# Neural Basis of Attentive Perceptual Organization

## Tai Sing Lee

*Carnegie Mellon University*

LHP: Lee

RHP: Attentive perceptual organization

Perceptual organization, broadly defined, is a set of visual processes that parses retinal images into their constituent components, organizing them into coherent, condensed, and simplified forms so that they can be readily interpreted and recognized. It generally includes many computational processes before object recognition, such as filtering, edge detection, grouping, segmentation, and figure-ground segregation. These processes are considered to be preattentive, parallel, and automatic (Treisman & Gelade, 1980), mediated by feedforward and intra-areal mechanisms (Palmer, 1999). Attention, on the other hand, is thought to be driven by figure-ground organizatiion rather than the other way around, even though some psychological evidence does suggest that later processes such as recognition and experience could influence earlier perceptual organization (Palmer, Neff, & Beck, 1996; Peterson & Gibson, 1991). The nature and the extent of top-down influence on perceptual organization thus remains murky and controversial. In this article, I will first sketch a theoretical framework to reason about the computational and neural processes underlying perceptual organization. This framework attempts to unify the bottom-up organizational processes and the top-down attentional processes into an integrated inference system. I will then discuss some neurophysiological experimental findings that lend strong support to these ideas.

THEORIES OF FEEDBACK

Marr's (1982) proposal that visual processing could be decomposed into a feedforward chain of relatively independent modules has had a strong influence on the vision community over the last twenty years. Neurophysiologists have focused on the detailed elucidation of single cells' properties and tuning in each cortical area, while computer scientists have attempted to formulate each computational module mathematically in isolation. Some psycholog-

2

ical evidence seem to suggest feedforward computations may be sufficient in normal scene analysis. Thorpe, Fize, and Marlot (1996) demonstrated that people and monkeys could perform categorization tasks very rapidly and that event-related potentials (ERP) relevant to decision making can be observed in the prefrontal areas within 150 ms, apparently leaving little time for computations to iterate up and down the visual hierarchy. Much of visual processing, they argued, must be based on essentially feedforward mechanisms. In fact, many successful object detection and recognition algorithms (Lades et al., 1993; Rowley, Beluja, & Kanade, 1998; Schneiderman & Kanade, 1998) are based only on feedforward algorithms. These algorithms typically use clustering or likelihood tests to classify patterns based on the configuration of responses of low-level features detectors, effectively bypassing the difficult perceptual organization problems. Hence, the dominant conceptual framework today on perceptual processing is still based on feedforward computations along a chain of computational modules (Palmer, 1999).

In recent years, it has become increasingly clear to computer scientists that many problems in perceptual organization are difficult to solve without introducing the contextual information of a visual scene (see Lee, Mumford, Romero, & Lamme, 1998). Psychologists and neural modelers have in fact long emphasized the importance of contextual feedback in perceptual processing (Dayan, Hinton, Neal, & Zemel, 1995; Grossberg, 1987; McClelland & Rumelhart, 1981; Mumford, 1992; Rao & Ballard, 1999; Ullman, 1994). Their arguments were inspired partly by psychological findings, and partly by theoretical considerations and the knowledge that there is an enormous amount of recurrent anatomical connections among the cortical areas (Felleman & Van Essen, 1991).

The exact nature of information being fed back to the earlier areas, however, is far being clear. There are three main proposals. The first suggests that feedback carries explicit hypotheses or predictions similar to model-based image rendering in computer graphics (Mumford, 1992; Mumford, 1996a). The higher order hypothesis could feed back to to suppress (or explain away) the earlier level descriptions that it can explain, as suggested in the predictive coding framework (Mumford, 1996a; Rao & Ballard, 1999). Alternatively, it could feed back to enhance (resonate with) the earlier representation that is consistent with it, facilitating perceptual completion, as suggested in the adaptive resonance/interactive activation framework (Grossberg, 1987; McClelland & Rumelhart, 1981).

In the second proposal, the information being fed back is more general and may be best understood in terms of top-down probabilistic priors in an inference framework (Grenander, 1976; Grenander, 1978; Grenander, 1981; Dayan, Hinton, Neal, & Zemel, 1995; Lee & Mumford, 2002; Tu & Zhu, 2002). Each area is endowed with its unique computational machinery and carries out its own special computation. The priors could be specific in the object domain but unspecific in spatial domain, or vice versa. They provide general guidance to influence, rather than to micro-manage, the lower level inference.

The third proposal has recently become popular in the neuroscience community. It is primarily a mechanism for implementing selective attention (for review, see Desimone & Duncan, 1995) and gain control (Prezybyszewski, Gaska, Foote, & Pollen, 2000). The mechanistic framework for attentional selection favored by neural modelers is called biased competition. Feedback in this framework serves to provide a positive bias to influence the competition at the earlier levels (Deco & Lee, 2002; Reynold, Chelazzi, & Desimone,

1999; Usher & Niebur, 1996).

Despite some superficial contradictions, these three proposals are in fact quite similar at a certain level. They reflect the concerns of three different communities: psycho/neural modeling, statistical/AI, and biological. Here, I attempt use a probabilistic inference framework rooted in the second proposal to reconcile and unify all these perspectives.

BAYESIAN INFERENCE IN THE VISUAL HIERARCHY

Visual processing may be conceptualized as what Helmholtz (Helmholtz, 1867; Palmer, 1999) called the unconscious inference, or, what is recently referred to as Bayesian inference (Knill & Richards, 1996). That is, we rely on contextual information and our prior knowledge of the world to make inferences about the world based on retinal data. Consider the image patch depicted in Fig. 13.1a. Seen alone, it is merely a collection of spots and dots. However, when placed in a larger scene context (Fig. 13.1b), the same image patch assumes a more specific and richer meaning. The image in Fig. 13.1b is still quite ambiguous. It will take unfamiliar viewers a few minutes before they perceive the object in the scene. However, once they are told that the picture depicts a Dalmatian dog sniffing the ground near a tree, the perception will start to crystallize in their minds. The spots and dots are transformed into the surface markings of the dog's body. Furthermore, if the same viewers were to see this image again in the future, their memory would help them see the dog instantly. This terrific example by R.C. James illustrates the important roles of both the global context and prior knowledge in perceptual inference.

From the Bayesian perspective, perceptual inference can be formulated as the computation to obtain the most probable causes of the visual scene by

finding the *a posteriori* estimate $S$ of the scene that maximizes $P(S|I, K)$, the conditional probability of a scene $S$ given a particular image $I$, and our knowledge of the world $K$. By Bayes' theorem, this is given by,

$$P(S|I, K) = \frac{P(I|S, K)P(S|K)}{P(I|K)}$$

where $P(I|S, K)$ is the conditional probability of the image given the scene hypothesis $S$ and the prior knowledge $K$. $S$ has a hierarchical description (e.g. edges, eyes, face, person), i.e. it is in fact a collection of hypotheses at different levels $S_i$ with $i$ indicating the level in the hierarchy. At a particular level $i$, one can think of prior knowledge to be captured by the possible hypotheses at the other levels, i.e. $P(S_{i-1}, P(S_{i+1}), P(S_{i+2})$, etc. If we assume that a cortical area talks primarily to an adjacent area, but not to more distant areas, then the hierarchy can be considered to be roughly Markovian, and the probability distribution of hypotheses at level $i$ can be factorized as

$$P(S_i|I, K) = P(S_{i-1}|S_i)P(S_i|S_{i+1})/Z,$$

where $Z$ is a normalization constant.

