PROOF COPY [A-8700] 005307JOA

Hierarchical Bayesian inference in the visual cortex

Tai Sing Lee

Room 115, Mellon Institute, Department of Computer Science, Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

David Mumford

Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912

Received October 23, 2002; revised manuscript received February 21, 2003; accepted February 26, 2003

Traditional views of visual processing suggest that early visual neurons in areas V1 and V2 are static spatiotemporal filters that extract local features from a visual scene. The extracted information is then channeled 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 loops in the cortex serve to integrate top-down contextual priors and bottom-up observations so as to implement concurrent probabilistic inference along the visual hierarchy. We suggest that the algorithms of particle filtering and Bayesian-belief propagation might model these interactive cortical computations. We review some recent neurophysiological evidences that support the plausibility of these ideas. © 2003 Optical Society of America

OCIS codes: 330.4060.

1. INTRODUCTION

In this paper we propose a Bayesian theory of hierarchical cortical computation based both on (a) the mathematical and computational ideas of computer vision and pattern theory and on (b) recent neurophysiological experimental evidence. We[1,2] have proposed that Grenander's pattern theory[3] could potentially model the brain as a generative model in such a way that feedback serves to disambiguate and "explain away" the earlier representation. The Helmholtz machine[4,5] was an excellent step toward approximating this proposal, with feedback-implementing priors. Its development, however, was rather limited, dealing only with binary images. Moreover, its feedback mechanisms were engaged only during the learning of the feedforward connections but not during perceptual inference, though the Gibbs sampling process for inference can potentially be interpreted as top-down feedback disambiguating low-level representations.[6] Rao and Ballard's predictive coding/Kalman filter model[7] did integrate generative feedback in the perceptual inference process, but it was primarily a linear model and thus severely limited in practical utility. The data-driven Markov chain Monte Carlo approach of Zhu and colleagues[8,9] might be the most successful recent application of this proposal in solving real and difficult computer vision problems by using generative models, though its connection to the visual cortex has not been explored. Here we bring in a powerful and widely applicable paradigm from artificial intelligence and computer vision to propose some new ideas about the algorithms of visual cortical processing and the nature of representations in the visual cortex. We will review some of our and others' neurophysiological experimental data to lend support to these ideas.

A prevalent view in the biological community on the role of feedback among cortical areas is that of selective attention modeled by biased competition.[10–12] Vision is still considered to be accomplished by a feedforward chain of computations.[13,14] Although these models give an apparently complete explanation of some experimental data, they use the sophisticated machinery of feedback pathways[15] in a rather impoverished way (as we shall illustrate), and they persist in viewing the computations in each visual area as predominantly independent processes. However, some of our recent neurophysiological evidence cannot be fully accounted for by biased competition models. Instead, we believe that they reflect underlying cortical processes that are indicative of a generative model. We will link these data to the proposed framework and explain how ideas of resonance[16,17] and predictive coding[2,4,5,7] can potentially be reconciled and accommodated in a single framework.

We have not offered a simulation to accompany our proposal, partly because many details remain to be worked out and partly because the choice of model is still quite unconstrained and any specific simulation provides only weak support for a high-level hypothesis like ours. For us, the strongest argument for this theory is the computational one: Work on robust computer vision systems has shown how hard it is to interpret images with simpler

1084-7529/2003/070001-15$15.00 © 2003 Optical Society of America

PROOF COPY [A-8700] 005307JOA

PROOF COPY [A-8700] 005307JOA

algorithms and has led to some key unifying principles for overcoming the exponential explosion stemming from combining conflicting interpretations of the many parts or aspects of an image. Bayesian-belief propagation[18–20] and particle filtering[21,22] are the most successful computer vision algorithms to date. The latter has been used for tracking moving objects in the presence of clutter and irregular motion (situations in which all other techniques have failed).[23,24] Its use is also developing rapidly in the robotics community, for example, for solving mapping and localization problems in mobile robots in a real-world scenario.[25] We see some very attractive features in both of these algorithms that might be implemented naturally by cortical neural networks. We believe that this theory provides a plausible and much more tightly coupled model of the processing in visual areas and especially in V1 and V2.

The hierarchical Bayesian framework provides an alternative perspective for understanding many recent neurophysiological findings, and the particle-filtering mechanisms point to a potentially new aspect of cortical processing. In writing this paper, we hope (1) to introduce to computer scientists a plausible brain model by using a hierarchical Bayesian framework and particle-filtering mechanisms, (2) to draw the attention of the neural modeling community to the possibility of a cortical algorithm based on particle filtering, and (3) to expose to neuroscientists these powerful paradigms in computer vision and artificial intelligence. In particular, we emphasize that inference is more general than competition and that feedback should not be conceived merely in terms of attentional selection or biased competition but could be more profitably conceived as mechanisms for biasing inference and computations along the visual hierarchy. Attentional selection corresponds to only a small subset of such priors. We will first sketch the general theoretical framework and then in subsequent sections review the experimental evidence that points in the direction of this theory.

2. BAYESIAN PERSPECTIVE ON CORTICAL COMPUTATION

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.[1,2,4,6,26,27] This idea can be traced back to the "unconscious inference" theory of perception by Helmholtz[28] and has recently been connected to the evolution of the perceptual systems.[29]

Recall that Bayes's rule proposes that with observations x_0, hidden variables x_1 to be inferred, and contextual variables x_h, a probabilistic description of their effects on one another 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., that 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 1 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 lateral geniculate nucleus (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 $P(x_0|x_1)$ by the feedback biasing probabilities $P(x_1|x_h)$ and maximizing the result 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, namely, the degree of compatibility of *every* possible set of features x_1 with the high-level data x_h. In short, this factor now includes *all* possible ways in which higher-level information about the scene may affect the V1 features x_1, i.e., the beliefs at V1. Figure 1 illustrates this idea in two ways. First, the data under high level of illumination make probable the low-level fact that certain

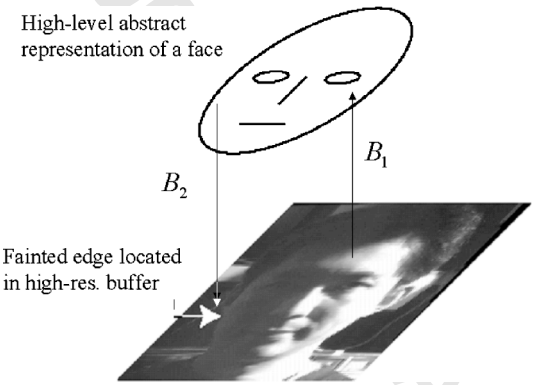

Fig. 1. V1 is reciprocally connected to all the expert visual modules either directly or indirectly. It therefore can serve as a high-resolution buffer to integrate various information together into a coherent percept. In this example of the high-resolution buffer, the bottom-up cues from the illuminated part of the face cause a face hypothesis to respond, which provides the contextual priors of the face to reexamine the data at the high-resolution buffer, locating the faint edge in the shadow as a part of the face.

PROOF COPY [A-8700] 005307JOA

PROOF COPY [A-8700] 005307JOA

areas of the image are in shadow. Second, the high-level knowledge of the identity of an individual suggests that a face should have certain proportions, as measured from the low-level data in V1. Both sets of information would go into the full explanation of the image.

This basic formulation can also capture the interaction among multiple cortical areas, such as V1, V2, V4, and the inferotemporal cortex (IT). Note that although feedback goes all the way back to the LGN and it is simple to include the LGN in the scheme, the computational role of the thalamic nuclei could potentially be quite different.[30] Hence we decide not to consider the various thalamic areas, the LGN, and the nuclei of the pulvinar, in this picture at present. The formalism that we introduce applies to any set of cortical areas with arbitrary connections between them. But for simplicity of exposition, we assume that our areas are connected like a chain. That is, we assume that each area computes a set of features or beliefs, which we now call x_{v1}, x_{v2}, x_{v4}, and x_{IT}, and we make the simplifying assumption that if, in the sequence of variables $(x_0, x_{v1}, x_{v2}, x_{v4}, x_{\mathrm{IT}})$, any variable is fixed, then the variables before and after it are conditionally independent. This means that we can factor the probability model for these variables and the evidence x_0 as

$$P(x_0, x_{v1}, x_{v2}, x_{v4}, x_{\mathrm{IT}}) = P(x_0|x_{v1})P(x_{v1}|x_{v2})P(x_{v2}|x_{v4})P(x_{v4}|x_{\mathrm{IT}})P(x_{\mathrm{IT}})$$

and make our model an (undirected) graphical model or Markov random field based on the chain of variables:

$$x_0 \leftrightarrow x_{v1} \leftrightarrow x_{v2} \leftrightarrow x_{v4} \leftrightarrow x_{\mathrm{IT}}.$$

