Efficient Coding of Visual Scenes by Grouping and Segmentation: Theoretical Principles and Biological Evidence

Tai Sing Lee
Computer Science Department
Center for the Neural Basis of Cognition
Carnegie Mellon University
Pittsburgh, PA 15213

Alan Yuille
Department of Statistics
University of California at Los Angeles
Los Angeles, CA 90095
yuille@stat.ucla.edu

In the Bayesian Brain: Probabilistic Approaches to Neural Coding. Ed. K. Doya, S. Ishii, A. Pouget, and R.P.N. Rao. MIT Press. Pp 145-188. January 2007.

Efficient coding of visual scenes by grouping and segmentation: theoretical predictions and biological evidence

Tai Sing Lee & Alan L. Yuille

Introduction

The goal of this chapter is to present computational theories of scene coding by image segmentation and to suggest their relevance for understanding visual cortical function and mechanisms. We will first introduce computational theories of image and scene segmentation and show their relationship to efficient encoding. Then we discuss and evaluate the relevant physiological data in the context of these computational frameworks. It is hoped that this will stimulate quantitative neurophysiological investigations of scene segmentation guided by computational theories.

Our central conjecture is that areas V1 and V2, in addition to encoding fine details of images in terms of filter responses, compute a segmentation of images which allow a more compact and parsimonious encoding of images in terms of the properties of regions and surfaces in the visual scene. This conjecture is based on the observation that neurons and their retinotopic arrangement in these visual areas can represent information precisely, thus furnishing an appropriate computational and representational infrastructure for this task. Segmentation detects and extracts coherent regions in an image and then encode the image in terms of probabilistic models of surfaces and regions in it, in the spirit of Shannon's theory of information. This representation facilitates visual reasoning at the level of regions and their boundaries, without worrying too much about all the small details in the image.

Figure 1 gives three examples which illustrate the meaning of higher level efficient encoding of scenes. Firstly, consider Kanizsa's (1979) famous illusory triangle (Figure 1a). It is simpler to explain it as a white triangle in front of, and partially occluding, three black circular discs rather than as three pac-mens which are accidentally aligned to each other. Indeed this simple explanation is what human perceive and, in fact, the perception of a triangle is so strong that we hallucinate the surface of the triangle as being brighter than the background and perceive sharp boundaries to the triangle even at places where there is no direct visual cues. Secondly, when confronted with the image shown in Figure 1b (Ramachandran 1988), we perceive it as a group of convex spheres mixed together with a group of concave indentations (e.g. an egg carton partly filled with eggs). This interpretation is more parsimonious than describing every detail of the intensity shading and other image features. Thirdly, at first glance, the image in Figure 1c (Gregory 1970) appears to be a collection of random dots and hence would not have a simple encoding. But the encoding becomes greatly simplified once the viewer perceives the Dalmation dog and can invoke a dog model. The viewer will latch on to this interpretation whenever he sees it again, underscoring the powerful interaction between memory and perception when generating an efficient perceptual description.

These three examples suggest that we can achieve a tremendous amount of data compression by interpreting images in terms of the structure of the scene. They suggest a succession of increasingly more compact and semantically more meaningful codes as we move up the visual hierarchy. These codes go beyond efficient coding of images based on Gabor wavelet responses (Daugman 1985, Lee 1996) or independent components (Olshausen and Field 1996, Bell and Sejnowski 1997, Lewicki and Olshausen 1999).

In this chapter, we will concentrate on image segmentation which is the process that partitions an image into regions, producing a clear delineation of the boundaries between regions and the labelling of properties of the regions. The definition of "regions" is a flexible one. In this chapter, we focus on early visual processing and so a region is defined to be part of an image that is characterized by a set of (approximately) homogeneous visual cues, such as color, luminance, or texture. These regions can correspond to 3D surfaces in the visual scene, or they can be parts of a 3D surface defined by (approximately) constant texture, albedo, or color (e.g. the red letters "No Parking" on a white stop sign). Based on a single image, however, it is often difficult to distinguish between these two interpretations. At a higher level of vision, the definition of region is more complex and can involve hierarchical structures involving objects and scene structures.

The approach we have taken stems from the following computational perspective about the function of the visual system. We hold it to be self-evident that the purpose of the visual system is to interpret input images in terms of objects and scene structures, so that we can reason about objects and act upon them. As thirty years of computer vision research has shown, interpreting natural images is extremely difficult. Tasks, such as segmenting images into surfaces and objects, appear effortlessly easy for the human visual system but, until recently, have proved very difficult for computer vision systems.

The principles of efficient coding and maximization of information transmission have been fruitful for obtaining quantitative theories to explain the behavior of neurons in the early stages of the visual system. These theories explain linear receptive field development, and various adaptation and normalization phenomena observed in these early areas (Atick and Redlich 1992, Dan et al. 1996, Olshausen and Field 1996, Simoncelli 2003). Moreover, the behaviors of the neurons in the early stages of the visual systems, such as the retina, LGN, and simple cells in V1, are reasonably well characterized and successfully modeled in terms of linear filtering, adaptation and normalization (Caradini et al. 2005). But there is a disconnect between this quantitative research in early sensory coding and the more qualitative, function-oriented research that has been directed to the extrastriate cortex.

How can these principles of efficient encoding be extended to hierarchical encoding of images in terms of regions, surfaces and objects? A more advanced way of encoding images is to organize the information in a hierarchy, in the form of a generative model (Mumford 1992). In this hierarchical theory, a description at the higher level can synthesize and predict ('explain away') the representations at the lower level. For example, the theory would predict that when a face neuron fires, the eye neurons and the month neurons will need to fire less. This can lead to an efficient encoding of images, as illustrated in Figure 1. Rao and Ballard's (1999) model, though limited in the type of visual interpretation it can perform, is a good illustrative example of this idea of predictive coding.

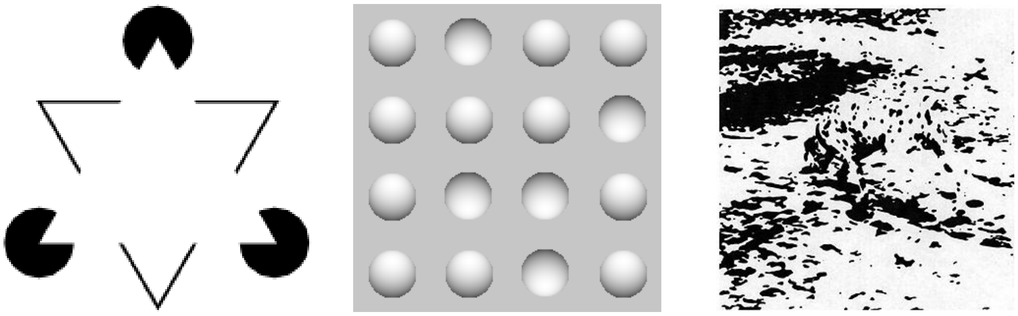
In reasoning about the functions of V1, it is natural to propose that this first visual area in the neocortex, which contains orders of magnitude more neurons than the retina, should participate in interpreting visual images rather than simply encoding and representing them. Decades of computer vision research on scene segmentation, surface inference and object recognition can potentially provide theoretical guidelines for the study of brain function. This is where interaction between neurophysiological investigation and computer vision research could prove to be valuable: computer vision can provide knowledge of natural images, and the design of mathematical theories and computational algorithms that work, while biology can provide insights as to how these functions are being solved in the brain.

We will now describe an important class of theories that perform image segmentation in terms of efficient encoding of regions (Geman and Geman 1984, Mumford and Shah 1985, Blake and Zisserman 1987, and Leclerc 1989). Next, we will briefly discuss a second generation of theories which represent the current state of the art in computer vision. The first class of models has been instrumental for motivating physiological experiments discussed in the second part of this chapter, while the second generation theories can provide insights for understanding some of the physiological observations not predicted by the first class of models.

These models are mathematically formulated in terms of probability models on graphs and, in particular, Markov Random Fields (MRF's) (Winkler 1995). These graphs, see Figure (3), consist of nodes whose activity represents image, or scene, properties (such as the presence or absence of a segmentation boundary). The connections between the nodes represent the likely probabilistic dependencies between the node activities. The graphical, or network, structure of these models makes it straightforward to implement them on networks reminiscent of the neural networks in the brain, as observed in Koch et al (1987) and Lee (1995). So these models can serve as useful guides for investigating the neural mechanisms which implement scene segmentation, as we will discuss later in this chapter.

An advantage of using computational theories of this type is that we can test their performance on natural images. If they fail to yield good segmentations, then we can design new theories which work better. This requirement that these theories yield acceptable segmentation results when applied to real-world images ensures that they are not merely "toy models". Indeed, the first generation of theories will be adequate for the types of stimuli used in this chapter but the second generation will be needed to perform segmentation fairly reliably on natural images. There is a growing literature on more advanced theories which, for example, include Gestalt laws for perceptual grouping, and even object specific knowledge.

We conjecture that the neural processes we describe in this chapter are representative of neural mechanisms that operate in other areas of the brain for performing other, higher level, tasks such as categorization and analogical association. In other words, we propose that these mechanisms are not unique to the segmentation task nor to the visual areas V1 and V2. Certainly the types of theories described here for scene segmentation might have very close analogies, and mathematical underpinnings, to the theories underlying other cognitive abilities such as language, reasoning, categorization, and other aspects of thinking.



(a) Kanizsa's Triangle (b) Ramachandran's Shapes (c) Gregory's Dalmatian Dog

Figure 1: Examples that illustrate images are interpreted to make descriptions simpler and shorter.

Computational Theories for Scene Segmentation

The earliest computational methods for scene segmentation were based on designing edge detectors to find the large changes in local intensity contrast which are associated with the boundaries of objects (e.g. Marr and Hildreth 1982). These methods involve convolving the image with first and second order derivative operators which are smoothed by Gaussian filters and then thresholding the resulting images to extract the edges. This can be implemented by local filters which are similar to the properties of cells in V1. These methods can be extended to allow the filters to be influenced by local context. For example, the Canny edge detector (Canny 1986) uses heuristics like hysteresis and non-maximum suppression to facilitate the completion of contours and to thin out multiple responses to the same edge. But overall, these methods are often unable to yield a complete segmentation of the scene because they do not take into account the global structure.