Let $I$ be the information output by the retina, then

$$P(S_{lgn}|I, K) = P(I|S_{lgn})P(S_{lgn}|S_{v1})/Z_1,$$

$$P(S_{v1}|I, K) = P(S_{lgn}|S_{v1})P(S_{v1}|S_{v2})/Z_2,$$

$$P(S_{v2}|I, K) = P(S_{v1}|S_{v2})P(S_{v2}|S_{v4})/Z_3,$$

etc., where $Z's$ are normalization constants for each of the distributions, and $S_{i=lgn, v1, v2, v4, it}$ describes the hypotheses generated at the respective area along the visual hierarchy.

6

For example, V1 receives input from the LGN and generates a set of hypotheses that might explain the LGN data. The generation is constrained by feedforward and intracortical connections specified by $P(S_{lgn}|S_{v1})$, i.e. how well each V1 hypothesis $S_{v1}$ can explain $S_{lgn}$. $S_{v2}$ are the hypotheses generated by V2 based on its input from V1 and feedback from higher areas. V2 communicates directly to V1, but not LGN. The feedback from V2 to V1 is given by the estimate that maximizes $S_{v2|I,K}$ weighted by feedback connections $P(S_{v1}|S_{v2})$, i.e. how well $S_{v2}$ can explain away $S_{v1}$. V1 is to find the $S_{v1}$ (at its level of interpretation) that maximizes $P(S_{v1}|I,K) = P(S_{lgn}|S_{v1})P(S_{v1}|S_{v2})/Z$.

This scheme can then be applied again to V2, V4, and IT recursively to generate the whole visual hierarchy. Perception corresponds to each of the cortical areas finding its best hypothesis $S_i$, constrained by the bottom-up and the top-down information. Each cortical area is an expert at inferring some aspects of the visual scene. Unless the image is simple and clear, each area normally cannot be completely sure of its conclusion and has to harbor a number of candidate proposals simultaneously, waiting for the feedback guidance and possibly a change in the input interpretation to select the best hypothesis. The feedforward input drives the generation of the hypotheses, and the feedback from higher inference areas provides the priors to help select the most probable hypothesis. Information does not need to flow forward and backward from V1 to IT in big loops, which would take too much time per iteration. Rather, successive cortical areas in the visual hierarchy can constrain each other's inference in small loops instantaneously and continuously in a Markov chain. The system, as a whole, could converge to an interpretation of the visual scene rapidly and almost simultaneously.

EFFICIENT CODING IN THE HIERARCHY

Hierarchical Bayesian inference can tie together rather nicely the three plausible proposals on the nature of feedback and its role on perceptual organization. In fact, it helps to reconcile some apparent contradictory predictions from the pattern theory (Grenander, 1978; Mumford, 1996a; Rao & Ballard, 1999) and the resonance theory (Grossberg, 1987; McClelland & Rumelhart, 1981). In the hierarchical Bayes framework, many levels of descriptions (organizations) can coexist in the visual hierarchy, with the highest level of explanations feasible most salient to visual awareness, or the cognitive/decision processes. The high level description feeds back to attenuate the saliency of the lower level descriptions, but should not annihilate them. This is an important, but subtle distinction between this theory with Mumford's earlier interpretation of the pattern theory (Mumford, 1992; Mumford, 1996a; Rao & Ballard, 1999). Most importantly, this top-down hypothesis also serves to eliminate the alternative hypotheses in the earlier level, suppressing more severely the responses of the neural ensembles that are representing the alternative hypotheses. Thus, the early-level hypothesis consistent with the higher level description is actually enhanced relative to the alternative hypotheses, as predicted by the resonance theory. In this way, this hierarchical Bayes engine contains both the explaining away element as well as the resonance element.

Let us use the Necker cube in Fig. 13.2a as an example. This line drawing can be immediately perceived as a cube, rather than a bunch of lines and junctions. The black dots, the most elementary level description, are organized into lines. The lines, their positions, and junctions are then organized into a 3D cube interpretation at the higher level. The 3D percept is the simplest description that explains all the evidences and is perhaps what first penetrates into our consciousness. The 3D cube interpretation is then fed back to early visual areas to attenuate the saliency of the representations

of line and edge elements, because they have been explained. For this simple picture, all one can observe is the explaining away, i.e., attenuation of early visual neurons' responses. Fig. 13.2b shows the same picture that is corrupted with spurious noises. In this case, the theory predicts that the higher order hypothesis of a cube will suppress all the noise elements more severely than the edge elements that are consistent with the cube hypothesis, enhancing (resonating) the consistent earlier representation in a relative sense. This scheme is in fact a generalized form of efficient and sparse coding (Barlow, 1961; Field, 1994; Lewicki & Olshausen, 1999; Olshausen & Field, 1995; Rao & Ballard, 1999 ). On the other hand, the consequence of resonance is that the higher level hypothesis can help to complete missing information at the earlier levels. One can imagine the V1 neurons at the location of the gap (location A) in Fig. 13.2c will be activated in order appropriately to be consistent with the 3D cube interpretation higher up! These theoretical predictions are precisely what we observed in the following two neurophyisological experiments in V1.

EVIDENCE IN THE EARLY VISUAL CORTEX

In the first experiment (Lee, Mumford, Romero, & Lamme, 1998), we studied how V1 neurons responded to different parts of a texture stimulus (Fig. 13.3a). While the monkey was fixating a red dot on the computer monitor, an image was flashed on the monitor and lasted for 330 ms. The image in this case was a texture strip subtending $4°$ of visual angle on a background of contrasting texture. The texture in each region was composed of small randomly positioned lines of uniform orientation. Texture contrast was defined as the difference in the orientation of the line elements. The stimulus was presented at a randomized series of sampling positions relative to the V1 cells' classical receptive fields so that the temporal response of the

neurons to different parts of the stimulus (0.5° steps over a 12° range) was measured one at a time. Fig. 13.3b shows the spatiotemporal response of a set of vertically oriented neurons to the stimulus in Fig. 13.3a. Several interesting observations can be made. First, the initial neuronal response (35-70 ms post-stimulus onset) was characterized by the response to local features, i.e. sensitivity to orientation of the line elements. Second, after the initial burst of response, there was a transient pause, followed by a more sustained response at a lower level. This phenomenon usually is considered an effect of intracortical inhibition, adaptation, or habituation. From a Bayesian framework, this decay, in response in the later period of V1 neurons' activity, could be considered as a part of the explaining away by the higher order description. Third, the response at the texture boundary was significantly higher than the responses within the texture regions. This relative enhancement could be considered as a consquence of the cells' resonance with the global percept of surface discontinuity. Fourth, orientation sensitivity of the cells was maintained at the later response, but the response was sustained at a very low-level, suggesting a reduction in saliency but not in coding specificity. Thus, the lower level representations did not completely disappear, but instead were maintained at a lower level. This is important because these activities might help to keep all irrelevant data for alternative hypotheses alive so that they might, at another moment, resurrect and support an alternative hypothesis (e.g., switching between the two percepts in the Necker cube). These observations are what one would expect from hierarchical Bayesian inference engine, though they can also potentially be explained by passive intracortical inhibition mechanisms (Stemmler, Usher, & Neibur, 1995; Li, 2001).