From this it follows that

$$P(x_{v1}|x_0, x_{v2}, x_{v4}, x_{\mathrm{IT}}) = P(x_0|x_{v1})P(x_{v1}|x_{v2})/Z_1,$$

$$P(x_{v2}|x_0, x_{v1}, x_{v4}, x_{\mathrm{IT}}) = P(x_{v1}|x_{v2})P(x_{v2}|x_{v4})/Z_2,$$

$$P(x_{v4}|x_0, x_{v1}, x_{v2}, x_{\mathrm{IT}}) = P(x_{v2}|x_{v4})P(x_{v4}|x_{\mathrm{IT}})/Z_4.$$

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

$$P(x_{v1}|x_0, x_{v2}, x_{v4}, x_{\mathrm{IT}}) = \phi(x_0, x_{v1})\phi(x_{v1}, x_{v2})/Z(x_0, x_{v2}),$$

$$P(x_{v2}|x_0, x_{v1}, x_{v4}, x_{\mathrm{IT}}) = \phi(x_{v1}, x_{v2})\phi(x_{v2}, x_{v4})/Z(v_{v1}, x_{v4}),$$

$$P(x_{v4}|x_0, x_{v1}, x_{v2}, x_{\mathrm{IT}}) = \phi(x_{v2}, x_{v4})\phi(x_{v4}, x_{\mathrm{IT}})/Z(x_{v2}, x_{\mathrm{IT}}),$$

where $Z(x_i, x_j)$ is a constant needed to normalize the function to a probability distribution. The potentials must be learned from experience with the world and constitute the guts of the model. This is a very active area in machine learning research.[4,6,8,19,20]

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 right-hand side of each of the above equations) and the top-down data feeding back (the second factor) [see Fig. 2(a)].

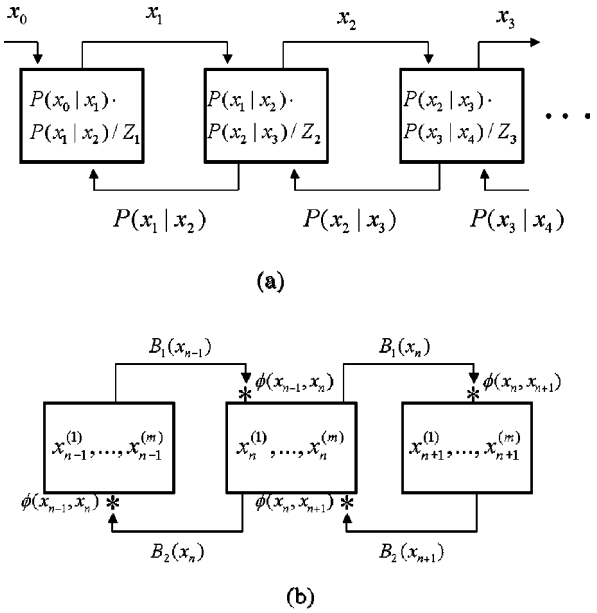

Fig. 2. (a) Schematic of the proposed hierarchical Bayesian inference framework in the cortex: The different visual areas (boxes) are linked together as a Markov chain. The activity in V1, x_1, is influenced by the bottom-up feedforward data x_0 and the probabilistic priors $P(x_1|x_2)$ fed back from V2. The concept of a Markov chain is important computationally because each area is influenced mainly by its direct neighbors. (b) An alternative way of implementing hierarchical Bayesian inference by 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 reflect 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 project 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 via either synchrony or rapid synaptic weight changes.

Each cortical area seeks to maximize by competition the probability of its computed features (or beliefs) x_i by combining the top-down and bottom-up data with use of the above formulas (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, at each point in time, a distribution of beliefs exist at each level. Feedback from all higher areas can ripple back to V1 and cause a shift in the preferred beliefs computed in V1, which in turn can sharpen and collapse the belief distribution in the higher areas. 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 long-latency responses in IT, on the other hand, will tend to reflect fine details and more-precise information about a specific object.

The feedforward input drives the generation of the hypotheses, and the feedback from higher inference areas

PROOF COPY [A-8700] 005307JOA

PROOF COPY [A-8700] 005307JOA

provides the priors to shape the inference at the earlier levels. Neither the feedforward messages nor the feedback messages are static: As the interpretation of an image proceeds, new high-level interpretations emerge that feed back new priors, and as low-level interpretations are refined, the feedforward message is modified. Such hierarchical Bayesian inference can proceed concurrently across multiple areas, so that each piece of information does not need to flow all the way forward to IT, return to 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 could constrain one another's inference in small loops rapidly and continuously as the interpretation evolved. One might hope that such a system, as a whole, would converge rapidly to a consistent interpretation of the visual scene incorporating all low-level and high-level sources of information; but there are problems, which we address next.

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 might 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.

A simple example of the problem would be a situation in which there are two competing interpretations of an image, A and B. Interpretation A incorporates values A_1 for the features in area 1 and values A_2 for the features in the higher area 2. Likewise, B gives the features values B_1 and B_2. Then A_1 and A_2 support each other through high values of $p(A_1|A_2)$ and $p(A_2|A_1)$ and together give a local maximum of the joint probability. B_1 and B_2 do the same. To decide between them, you must compare the joint probability $p(A_1, A_2)$ with $p(B_1, B_2)$ and choose the larger (which is usually much larger). This example shows how statistical inference involves much more than competition among neurons in an area—it is competition, but a global competition involving all aspects, low and high, of each interpretation. Such competition is ubiquitous in real images, although we are usually unaware of it as we make these inferences unconsciously in 100 ms or so. Some of the most striking examples are in the work of Adelson and collaborators,[31,32] where scenes are constructed involving tilted surfaces, shadows, and corners that have multiple interpretations but only one that optimally integrates the low-level data with our high-level knowledge of lighting and geometry.

The only remedy that has been found in the computational literature 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 rapidly in the computer vision and artificial intelligence communities.[21] The essential idea is to compute, for each area, not one guess for the true value of its set of features x_i but a moderate number of guesses (e.g., there could be n sets of values for the features in visual area i $\{x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)}\}$). In addition, these are assigned weights $w_{i,1}, w_{i,2}, \ldots, w_{i,n}$ in such a way 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 happens in perception), this is quite possibly the best way to deal with distributions on it, known to probabilists as using a "weak approximation." This idea would seem on the surface to be related to the ideas of Zemel[33] and Eliasmith and Anderson[34] on population coding and decoding. These authors also seek to represent probability distributions on perceptual variables by the joint activity of populations of neurons. But their main aim is to overcome or use the coarse and/or nonlinear tuning of single neurons, whereas our aim is to deal with the huge dimensional space of the *joint* distribution on all perceptual variables represented in a given area. Here we are ignoring the issues posed by this coarse tuning. It is important to note that particle filtering is *not* an issue of using any specific kind of probability model (e.g., a mixture model) but rather of what kind of algorithm is used for statistical learning and inference with the model.

The particle-filtering technique has produced the most successful computer vision programs to date for tracking moving objects in the presence of clutter and irregular motion.[23,24] It has also found widespread application in solving mapping and localization in mobile robots.[25] 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 into the current set of particles and their weights. One can also go backward 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 except that using particles allows one to overcome explosions in the number of states of the system. In the vision situation, information flow progresses both along the time axis and along the visual hierarchy, starting with local elementary image features and progressing to more-global and more-abstract features in the higher areas. The recurrent interaction across the hierarchy helps to collapse the hypothesis space over time. The algorithm should work at all levels simultaneously, communicating by what is called message passing or belief propagation in each cycle of computation.[19] More formally, one has a set of particles $\{x_i^{(1)}, \ldots, x_i^{(n)}\}$ at each level i, bottom-up messages $B_1(x_i^{(j)})$, and top-down messages $B_2(x_i^{(j)})$, and one alternates between propagating the messages up and down through

PROOF COPY [A-8700] 005307JOA

PROOF COPY [A-8700] 005307JOA

$$B_1(x_i^{(j)}) = \max_k [B_1(x_{i-1}^{(k)})\phi(x_{i-1}^{(k)}, x_i^{(j)})],$$

$$B_2(x_i^{(j)}) = \max_k [B_2(x_{i+1}^{(k)})\phi(x_i^{(j)}, x_{i+1}^{(k)})],$$

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

$$w_{i,j} = B_1(x_i^{(j)})B_2(x_i^{(j)})/(\text{normalizing factor } Z_i).$$

A schematic of this forward/backward algorithm is shown in Fig. 2(b). Note the 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 or context has been incorporated by the belief propagation so far. Algorithms 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[19,20,22] and constitute one of the most promising techniques for statistical inference in such large, multilayered domains.

In such an algorithm neurally plausible? Here we give some ideas but no details, which we hope to work out more fully in a subsequent paper. 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 hypothesize 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 in which the "max" is replaced by some weighted average) could be given by integration of inputs in the remote neurons. The recent work of Arathorn[35] on the use of feedback synthesis to rapidly weigh and collapse the space of feedfoward hypotheses has a similar flavor, even though probabilistic formulations are not used explicitly in his map-seeking circuits.