Global approaches to image segmentation began with a class of models that were developed independently in the 1980's (Geman and Geman, 1984, Mumford and Shah 1985, and Blake and Zisserman 1987). These were based on designing a global criterion for segmenting images into regions. The models assumed that regions have smoothly varying intensity patterns separated by boundaries where the intensity changes significantly. This corresponds to scenes containing smoothly varying spatial geometry separated by discontinuities (and with no texture or albedo patterns). We will describe later how these models can be extended to deal with more general images that include texture, albedo and shading patterns.

We discuss these models using Leclerc's perspective which formulates scene segmentation as an inference problem in terms of efficient encoding (Leclerc 1989). This approach is based on the Minimum Description Length (MDL) principle (Risannen 1987).

The computational goal is to choose the representation W of the regions best fits the image data D, or equivalently, which best encodes the data. In Bayesian terms, we seek to perform Maximum a Posteriori (MAP) estimation by maximizing the *a posteriori* distribution $P(W|D)$ of the representation conditioned on the data. By Bayes theorem, we can express this in terms of the likelihood function $P(D|W)$ and the prior $P(W)$ as follows:

$$P(W|D) = \frac{P(D|W)P(W)}{P(D)}.$$

The likelihood function $P(D|W)$ specifies the probability of observing data D if the true representation is W and $P(W)$ is the prior probability of the representation (before the data). In the weak-membrane model, the likelihood function is a simple noise distribution and the prior encodes assumptions that the image is piecewise smooth and the boundaries are spatially smooth (see next section for details).

In order to relate MAP estimation to efficient encoding, we take the logarithm of Bayes rule $\log P(W|D) = \log P(D|W) + \log P(W) - \log P(D)$. $P(D)$ is constant (independent of W), so MAP estimation corresponds to *minimizing* the encoding cost:

$$-\log P(D|W) - \log P(W)$$

We now interpret this in terms of minimal encoding. By information theory (Shannon 1948) the number of bits required to encode a variable X which has probability distribution $P(X)$ is $-\log P(X)$. The term $-\log P(W)$ is the cost of encoding the interpretation W. The term $-\log P(D|W)$ is the cost of encoding the data D conditioned on interpretation W. This cost will be 0 if the interpretation explains the data perfectly (i.e. $P(D|W)=1$). But usually the interpretation will only partially explain the data and so $-\log P(D|W)$ is called the residual (see the detailed example below).

Observe that the encoding depends on our choice of models $P(W|D)$ and $P(W)$. Different models will lead to different encoding, as we will describe later.

The Weak-Membrane model

The weak-membrane model deals with images where the intensity varies smoothly within regions, but can have sharp discontinuities at the boundaries of regions. We introduce it by the energy functional proposed by Mumford and Shah (1985). This model was defined on a continuous image space, rather than on a discrete lattice (hence the term "functional" rather than "function). But, as we will show, it can be reformulated on a lattice. (There are differences between the Mumford and Shah model and closely related models by Geman and Geman (1984) and by Blake and Zisserman (1987). The most important difference is that the Mumford and Shah model is guaranteed to segment the image into closed regions, while this is only strongly encouraged by the other models).

The weak-membrane model represents the image by variables (u, B), where B is the set of boundaries between regions and u is a smoothed version of the input image d. More precisely, the image is described by intensity values $d(x, y)$ specified on the image space (x, y). The model assumes that the intensity values are corrupted versions of (unobserved) underlying intensity values $u(x, y)$. The underlying intensity values are assumed to be piecewise smooth, in a sense to be described below.

We formulate this problem by describing $-\log P(D|u, B) - \log P(u, B)$ directly. We write this as $E(u, B)$:

$$E(u,B)=\int\int_R \frac{(u(x,y)-d(x,y))^2}{\sigma_d^2}dxdy+\int\int_{R-B}\frac{1}{\sigma_u^2}\|\nabla u\|^2dxdy+\alpha B$$

where R is the image domain, B denotes the set of boundary locations boundaries and $R-B$ indicates the entire image domain minus the boundary locations.

The first term in $E(u, B)$ is the data term, or $-\log P(D|u, B)$, where the distribution of the residues $P(D|u, B)$ is chosen to be Gaussian white noise with standard deviation σ_d. In other words, $P(D|u,B)=\prod_{x,y}P(d(x,y)|u(x,y))$, where $P(d(x,y)|u(x,y))=\frac{1}{\sqrt{2\pi\sigma_d^2}}e^{-(u(x,y)-d(x,y))^2/(2\sigma_d^2)}$. The double integrals over x and y is simply the continuous version of summing over all the pixels in the image domain R.

The second term is a smoothness prior, which assumes the variation on the estimated image intensity to be smooth within each region. Intuitively, the local variation following a Gaussian distribution, i.e. $P(\nabla u:(u,B))=\frac{1}{\sqrt{2\pi\sigma_u^2}}e^{-(\frac{\nabla u}{2\sigma_u})^2}$, where σ_u is the standard deviation of this distribution, but this is an oversimplification (for true interpretation, see Winkler 1995). Observe that when σ_u is very small, then the energy function enforces regions to have constant intensity. This smoothness term is deactivated (discounted) at the boundary B so the integral is over $R-B$. The intuition is that when the local image gradient is too steep, then it will be better to put in a boundary.

The third term is a penalty on the length of the boundaries. This is needed to prevent the image from breaking into too many regions or from creating regions with wildly zigzagging boundaries. The sum of the second and third term yields $-\log P(u, B)$ (technically this prior is improper, see Winkler 1995, but this is not significant for this chapter.)

To develop algorithms to minimize the energy function, it is convenient to reformulate it, in the style similar but not identical to that of Ambrosio and Tortorelli (1990), so that the boundaries B are replaced by line process variables $l(x, y)$ which take on values in $[0, 1]$:

$$\begin{aligned}E_p[u,l|d]=\int\int_R \frac{(u(x,y)-d(x,y))^2}{\sigma_d^2}dxdy+\int\int_R(1-l(x,y))^2\frac{|\nabla u(x,y)|^2}{\sigma_u^2}dxdy\\ +\alpha\int_R\{p|\nabla l(x,y)|^2+p^{-1}l(x,y)^2/4\}dxdy. \qquad (1)\end{aligned}$$

The line process variables take value $l(x, y) \approx 1$ at the boundaries, thereby cutting the smoothness constraint. It can be shown that, in the limit as $p \mapsto 0$ the minimization of this corresponds to Mumford and Shah (and the line process

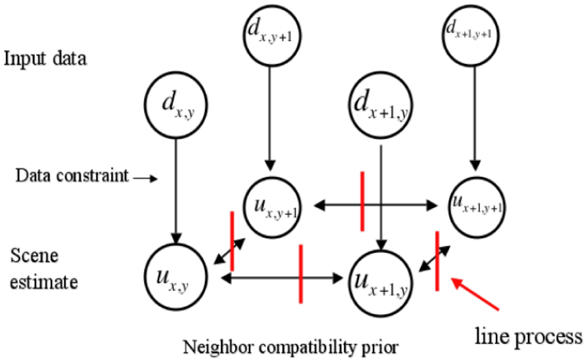


Figure 2: A locally connected network for scene segmentation: A node (x, y) has value $u_{x,y}$. It receives direct input from the data $d(x, y)$. It is also connected to its immediate neighbors (or more generally, to nodes in a local neighborhood) by symmetrical connections which embody the prior which enforces the estimated intensity to be piecewise smooth.

variables take values either 0 or 1). The advantages of the second formulation (1) are: (i) that it is easier to find algorithms for it than for Mumford and Shah, and (ii) it can be directly discretized and implemented on a grid.

The weak-membrane model can be generalized to perform 3D surface inference (Belhumeur 1996) and to the coupled-membrane model (Lee et al. 1992) for texture segmentation. In natural images, many regions are defined by texture properties (e.g. stimuli in Figure 12). The image intensities within each region are not smooth, but the texture properties are. To do this, we set the input $d(x, y)$ to be $WI(\sigma, \theta, x, y)$ which is a image of 4 continuous variables obtained by wavelet transform (which provides a simple local measure of texture). Each wavelet channel is fitted by a 2D weak-membrane but each of these membranes is coupled to its nearest neighboring membranes in the (σ, θ) domain. (The algorithm involves anisotropic diffusion which takes place in 4D but breaks are allowed in x and y only). This model has been shown to be effective in segmenting some texture images, tolerating smooth texture gradients (Lee et al. 1992).

One justification for models like the Mumford-Shah formulation comes from the study of statistics of natural images. Zhu and Mumford (1997) used a statistical learning approach to learn models for $P(d|u, B)$ and $P(u, B)$ from natural images, and their results are similar to the Mumford-Shah model (though with interesting differences). It is interesting to contrast how Zhu and Mumford use images statistics to learn the prior with how receptive fields can be learnt from similar statistics using efficient encoding principles.