In the second experiment (Lee & Nguyen, 2001), we used a similar paradigm to examine how V1 neurons responded to the subjective contour of a sub-

jective Kaniza square (Fig. 13.4a and Fig. 13.4b). Over successive trials, the illusory contour was placed at different locations relative to the center of the receptive field, 0.25° apart, spanning a range of 2.25°, as shown in Fig 13.4b. Figures 13.4c-13.4j display examples of other control stimuli also tested in the experiment. In each trial, while the monkey fixated, a sequence of four stimuli was presented. The presentation of each stimulus in the sequence lasted for 400 ms. In the presentation of the Kanizsa figure, four circular discs were first presented, and then they were abruptly turned into four partial discs, creating the illusion of a subjective square appearing in front of the four circular discs. Fig. 13.5a shows that the illusory contour response of a multiple-unit occurred at precisely the same location of a real contour response, indicating the spatial precision of the response. Fig. 13.5b and Fig. 13.5c compare the temporal evolution of this unit to the illusory contour and the responses to a variety of real contours and controls. This unit responded significantly more to the illusory contour than to the amodal contour or to any of the rotated disc configurations (controls) (see Lee & Nguyen, 2001). For this neuron, as well as for the V1 populations in the superficial layer, the temporal onset of the response to illusory contour was at about 100 ms after the abrupt onset of the illusory square, while the onset of the responses to the real contours occurred at about 40-50 ms (Fig. 13.5c and Fig. 13.5d). The response of V2 neurons to the same illusory contour was much earlier, at about 65 ms (Fig. 13.5e).

These observations suggest that V2 could be detecting the existence of an illusory contour first by integrating information from a more global spatial context, forming a hypothesis of the boundary of the Kanizsa square. V2 neurons, because of their larger receptive fields, could not provide a spatially precise representation of the sharp illusory contour. They can only inform the existence of a boundary at a rough location. V1 is recruited to compute

11

and represent the precise location and orientation of the contour because it has the machinery for computing and representing precise curvilinear geometry efficiently or 'sparsely'. The hypothesis of a Kanizsa square and its supporting illusory boundary are represented simultaneously by many visual areas, such as V1, V2, and even IT where the concept of a square is represented. From this perspective, V1 does not simply perform filtering and edge detection and then forward the results to the extrastriate cortex for further processing (Hubel & Wiesel 1962). Rather, it is an integral part of the visual system that continues to participate in all levels of visual reasoning insofar as the computations that require spatial precision and high resolution details provided by V1. This is the basic rationale underlying the high-resolution buffer hypothesis that Mumford and I (Mumford, 1996b; Lee, Mumford, Romero, & Lamme, 1998) proposed a few years ago – a view now shared by many others (e.g. Bullier, 2001).

## ATTENTION AS TOP-DOWN PRIORS

The hierarchical Bayesian framework discussed above is appropriate for conceptualizing perceptual organization or interpretation of the input image. The feedback carries the contextual priors generated by the higher level description, directly related to what biologists call contextual processing (see Albright & Stoner, 2001 for review). Attention is another type of feedback that places priority or value in the information to be analyzed and makes perception purposeful. In fact, the dominant view in the biological community on the functional role of feedback is the mediation of selective attention (for reviews, see Desimone & Duncan, 1995; Itti & Koch, 2001). Since both attention and contextual priors utilize the same recurrent feedback pathways, it might be reasonable to consider attention in terms of priors and unify the two functions in a single framework.

People usually think of attention in terms of spatial attention, a spotlight that 'illuminates' a certain location of visual space for focal visual analysis (Helmholtz, 1867; Treisman & Gelade, 1980). Attentive processing is usually considered a serial process that requires moving the spotlight around in the visual scene to select the location to be analyzed. There are in fact many types of attention. Feature or object attention is involved when we are searching for a particular feature or object in a visual scene (James, 1890). In spatial attention, selection is focused on the spatial dimension and dispersed (parallel) in the feature dimension; while in feature attention, the selection is focused on the feature dimension and dispersed (parallel) in the spatial dimension. A generalization of feature attention is object attention, in which a configuration of features belonging to an object is searched. It was believed that conjunctive search operates in a serial mode (Treisman & Sato, 1990; Wolfe, 1998).

In recent years, a number of neurophysiological studies have shown that attention can modulate visual processing in many cortical areas (Desimone & Duncan, 1995) and even the receptive fields of neurons (Connor et al., 1997; Tolias et al., 2001). The popular model for explaining the mechanism of attention is called biased competition (Deco & Lee, 2002; Desimone & Duncan, 1995; Duncan & Humphreys, 1989; Reynolds et al., 1999; Usher & Neibur, 1996). The basic idea is that when multiple stimuli are presented in a visual field, the different neuronal populations within a single cortical area activated by these stimuli will engage in competitive interaction. Attending to a stimulus at a particular spatial location or attending to a particular object feature, however, introduces a bias to influence the competition in favor of the neurons at the attended location and at the expense of the other neurons. The biased competition mechanism, formulated in terms of differential equations with roots in the connectionist models, has also been

used in several models for explaining attentional effects in neural responses observed in the inferotemporal cortex (Usher & Niebur, 1996) and in V2 and V4 (Reynolds, Chelazzi, & Desimone, 1999).

Conceptually, biased competition can also be formulated into a probabilistic framework as follows. Recall that the hierarchical Bayesian inference in the visual system can be described as the process for finding the scene variables $S_i$ $\forall i$ that maximizes the joint probability,

$$
\begin{aligned}
P(I, S_{lgn}, S_{v1}, ..., S_{it}) &= P(I|S_{lgn})P(S_{lgn}|S_{v1})P(S_{v1}|S_{v2}) \\
&\quad \cdot P(S_{v2}|S_{v4})P(S_{v4}|S_{it})P(S_{it}),
\end{aligned}
$$

where $P(S_{it})$ is the prior on the expected frequency of the occurrence of various object categories.

Top-down object attention can be incorporated in this framework by including the prefrontal areas (area 46) in the hierarchy as follows,

$$
\begin{aligned}
P(I, S_{lgn}, ..., S_{a46v}) &= P(I|S_{lgn})P(S_{lgn}|S_{v1})P(S_{v1}|S_{v2}) \\
&\quad \cdot P(S_{v2}|S_{v4})P(S_{v4}|S_{it})P(S_{it}|S_{a46v})P(S_{a46v}),
\end{aligned}
$$

where ventral area 46 (a46v) is the area for executive control that will determine what object to look for and what object to remember. It integrates a large variety of contextual information and memory from the hippocampus, basal ganglians, cingulate gyrus, and many other prefrontal areas to make decisions and resolve conflicts (Miller & Cohen, 2001). It sets priority and endows value to make visual object processing purposeful and adaptive.

14

Because the hierarchy is reciprocally connected, this implies that attention, higher order contextual knowledge, and behavioral experience should be able to penetrate back to the earliest level of visual processing, at least as early as V1 and LGN. This was precisely what we observed in the following experiments.

BEHAVIORAL EXPERIENCE AND TASK DEMANDS

In this experiment (Lee, Yang, Romero, & Mumford, 2002), my colleagues and I studied the effect of higher order perceptual construct such as 3D shape and behavioral experience on the neural processes in the early visual cortex (V1 and V2). We used a set of stimuli that allowed the dissociation of bottom-up low-level stimulus contrast from top-down higher order perceptual inference (Fig. 13.6). The stimuli included a set of shape from shading stimuli, which have been found to pop out readily (Braun, 1993; Ramachandran, 1988; Sun & Perona, 1996) and a set of two-dimensional contrast patterns, which do not pop out spontaneously, even though the latter have stronger luminance contrast and evoke stronger bottom-up raw responses in V1 neurons (see Fig. 13.6). The stronger pop-out of shape from shading stimuli in this case has been attributed to their 3D interpretation. Therefore, if we see a neural correlate of this stronger pop-out modulation due to shape from shading in V1, it would be a clear case of top-down modulation due to higher order percepts.