The particle itself might need to be represented by the concerted activity of an ensemble of neurons, which could be bound by timing (e.g., synchrony)[36–38] or by synaptic weights after short-term facilitation.[39] Note that the top-down messages can utilize the same recurrent excitatory mechanism that has been proposed for implementing biased competition for attentional selection.[10–12] In fact, visual attention itself could be considered a special case in this framework. The recurrent excitatory connections across the multiple modules in the visual hierarchy allow the neurons in different areas to link together to form a larger hypothesis particle by firing concurrently and/or synchronously. Since its implementation requires that groupings of mutually reinforcing alternative values of features in different areas be formed, this algorithm might be linked to the solution of the binding problem.

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 furnished by the retina and the LGN. In reality, the LGN is more directly connected to the retina and is the lowest level of the hierarchy which receives feedback. The LGN could serve as a high-resolution pointillistic buffer and V1 as a high-resolution geometric buffer. Neurally, this is reflected by 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 first reflect the best guesses of x_{v1} depending only on the local visual stimulus and then subsequently the progressive modification of these values based on feedback as higher-level aspects of the stimulus are recognized or guessed at. 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,[13] V1 is better described as the unique high-resolution buffer in the visual system for geometric calculations.[40,41]

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.[42] 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,[43] and the receptive fields in IT tend to cover a large portion of the visual field.[44] This dramatic increase in receptive-field size leads to a successive convergence of visual information necessary for extracting invariance and abstraction (e.g., translation and scaling), 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 among the areas would allow the areas to constrain one an other's computation. This perspective dictates that the early visual areas do not merely perform simple filtering[45] or feature extraction operations.[42] 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 likely 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[1,4,7,16,17] primarily on the basis of psychological literature. However, it has not been until recently that solid neurophysiological evidence has started to emerge to champion this idea.

PROOF COPY [A-8700] 005307JOA

PROOF COPY [A-8700] 005307JOA

3. EXPERIMENTAL EVIDENCE

When the high-resolution-buffer hypothesis was first proposed,[40,41] it was primarily conjectural and based on data that are open to multiple interpretations.[41,46,47] However, recent findings from various laboratories on contextual modulation of neural activities in V1[48–54] lend support to the high-resolution-buffer hypothesis and, more generally, to the hierarchical-inference framework.

A. Timing

First, the timing studies of Thorpe's laboratory[14] clearly show that high-level visual judgments (e.g., whether an image contains an animal or not) could be computed within 150 ms. Thorpe's 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 poststimulus. Thus a rather complete analysis including very-high-level abstract judgments seems to be formed in 150 ms. On the other hand, transcranial magnetic stimulation (TMS) studies from Shimojo's laboratory[50] 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, during a period that overlaps 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.

Although data from Thorpe's laboratory[14] and Schall's laboratory[55] indicate that perceptual decision signals appear in the prefrontal cortex at ~150 ms poststimulus onset, this does not necessarily mean that object recognition is done on a feedforward and one-pass basis. In hierarchical Bayesian inference, the coupling could be 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[55] showed that when a monkey has to choose a target among multiple distractors in a conjunctive search task, the neural signal at the target location starts to differentiate from the signals at distractor locations at approximately 120–150-ms poststimulus onset. In a similar experiment, we[49] found the differentiation between target and distractors appear within the same time frame in early visual areas (V1 and V2), at approximately 100–120-ms poststimulus onset, suggesting that computation in the cortex is rather concurrent. It is thus conceivable that through the continuous dynamics of the cortical interaction, the whole hierarchy could converge to a single hypothesis with 60–80 ms of interaction.

B. Scale of Analysis

Lamme[46] found that a V1 neuron (receptive field size <0.8 deg) fires more rigorously when its receptive field is inside a 4-deg-diameter figure than when it is in the background, as if the neuron is sensitive to an abstract construct of figure–ground. The initial response of the neuron is sensitive mainly to local features, and only 40 ms later does it become sensitive to the figure–ground context.[41,47] Thus the early visual neurons' computation seems to progress in a local-to-global (fine-to-coarse) manner. On the other hand, recordings in IT have shown that higher-level neurons behave in the opposite way.[56] In response to images of human faces, the initial responses of the neurons contain information on a coarser scale (such as gender of the face), and the later responses contain finer details, such as the specific information about an individual, suggesting that IT's computations progress in a coarse-to-fine manner. These observations are consistent with the idea 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, whereas 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. One may imagine that the higher-level areas in the case illustrated in Fig. 1 can instantly "recognize" the face image on the basis of 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 could not be detected under the same illumination condition, we would be alarmed and might form a different interpretation about what we actually saw.

Not every computation has to work its way all the way back to V1. Kosslyn *et al.*[57] showed that in fMRI studies a subject's V1 lights up differentially only when he or she is asked to imagine things or perform tasks that involve 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, feedforward computation might be sufficient. All the experiments that managed to demonstrate a high-level or attentional effect in V1 seemed to require the monkeys to utilize information about fine details in their tasks. In Roelfsema and colleagues' experiment,[53] for example, the monkey was asked to trace one of the two curves displayed on the screen. They 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.

Attention effect in V1 usually becomes evident only when the scene is ambiguous. Motter[58] found that it is very difficult to demonstrate "attentional modulation" (i.e., top-down effect) when there is only a 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 atten-

PROOF COPY [A-8700] 005307JOA

PROOF COPY [A-8700] 005307JOA

tion to a particular location often results in the removal of the inhibition imposed by the surround on that location. Gilbert and colleagues[54] demonstrated an attentional effect in V1 only after the monkeys were trained to perform a vernier acuity test—aligning two small vertical lines. All these findings suggest that when the monkeys perform tasks that require the discrimination of fine features, feedback can penetrate back to V1. Interestingly, Kamitani and Shimojo[50] found that when different spatial-frequency gratings were used as stimuli in their TMS experiment, the optimal range of TMS delay was systematically increased as the spatial frequency increased, indicating that a finer-resolution analysis might indeed require earlier visual areas. Ideas similar to this have also been proposed recently by Hochstein and Ahissar in their reverse hierarchy theory[59]; they argue, on the basis of their psychophysical observations, that vision at a glance can be accomplished by feedforward computation but vision with scrutiny involves coupling to early visual areas.

C. Interactive Hierarchy

Although the above experiments demonstrate the emergence of attentional effect in early visual areas during high-resolution analysis, it is unclear to what degree feedback is involved in normal perceptual processing. In a hierarchical-inference framework, feedback should be quite automatic. We[41,48,49] have conducted a series of experiments to investigate the role of V1 and V2 in complex perceptual processes that likely involve interaction among multiple cortical areas. Here we will focus on the experiments on contour completion and shape from shading.

Since the time of Hubel and Wiesel,[42] it has been hypothesized that V1 is involved in edge detection. We[41] demonstrated earlier that while the initial responses of V1 neurons are characterized by the filtering of the local texture elements, the later part of their responses are correlated with more-abstract global boundaries. Studies from Gilbert's laboratory[60] found that an additional bar along the longitudinal direction outside the receptive field could exert a facilitatory effect on a V1 neuron. Such findings have inspired a set of models based on V1 circuitries for contour continuation.[29,61–64] In addition, a number of experiments (e.g., Ref. 65) found that additional bars on the two sides of the neuron's longitudinal axis tend to suppress the response of a neuron, reminiscent of the nonmaximum suppression in edge detection. Nevertheless, there is no direct evidence for contour completion in V1. Neural correlates of illusory contour from the Kanizsa triangle type have been observed only in V2 but not in V1; this finding has also been used to argue for a feedforward scheme of computation.[66]

The high-resolution-buffer hypothesis suggests that V1 has the ideal machinery for computing geometrical curvilinear structures, as illustrated in Roelfsema and colleagues' curve-tracing experiment.[53] In light of these considerations, we[48] reexamined the issue of neural responses to illusory contours in area V1 and V2, using a