A Computational Algorithm for the Weak-Membrane Model

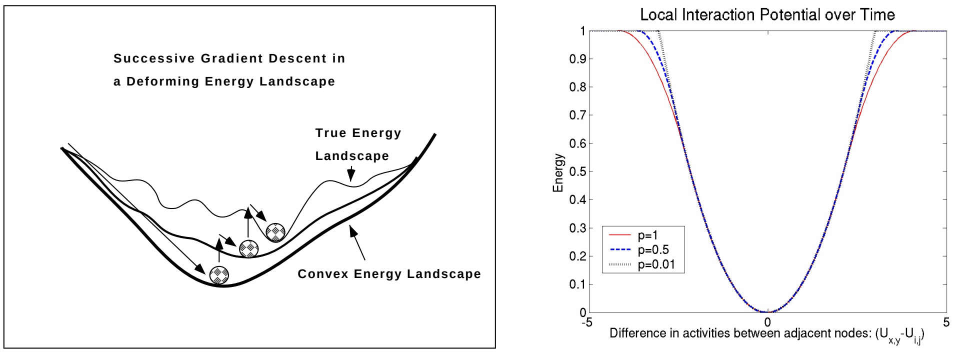
When discretized on a grid, the weak-membrane model can be implemented in a network structure as shown in Figure 2. Such a network contains a layer of input observation $d(x, y)$, a layer of hidden nodes $u(x, y)$, and a set of line processes (or boundary processes) $l(x, y)$. The u nodes can communicate (passing messages) to each other, but their communication can be broken when the line process between them become active. Thus, this is an interacting system of two concurrent processes.

For the weak-membrane implemented in a Markov network or Markov random field (Geman and Geman 1984), the local connections between the nodes enforce the smoothness constraint, which make the states of the adjacent nodes vary as little as possible, subject to other constraints such as faithfulness to the data (the data term in the energy functional). More generally, Markov random fields can implement any form of compatibility constraint and do not have to be local connections (Winkler 1995).

We now describe algorithms that can find the solution of the weak-membrane model using such a network. Finding the solution that minimizes the energy functional of the weak-membrane model is not trivial, as there are many local minima due to the coupling of the two processes.

The simplest way of finding the scene estimate function u and partition boundary B that would minimize this class of energy functionals is to search through all possible sets of regions, calculating the cost for each set and choosing the set with the smallest cost. But the number of possible sets of regions is prohibitively large for image of any reasonable size, making an exhaustive search infeasible. Simple gradient descent descent algorithms will fail because the interaction with line processes makes the system highly non-convex and gives it a lot of local minima to trap the algorithm into inferior or incorrect solution.

The best algorithms that have emerged so far belong to a class of optimization techniques generally known as contin-



(a) (b)

Figure 3: (a) The energy landscapes of a sequence of deforming energy functionals and the successive gradient descent of a 'ball' in this sequence of landscapes: the system first converges to the global minimum of the convex landscape, which is then used as the starting point for the next descent in the new energy landscape. This strategy of successive gradual relaxation will allow the system to converge to a state that is close to a global minimum of the original energy functional. This strategy can be considered a coarse-to-fine search strategy in the scale-space of energy landscape. (b) The transformation of the approximating energy functional, from the convex one back to the original one, is achieved simply by modifying the local interaction potential as a function of p as shown here. The local interaction potential dictates the contribution of the second and the third terms in the Mumford-Shah energy functional. At $p = 1$, the local interaction potential transits smoothly from the quadratic function $\lambda^2(\triangle u)$ to the α, which corresponds to a global convex energy function. As p decreases, the transition becomes more and more abrupt, and converge to the local interaction potential prescribed by the original energy functional as p becomes very small (see Blake and Zisserman 1987 for details).

uation methods (Blake and Zisserman 1987, Ambrosio and Tortorelli 1990). The basic idea of the continuation method is to embed the energy function in a family of functions $E_p(u, l)$ with the continuation parameter p. At large p, the energy function $E_p(u, l)$ is convex and will have only one single minimum. As p approaches zero, the energy functional will transform gradually back to the original function which can have many local minima. The strategy is to minimize the energy at large p and then track the solution to small values of p. More precisely, we initialize p^0, select random initial conditions for (u, l) and perform steepest descent to find a minimum (u^0, l^0) of $E_{p^0}(u, l)$. Then we decrease p to p^1, and perform steepest descent on $E_{p^1}(u, l)$ using (u^0, l^0) as initial conditions, and so on. This approach is not guaranteed to find the global minumum of $E(u, B)$, but empirically it yields good results provided the initial value of p^0 is sufficiently large. The dynamics of the algorithm with the transformation of the energy landscape is shown in Figure 3.

The steepest descent equations for the functional reformulation of weak-membrane model as in equation 1 are given by:

$$\frac{du(x,y,p,t)}{dt} = r_u\{-u(x,y,p,t) + d(x,y) + \nabla\cdot[\frac{\sigma_d^2}{\sigma_u^2}\nabla u(x,y,p,t)(1-l(x,y,p,t))^2]\} \tag{2}$$

$$\frac{dl(x,y,p,t)}{dt} = r_l\{\alpha p \nabla^2 l(x,y,p,t) + \frac{1}{\sigma_u^2}(1-l)\|\nabla u(x,y,p,t)\|^2 - \frac{\alpha l(x,y,p,t)}{2p}\}. \tag{3}$$

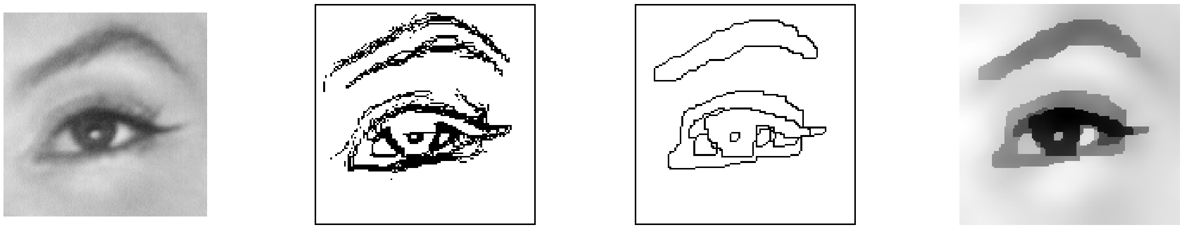
The parameters r_u and r_l are positive rate constants which control the rate of descent. At each stage with a particular p, which changes slowly from 1 to 0, the system relaxes to an equilibrium, i.e. $\frac{du}{dt}$ and $\frac{dl}{dt}$ are driven to 0.

In these equations, u follows a nonlinear diffusion process that is sensitive to the values of the line process. The line process at each location is a continuous variable, indicating that the boundaries are soft during the early stages of the algorithm and only become sharp at the end.

Figure 4 shows the outcome of such algorithm given an input image (a). Minimizing $E(u, B)$ yields a map of boundary process B which, during the first stages of the algorithm, resembles an artist's sketch (b), and then develops into crisper boundaries (c) which partition the image into a set of piecewise-smooth regions (d).

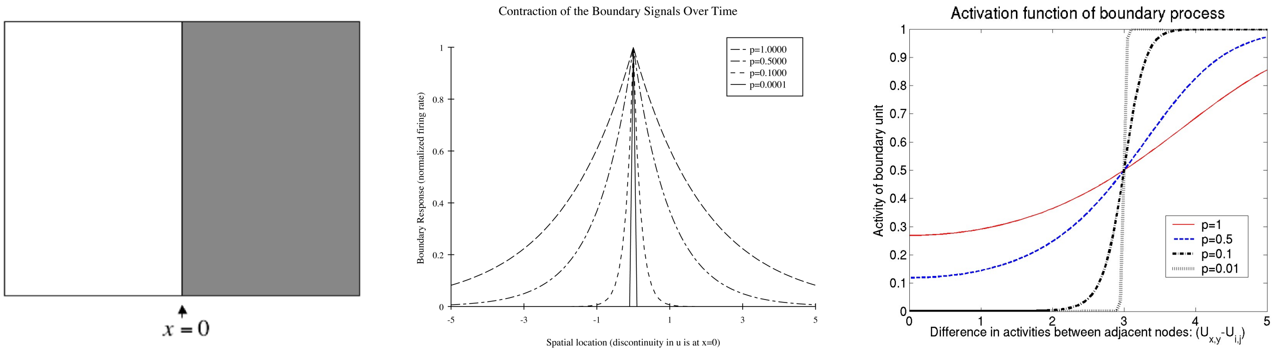
The gradual sharpening of the boundaries can be seen in the responses of a collection of line processes (over space) to a step edge at $x = 0$ (Figure 5a). That is, the gradient along y is all zero, and the gradient along x is all zero except at $x = 0$, which is strong enough to activate $l(x = 0)$ fully, so that $l(x = 0) = 1$. Therefore the second term in Equation 3 just vanishes, and at the equilibrium of each p relaxation stage (i.e. $\frac{dl}{dt} = 0$), the dynamical equation (Equation 3) yields

$$\alpha p \frac{\partial^2 l}{\partial x^2} = \frac{l\alpha}{4p} \Longrightarrow l = e^{-\frac{|x|}{2p}} \tag{4}$$



(a) Input image (b) Initial edge map (c) Final edge map (d) Final surface cue map

Figure 4: Segmentation of an image by the weak-membrane model. Results are encoded in two maps: the boundary map and the region map. (a) The input image. (b) The initial response of the edge (the line process $l(x, y)$) map resembles an artist's sketch of the scene, with uncertainty about the boundary. (c) The final response of the edge map B shows crisper boundaries. (d) The final estimate of the piecewise smooth intensity values $u(x, y)$ is like a smoothed and sharpened image, as if make-up had been applied. Adapted from Lee (1995).