To evaluate the impact of behavior on early visual processing, we also divided the experiment into two stages, a pre-behavior stage and a post-behavior stage. In both stages, the monkeys performed the same fixation task, i.e., fixating a red dot on the screen during stimulus presentation. In the pre-behavior stage, the monkeys had not used the stimuli in their behavior. In

the post-behavior stage, the monkeys had utilized the stimuli in their behaviors for a period of time. Specifically, they had been trained to detect the oddball of the various types and make a saccadic eye movement to it. Interestingly, V1 neurons were significantly sensitive to perceptual pop-out modulation in the post-behavior stage, but not in the pre-behavior stage. Pop-out modulation was defined by the enhancement of the neuronal responses to the oddball condition relative to the uniform condition, while the stimulus on the receptive field of the neurons was kept constant (Fig. 13.6; Fig. 13.7). Furthermore, the pop-out modulation in V1, and similarly in V2, was a function of the stimuli, directly correlated with the subjective perceptual pop-out saliency we perceive in the stimulus. Fig. 13.7 shows that the lighting from above [LA] and lighting from below [LB] oddballs pop out strongly, the lighting from left [LL] and right [LR] oddballs pop out moderately, and the 2D contrast oddballs do not pop out at all (Ramachandran, 1988). It also shows a strong correlation between the behavioral performance (reaction time and percentage correct) of the monkeys and the neural pop-out modulation in V1. Thus, the neural modulation we observed in V1 could be considered a neural correlate of perceptual saliency.

Apparently, the pop-out detection task forced the monkeys to see the stimulus more clearly and to precisely localize the pop-out target in space. This is a task that would engage V1's machinery according to the high-resolution buffer hypothesis. Even though the monkeys were required only to fixate during the experiment, having practised the pop-out detection task for two weeks apparently had made the early pop-out processing more or less automatic. On the other hand, the pop-out effect could be greatly attenuated when the monkeys were asked to perform a very attention demanding conflicting task. The pop-out signals emerged in V1 and V2 at roughly the same time (95 ms for V2 and 100 ms for V1). Interestingly, Bichot and

Schall (1999) also found that the target selection/decision signal emerged in the frontal eye field during a visual search task at about the same time frame, around 100-130 ms, supporting the intuition that interactive computation may not take place in a step-wise linear fashion iteratively, but may occur interactively and concurrently between adjacent areas in the brain. The cycle time is much shorter under continuous dynamics. From this point of view, the 150 ms time frame reported by Thorpe, Fize, and Marlot (1996) is quite sufficient for the whole hierarchy to settle down to a perceptual interpretation of the visual scene.

Perceptual pop-out has been thought to be an operation that is parallel, automatic, and preattentive. The findings discussed suggest that attention may be involved in the normal operation for early perceptual organization such as pop-out and grouping. The idea that attention may play a role in this parallel computation might seem to be at odds with conventional notions. However, recent psychological studies (Joseph, Chun, & Nakayama, 1997) suggested that attention may be critical for the detection of preattentive features and may, in fact, be necessary for overt perception of these stimulus features. These data suggest a stronger link between perceptual organization and attention, as well as behavioral experience.

OBJECT-BASED SPATIAL ATTENTION

Granted the signal we observed in the last experiment was related to perceptual saliency, what is the advantage of having the attentional highlighting signals going all the way to V1? Two observations help to reveal the important role of V1 in this computation. First, in the shape from shading pop-out experiment, we found that the pop-out enhancement could be found only at exactly the location of the oddball, but not at the locations right next to it,

indicating that the highlighting effect is both spatially specific and stimulus specific. Second, when we examined the responses of V1 neurons to different parts of a texture square in a contrasting background (Fig. 13.8), my colleagues and I (Lee, Mumford, Romero, & Lamme, 1998) found that V1 neurons' responses were enhanced when their receptive fields were placed inside a texture-defined figure relative to when they were placed in a textured background and that this enhancement was uniform within the spatial extent of the figure, just as Lamme (1995) discovered earlier. Further, we also found that this highlighting signal is spatially bounded by the boundary response of the neurons (Fig. 13.8). Cortical organization beyond V1 is segregated according to more abstract attributes and is less topologically precise. Only in V1 could one find a spatially precise grid-like spatial topology, in Ullman's (1984) terms, to color the surface of an object clearly and precisely. This highlighting or coloring operation through attention might be the neural basis of object-based spatial attention (Behrmann, Zemel, & Mozer, 1998; Olson, 2001).

INTEGRATION OF OBJECT AND SPATIAL INFORMATION

The computation of perceptual pop-out saliency of the shape from shading stimuli is likely a product of three types of computation, bottom-up saliency, shape recognition (stimulus-evoked object attention), and spatial localization (stimulus-evoked spatial attention). It requires the interaction of the early visual areas with both the dorsal stream (e.g. LIP) and the ventral stream (e.g. IT, see Logothetis, 1998). The top-down object attention and spatial attention feedback from both streams, coupled with intracortical contextual computation, produce the spatially precise higher order perceptual saliency effect.

18

So far I only talked about the hierarchical inference of object forms, but the inference of space could also be formulated in the same way, only with a change of variable from $S$ to $S'$ to indicate the spatial aspects of the information and the assumption that spatial attention is initiated by an input from dorsal area 46.

$$
\begin{aligned}
P(I, S_{lgn}, ..., S'_{po}, ..., S'_{a46d}) \;=\; & P(I|S_{lgn})P(S_{lgn}|S_{v1})P(S_{v1}|S'_{v2}) \\
& \cdot P(S'_{v2}|S'_{v3})P(S'_{v3}|S'_{po}) \\
& \cdot P(S'_{po}|S'_{a46d})P(S'_{a46d}),
\end{aligned}
$$

The scene variables $S'_i$ in this case concern the spatial position encoding and spatial coordinate transforms. For simplicity, let us assume that the cross-talk between the higher areas in the different streams is relatively weak, then the activity of V1 is given by $S_{v1}$ that maximizes,

$$
P(S_{v1}|S_{lgn}, S_{v2}, S'_{v2}, ...) = P(S_{lgn}|S_{v1})P(S_{v1}|S_{v2})P(S_{v1}|S'_{v2})/Z
$$

In the cortex, the what and where pathways are segregated into the ventral and the dorsal streams respectively (Ungerleider & Mishkin, 1982). Their recurrent interaction in V1 therefore can integrate the spatial and object information. More generally, different aspects of information from each hypercolumn in V1 are channelled to visual modules or modular streams for further processing, and the feedback from these extrastriate modules to V1 carries the invariant information they inferred. V1, as the high-resolution buffer, is where all the higher order information can come back together to the same spatial locus to re-integrate all the 'broken' features into a unified percept.

Deco and I (Deco & Lee, 2002) developed a neural dynamical model, which could be considered as a deterministic approximation (Abbott, 1992; Amit & Tsodyks, 1991; Wilson & Cowen, 1997) of the statistical inference framework, to explore the possibility that V1 can serve as a buffer to coordinate the interaction between the dorsal stream and the ventral stream and to achieve feature integration in conjunctive visual search (Deco & Lee, 2002). For simplicity, the model contains only three modules. The V1 module is directly and reciprocally connected to the IT module, which encodes object classes, and to the PO module, which encodes spatial location. The prefrontal areas can exert a top-down bias to a particular neuronal pool in IT to initiate object attention (what to look for) or to a particular neuronal pool in PO to initiate spatial attention (where to look at). V1 is modeled with a two-dimensional grid of hypercolumns, each with 24 pools of complex cells (8 orientations and 3 scales). PO is modeled by a two-dimensional grid of nodes. Each node (neuronal pool) indicates a particular spatial location and is connected to a small spatially contiguous subset of V1 hypercolumns in a reciprocal manner. Each IT neuron represents a particular object and is connected reciprocally to every V1 neuron. The pattern of connection is symmetrical and is learned by Hebbian learning, but the feedback weights are weaker than the feeforward weights (set to 60 percent) (see Deco & Lee, 2002 for details). Within each module, there are inhibitory neurons to mediate competition.