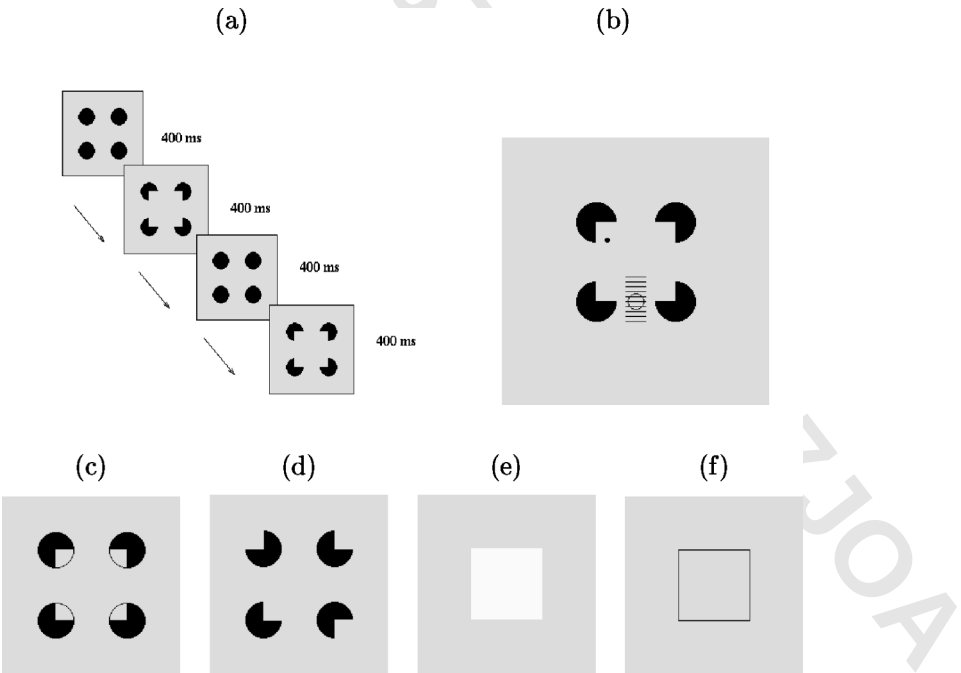

Fig. 3. Selected stimuli in the subjective contour experiment. (a) Example of a stimulus presentation sequence in a single trial. (b) Kanizsa square with illusory contour. Receptive field of the tested neuron was "placed" at ten different positions across the illusory contour, one per trial. (c) Amodal contour stimulus; the subjective contour was interrupted by intersecting lines. (d) One of the several rotated partial disk 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, used as control to assess the neuron's sensitivity to the spatial location of the real contour as well as to comparing the temporal responses between real and illusory contours. See Lee and Nguyen[48] for details.

PROOF COPY [A-8700] 005307JOA

PROOF COPY [A-8700] 005307JOA

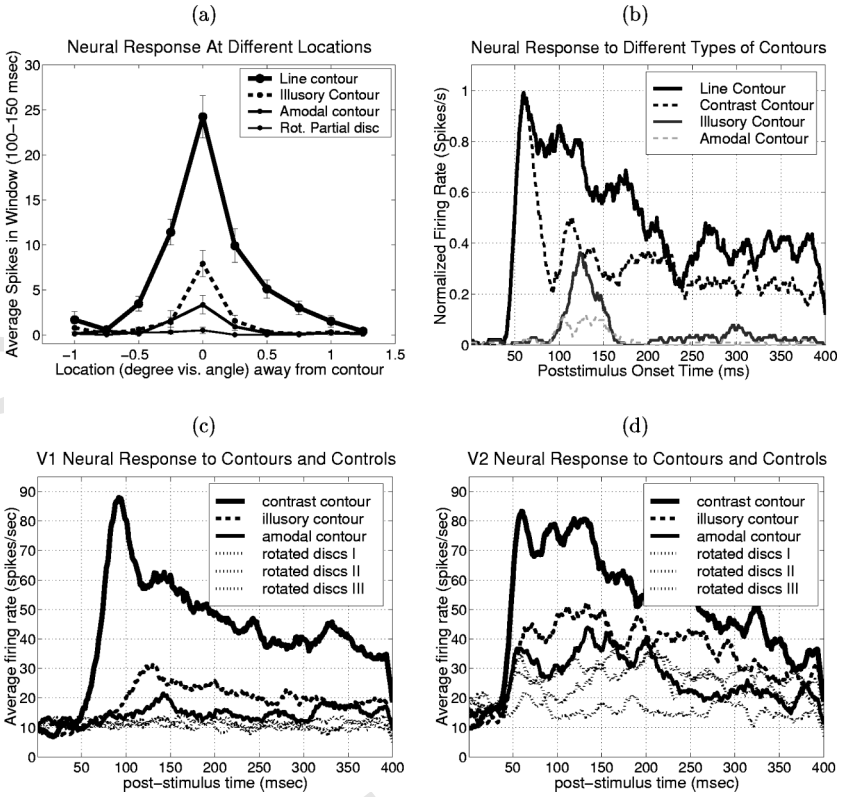

Fig. 4. (a) Spatial profile of a V1 neuron's response to the contours of both real and illusory squares, in a temporal window 100–150 ms after stimulus onset. The real or illusory square was placed at different spatial locations relative to the receptive field of the cell. This cell responded to the illusory contour when it was at precisely the same location where a real contour evoked the maximal response from the neuron. It 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 partial disks were rotated. (b) Temporal evolution of the cell's response to the illusory contour compared with 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, ~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. The results show that V2 responds to illusory contour earlier than V1. See Lee and Nguyen[48] for details.

static display paradigm that allowed us to monitor the temporal evolution of responses to specific stimulus locations. We found that neurons in V1 do indeed respond to illusory contours, e.g., completing the contour induced by the partial disks shown in Fig. 3, although at a latency greater than that in V2.

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 the different locations relative to the illusory contour was recorded (Fig. 3). At the beginning of the experiment, consistent with Von der Heydt's earlier report,[66] we found that V1 neurons in fact do not respond to the illusory contours. We then realized that because the partial disks (pac men) were shown in the periphery, the monkey might simply be seeing the on and off flashing of partial disks on the screen without perceiving the illusory square. We took several measures to enhance the monkey's attention to the illusory square. First, we placed the fixation spot inside the illusory square, so that the monkey was looking at the illusory square. Second, we presented the stimuli in a sequence: Four black circular disks appeared first for 400 ms and then turned into the partial disks, 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 served 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 session after this shift in paradigm, we started to find V1 neurons responding to the illusory contour in the stimulus (Fig. 4).

The neural correlate of the illusory contour signal emerged in V1 neurons at precisely the same location where a line or luminance contrast elicited the maximum response from the cell [Fig. 4(a)]. The response to the illusory contour was delayed relative to the response to the real contours by 55 ms [Fig. 4(b)], emerging ~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 partial disks were rotated. At the population level, we found that sensitivity to illusory contours emerged at 65 ms in V2, 35 ms ahead of V1 [Fig. 4(c) and 4(d)]. A possible explanation is that V2 detects the existence of an illusory contour by integrating information from a spatially more global context and then generates a prior $P(x_{v1}|x_{v2})$ to constrain the contour inference in V1. The resulting con-

PROOF COPY [A-8700] 005307JOA

PROOF COPY [A-8700] 005307JOA

tour supports the hypothesis particle that maximizes $P(x_0, x_{v1}, x_{v2}, x_{v4}, x_{IT})$ which is the product of a cascade of feedback priors and bottom-up hypotheses.

Contour completion is an excellent example of particle filtering. In toy examples, there is one and only one completion of the local edges into global contours. In natural scenes, however, there are thousands of edge fragments that may potentially be part of very salient global contours. The strongest extended contours can emerge as a result of the competition between partial hypotheses, each of which has linked up some of the edge fragments and seeks confirming evidence either by linking onto longer contours or by binding to emerging objects, missing pieces being explained by occlusion, shadows, etc. The actual implementation of this contour completion process in V1 might be similar to Williams and Jacobs's[62] stochastic random-walk model for contour continuation, except that it also contains many hierarchical layers of computations involving higher-level information such as emerging objects and their occlusion as well.

The illusory contour result supports the idea of a generative model, but the generative process in this case could be mediated by horizontal connections and not necessarily by feedback. To firmly demonstrate that higher-order perceptual constructs could exert an influence in the early visual areas, we studied a set of shape-from-shading (SFS) stimuli that likely involve the interaction between high-level three-dimensional (3D) inference and low-level parallel processing. When viewing the display shown in Fig. 5(a), we perceive a set of convex shapes automatically segregating from a set of concave shapes. These interpretations of 3D shapes emerge purely from the shading information, assuming lighting comes from above. If our assumption is changed to lighting from below, perceptually the convex shapes can be seen as concave while the concave shapes can be seen as convex. These two interpretations can alternate in perceptual rivalry as in the Necker cube illusion. Ramachandran[67] 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 black-and-white (BW) contrast elements, but without a 3D interpretation, does not readily segregate into groups [Fig. 5(b)]. These pairs of stimuli are therefore ideal for probing the interaction between high-level interpretation (3D inference) and low-level parallel processes.

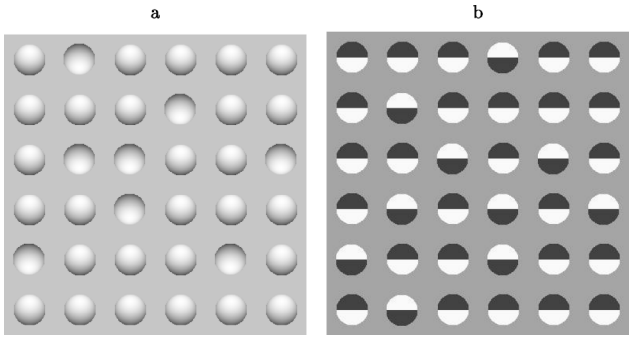

Fig. 5. Ramachandran[67] showed that SFS stimuli produced instantaneous segregation, whereas BW contrast stimuli did not. Given the main distinction between the two types of stimuli is that only the SFS stimulus elements in (a) but not those in (b) afford 3D interpretation; 3D information must directly influence the early parallel processes of perceptual grouping.