(a) (b) (c)

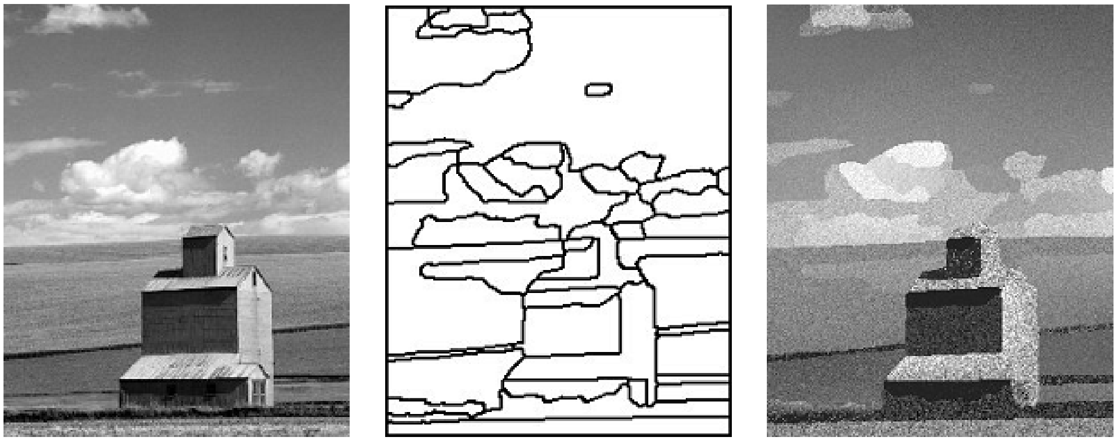
Figure 5: (a) An image with a luminance step edge. (b) In response to the step edge, initially, line processes near the edge will respond, resulting in a broader spatial response profile among these population of nodes. Over time, as p decreases, the boundary responses start to contract spatially to the exact location of the step edge. The activity at each location represents the response of a line process at that location. The curve does represents the responses of a population of identical boundary processors distributed over space, or equivalently the spatial response profile of a boundary unit to different part of the image with a step edge. As p decreases, the spatial response envelope becomes narrower and narrower (Lee 1995). (c) Another perspective on the continuation method is that the activation function of the line process becomes steeper as p decreases (Geiger and Yuille 1991).

$\nabla^2 l$ controls the lateral interaction between the boundary signals l. As $p \to 0$, the boundary signal l in the surrounding is gradually suppressed, resulting in the contraction to a sharp boundary as shown in Figure 5b.

Observing that a locally connected network is quite compatible with the known anatomical connections in the primary visual cortex (V1), Koch et al. (1987) first proposed a neural circuit for implementing a variant of the weak-membrane model in V1 for the purpose of depth estimation (the same model was applied to segmenting images by Geiger and Yuille 1991). Lee (1995) explored the implementation of the coupled-membrane model (Lee et al. 1992) with V1 circuitry further based on the energy functional stated in equation 1 and the descent equations described here. This circuit takes data input in the form of V1 complex cell responses, and might be considered more 'neurally plausible'. It was observed that the two concurrent and interacting processes of region inference and boundary detection implicated by the descent equations are very similar in spirit to a neural model proposed for V1 earlier by Grossberg and Mingolla (1985).

Generalizations of the Weak Membrane Model

The main limitation of the weak-membrane model is that it uses very simple assumptions about the intensity properties within regions. We have briefly described how this can be generalized to cases where the regions have smooth texture patterns. But most natural images have richer properties. Some surface regions are characterized by smooth texture properties, others have smooth intensities, and yet other have shaded intensity properties which depend on the tilt and slant of surfaces. The piecewise smooth intensities, assumed by the weak-membrane model, are suitable for the 3D



(a) Input image (b) Segmentation of the image (c) Synthesis of the image

Figure 6: An illustration of how region competition model can encode an image economically by encoding each region with a set of boundary elements and two numbers – the mean and the variance of the intensity values in each enclosed region. Different models of regions will compete to explain the image. (a) Input image. (b) Segmentation boundary. (c) Representation of image based on simple region models. Results produced by the Data-Driven Monte Carlo algorithm for the region competition class of models (Tu and Zhu 2002). Courtesy of Tu and Zhu.

surfaces which are approximately flat but fail to accurately model images of more complicated surfaces, such as spheres.

We now briefly describe some second generation models of image segmentation. These are significantly more effective than the weak-membrane model when evaluated on images with ground truth (as provided by Berkeley dataset, see Martin et al 2001). We will discuss three different aspects of these methods that might be relevant to the interpretation of cortical mechanisms.

Firstly, we can generalize the weak-membrane model by observing that natural images are very complex. The images of object surfaces are characterized by a combination of multiple factors such as textures, color, shapes, material properties, and the lighting conditions. Each region is described by its own model, for example shading or texture. The task of segmentation is now much harder. We have to determine the segmentation boundaries while simultaneously determining the appropriate model for each region.

We can formulate this as a generalization (Zhu et al. 1995, Zhu and Yuille 1996) of the weak membrane model. We refer to these as region competition models.

$$E((R_r), n, (a_r), (\theta_r)) = \sum_{r=1}^{n} \int\int_{R_r} \{-\log P(d(x,y)|a_r, \theta_r)\} dxdy - \sum_{r=1}^{n} \log P(a_r, \theta_r) + \frac{\alpha}{2} \sum_{r=1}^{n} |\partial R_r| + cn.$$

This includes a set of generative models $P(d(x,y)|a_r, \theta_r)$ which are indexed by model type index variable a_r (corresponding to texture, shading) and a variable θ_r corresponding to the parameters of the model. This corresponds to encoding the image in terms of a richer language which allows different regions to be encoded by different models (a texture model, or a shaded model, or an alternative). Figure 6 illustrates coding regions of a segmented image with relatively simple models of regions. Regions can be encoded as one of three types: (i) a Gaussian model encoding the mean and variance of the intensity in the region, (ii) a shading model where the image intenesity follows a simple parameterized form, and (iii) a simple texture/clutter model. The segmentation thus encodes an image in terms of a set of boundaries as well as a set of region model codes (e.g. containing two values for each region the algorithm decides to encode as Gaussians). From such a representation, an approximation of the original image can be synthesized (Figure 6c). More sophisticated generative models will give increasingly realistic synthesized images.

There is no limitation to the class of models that can be used. Recent work has generalized this approach to include models of objects, such as faces and text, so that segmentation and object recognition can be performed in the same unified framework (Tu et al 2005).

This generalized model requires a significantly more complicated algorithm than the weak-membrane model. It needs processes to select the models and to 'move' the boundaries. The movement of boundaries is dictated by a diffusion process on region properties similar to that for the weak-membrane model. We will discuss this further at the end of this section.

Secondly, there is a second class of models which are, in a sense, complimentary to the region competition model. We will refer to these as affinity-based models.

This family of approaches uses affinity weights w_{ij} between different image pixels v_i and v_j. These affinity functions are based on properties measured from the image and are designed so that pixels in homogeneous image regions have high affinity ($w_{ij} \approxeq 1$) with each other. For example, we can obtain a model with similar properties to the weak-membrane by setting $w_{ij} = e^{-\alpha|d_i - d_j|}$, where d_i is the intensity at lattice site v_i in the image. This construct is naturally regarded as a graph with the image pixels constituting the node set, and the weights between them as the edge set (here the boundaries occur when w_{ij} is small).

Given such a graph, a sensible goal is to label the nodes such that intra-region affinity is maximized (defined by the labels), while minimizing the inter-region affinity. Given the label set $\{l_1, ..., l_k\}$, we assign a label to each image pixel so that pixels with the same labels define a region. Finding such a labeling can be formalized in the following minimization (Yu an Shi 2003):

$$E(m) = \min_{m \in \mathcal{P}(n,k)} \; : \; \frac{1}{k} \sum_{p=1}^{k} \frac{\sum_{i<j} w_{ij}(m_{ip} - m_{jp})^2}{\sum_{i<j} w_{ij}(m_{ip}^2 + m_{jp}^2)}$$

where n is the number of pixels, k the number of labels, and $\mathcal{P}(n,k)$ denotes the set of $n \times k$ indicator matrices. An indicator matrix m satisfies the following constraints $m(i,p) \in \{0,1\}$ and $\sum_{p=1}^{k} m(i,p) = 1$. In this application the node indices i and j are taken as the indices of the image pixels. Accordingly, the indicator matrix takes the value $m_{ip} = 1$ when the i^{th} pixel is assigned the p^{th} label l_p, otherwise $m_{ip} = 0$.

The objective function falls into the class of NP-hard problems. Subsequently, a variety of algorithms have been developed to find nearly optimal solutions in polynomial time. One class of algorithms relaxes the discrete constraints on m to a continuous values transforming the minimization into a generalized eigenvalue problem (Shi and Malik 2000). The obtained continuous solutions are then rounded to discrete values producing the cut. Another interesting class uses hierarchical grouping and is fast and effective (Sharon et al 2001). A third class of algorithm iteratively modifies the affinity matrix, using the eigenvectors as soft membership labels, until the graph eventually disconnects, producing reasonable segmentations (Tolliver et al. 2005).

This type of approach is appealing for several reasons. It simplifies the problem formulation as it does not require us to specify models. It enables a richer set of connections, defined by the affinities, than those implemented by the weak-membrane model (which essentially a Markov random field with the nearest neighbors, see Figure 2). These rich connection patterns are consistent with the known anatomy of V1. But these models are still being developed and detailed biological predictions have not yet been worked out. Finally efficient computation of approximate solutions is possible.

Interestingly, there may be interesting connections between these types of theory and the algorithms used to implement the region competition models. In the Data-Driven MCMC algorithm, Tu and Zhu (2002) uses “proposals” to activate the models. These proposals can be generated based on grouping pixels into regions based on affinity cues and then evaluating these groupings by accessing the models. From this perspective, the affinity-based methods can be used as sophisticated ways to generate proposals. In this way, it may be possible to combine these two approaches.

Thirdly, we have so far treated segmentation in terms of image intensity properties only. But the extended class of models, in principle, could enable us to integrate segmentation with the estimation of 3D shape properties such as 3D geometry or Marr’s 2.5D sketch (Marr 1982).

The inference of surface geometry, integrated into the segmentation framework, can be illustrated by a simple example where the image intensity corresponds to a shaded surface. This requires a model for how the image intensity has been generated from the surface of an object by light reflected from it. We assume the standard Lambertian model which is characterized by a reflectance function $R_{\vec{s}}(\vec{n}) = \vec{n}(x,y) \cdot \vec{s}$, with $\vec{n}(x,y)$ being the surface gradient at position (x,y) and $\vec{s}$ the light source (we assume a single light source here). We also assume that the light source $\vec{s}$ is known (there are techniques for estimating this). It is convenient to express the surface normal $\vec{n}(x,y)$ in terms of the surface slant (f,g) in the x and y directions respectively ($\vec{n}(x,y) = \frac{1}{\sqrt{1+f_x^2(x,y)+f_y^2(x,y)}}(-f_x(x,y), -f_y(x,y), 1)$).

