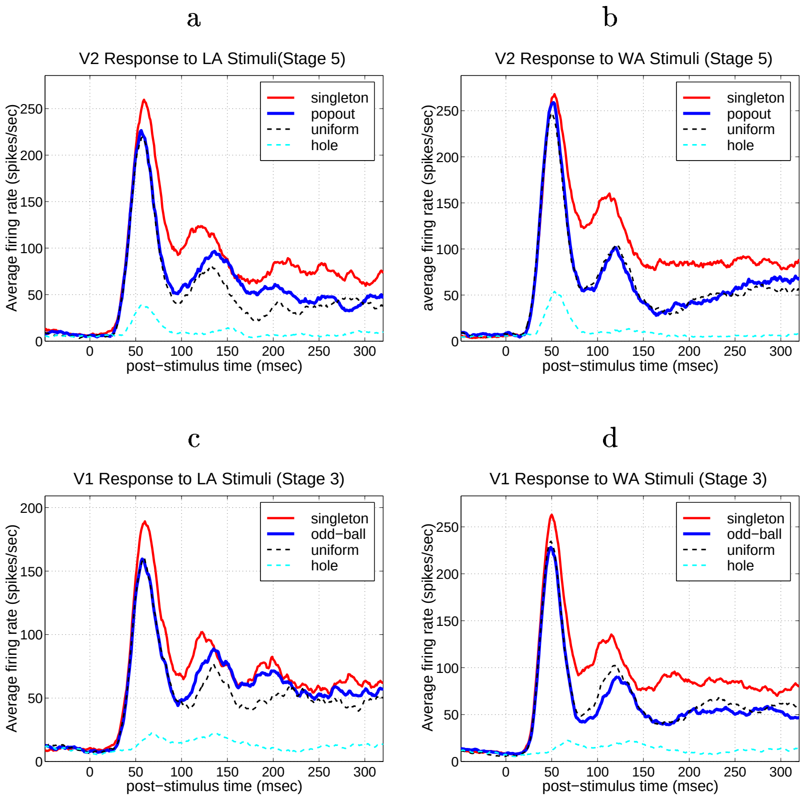

LA singleton
LA oddball
LA uniform
LA hole
LB singleton
LB oddball
LB uniform
LB hole
LL oddball
LR oddball
WA oddball
WB oddball
a
b

Figure 7:


a
V2 Response to LA Stimuli(Stage 5)
b
V2 Response to WA Stimuli (Stage 5)
c
V1 Response to LA Stimuli (Stage 3)
d
V1 Response to WA Stimuli (Stage 3)
singleton
popout
odd−ball
uniform
hole
Average firing rate (spikes/sec)
average firing rate (spikes/sec)
post−stimulus time (msec)
0
50
100
150
200
250
300