When a single object is presented to the retina, a local region of V1 neurons will be activiated, which will activate a number of IT neurons and a number of PO neurons. Competition within IT and within PO will rapidly narrow down a winner cell in IT (corresponding to recognizing the identity of the presented object) and a winner cell in PO (corresponding to localizing the object in space). The coactivation of the specific pools of neurons in the

20

three modules corresponds to the unified percept of identity, location, and features of the presented object.

Visual search (object attention) is initiated by introducing a top-down positive bias to an IT neuron, presumably from the prefrontal cortex (Rao, Rainer, & Miller, 1997). The IT neuron will project a top-down template of subthreshold activation to V1. Any subset of V1 hypercolumns whose response patterns match that of the top-down template will be selectively enhanced, as in resonance. These enhanced neurons exert a suppressive effect on other V1 neurons and provide a stronger bottom-up bias to the PO neuron corresponding to that location. Initially a number of PO neurons scattered in space will be activated by bottom-up V1 input. Over time, the lateral inhibition in PO will narrow the activation to a very localized set of neurons in PO, indicating the localization of the searched object. Interestingly, this model can produce the effect of both the parallel search and the serial search using one single mechanism. When the system is instructed to search for an E in a field of X's, the time for the PO to converge to a single location is independent of the number of X's, corresponding to parallel search. When the system is instructed to search for an E in a field of F's, the search time increases linearly with the number of distractors F. This is because when the target and the distractor are similar, the responses in the V1 hypercolumns (at least in feature level) to each are very similar, causing a confusion that requires constant interaction between V1 and IT to gradually resolve. Surprisingly, for reasons we do not completely understand, the time required to search for an ambiguous conjunctive target is linearly proportional to the number of distractors, as in serial search. Apparently, the phenomena of serial and parallel search is a stimulus-dependent effect that emerges from the same parallel mechanism (see Deco & Lee, 2002 for details).

21

Spatial attention can be introduced by providing a bias (spatial prior) to a particular neuronal pool in PO. The interaction between PO and V1 acts in very much the same way except that the top-down bias from PO is spatial rather than featural. PO extracts only the spatial information from V1 to focus the competition on the spatial domain. This facilitates the convergence, or localization, of the winner in a visual search task. In our model, spatial attention serves a very useful purpose: the gating of information from V1 to IT is accomplished simply by PO's modulation of V1 activities rather than by modulating the feedforward connection weights as in other models (Olshausen, Andersen, & Van Essen, 1993; Reynolds, Chelazzi, & Desimone, 1998). The system can serially channel visual information from different V1 hypercolumns to IT for object recognition simply by allocating the top-down spatial prior to different neuronal pools in PO. Another interesting observation is that a relatively small top-down modulation in V1 from PO is sufficient to relay a bias to IT to produce a winner in IT. This suggests that even though the top-down effect in the early visual areas is small, it could still be effective in coordinating the communication and information integration among multiple visual streams. Simple as it is, the model is sufficient to illustrate how the early visual cortex can coordinate and organize parallel and distributed computations in the visual system from a dynamical system perspective.

CONCLUSION

In this article, I propose that hierarchical probabilistic Bayesian inference, when coupled with the concept of efficient coding, provides a reasonable framework for conceptualizing the principles of neural computations underlying perceptual organization. I have described a number of recent experimental findings from my laboratory providing evidence in support of this

framework. Evidence from other laboratories (Albright & Stoner, 2002; Crist & Gilbert, 2001; Grosof, Shapley, & Hawken, 1993; Haenny & Schiller, 1988; Hupe et al., 1998; Lamme, 1995; Itto & Gilbert, 1999; Motter, 1993; Murray et al., 2002; Ress, Backus, & Heeger, 2000; Roelfsema, Lamme, & Spekreijse, 1998; von der Heydt, Peterhans, & Baumgarthner, 1984; Zhou, Friedman, & von der Heydt, 2000; Zipser, Lamme, & Schiller, 1996) on the top-down contextual influences in the early visual areas also support the basic premise of this theory.

The hierarchical Bayes framework proposed here can reconcile some apparent contradictions between the predictive coding theory (Mumford, 1996a; Rao & Ballard, 1999) and adaptive resonance theory (Grossberg, 1987) in that it contains both the 'explaining away' as well as the 'resonance' components. Here, the top-down feedback of a higher level hypothesis does attenuate the saliency of the earlier representations that support it on the one hand, but also suppresses even more severely the alternative evidence and hypotheses. This framework also unifies top-down attention and bottom-up perceptual inference into a single hierarchical system. Attention can be considered as a variety of top-down priors (spatial, object, feature) for influencing the perceptual inference at the earlier levels.

Perceptual organization, such as grouping, segmentation, and figure-ground segregation, however, involves far more sophisticated computations than competitive interactions (e.g., August & Zucker, 2000; Blake & Zisserman, 1987; Grossberg, Mingolla, & Ross, 1994; Konishi, Yuille, Coughlan, & Zhu, 2002; Lee, 1995; Shi & Malik, 2000; Tu & Zhu, 2002; Weiss, Simoncelli, & Adelson, 2002; Williams & Jacobs, 1997; Yu, Lee, Kanade, 2002; Yu, 2003; Yuille & Bulthoff, 1996; Zhu, Lee, & Yuille, 1995). The hierarchy idea proposed here is deeply connected with the idea of compositional hierarchy

23

from Bienenstock, Geman, and Potter (1997). With the addition of recurrent feedback of contextual and attentional priors, the proposed hierarchical framework provides a broader view on the nature of cortical computations of perceptual organization. Elucidating the varieties of Bayesian priors as a function of task demands and environmental statistics is important for understanding the computational and neural basis of attentive perceptual organization.

REFERENCE

Abbott, L. (1992). Firing rate models for neural populations. In O. Benhar, C. Bosio, P. Giudice, & E. Tabet (Ed.), *Neural networks: From biology to high energy physics*. Pisa: ETS Editrice.

Albright, T. D., & Stoner, G. R. (2002). Contextual influences on visual processing. *Annual Review of Neuroscience, 25,* 339–379.

Amit, D., & Tsodyks, M. (1991). Quantitative study of attractor neural network retrieving at low spike rates: I. Substrate spikes, rates and neuronal gain. *Network, 2,* 259–273.

August, J., & Zucker, S. W. (2000). The curve indicator random field: curve organization via edge correlation. In K. Boyer & S. Sarka (Ed.), *Perceptual organization for artificial vision systems* (pp. 265–288). Boston, MA: Kluwer Academic.

Barlow, H. B. (1961). Coding of sensory messages. In W.H. Thorpe, & O.L. Zangwill (Ed.), *Current problems in animal behavior* (pp. 331–360). Cambridge, UK: Cambridge University Press.

Bienenstock, E., Geman, S., & Potter, D. (1997). Compositionality, MDL priors, and object recognition. In M.C. Mozer, M.I. Jordan, & T. Petsche (Ed.), *Advances in Neural Information Processing Systems, 9* (pp. 838–844). Cambridge, MA: MIT Press.