To study the neural basis of higher-order pop-out, we used a paradigm developed originally by Knierim and Van Essen,[68] who had demonstrated that V1 is sensitive to pop-out that is defined by an oriented bar. Here we[49] studied how V1 and V2 neurons respond to SFS stimuli and the neural correlates of their pop-out saliency due to 3D interpretation. We tested the responses of V1 and V2 neurons when the center of a stimulus element was placed on top of their receptive fields. Typically, the receptive field is less than 0.7 deg, while the diameter of the stimulus element is 1 deg visual angle. When comparing the neuronal responses to the BW stimulus with the responses to the SFS stimulus (Fig. 6), we found that V1 neurons are very sensitive to contrast and respond better to the BW stimulus than to 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 the later part of their responses [Fig. 7(a) and 7(b)]. This shows that the V2 neurons might be more interested in a representation of 3D surface that is more abstract than the bottom-up luminance contrast. Furthermore, we found that whereas both V1 and V2 neurons did not exhibit pop-out responses for the BW stimulus, V2 but not V1 neurons did exhibit the pop-out response for the SFS stimulus in a passive fixation task from the very beginning. The pop-out response is defined by the ratio of the response of the neuron to the oddball condition over its response to the uniform condition. In these two conditions, the stimulus on the receptive field is the same, but the surrounding contexts are different [Figs. 7(a) and 7(b)]. The fact that V2 exhibits a preattentive pop-out response to shape from shading further argues for the possibility that V2 neurons might be representing 3D shape primitives, allowing parallel pop-out computation in V2 through lateral inhibition. Recently, Von der Heydt's laboratory[69] found that V2 neurons are indeed sensitive to convex shapes defined by both shape from shading and random-dot stereogram, providing more direct evidence in support of this idea.

Interestingly, although V1 neurons were not sensitive to the SFS pop-outs at the beginning, they became sensitive to them after the monkeys utilized them in a task that required them to detect the location of the pop-out target. On the other hand, even though the monkeys could detect the oddball in the BW stimulus, albeit at much slower speed, their V1 and V2 neurons exhibited the pop-out effect for the SFS stimulus but not for the BW stimulus. As a population, the SFS pop-out emerged in V2 at ~95-ms poststimulus onset and in V1 at 100 ms (Fig. 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.[50]

What is the purpose of these higher-order pop-out saliency signals to penetrate back to V1? One possible clue is our observation that the pop-out signal was spatially precise in V1—that it could be observed only on the target but not on the distractor elements right next to it. We take this to mean that when the monkeys have to detect

PROOF COPY [A-8700] 005307JOA

PROOF COPY [A-8700] 005307JOA

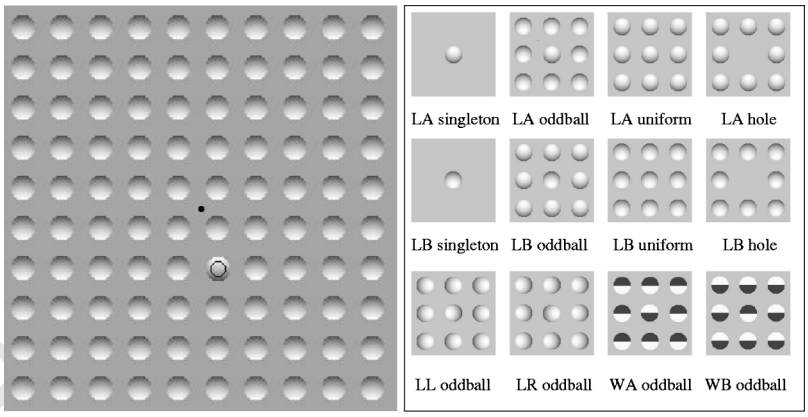

Fig. 6. Higher-order perceptual pop-out. (a) A typical stimulus display was composed of ten × 10 stimulus elements. Each element was 1° visual angle in diameter. The diameter of the classical receptive field of a typical cell at the eccentricities tested ranged from 0.4° to 0.8° visual angle. Displayed is an example of a lighting from above (LA) 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) There are six sets of stimuli. The SFS stimulus elements include LA and Lambertian sphere with lighting from above, below, left and right (LB, LL, and LR, respectively). The BW stimulus elements include white above (WA) and white below (WB). 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 and the oddball conditions for the other four sets. The center element in the iconic diagram covers the receptive field of the neuron in the experiment. The surround stimulus elements were placed outside the receptive field of the neuron. The key comparison was made between the oddball condition and the uniform condition, while the singleton and the hole conditions were controls. See Lee *et al.*[49] for details.

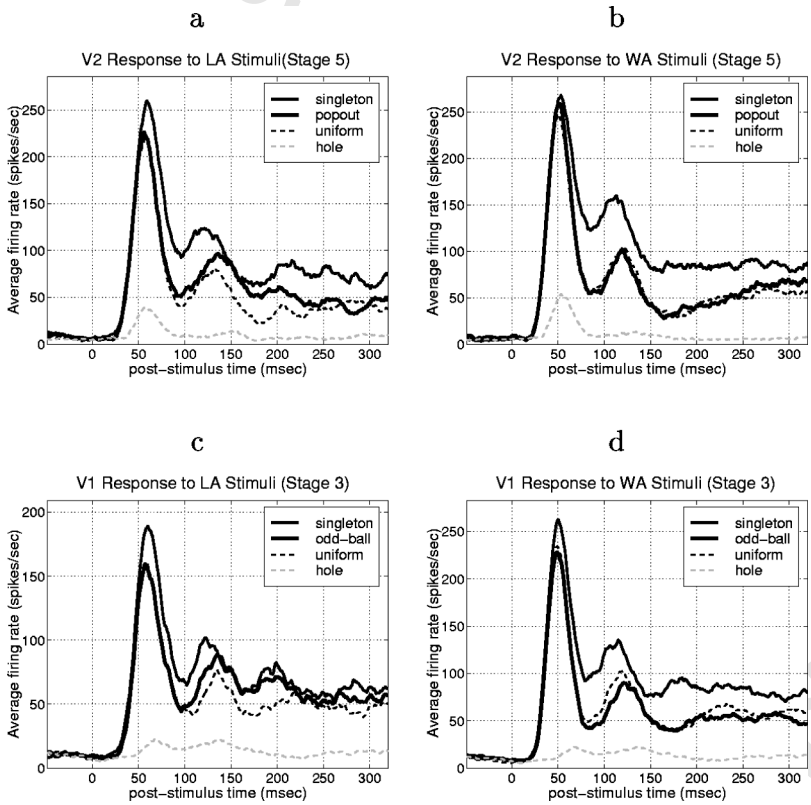

Fig. 7. Temporal evolution of the average population response of 22 V2 units and 30 V1 units from a monkey to the LA set and the WA set in a stage after the monkey had utilized the stimuli in its behavior. Each unit's response was first smoothed by a running average within a 15-ms window and then averaged across the population. A significant difference (pop-out response) was observed between the population average response to the oddball condition and that to the uniform condition in the LA set for both V2 and V1 [(a), (c)] neurons, starting at 100-ms poststimulus onset. No pop-out response was observed in the WA set [(b), (d)]. See Lee *et al.*[49] for details.

the location of a small target, a spatially precise signal needs to be established with the aid of the high-resolution buffer. In addition, the fine interpretation of the 3D shape might also involve constant interaction with V1. The pop-out signals can be attenuated when the monkeys are required to do another attention-demanding task. However, the signals could not simply be attributed to attention alone, because we can observe them in the mon-

PROOF COPY [A-8700] 005307JOA

PROOF COPY [A-8700] 005307JOA

keys even after they have not performed the oddball detection task for over a year. It seems that this coupling between V2 and V1 had increased in strength with practice and become more or less automatic.

We suspect that the enhancement signal observed here is very similar to Lamme's[46] figure–ground enhancement effect observed in the texture figures. In that experiment, V1 neurons' responses to different parts of the texture stimuli along a horizontal line across the center of the stimuli were studied. The stimuli include 4-deg-wide textured squares in a background defined by orthogonal textures [Figs. 8(a) and 8(b)]. When the responses to the two complementary stimuli were summed at the corresponding locations, the response inside the figure was found to be stronger than the response outside the figure. When this experiment was repeated, we[41] found that there was a general uniform enhancement within the figure, which abruptly terminated at its boundary [as shown in Fig. 8(d)], even though the magnitude of the effect was ~15%—significantly weaker than observed earlier.[46,47] [Note that when preferred orientation of the cells was parallel to that of the texture boundary, a very strong boundary effect was found to be superimposed on the interior enhancement effect, as shown in Fig. 8(e)]. The enhancement effect within an object's surface is reminiscent of the "coloring" operation in Ullman's[70] visual routines. Coloring an object precisely within the boundary of a surface requires the spatial precision provided by the high-resolution buffer.