We can use a classic shape-from-shading method due to Horn (1985) as one of the models in region competition. There are more sophisticated models on shape from shading. But this is the simplest model that we can illustrate the principle of surface inference. The cost function is of form:

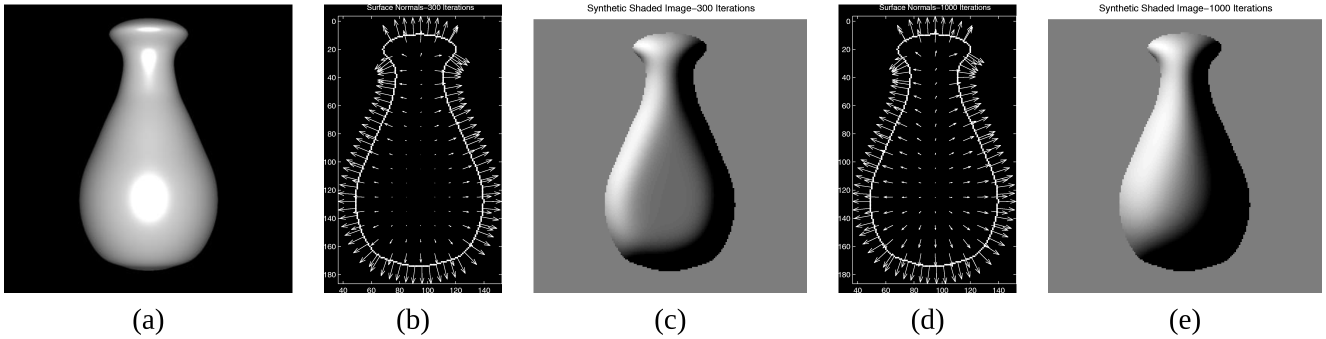


Figure 7: Surface interpolation by propagation of surface orientation from the boundary using locally connected network using Horn's algorithm. (a) Input image. (b) Initial estimate of surface as represented by the needle map. The needle points in the direction of surface normal. (c) A rendering of the surface as represented by (b) assuming lighting from the left, to illustrate the initial estimate of the surface. (d) Final estimate of surface orientations at each location and (e) its shaded rendering. This illustrates the propagation of surface orientation information from the border to the interior of the surface over time.

$$E(f,g:x,y) = \int\int_{\Omega} \frac{(d(x,y) - R_s(f,g))^2}{\sigma_d^2} dxdy + \frac{1}{\sigma_s^2} \int\int_{\Omega} ((f_x^2 + f_y^2) + (g_x^2 + g_y^2))dxdy$$

where Ω is a subregion of the image, $d(x,y)$ is the intensity of the image at location (x,y). The first term is the standard Gaussian noise model. The second term is a smoothness prior on the surface orientation.

An additional constraint on the surface normal $\vec{n}(x,y)$ is available at the occlusion border $\partial\Omega$, where the surface normal is always perpendicular to the line of view. This provides boundaries conditions to start a surface interpolation process. Figure 7 illustrates how the surface orientation information, as represented by the needle map that indicate the direction of surface normals, can propagate in from the occlusion border to the surface interior during the process of surface interpolation. We would like to draw your attention to this propagation of signals from the border, as such propagation from border to the interior has also been observed in neuronal activities in V1 as we shall discuss. This model provides one potential interpretation (among many) of such a signal being either a part of or at least a reflection of an on-going surface inference process.

As in all the region competition theory (as well as the weak-membrane model), the boundary of the region Ω must be detected at the same time as inference of the regional properties – in this case, the 3D surface orientation represented by the functions f, g.

It is important to remember that many other models of regions and boundaries have also been developed for the region competition theory, from simple models that describe each region by the mean and the variance of its intensity values within that region, or by covariance of the first derivatives along different directions for encoding texture (e.g. Zhu et al. 1995, Zhu and Yuille 1996), to more advanced models of textures (Zhu and Mumford 1997, Zhu et al. 2005) and even objects (Tu et al. 2005).

In summary, the second generation of theories requires richer modeling for the types of intensity patterns that occur in real images. This can be considered to be a richer class of description of the image which can involve 3D surface shape and even models of objects. Such theories require complex inference algorithms because the segmentation of the image into regions must be performed while simultaneously determining what type of regions they are, and what are their regional properties. Different models cooperate and compete to explain different regions in the image. These inference algorithms can be helped by the use of proposals, some of which may be provided by the output of the affinity-based models. For more details on these class of models, see (Tu and Zhu 2002, Tu et al 2005). For related theories, from a different perspective see (Borenstein and Ullman 2001, Yu and Shi 2004).

Three aspects of the second generation of segmentation models may relate to the physiological experiments described in the biological section. Firstly, the algorithms for the region competition methods (with multiple models) have aspects which are similar to the diffusion methods used for the weak-membrane. The difference is that diffusion is only one of several components of the algorithm. Secondly, these algorithms make a natural use of bottom-up and top-down processing where the models are accessed by recurrent connections to different visual areas. Thirdly, this approach might also fit naturally in a hierarchical coding scheme which can potentially compress the total coding length by using higher level (more abstract) description to replace or explain away the redundant information in the low level representations as suggested by Mumford (1992).

Biological Evidence

In this section, we will discuss some neurophysiological experiments which suggest that the primary visual cortex is involved in image segmentation. We start by enumerating the predictions of the computational models, particularly those of the weak-membrane model for segmentation, then compare them to the experimental results.

1. There exists a dual representation of region and boundary properties. This is possibly in two distinct groups of neurons, one set coding region properties, while the other set coding boundary location.

2. The processes for computing the region and the boundary representations are tightly coupled, with both processes interacting with and constraining each other.

3. During the iterative process, the regional properties diffuse within each region and tend to become constant. But these regional properties do not cross the regional boundaries. For the weak membrane model, such spreading can be described as nonlinear diffusion, which propagates at roughly constant speed. For more advanced models, the diffusion process may be more complicated and involve top-down instantiation of generative models.

4. The interruption of the spreading of regional information by boundaries results in sharp discontinuities in the responses across two different regions. The development of abrupt changes in regional responses also results in a gradual sharpening of the boundary response, reflecting increased confidence in the precise location of the boundary.

5. In the continuation method, there is additional sharpening of the boundary response. This is modulated by a global parameter p, which increases the sensitivity of all boundary-sensitive neurons (Figure 3).

The computational models say nothing about where the representation, and the processes required to compute them, occur in the visual cortex. To relate these models to cell recordings requires making additional conjectures that we will now state.

We conjecture that segmentation computation is embodied in V1, with information about boundaries and regions explicitly represented there. However, not all the computations required for segmentation need take place in V1. For example, it is not clear, based on existing evidence, whether surfaces are represented in V1. There is evidence, however, on the sensitivity of V2 neurons to surface properties such as relative depth (Thomas et al. 2002), border-ownership (Zhou et al. 2000), da Vinci stereo (Bakin et al. 2000), and pop-out due to shape from shading (Lee et al. 2002). Thus, it seems more likely that surface inference takes place in V2, but the process can be coupled to the segmentation process in V1 through recurrent connections.

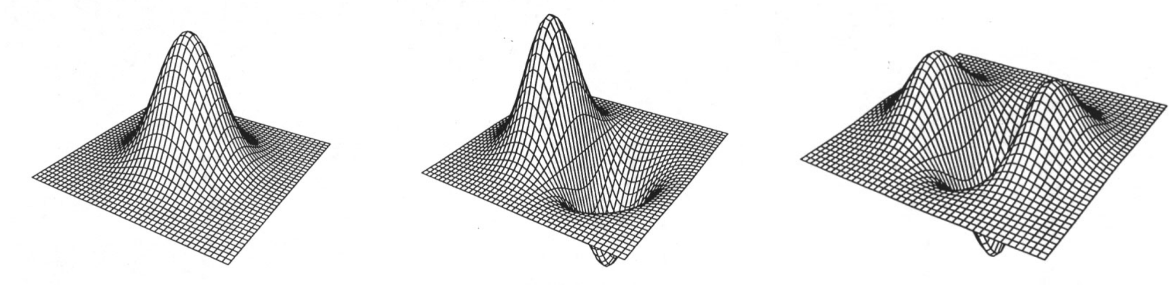
This conjecture is related to the *high-resolution buffer* theory of V1 (Lee et al. 1998, Lee and Mumford 2003) partially inspired by some of the experiments described in this section. In this theory, V1 acts as a high-resolution computational buffer which is involved in *all* visual computations that require high spatial precision and fine scale detail, since V1 is the only visual area that contains small receptive fields. Some of these computations are performed in V1 directly, while others are performed by recurrent connections between V1 and the other visual areas in the extrastriate cortex. Processing in V1 detects salient regions which are enhanced by a recurrent feedback mechanism very much like the adaptive resonance or interactive activation mechanisms hypothesized by the neural modeling community (Grossberg 1987, McClelland and Rumelhart 1981, Ullman 1994, see also Yuille and Grzywacz 1998 for a motion perception theory utilizing recurrent feedback, Deco and Lee 2004 for a model for integrate what and where in the high-resolution buffer using on recurrent biased competition). In particular, Lee et al. (1998) have argued that segmentation in V1 cannot be complete and robust without integrating with other higher order computations, such as object recognition and shape inference. On the other hand, higher order computations might not be robust without continuous consultation and interaction with the high-resolution buffer in V1. Thus, in their view, visual computations such as scene segmentation should involve the whole visual hierarchy utilizing the feedforward and recurrent feedback connections in the visual cortex. This relates to some of the second generation computational models.

We now present neurophysiological findings that show experimental support for several of these computational predictions. In particular, we discuss evidence for: (1) the gradual sharpening of the response to boundaries, (2) the simultaneous spreading of regional properties and (3) the development of abrupt discontinuities in surface representations across surfaces.