Behrmann, M., Zemel, R. S., & Mozer, M. C. (1998). Object-based attention and occlusion: evidence from normal participants and a computational model. *Journal of Experimental Psychology: Human Perception and Performance, 24*(4), 1011–1036.

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

Blake, A., & Zisserman, A. (1987). *Visual reconstruction*. Cambridge, MA: MIT Press.

Braun, J. (1993). Shape from shading is independent of visual attention and may be a "texton". *Spatial Vision, 7*(4), 311–322.

Bullier, J. (2001). Integrated model of visual processing. *Brain Research Review, 36*(2-3), 96–107.

Crist, R. E., Li, W., & Gilbert, C. D., (2001). Learning to see: experience and attention in primary visual cortex. *Nature Neuroscience, 4*(5), 519–525.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation, 7*(5), 889–904.

Deco, G., & Lee, T. S. (2002). A unified model of spatial and object attention based on inter-cortical biased competition. *Neurocomputing, 44-46*, 769–774.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review Neuroscience, 18*, 193–222.

Duncan, J., & Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review, 96*, 433–458.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1,* 1–47.

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computing, 6,* 559–601.

Grenander, U. (1976). *Lectures in pattern theory I: Pattern analysis.* New York: Springer-Verlag.

Grenander, U. (1978). *Lectures in pattern theory II: Pattern synthesis.* New York: Springer-Verlag.

Grenander, U. (1981). *Lectures in pattern theory III: Regular structures.* New York: Springer-Verlag.

Grosof, D. H., Shapley, R. M., & Hawken, M. J. (1993). Macaque V1 neurons can signal 'illusory' contours. *Nature, 365*(6446), 550–552.

Grossberg, S. (1987). Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science, 11*, 23–63.

Grossberg, S., Mingolla, E., & Ross, W. (1994). A neural theory of attentive visual search: interactions at boundary, surface, spatial and object atttention. *Psychological Review, 101*(3), 470-489.

Haenny, P. E., & Schiller, P. H. (1988). State dependent activity in monkey visual cortex. I. Single cell activity in V1 and V4 on visual tasks. *Experimental Brain Research, 69*(2), 225–244.

Helmholtz, H. V. (1867). *Handbook of physiological optics.* Leipzig: Voss.

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

Hupe, J. M., James, A. C., Payne, B. R., Lomber, S. G., Girard P., & Bullier, J. (1998). Cortical feedback improves discrimination between figure and background

by V1, V2 and V3 neurons. *Nature, 394*(6695), 784–787.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Review Neuroscience, 2*(3), 194-203.

Ito, M., & Gilbert, C. D. (1999). Attention modulates contextual influences in the primary visual cortex of alert monkeys. *Neuron, 22,* 593–604.

James, W. (1890). *The principles of psychology.* New York: Henry Holt.

Joseph, J. S., Chun, M. M., & Nakayama, K. (1997). Attentional requirements in a 'preattentive' feature search task. *Nature, 387*, 805-807.

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

Konishi, S. M., Yuille, A. L., Coughlan J. M., & Zhu, S. C. (In press) Statistical Edge Detection: Learning and Evaluating Edge Cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Lades, M., Vorbruggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, R. P., Konen W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions in Computers, 42*(3), 300–311.

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

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

Lee, T. S. (1996). Image representation using 2D Gabor wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(10), 959–971.

Lee, T. S., Mumford, D., Romero, R., & Lamme, V. A. F. (1998). The role of the primary visual cortex in higher level vision. *Vision Research, 38*(15-16), 2429–2454.

27

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

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

Lee, T. S., & Mumford, D. Hierarchical Bayesian inference in the visual cortex. Manuscript submitted to *Journal of Optical Socity of America, A.*

Lewicki, M. S., & Olshausen, B. A. (1999). Probabilistic framework for the adaptation and comparison of image codes. *Journal of Optical Society of America, A, 16*(7), 1587–1601.

Li, Z. (2001). Computational design and nonlinear dynamics of a recurrent network model of the primary visual cortex. *Neural Computation 13*(8), 1749–1780.

Logothetis, N. K. (1998). Object vision and visual awareness. *Current Opinions in Neurobiology, 8*(4), 536–544.

Marr, D. (1982). *Vision*. NJ: W. H. Freeman & Company.

McClelland, J. L., & Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception. Part I: an account of basic findings. *Psychological Review, 88,* 375–407.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review Neuroscience, 24,* 167–202.

Motter, B. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *Journal of Neurophysiology, 70,* 909–919.

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

Mumford, D. (1996a). Pattery theory: A unifying perspective. In D. C. Knill, & W. Richards (Ed.) *Perception as Bayesian inference* (pp. 25–62). Cambridge, UK: Cambridge University Press.

Mumford, D. (1996b). Commentary on the article by H. Barlow. In D. C. Knill, & W. Richards (Ed.) *Perception as Bayesian inference* (pp. 451–506). Cambridge, UK: Cambridge University Press.

Murray, S. O., Kersten, D., Olshausen, B. A. Schrater, P., & Woods, D. L. (2002). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Science, USA 99*, 15164-15169.

Olshausen, B. A., Andersen, C., & Van Essen, D. (1993). A neural model for visual attention and invariant pattern recognition. *Journal of Neuroscience, 13*(11), 4700-4719.

Olshausen, B. A., & Field, D. J. (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381,* 607-609.

Olson, C. R. (2001). Object-based vision and attention in primates. *Current Opinion in Neurobiology, 11*(2), 171–179.

Palmer, S. (1999). *Vision Science: Photons to Phenomenology.* Cambridge, MA: MIT Press.

Prezybyszewski, A. W., Gaska, J. P., Foote, W., & Pollen, D. A. (2000). Striate cortex increases contrast gain of macaque LGN neurons. *Visual Neuroscience, 17,* 485–494.

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

Rao R., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87.

Rao, S.C., Rainer, G., & Miller, E. (1997). Integration of what and where in the primate prefrontal cortex. *Science, 276*, 821-824.

Ress, D., Backus, B.T., & Heeger, D.J. (2000). Activity in primary visual cortex predicts performance in a visual detection task. *Nature Neuroscience, 3*(9):940-5.

Reynolds, J., Chelazzi, L., & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience, 19,* 1736–1753.

Roelfsema, P. R., Lamme, V. A. F., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature, 395*(6700), 376–81.

Rowley, H., Baluja, S., & Kanade, T. (1998). Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(1), 23–38.

Schneiderman, H., & Kanade, T. (1998). Probabilistic modeling of local appearance and spatial relationships for object recognition. *IEEE Conference on Computer Vision and Pattern Recognition,* 45-51.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(8), 888–905.

Stemmler, M., Usher, M., & Niebur, E. (1995). Lateral interactions in primary visual cortex: A model bridging physiology and psychophysics. *Science, 269,* 1877–1880.

Sun, J., & Perona, P. (1996). Early computation of shape and reflectance in the visual system. *Nature, 379,* 165–168.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381,* 520–522.

Tolias, A. S., Moore, T., Smirnakis, S. M., Tehovnik, E. J., Siapas, A. G., & Schiller, P. H. (2001). Eye movement modulate visual receptive fields of V4 neurons. *Neuron, 29*(3), 757–67.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12,* 97–136.

Treisman, A., & Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 459–478.

Tu, Z. W., & Zhu, S.C. (2002). Image segmentation by Data-Driven Markov Chain Monte Carlo. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 24*(5), 657–673.

Ullman, S. (1984). Visual routines. *Cognition, 18,* 97–159.