The beliefs on 3D shape from V2 might provide the necessary priors to modulate the parallel pop-out computation and the precise localization of the pop-out target in V1. The data suggest that these priors contain not only 3D information but also the information on saliency and behavioral relevance.[49] When we changed the top-down priors, for example, by manipulating the presentation frequency of the different oddball stimuli, the monkey's reaction time and behavioral accuracy improved for the more-frequent stimuli. The change in the behavioral performance of the monkeys was often accompanied by a parallel change in the relative pop-out strength in the neural signals. These interactions among statistics of stimuli, behavioral experience, and neural processing are characteristic of a hierarchical Bayesian-inference framework.

Hierarchical inference is most evident when stimuli are ambiguous and the correct interpretation requires integration of multiple contextual factors for disambiguation. In the case of Kanizsa square, there are several possible hypotheses for explaining the bottom-up data. The brain seems to choose the simplest explanation: that a white square is occluding four black circular disks, even at the extra expense of hallucinating a subjective contour at locations where there is really no visual evidence for it. It

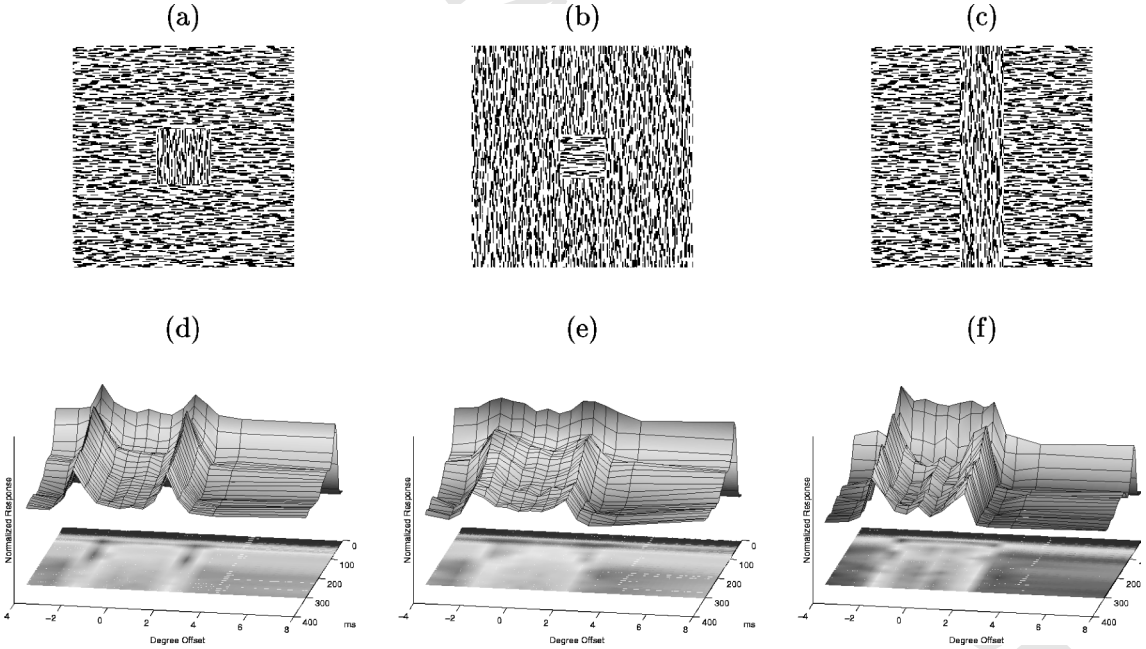

Fig. 8. (a)–(c) Three examples of the texture stimuli used in the experiment. Different parts of the stimuli, along a horizontal line across the middle of the square or across the strip, were placed on the receptive field of the recorded neuron over successive trials. The width of the square or the strip is 4°. (d), (e). Spatiotemporal evolution of the summed response of a population of V1 neurons to the texture squares. The summed response was obtained by adding the response to stimulus (a) and the response to stimulus (b) at the corresponding locations. This addition eliminates the effect due to orientation tuning and reveals a signal that enhances the figure's interior relative to the background. The spatial offset is the distance in degrees of visual angle from the center of the square or the strip; hence −2° and 2° offsets represent the boundary locations. When the neurons' preferred orientation was orthogonal to the vertical texture boundaries, the enhancement was relatively uniform within the figure [(d)]. When the neurons' preferred orientation was parallel to that of the texture boundaries, a very strong boundary signal was superimposed on the interior enhancement (coloring) signal. (e) Population-averaged response of a set of vertical neurons to stimulus (c). The initial response was characterized by a burst, whose magnitude was correlated with sensitivity to local feature orientation, followed by a more sustained response at a lower level. The response at the boundary is significantly higher than the response at the interior. These phenomena underscore the interplay among resonance, competition and "explaining away." See Lee *et al.*[41] for details.

PROOF COPY [A-8700] 005307JOA

PROOF COPY [A-8700] 005307JOA

is only in this ambiguous situation that one can see a feedback effect in V1. In the SFS experiment, it is the need to finely localize the pop-out stimulus that drives the processing back to V1. An experiment by Bullier's laboratory[51] showed that the effect of feedback was most evident in V1 only when the stimuli were of low visibility, low saliency, and high ambiguity.

D. Multiple Hypotheses

Is there any neurophysiological evidence that is suggestive of particle filtering in the cortex? A hallmark of particle filtering is that multiple hypotheses are kept alive during the computation so that the system does not need to jump to conclusions and is capable of entertaining other possibilities simultaneously. In general, demonstrating particle filtering in cortex requires ambiguous visual images in which competing structures are present and simultaneous recording of activity of assemblies of neurons in the same and different areas. Such data are not available at present.

One line of evidence that potentially supports such an idea is the binocular-rivalry experiment from Logothetis's laboratory.[71] When two different images are presented to the two eyes, we often only see one image at a time, and the two images fluctuate over time. This has been 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.[71] A curious fact is that almost all the *relevant* IT neurons (i.e., neurons that can distinguish the two images when monocularly presented based on features within their receptive fields) respond consistently with perception, whereas only 10% of the *relevant* V1 neurons and 20% of the V2 neurons responded in accordance with the percept. This gradual increase in the percentage of neurons whose responses are correlated with current state of perception along the visual hierarchy has also been observed in both our illusory contour and SFS experiments.[48,49] The gradual increase in the neural correlate of perception along the visual hierarchy has been taken to mean that V1 is less "conscious" than IT. From the point of view of particle filtering, this could imply less than 10% of *each type* of neuron in V1 are involved in representing a particular hypothesis or perceptual state, whereas the other 90% are involved in the representation of the alternative hypotheses. On the other hand, neurons in IT are more hypothesis specific in that most of them do not care about a particular hypothesis, but for those who do, they respond consistently with the current perceptual state. When IT or prefrontal cortex is tired of or satisfied with one hypothesis, the remaining hypotheses that have been kept alive lower in the visual hierarchy will emerge to offer alternative explanations to the data.

E. Resonance and Predictive 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 have been striking experiments recently that show relative suppression of low-level activity when an integrated simple high-level percept can explain the low-level data. Murray and colleagues[72] using fMRI on human subjects showed that when similar sets of stimuli were presented—one relatively complex two-dimensional (2D) pattern and one with a simple 3D 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. They found a correlation between the times in which the subject reported seeing the 3D percept and the times in which V1 activity decreased. Furthermore, Roe and her colleagues[73] also found that in their optical imaging and single-unit experiments, V1 neurons' responses were suppressed but V2 neurons' activities were enhanced when an illusory contour defined by abutted sine-wave gratings was presented. These experiments support our earlier proposal[1] and related ideas[7,16,74] that top-down generative signals could explain away the earlier evidence based on efficient coding consideration. In that proposal, certain bottom-up pathways carried error signals indicating when there was a 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, we would like to propose a completely different interpretation of Murray *et al.*'s[72] results here, which uses the theory of multiple hypotheses or particles. In a situation in which complex data are 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, now enhanced by feedback from higher areas. This single enhanced grouping contains, in total, less neural activity than the multitude of competing but nonenhanced groupings.