Edge and boundary representations and their spatial temporal evolution

We first review evidence that V1 neurons represent edge and boundary locations. The early experiments by Hubel and Wiesel (1972) showed that neurons in V1 are sensitive to oriented intensity edges. Indeed Hubel and Wiesel conjectured that the neurons were edge detectors because of their orientation selectivity.

Detailed analysis of the receptive fields of simple cells, using the reverse correlation method, showed that they could be modeled by linear Gabor filters (Daugman 1985; Jones and Palmer 1988). The receptive fields at different scales resemble scaled and rotated version of each other, which means they can be considered to be two-dimensional Gabor wavelets (Daugman 1985, Lee 1996). In the theoretical neuroscience community, the Gabor filter interpretation has been popular because Gabor filters achieve the limit of representing information with maximum resolution in the conjoint space of space-time and frequency. Recently, it has been shown that they can be derived from the statistics of natural images as efficient codes for natural images based on independent component analysis (Olshausen and Field 1996). Although such rationalization about efficient representation is interesting intellectually, we should not lose sight of the real functional purposes of simple and complex cells as edge detectors, as proposed by Hubel and Wiesel (1972). Young (1987) has long championed that simple cells can be described equally well in terms of the first and second order 2D Gaussian derivative operators (see Figure 8). The odd-symmetric Gabors are sensitive to intensity edges and the even-symmetric Gabors to bars (such as peaks and valleys). In fact, Bell and Sejnowski (1997) have aptly argued that the independent components likely arise from the structures of edges in natural images.



(a) $G_0(x)G_0(y)$ (b) $G_1(x)G_0(y)$ (c) $G_2(x)G_0(y)$

Figure 8: Graphs of (a) a 2D Gaussian, (b) the first x derivative of a 2D Gaussian which resembles the odd-symmetric Gabor filter/wavelet, (c) the second x derivative of a 2D Gaussian which resembles the even-symmetric Gabor filter/wavelet. Adapted from Young (1987).

In summary, simple cells can serve as edge and bar detectors (Hubel and Wiesel 1972), with their maximum responses used to signal the location of the boundary (Canny 1986). But, considered as derivative of Gaussian filters, they can also perform some of the other mathematical operations required by the computational models. The complex cells, which are not sensitive to the polarity of the luminance contrast at edges, would be particularly suitable for representing borders or boundaries of regions. The hypercomplex cells could serve as derivative operators which act on complex cells' responses to detect texture boundaries (see Lee 1995).

We now turn to evidence of non-local interactions associated with computational models of edge detection. Studies by Kapadia et al. (2000) showed that the activity of a V1 neuron to a bar within its receptive field can be enhanced by the presence of other bars outside the receptive field of the neuron, provided these bars are aligned to form a contour (longitudinal facilitation). Conversely, the neuron's response is suppressed if these bars are parallel to the bar within the receptive field (lateral inhibition). One manifestation of longitudinal facilitation is that V1 neurons looking at the short gaps in boundaries of the Kanizsa figures (see Figure 1a) have been found to respond after a certain delay (100 ms after stimulus onset versus 40 ms required for the response to a real edge) as shown in the study by Lee and Nguyen (2001) (see Grosof et al. 1993, Sheth et al. 1996 for other subjective contour effects in V1, and von der Heydt et al. 1984 for classic results in V2). Lee and Nguyen (2001) also found that V2 neurons responded earlier to the same stimuli. This raises the possibility that the illusory contour response found in V1 is in part due to feedback influence from V2, and in part carried out by the horizontal collaterals within V1.

Longitudinal facilitation and lateral inhibition are consistent with the mechanisms in basic edge detection models (Canny 1986) and contour completion models (Nitzberg et al. 1993, Williams and Jacobs 1997)). These mechanisms are also embodied in the weak-membrane model as described. The continuation method for the weak-membrane model further gives a quantitative prediction on the temporal evolution of the boundary process. Specifically, in response to a simple luminance border (as shown in Figure 5), the neural response to boundaries gets sharper over time as the

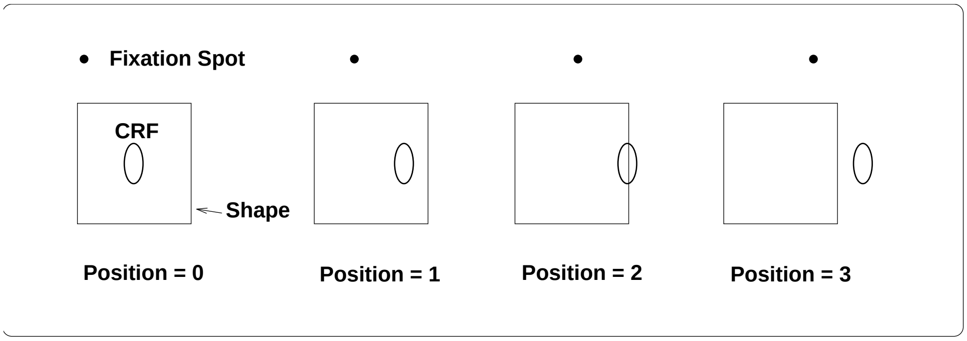


Figure 9: The spatial sampling scheme. The monkey fixates on the solid dot on upper left. In successive trials, the image on the monitor is displaced relative to the fixation point of the monkey, or equivalently, relative to the recorded neuron's classical receptive field (CRF). This gives a spatial sampling of the neural response to different parts of the stimulus.

continuation parameter p decreases. (Predictions 4 and 5).

Analysis of data obtained from the experiment described in Lee and Nguyen (2001) provides some evidence in support of the boundary contraction prediction. In that experiment, they used used a sampling paradigm to examine the spatiotemporal responses of neurons to a visual stimulus consisting of a square region whose boundaries were defined by a variety of cues (this sampling paradigm is used for many of the experiments reported in this chapter). In this paradigm, the monkey fixated on a dot on the computer monitor while the visual stimulus was presented for a short period of time (typically 400 ms for each trial). In different trials, the image was shifted spatially to a different location so that the receptive field of the neuron overlapped with different parts of the stimulus. This finely sampled the spatial response of the neuron at an interval of $0.25°$ as illustrated in Figure 9. During each presentation, the stimulus remained stationary on the screen so that the temporal evolution of the neural responsess at each location could be monitored and studied.

Figure 10 shows the response of a cell to the different locations around the borders of three types of stimuli. The normalized responses at three different time windows reveal that the half-height width of the spatial response profile around the boundary decreases over time, which is consistent with the boundary sharpening prediction. However, the absolute response of the neuron also decayed over time. This raises the possibility that this sharpening could potentially be simply be an "iceberg effect" where the response profile stays the same but simply sinks over time uniformly across space. The iceberg tip that remains visible, when normalized, might appear to be sharper. This seems unlikely, however, firstly because firing rate adaptation tends to be proportional to the firing rate (i.e. higher firing rates will adapt by a larger amount), so pure adaption would flatten out the response profile; and secondly because the reduction of half-height width of the response profile can be observed even when absolute response of the neurons remain the same. Figure 11 shows the distribution of the half-height widths of the spatial response profiles of a population of neurons at three different time windows post-stimulus onset, which demonstrates a statistically significant effect of boundary contraction. Further experiments are needed to confirm this contraction of the boundary response, for example, by testing the neurons' responses to noisy and ambiguous figures which, we conjecture, should exaggerate and prolong the sharpening process.

It is evident from these data that a neuron's response often decays over time. This is traditionally considered to be an adaptation effect (the initial burst can partly be attributed to the neuron's temporal receptive field responding to the sudden onset of the stimulus). This adaptation is not a property of the weak-membrane model considered. Adaptation is not necessarily due to the neurons losing steam metabolically, since 'habituating' neurons are capable of increasing their firing if a new and different stimulus suddenly appears (Miller and Desimone 1994). For example, when the global orientation of the contour is more consistent to the neurons' orientation preference than the local texture, the later responses of the neurons are found to be stronger than the initial responses (Figure 10 in Lee et al. 1998). The adaptation effect can potentially be understood as a 'predictive coding' or 'explaining away' effect proposed by Mumford (1992) and Rao and Ballard (1999). According to that theory, when there is a good high-level interpretation of an image region, as represented in a higher visual area, V1's responses are attenuated because they are partially explained away or replaced by the higher order, presumably simpler description. V1, as a high-resolution buffer, is still needed to represent the residual information between the high-level prediction and the image stimulus. This information will include the local texture and color and disparity, because only V1 can represent such fine details in high resolution. Furthermore, not all predictive coding requires a top-down feedback mechanism. Recurrent center-surround or lateral inhibition mechanisms within V1 can also perform 'predictive coding', just as the center-surround structure of the retinal receptive field has been considered a