Ullman, S. (1994). Sequence seeking and counterstreams: A model for bidirectional information flow in the cortex. In C. Koch & J. Davis (Ed.), *Large-scale theories of the cortex* (pp. 257-270). Cambridge, MA: MIT Press.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle (Ed.), *Analysis of visual behavior* (pp. 549–586). Cambridge, MA: MIT Press.

Usher, M., & Niebur, E. (1996). Modeling the temporal dynamics of IT neurons in visual search: a mechanism for top-down selective attention. *Journal of Cognitive Neuroscience, 8,* 311–327.

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

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal

percepts. *Nature Neuroscience, 5*(6), 598-604.

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

Wilson, H., & Cowan, J. (1972). Excitatory and inhibitory interactions in localised populations of model neurons. *Biological Cybernetics, 12,* 1–24.

Wolfe, J. M. (1998). Visual search: A review. In H. Pashler (Ed.), *Attention* (pp. 13-77). London: University College London Press.

Yu, S., Lee, T. S., & Kanade, T. (2002). A Hierarchial Markov Random Field Model for Figure-ground Segregation. *Lecture Notes in Computer Science 2134,* 118–133.

Yu, S. (2003). *Computational models of perceptual organization.* Ph.D. thesis, Robotics Institute, Carnegie Mellon University.

Yuille, A. L., & Bulthoff, H. H. (1996). Bayesian decision theory and psychophysics. In D. C. Knill & W. Richards (Ed.), *Perception as Bayesian inference* (pp. 123-162). Cambridge, UK: Cambridge University Press.

Zhou, H., Friedman, H. S., & von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *Journal of Neuroscience, 20*(17), 6594–6611.

Zhu, S. C., Lee, T. S., & Yuille, A. (1995). Region competition: unifying snakes, region growing and MDL for image segmentation. *Proceedings of the Fifth International Conference in Computer Vision*, 416-425.

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

Figure captions

Figure 13.1

(A) An image patch of spots and dots. (B) The image patch situated in a particular scene, as originally designed by R.C. James.

Figure 13.2

(A) A Necker cube. (B) A cube with background noises. (C) A cube with missing information.

Figure 13.3

(A) A texture strip with width of $4°$ visual angle. The strip is composed of short vertical line segments, and the background is composed of short horizontal line segments. (B) Spatiotemporal average response of a population of 14 V1 vertical neurons to a texture strip stimulus at different positions along a horizontal sampling line across the strip. The abscissa is the distance in visual angles from the RF center to the center of the strip. The texture boundary is located at -2.0 and 2.0 degree visual angles away from the center. The responses to the texture stimuli were initially uniformly high within the strip and low outside the strip corresponding to the the vertical orientation tuning of the cells. At 60 ms after stimulus onset, boundary signals started to develop at the texture boundaries. In the later stage, the responses in general were lower than the initial responses, but the responses at the boundaries were sharper and stronger relative to the rest of the image (see Lee, Mumford, Romero and Lamme (1998) for details).

Figure 13.4

(A) A figure of illusory contour and the ten different parts (marked by lines) that were placed over the receptive field of a horizontally oriented neuron over successive trials during an experiment. (B) In a typical trial, the stimulus was presented in a sequence, 400 ms for each step. The first step displayed four circular discs, which then turned into four partial discs in the second step. The abrupt onset of the illusory square captures the attention of the monkey and makes the illusory square more salient. The third and fourth steps repeated the first and second steps. In each trial, the response of the cell to one location of the figure was examined. (C) Some examples of other stimuli that were also examined as controls. From "Dynamics of Subjective Contour Formation in the Early Visual Cortex," by T.S Lee, and M. Nguyen, 2001, PNAS 98(4), 1907–1911. Copyright 2001 by PNAS. Adapted with permission.

Figure 13.5

(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 at the 10 different locations relative to the illusory contour. This cell responded to the illusory contour when it was at precisely the same location (x = 0) 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 partial discs men were rotated. (B) Temporal evolution of this cell's response to the illusory contour, the amodal contour and the various rotated corner disc controls at the location where the real contour elicited the maximum response. The response to the illusory contour emerged at about 100 ms after the illusory square appeared. The cell responded slightly to the amodal contour and did not respond to any of the rotated corner discs. (C) The cell's

response to the illusory contour compared to its response to the real contours of a line square, or a white square. The onset of the response to the real contours was at 45 ms, about 55 ms before the illusory contour response. (D) Population averaged temporal response of 50 V1 neurons in the superficial layer to the real and illusory contours. (E) Population averaged temporal response of 39 V2 neurons in the superficial layer to the real and illusory contours. From "Dynamics of Subjective Contour Formation in the Early Visual Cortex," by T.S Lee, and M. Nguyen, 2001, <u>PNAS 98</u>(4), 1907–1911. Copyright 2001 by PNAS. Adapted with permission.

Figure 13.6

(A) A typical stimulus display was composed of 10 x 10 stimulus elements. Each element was 1° visual angle in diameter. The diameter of the classical receptive field (RF) of a typical cell at the eccentricities tested ranged from 0.4° to 0.8° 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) There are six stimulus sets. Each stimulus set had four conditions: singleton, oddball, uniform, and hole. Displayed are the iconic diagrams of all the conditions for the LA set, 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. The singletons measured the neuronal response to direct stimulation of the RF alone; the holes measured the response to direct stimulation of the extra-RF surround

only. From "Neural Activity in Early Visual Cortex Reflects Experience and Higher Order Perceptual Saliency," T.S. Lee, C. Yang, R. Romero, and D. Mumford, 2002, Nature Neuroscience, 5(6), 589–597. Copyright by Nature Neuroscience. Adapted with Permission.

Figure 13.7:

Temporal evolution of the normalized population average response of 45 V1 units from a monkey to the LA set (A) and the WA set (B) at the post-behavior stage. Significant pop-out response was observed in LA (as well as LB, LL, and LR) starting at 100 msec after stimulus onset. No pop-out response was observed for WA (or WB). (C) Mean pop-out modulation ratios of 45 units for all six stimulus sets. Pop-out enhancements were statistically significant for stimuli LA, LB, LL, and LR, but not for WA and WB. The pop-out modulation is computed as $(A-B)/(A+B)$, where A was the response to the oddball condition and B was the response to the uniform condition. (D) Correlation between reaction time and V1 pop-out modulation for the six sets of stimulus. Data from three different stages (hence 18 points) are plotted. A significant negative correlation was observed between reaction time and pop-out modulation ratio. From "Neural Activity in Early Visual Cortex Reflects Experience and Higher Order Perceptual Saliency," T.S. Lee, C. Yang, R. Romero, and D. Mumford, 2002, Nature Neuroscience, 5(6), 589–597. Copyright by Nature Neuroscience. Adapted with Permission.

Figure 13.8

Object-based spatial attention effect. V1 neurons' responses were found to be enhanced inside the figure relative to outside the figure. (A) and (B) show figures defined by texture contrast. (C) illustrates the placement of the receptive field when the neuron's preferred orientation is vertical. There

are two sampling schemes: parallel sampling, when the preferred orientation of the cells is aligned with the orientation of the texture boundary; and orthogonal sampling, when the preferred orientation of the cells is orthogonal to the orientation of the texture boundary. (D) shows the population averaged response of 45 neurons in the parallel sampling scheme with the response to (A) and the response to (B) summed at each spatial location. (E) shows the population averaged response of 16 neurons in the orthogonal sampling scheme. In the parallel sampling scheme, we found a persistent and large texture edge response, which was absent in the orthogonal sampling scheme, suggesting that cells were sensitive to the orientation of the texture boundaries. The response inside the figure showed a definite enhancement at about the 15 percent level for both schemes. The response inside the figure in the orthogonal sampling scheme appeared as a plateau. The normalization exaggerated or dramatized the 15 percent enhancement effect. From "The Role of the Primary Visual Cortex in Higher Level Vision," by T.S. Lee, D. Mumford, R. Romero, and V.A.F. Lamme, 1998, Vision Research, 38, 2429–2454. Copyright by Elsevier Science. Adapted with permission.