This key insight might help to reconcile the concepts of the resonance and the explaining-away phenomena in generative models. Higher-order description not only explains away the data that are consistent with it but more violently suppresses the low-level data that are noise or are supporting alternative hypotheses. Note that the feedback never completely eliminates the low-level responses, as there are features and cues in the earlier representation that are not captured by the higher ones. This perspective recasts the findings in the texture experiment (Fig. 8) in a different light. Figure 8(f) shows the spatiotemporal response of V1 vertical neurons to a texture-strip stimulus [Figure 8(c)]. Several observations are particularly interesting. First, the initial neuronal response (35–70-ms poststimulus onset) was characterized by the response to local features, i.e., sensitivity to orientation of the line elements, but the later responses (80 ms onward) emphasized the responses at the texture boundary. The suppression of the interior response relative to the boundary might mean that the redundant information in the interior of the figure is being explained away by the surround. The maintained enhanced response at the the boundary response might arise from

PROOF COPY [A-8700] 005307JOA

PROOF COPY [A-8700] 005307JOA

resonance with the global boundary representation in higher areas. Second, there is a general adaptation process that results in lower and sustained activities across the whole image in the later phase of the response (even at the boundary location). Such adaptation is also evident in V1 and V2 neuronal activities in the later stage of their responses in many other different scenarios (Figs. 4 and 7). Traditionally, such adaptation is attributed to the nonlinear neural dynamics, lateral inhibition, or synaptic depression mechanisms, but certain adaptation phenomena potentially can be interpreted as the explaining away of the earlier representation by higher-order representations as a result of feedback. The idea that feedback collapses the particle distribution means that it explains away inconsistent evidence more severely than the consistent evidence at the lower level, and the net outcome is *relative* resonance.

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 based on attention modulation and biased competition. From that perspective, visual processing is still primarily a series of feedforward computations, except that the computation and information flow are regulated by selective attention.[10] On the other hand, within the neural modeling community, there have been a number of models or theories[1,4,6,7,16,17] with increasing sophistication, emphasizing that the feedback from higher-order areas might directly or indirectly serve as contextual priors for influencing lower-level inference. Here we suggest that these ideas could be formulated in the form of a hierarchical Bayesian system and that ideas from Bayesian belief propagation[19] and particle filtering[21,23,25] 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. The top-down priors can reshape the probabilistic posterior distribution of the various hypotheses at each level by recurrent feedback.

We reviewed a number of recent neurophysiological findings that are highly suggestive of such a hierarchical inference system and, in particular, suggestive of the unique role of the primary visual cortex as a high-resolution buffer in this hierarchy. The effect of feedback is often subtle and often 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 evidence that is not necessarily consistent with the current dominant hypothesis. As a result, only a smaller percentage of early visual neurons in each class are correlated with the particle that supports the current perceptual state.

Central to our framework is the forward/backward mechanism that is embodied conceptually in many existing neural models.[1,4,7,16,17] Here we attempt to reconcile a subtle, but important, difference between two competing schools of thought. In the adaptive-resonance[16] or interactive-activation models,[17] an active global concept will feed back to enhance the neural activities in the early areas that are consistent with the global percept. These ideas are supported by numerous neurophysiological experiments that show that higher-order information can enhance early visual responses.[46,48,49] On the other hand, the efficient-coding[1] and the predictive-coding models[7] emphasize that feedback serves to suppress the activities in the early areas as a way of "explaining away" the evidence in the earlier areas. This idea is supported particularly by some recent imaging experiments.[7,72] In the latter class of models, only error residues are projected forward to the higher areas. In our current proposal, the computation of beliefs is based on both bottom-up and top-down messages. A particle is then an ensemble of both deep and superficial cells whose strength as an ensemble (the binding strength via synchrony or rapid synaptic weight changes) is something like the weight of the particle. The two schools of thought can be reconciled by simply understanding that both inconsistent evidence and consistent evidence in earlier areas are being explained away, but the effect is more severe on the inconsistent data, resulting in a relative enhancement of the consistent data as a result of resonance. Therefore resonance, competition, and predictive coding are all key components in this framework. During perceptual inference, beliefs are propagated up and down to collapse the hypothesis space. When an interpretation is reached, the residues, the parts not explained, will start to attract attention to initiate further processing.

Recent Bayesian-belief propagation and particle-filtering models keep particles for the forward and backward streams separate and noninteracting until the last step for mathematical reasons,[20] but it might be beneficial to combine top-down and bottom-up information as soon as possible to form particles that reflect both bottom-up and top-down information, as we have suggested here. Although the precise computational and neural implementation of many aspects of Bayesian-belief propagation and particle filtering is not entirely clear, we think that the parallel between recent artificial intelligence work on Bayesian-belief propagation and particle-filtering and recent neurophysiological findings in the visual cortex are striking and should not be ignored. This paper summarizes our thoughts on their likely connections and aims at stimulating more-precise experimental research along this line. We expect that these ideas will grow exponentially in the next few years in the computational vision and biological vision communities and might revolutionarize how we think about neural and computational processes underlying vision.

ACKNOWLEDGMENTS

Tai Sing Lee (tai@cs.cmu.edu) was supported by National Science Foundation CAREER grant 9984706 and National Institutes of Health grants MH64445 and EY08098. David Mumford (David_Mumford @brown.edu) was supported by National Science Foundation grant DMS-0074276 and Burroughs Wellcome Foundation grant 2302.