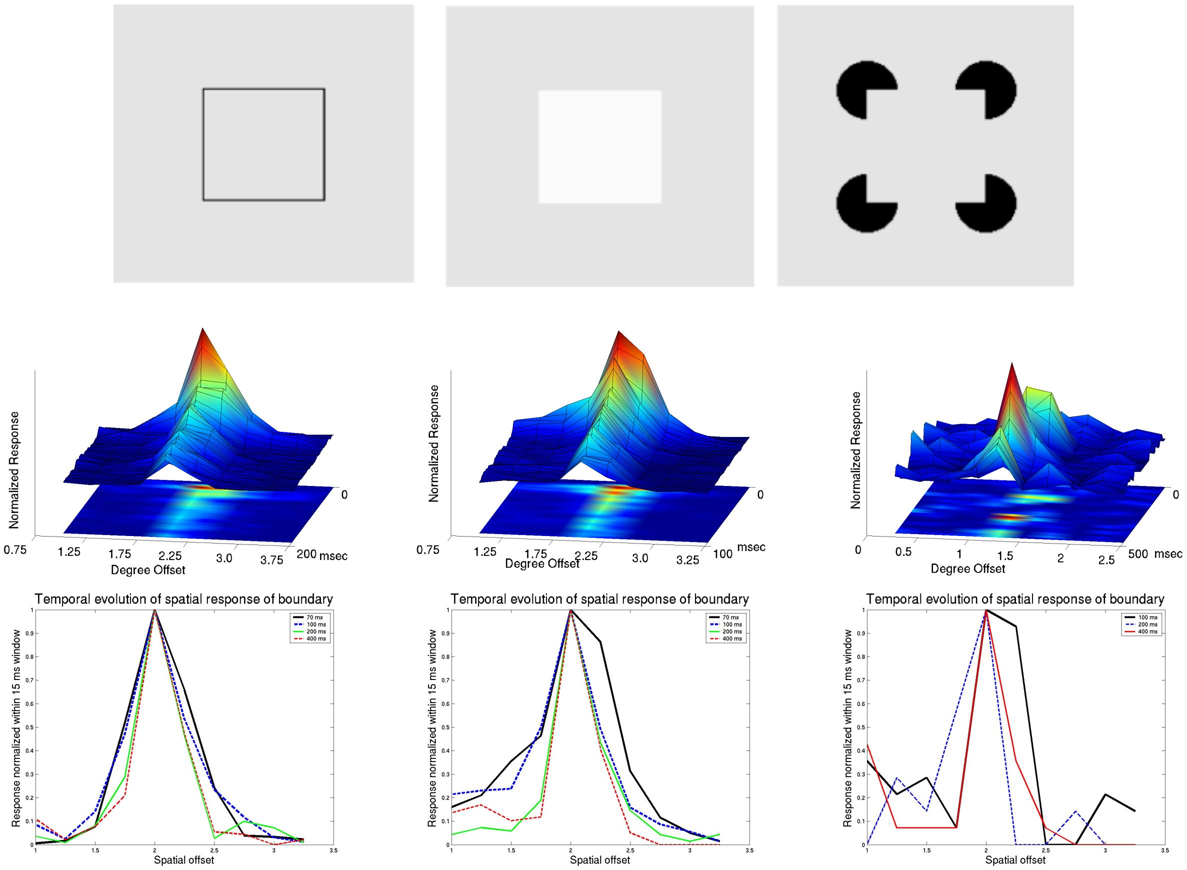
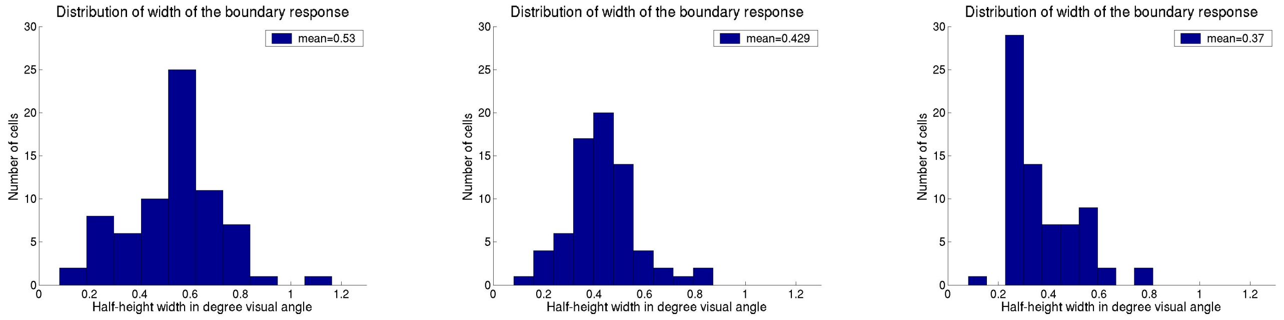


Figure 10: First row: three stimuli tested. Second row: A V1 neuron's spatiotemporal response to boundaries sampled at $1.5°$ for a line boundary, a luminance boundary and a subjective boundary (left to right panels). A gradual contraction of the spatial responses of the neurons to the boundary (at $2°$) can be observed for the line and the luminance border. Third row: When the peaks of the spatial response profiles were normalized to the same heights at different time windows, this reduction of the half-height width of the spatial response profile, a measure of sharpness of the boundary representation, becomes more evident.



0-80 ms 80-200 ms 200-400 ms

Figure 11: The distributions of half-height widths in the spatial response profiles of a population of V1 neurons in response to the boundary of the luminance square at different time periods after stimulus onset. A successive reduction in the average boundary widths of the spatial response profile can be observed over time. The mean widths of boundary at three different time periods are 0.53, 0.42 and 0.37 degree visual angles, with standard error equal to 0.022, 0.016, 0.016 respectively. The reduction of the boundary widths is statistically significant. Note that these boundary widths are necessarily over-estimated as the sampling resolution was 0.25 degree visual angles, and the eye movement fixation window was about 0.5 degree window. Both factors would have increased the spatial width of the boundary response.

form of predictive code.

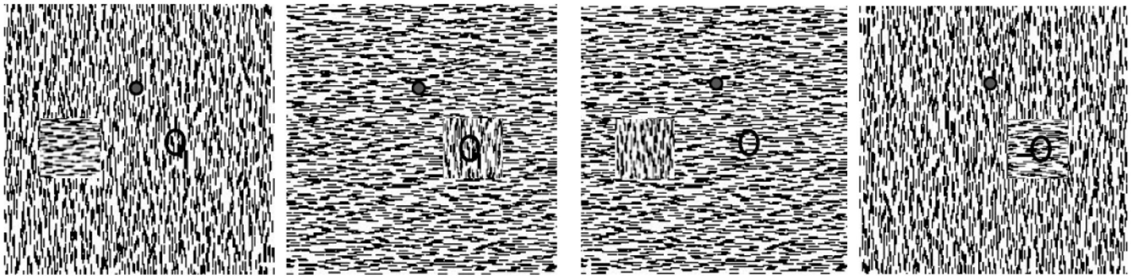
Segmentation of texture figures

The first study that systematically explored the responses of V1 neurons to a figure against the background in a visual scene was performed by Lamme (1995). In his experiments, there is a square region containing one type of texture surrounded by a background region with a different texture (Figure 12). This stimulus is ambiguous, however, in terms of physical interpretation: it could be seen as a square foreground figure in front of a textured background, or alternatively, a window (background) on a textured wall. The simplest interpretation might be that there is a single square region with an albedo (pattern) discontinuity relative to the rest of the stimuli (e.g. a patch of texture cloth sewn into a hole in a cloth with a different texture). Since the cues embodied in the test images literally cannot distinguish between these interpretations, caution must therefore be taken not to over-interpret the results. With this caveat, we can agree that the common perceptual interpretation of this square is that of a foreground figure based on its many converging figure-ground organizational cues (Palmer 1999) such as smallness, convexity and compactness.

Using a spatial sampling paradigm similar to the one we have described earlier for Lee and Nguyen's (2001) study, Lamme (1995) examined neuronal responses at different spatial locations of the image relative to the texture square. In particular, he was interested in comparing the four conditions shown Figure 12. His findings are startling. First, he found that V1 neurons respond much more strongly (on the order of 40 to 100 percent) when their receptive fields are inside the figure than when they are in the background (i.e. $B > A$, $D > C$). This occurred even though the size of the figure, ($4^\circ \times 4^\circ$), was much larger than the classical receptive field of the neurons (which is typically $0.6^\circ - 1^\circ$ for RFs from 2-4 degrees eccentricity). Second, this enhancement was uniform within the figure and terminated abruptly at the boundary.

It must be asked whether Lamme's results are significantly different from previous findings on surround suppression, which result from the well-known center-surround interactions in the primary visual cortex. It is well known that V1 neurons exhibit a phenomenon known as iso-orientation surround inhibition. That is, a neuron that prefers vertical orientation, in response to a vertical bar or vertical sinewave grating in its receptive field, will be inhibited by a surrounding vertical bars or grating (e.g. Maffei and Fiorentini 1976, Knierim and Van Essen 1992, Li and Li 1994). A cell inside a compact region of texture will receive less inhibition than a cell located in a large region of similar texture. But classical iso-orientation surround suppression theory has not anticipated an uniform enhancement within the figure, nor an abrupt discontinuity of enhancement response at the border.

Subsequent studies showed that the figure enhancement effect was weaker than Lamme described (about 15 percent enhancement, Lee et al. (1998)) or even less (Rossi et al. (2001)) for a $4^\circ \times 4^\circ$ texture square figure (see also, Marcus and Van Essen 2001). Lee et al. (1998) nevertheless confirmed that the enhancement was indeed uniform within the figure, with an abrupt discontinuity at the boundary, as shown in Figure 13a. This uniform enhancement response was obtained only when the cell's preferred orientation was *not* parallel to that of the border that it encountered along the spatial sampling line. When the cell's preferred orientation was parallel to that of the border, a significant response was observed at the border which can overshadow the small uniform enhancement observed within the texture figure (see Figure 13b).



A: Figure outside RF
Vertical texture on RF

B: Figure over RF
Vertical texture on RF

C: Figure outside RF
Horizontal texture on RF

D: Figure over RF
Horizontal texture on RF

Figure 12: Lamme measured V1 neurons' responses to four conditions. In (A) and (B) the receptive field of the neuron is stimulated by vertical texture stimuli, but the receptive field is outside the figure in condition (A) and inside the figure in (B). Cases (C) and (D) are similar, except that now the receptive field of the neuron is now stimulated by horizontal texture. Lamme found that neurons' activity is enhanced when its receptive field is inside the figure compared to when it is outside the figure. The solid dot indicates the fixation spot, while the ellipse indicates the receptive field of the neuron.