Figure 13.9

A schematic diagram of the model. The simplified model contains three modules: the early visual module (V1), the ventral stream module (IT), and the dorsal stream module (PO). The V1 module contains orientation-selective complex cells and hypercolumns as in the primary visual cortex (V1). The IT module contains neuronal pools coding for specific object classes as in the inferotemporal cortex. The PO module contains a map encoding positions in spatial coordinates as in the parietal occipital cortex or posterior parietal cortex. The V1 module and the IT module are linked with symmetrical connections developed from Hebbian learning. The V1

module and the PO module are connected with symmetrically localized connections modeled with Gaussian weights. Competitive interaction within each module is mediated by inhibitory pools. Connections between modules are excitatory, providing biases for shaping the competitive dynamics within each module. Convergence of neural activation to an individual pool in the IT module corresponds to object recognition. Convergence of neural activation to a neuronal pool in PO corresponds to target localization. The V1 module provides the high-resolution buffer for the IT and the PO modules to interact. From "An Unified Model of Spatial and Object Attention Based on Inter-cortical Biased Competition," by G. Deco and T.S. Lee, 2002, Neurocomputing, 44-46, 769–774. Copyright by Elsevier Press. Adapted with permission.

Figure 13.10

(A) A Paris scene. (B) In a *visual search* task, the system functions in the object attention mode. For example, when the scene was presented to the retina, a top-down bias is imposed by the prefrontal area to a tower neuronal pool in the IT. This bias, when combined with the bottom-up excitation from V1, enables the tower neuron to dominate over the neurons encoding the other objects, such as the sculpture. (C) and (D) show that the activation of the PO map and the V1 map have converged to a single spatial locus at 160 ms after stimulus onset, indicating that the target (tower) has been localized. (E) The letter L can be instantly detected in a field of X's but not in a field of T's by humans. (F) The model system can detect L in a field of X in constant time, but the time required to detect L in a field of T's increases linearly with the number of T's. This shows that serial search and parallel search is in fact implemented by a single object-attention mechanism.

Figure 1:

A

B

Figure 2:



A                         B                         C
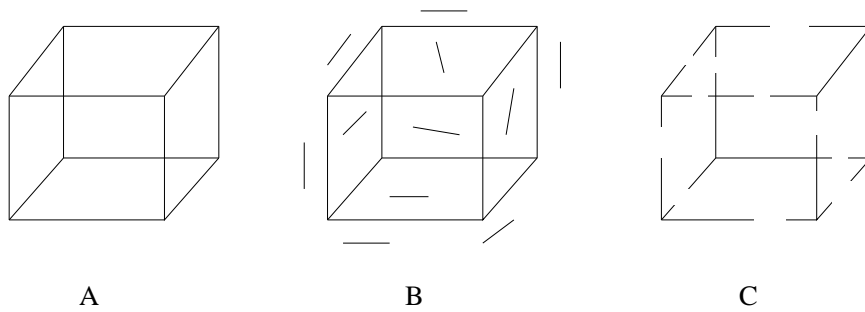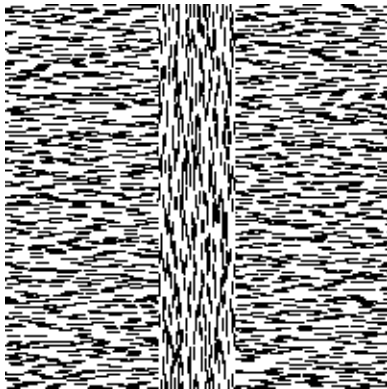
Figure 3:
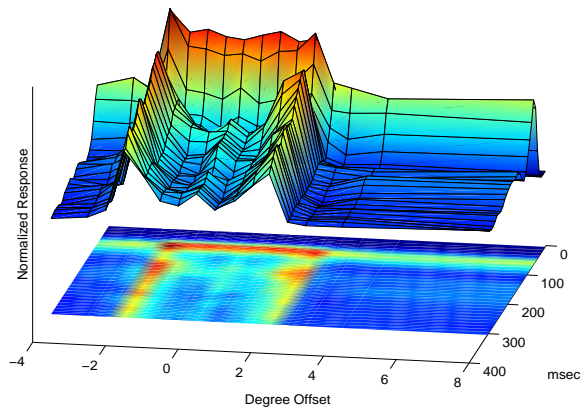


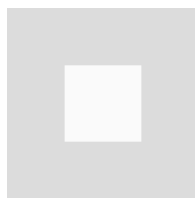A. Texture strip                    B. Spatiotemporal response

Figure 4:
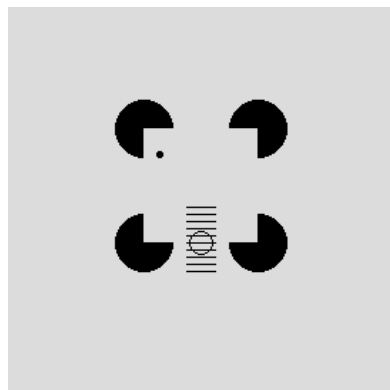


A
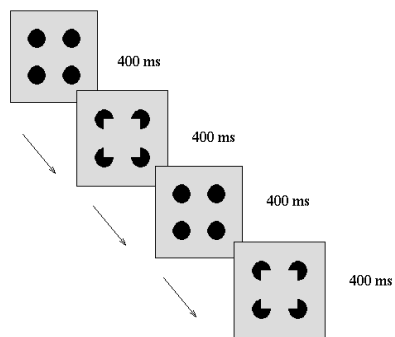


B
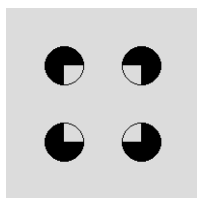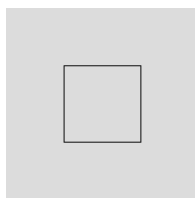


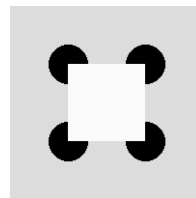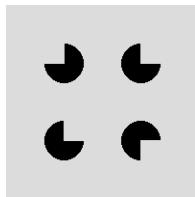C          D          E          F
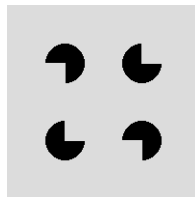
G          H          I          J

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

44

Figure 7:



V1 Response to LA Stimuli (Stage 3)

V1 Response to WA Stimuli (Stage 3)

A

B

Monkey A Neural Pop−out Modulation

Correlation with Reaction Time: Monkey A

C

D

Figure 8:



A



B



C



D



E

46

Figure 9:



**Dorsal–Stream Module (PO)**     **Ventral–Stream Module (IT)**

Space  Position Map                    Object class map

Top–down spatial bias                   Top–down object bias

Each  PO  neuron is
connected to neurons in
several hypercolumns
birectionally with
Gaussian weights.

Each  IT  neuron is connected
to every  V1 neuron bidirectionally.

**Early Visual Module
(V1)**

Hypercolumns  in retinotopic coordinate

weak excitatory connection          strong excitatory  connection

A



B



C



D



E



F