PROOF COPY [A-8700] 005307JOA

PROOF COPY [A-8700] 005307JOA

REFERENCES

1. D. Mumford, "On the computational architecture of the neocortex II," Biol. Cybern. **66**, 241–251 (1992).
2. D. Mumford, "Pattern theory: a unifying perspective," in *Perception as Bayesian Inference*, D. C. Knill and W. Richards, ed. (Cambridge U. Press, Cambridge UK, 1996), pp. 25–62.
3. U. Grenander, *General Pattern Theory* (Oxford U. Press, Oxford, UK, 1993).
4. G. Hinton, P. Dayan, B. Frey, and R. Neal, "The wake-sleep algorithm for unsupervised neural networks," Science **268**, 1158–1161 (1995).
5. P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel, "The Helmholtz machine," Neural Comput. **7**, 889–904 (1995).
6. M. S. Lewicki and T. J. Sejnowski, "Bayesian unsupervised learning of higher order structure," in *Advances in Neural Information Processing Systems 9*, M. Mozer, M. Jordan, and T. Petsche, eds. (MIT Press, Cambridge, Mass., 1997), pp. 529–535.
7. R. Rao and D. Ballard, "Dynamic model of visual recognition predicts neural response properties in the visual cortex," Neural Comput. **9**, 721–763 (1997).
8. C. E. Guo, S. C. Zhu, and Y. N. Wu, "Visual learning by integrating descriptive and generative models," Int. J. Comput. Vision (to be published).
9. Z. W. Tu and S. C. Zhu, "Image segmentation by data-driven Markov chain Monte Carlo," IEEE Trans. Pattern Anal. Mach. Intell. **24**, 657–673 (2002).
10. R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," Annu. Rev. Neurosci. **18**, 193–222 (1995).
11. M. Usher and E. Niebur, "Modeling the temporal dynamics of IT neurons in visual search: A mechanism for top-down selective attention," J. Cognit. Neurosci. **8**, 311–327 (1996).
12. G. Deco and T. S. Lee, "A unified model of spatial and object attention based on inter-cortical biased competition," Neurocomputing **44–46**, 769–774 (2002).
13. D. Marr, *Vision* (Freeman, San Francisco, Calif., 1983).
14. R. VanRullen and S. Thorpe, "Is it a bird? Is it a plane? Ultra-rapid visual categorization of natural and artificial objects," Perception **30**, 655–668 (2001).
15. D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex," Cereb. Cortex **1**, 1–47 (1991).
16. G. Carpenter and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern recognition," Machine Comput. Vision Graphics Image Process. **37**, 54–115 (1987).
17. J. L. McClelland and 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).
18. J. Pearl, *Probabilistic Reasoning in Intelligent Systems* (Morgan Kaufmann, San Mateo, Calif., 1988).
19. J. Yedidia, W. T. Freeman, and Y. Weiss, "Understanding belief propagation and its generalization," presented at the International Joint conference on Artificial Intelligence (IJCAI 2001), Seattle, Washington, August 4–10, 2001.
20. E. Sudderth, A. Ihler, W. Freeman, and A. Willsky, "Nonparametric belief propagation," MIT Artificial Intelligence (AI) Laboratory Memo No. 20 (MIT AI Lab, Cambridge, Mass., 2002).
21. A. Doucet, N. de Freitas, and N. Gordon, eds., *Sequential Monte Carlo Methods in Practice* (Springer-Verlag, New York, 2001).
22. M. Isard, "Pampas: real-valued graphical models for computer vision," Proc. Comput. Vision Pattern Recog. 2003 (to be published).
23. M. Isard and A. Blake, "ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework," in *Lecture Notes in Computer Science 1406*, H. Burkhardt, B. Neumann, ed. (Springer-Verlag, Berlin, 1998), pp. 893–908.
24. A. Blake, B. Bascle, M. Isard, and J. MacCormick, "Statistical models of visual shape and motion," Proc. R. Soc. London, Ser. A **356**, 1283–1302 (1998).
25. S. Thrun, D. Fox, W. Burgard, and F. Dellaert, "Robust Monte Carlo localization for mobile robots," Artif. Intell. **101**, 99–141 (2001).
26. D. C. Knill and W. Richards, ed. *Perception as Bayesian Inference* (Cambridge U. Press, Cambridge, UK, 1996).
27. T. S. Lee, "A Bayesian framework for understanding texture segmentation in the primary visual cortex," Vision Res. **35**, 2643–2657 (1995).
28. H. V. Helmholtz, *Handbuch der physiologischen Optik* (Voss, Leipzig, Germany 1867).
29. W. S. Geisler, R. L. Diehl, "Bayesian natural selection and the evolution of perceptual systems," Philos. Trans. R. Soc. London, Ser. B **357**, 419–448 (2002).
30. D. Mumford, "On the computational architecture of the neocortex I," Biol. Cybern. **65**, 135–145 (1991).
31. E. Adelson and A. Pentland, "The perception of shading and reflectance," in *Perception as Bayesian Inference*, D. Knill and W. Richards, eds. (Cambridge U. Press, Cambridge, UK, 1996), pp. 409–423.
32. P. Sinha and E. Adelson, "Recovering reflectance in a world of painted polyhedra," in *Proceedings of the 4th International Conference on Computer Vision* (IEEE Computer Society Press, Los Alamitos, Calif., 1993), pp. 156–163.
33. R. Zemel, "Cortical Belief Networks," in *Computational Models for Neuroscience*, R. Hecht-Neilsen and T. McKenna, eds. (Springer-Verlag, New York) (to be published).
34. C. Eliasmith and C. H. Anderson, *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems* (MIT Press, Cambridge, Mass., 2002).
35. D. W. Arathorn, *Map-Seeking Circuits in Visual Cognition: a Computational Mechanism for Biological and Machine Vision* (Stanford U. Press, Palo Alto, Calif., 2002).
36. C. M. Gray, "The temporal correlation hypothesis of visual feature integration: still alive and well," Neuron **24**, 31–47 (1999).
37. E. Bienenstock, S. Geman, and D. Potter, "Compositionality, MDL priors, and object recognition," in *Advances in Neural Information Processing Systems*, M. C. Mozer, M. I. Jordan, and T. Petsche, eds. (MIT Press, Cambridge, Mass., 1997), Vol. 9, pp. 838–844.
38. C. v. D. Malsburg, "The what and why of binding: the modeler's perspective," Neuron **24**, 95–104 (1999).
39. P. J. Sjostrom and S. B. Nelson, "Spike timing, calcium signals and synaptic plasticity," Curr. Opin. Neurobiol. **12**, 305–314 (2002).
40. D. Mumford, "Commentary on banishing the homunculus by H. Barlow," in *Perception as Bayesian Inference*, D. C. Knill and W. Richards, eds. (Cambridge U. Press, Cambridge, UK 1996), pp. 501–504.
41. T. S. Lee, D. Mumford, R. Romero, and V. A. F. Lamme, "The role of the primary visual cortex in higher level vision," Vision Res. **38**, 2429–2454 (1998).
42. D. H. Hubel and T. N. Wiesel, "Functional architecture of macaque monkey visual cortex," Proc. R. Soc. London, Ser. B **198**, 1–59 (1978).
43. R. Gattass, A. P. Sousa, and C. G. Gross, "Visuotopic organization and extent of V3 and V4 of the macaque," J. Neurosci. **8**, 1831–1845 (1988).
44. C. G. Gross, "Visual function of inferotemporal cortex," in *Handbook of Sensory Physiology*, 7/3B, L. R. Jung, ed. (Springer-Verlag, Berlin, 1973), pp. 451–482.
45. R. L. De Valois and K. K. De Valois, *Spatial Vision* (Oxford U. Press, New York, 1988).
46. V. A. F. Lamme, "The neurophysiology of figure-ground segregation in primary visual cortex," J. Neurosci. **15**, 1605–1615 (1995).
47. K. Zipser, V. A. F. Lamme, and P. H. Schiller, "Contextual modulation in primary visual cortex," J. Neurosci. **16**, 7376–7389 (1996).
48. T. S. Lee and M. Nguyen, "Dynamics of subjective contour formation in the early visual cortex," Proc. Natl. Acad. Sci. USA **98**, 1907–1911 (2001).
49. T. S. Lee, C. Yang, R. Romero, and D. Mumford, "Neural activity in early visual cortex reflects behavioral experience

PROOF COPY [A-8700] 005307JOA

PROOF COPY [A-8700] 005307JOA

and higher order perceptual saliency," Nat. Neurosci. **5**, 589–597 (2002).
50. Y. Kamitani and S. Shimojo, "Manifestation of scotomas by transcranial magnetic stimulation of human visual cortex," Nat. Neurosci. **2**, 767–771 (1999).
51. J. M. Hupe, A. C. James, B. R. Payne, S. G. Lomber, P. Girard, and J. Bullier, "Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons," Nature **394**, 784–787 (1998).
52. H. Super, H. Spekreijse, and V. A. F. Lamme, "Two distinct modes of sensory processing observed in monkey primary visual cortex (V1)," Nat. Neurosci. **4**, 304–310 (2001).
53. P. R. Roelfsema, V. A. F. Lamme, and H. Spekreijse, "Object-based attention in the primary visual cortex of the macaque monkey," Nature **395**, 376–381 (1998).
54. M. Ito and C. D. Gilbert, "Attention modulates contextual influences in the primary visual cortex of alert monkeys," Neuron **22**, 593–604 (1999).
55. N. P. Bichot and J. D. Schall, "Effects of similarity and history on neural mechanisms of visual selection," Nat. Neurosci. **2**, 549–554 (1999).
56. Y. Sugase, S. Yamane, S. Ueno, and K. Kawano, "Global and fine information coded by single neurons in the temporal visual cortex," Nature **400**, 869–873 (1999).
57. S. Kosslyn, W. L. Thompson, I. J. Kim, and N. M. Alpert, "Topographical representations of mental images in primary visual cortex," Nature **378**, 496–498 (1995).
58. B. C. Motter, "Focal attention produces spatially selective processing in visual cortical areas V1, V2, V4 in the presence of competing stimuli," J. Neurophysiol. **70**, 909–919 (1993).
59. S. Hochstein and M. Ahissar, "View from the top: hierarchies and reverse hierarchies in the visual system," Neuron **36**, 791–804 (2002).
60. M. K. Kapadia, G. Westheimer, and C. D. Gilbert, "Spatial distribution of contextual interactions in primary visual cortex and in visual perception," J. Neurophysiol. **84**, 2048–2062 (2000).
61. J. August and S. W. Zucker, "The curve indicator random field: curve organization via edge correlation," in *Perceptual Organization for Artificial Vision Systems*, K. Boyer and S. Sarka, eds. (Kluwer Academic, Boston, Mass., 2000), pp. 265–288.
62. L. Williams and D. Jacobs, "Stochastic completion fields: a neural model of illusory contour shape and saliency," Neural Comput. **9**, 837–858 (1997).
63. J. Braun, "On the detection of salient contours," Spatial Vis. **12**, 211–225 (1999).
64. Z. Li, "A neural model of contour integration," Neural Comput. **10**, 903–940 (2001).
65. R. T. Born and R. B. H. Tootell, "Single-unit and 2-deoxyglucose studies of side inhibition in macaque striate cortex," Proc. Natl. Acad. Sci. USA **88**, 7071–7075 (1991).
66. R. von der Heydt, E. Peterhans, and G. Baumgarthner, "Illusory contours and cortical neuron responses," Science **224**, 1260–1262 (1984).
67. V. S. Ramachandran, "Perception of shape from shading," Nature **331**, 163–166 (1988).
68. J. J. Knierim and D. C. Van Essen, "Neuronal responses to static texture patterns in area V1 of the alert macaque monkey," J. Neurophysiol. **67**, 961–980 (1992).
69. F. T. Qiu, R. Endo, and R. von der Heydt, "Selectivity for structural depth in neurons of monkey area V2," presented at the 30th Annual Meeting of the Society of Neuroscience, New Orleans, Louisiana November 4–9, 2000.
70. S. Ullman, "Visual routines," Cognition **18**, 97–159 (1984).
71. N. K. Logothetis, "Object vision and visual awareness," Curr. Opin. Neurobiol. **8**, 536–544 (1998).
72. S. Murray, D. Kersten, B. Olshausen, P. Schrater, and D. Woods, "Shape perception reduces activity in human primary visual cortex," Proc. Natl. Acad. Sci. USA **99**, 15164–19169 (2002).
73. 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," Cereb. Cortex **11**, 648–665 (2001).
74. C. Koch and T. Poggio, "Predicting the visual world: silence is golden," Nat. Neurosci. **2**, 9–10 (1999).

PROOF COPY [A-8700] 005307JOA

PROOF COPY [A-8700] 005307JOA