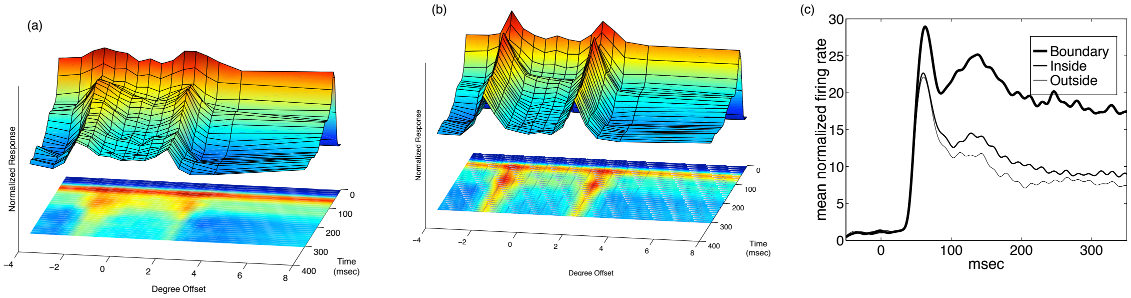


Figure 13: Population averaged combined responses (30 cells recorded individually) to texture figures in a contrasting background. Panel (a): Spatiotemporal responses when the preferred orientation of the cells is orthogonal to the orientation of the border along the sampling line. The combined response inside the figure is uniform within the figure and greater than that in the background. Panel (b): Spatiotemporal responses when the preferred orientation of the cells is parallel to the border. In addition to the enhancement of response inside the figure relative to that in the background, a strong boundary response can be observed. The X-axis indicates the offset of the RF location relative to the center of the square, which is $4^\circ \times 4^\circ$ in size. The combined response in the figure in both cases was enhanced relative to the background at about 80 msec after stimulus onset. Panel (c) gives the population PSTHs (peri-stimulus time histograms) of the neurons at three locations (on the boundary, in the figure and in the background) and shows that the elevation of the response at the boundary is over three times larger than the elevation of response inside the figure (both relative to the background). The enhancement inside the figure emerged at about 80 ms for texture figures.

Figure 13c plotted the temporal evolution of combined responses at three different locations (inside the figure, in the background, and at the boundary) to compare the magnitude of the boundary effect to the figure enhancement effect. It shows that the enhancement effect (the difference between the response inside the figure and the response outside the figure) emerged at about 80 ms after the stimulus onset, after the initial burst of responses of the neurons, and that the boundary response, when the preferred orientation of the neurons is parallel to that of the texture boundary, is 3-4 times larger than the 'figure enhancement' response. (Lee et al. 1998).

Note that in the spatiotemporal response profiles shown in Figure 13, the responses of each neuron to the two complementary cases (i.e. a vertically-textured figure in a horizontally-textured background versus a horizontally-texture figure in a vertically-textured background) are summed together at each location for each point in time. For example, the combined response within the figure is obtained by summing the response for conditions B and D in Figure 12, while the combined response in the background is the summation of the responses to A and C. By adjusting the position of the stimuli relative to the neurons, a spatiotemporal profile of the neuron's response to the images was obtained. If not summed, a vertical neuron will naturally respond more inside a vertically-textured figure simply because of its orientation tuning. Summing the response helps to isolate aspect of response that is due to figure-ground context or compactness of the region, rather than the orientation tuning of the cell (see Lee et al. 1998 for further details).

How do these experimental findings relate to the computational models described earlier? Clearly several aspects of these results are consistent with the first class of computational segmentation models. There are uniform responses within each region with a sharp discontinuity at the border (prediction 3). There are responses to the boundaries (prediction 4). But there is a significant discrepancy, since none of the segmentation models predict the delayed enhancement within the figure. Some additional mechanisms not included in the weak-membrane model must also be involved.

Lamme (1995) interpreted the enhancement effect as a signal for figure-ground segregation (i.e. a signal that can contrast a figure against the background). He argued that since figure-ground segregation is a high order perceptual construct, the delayed enhancement effect observed in V1 is likely a consequence of feedback from the extrastriate cortex. He has subsequently reported experiments showing that the enhancement effects disappeared when the extrastriate cortex was lesioned (Lamme et al. 1997) or when the monkeys were under anesthesia (Lamme et al. 1998) – see also Hupe et al. (1998). The temporal delay in enhancement as evident in the PSTH (peri-stimulus time histograms) is also consistent with the idea of feedbback, as signals from even IT would be able to propagate back to V1 within 80 ms after stimulus onset (or 40 ms after the V1 initial response) (Zipser et al. 1996). However, it is still possible that the enhancement could be computed 'bottom-up' with only feedforward and recurrent interaction within V1.

The segmentation of luminance and color figures

To demonstrate that the enhancement effect is more general, and not limited only to texture stimuli, Lee et al. (1998) examined the responses of V1 neurons to luminance figure stimuli (a dark figure in a bright background, a bright figure in a dark background, or a gray figure in a texture background) and found a similar but stronger enhancement effects (in terms of modulation ratio or percentage enhancement) for the luminance figures, even though the cells' absolute responses were much less because there were no oriented features inside their receptive fields. The enhancement was found to decrease as the size of the figure increased, but remained significant for figures as large as 7 degrees in diameter, far bigger than the size of the classical receptive field of the neurons.

Of particular interest were the responses to the grey figure in a texture background. In this test, the entire image was initially grey and then the background region was given a texture. Nevertheless, enhancement was observed even though there was no change in the stimulus within the receptive field of the neuron (only the surround was updated). Rossi et al (2001) repeated this experiment and found a progressive delay in the onset time of responses inside the grey figure as the size of the figure increased, as would be expected if the signal was propagated in from the border. They however did not believe the 'enhancement' is necessarily related to figure-ground organization.

Similar observations have also been made for equiluminance color figures (Yan and Lee 2001). In this experiment, the entire screen was first set to a particular color, and then 300 ms later, a disc was made to appear centered at the receptive field of the neuron by changing the color of the background to the opponent color. This ensures, as for the grey figure case, that there was no change in the stimulus within or near the receptive field when the figure appears. Hence the transient response due to sudden color onset in the receptive field can be dissociated from the response due to the appearance of the figure due to background update. Figure 14 shows a typical neuron's temporal responses at the center of discs of 6 different diameters, ranging from 1 degree (border close to the receptive field) to 14 degrees (border far away from the receptive field). Even though the classical receptive field of the neuron (measured as the so-called minimum responsive area) was quite small, about 0.8° in diameter, the neuron responded to the appearance of a color contrast border 7 degrees radius away! As the disc diameter became larger, the onset of the neural responses was progressively delayed, roughly in the order of 20 ms per degree. This result is consistent with Rossi et al. (2001)'s observation on the luminance figure, and the hypothesis that signals are propagating (or diffusing) from the border to the interior surface of the figure. However, it is also possible the enhancement effect arises simultaneously in cells across the entire surface with a delay that is proportional to the size of the figure.

To resolve whether the progressive delay in onset of the enhancement is due to an increase in distance away from the border, or simply due to a larger figural size, the spatial sampling paradigm was applied to monitor the temporal emergence of the enhancement signals at different locations relative to the border for the chromatic stimuli. A red figure in a green background or vice versa, as shown in Figure 16, were tested. The findings as described below were consistent with the border propagation hypothesis than the simultaneous emerging hypothesis (see Zipser et al. 1996 for the texture stimuli). In addition, the data showed that there were three major classes of neurons. One class of neurons responded primarily to the boundaries, while a second class responded well inside the figure, even though there are no oriented features within their receptive fields. The third class responded well at the boundaries as well as inside the surface. Figure 16 shows three typical neurons which illustrate these three types of behavior.

Neuron 1 (row 1) was a cell that responded both to regions and boundaries. This cell responded much more strongly

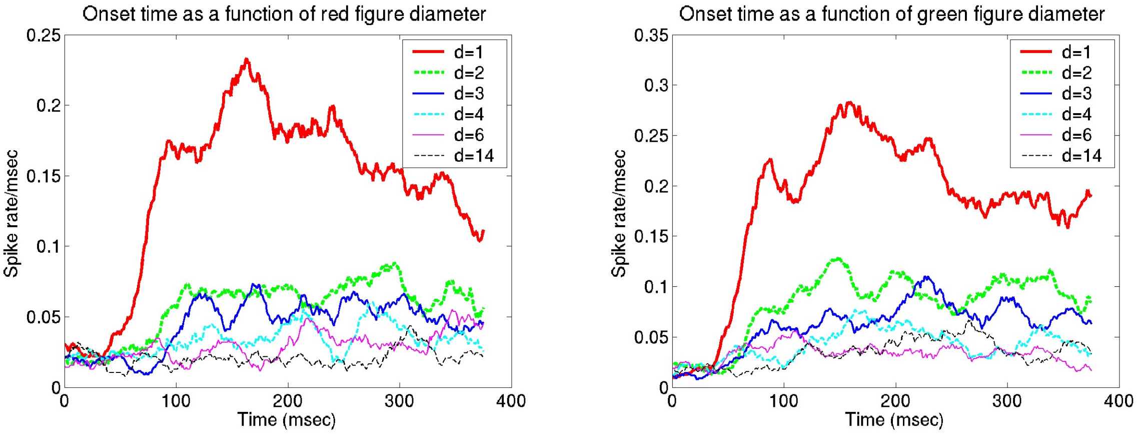


Figure 14: The PSTHs (peri-stimulus time histograms) of a neuron in response to the onset of a color change in the disk surround that makes visible a disc figure of 1,2,3,4,6,14 degrees diameter centered at the receptive field of the neuron. The interior response was shown to be progressively delayed relative to the onset of response near the border (i.e. the RF is very close to the border when the figure is only 1 degree in diameter.

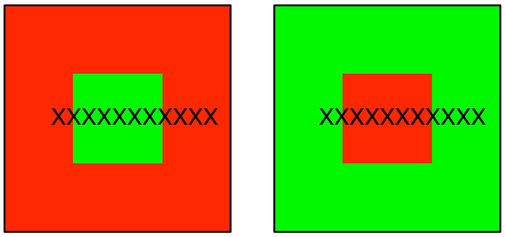


Figure 15: The static presentation paradigm: the receptive field sampling positions in the red square and the green square are labelled here by crosses.

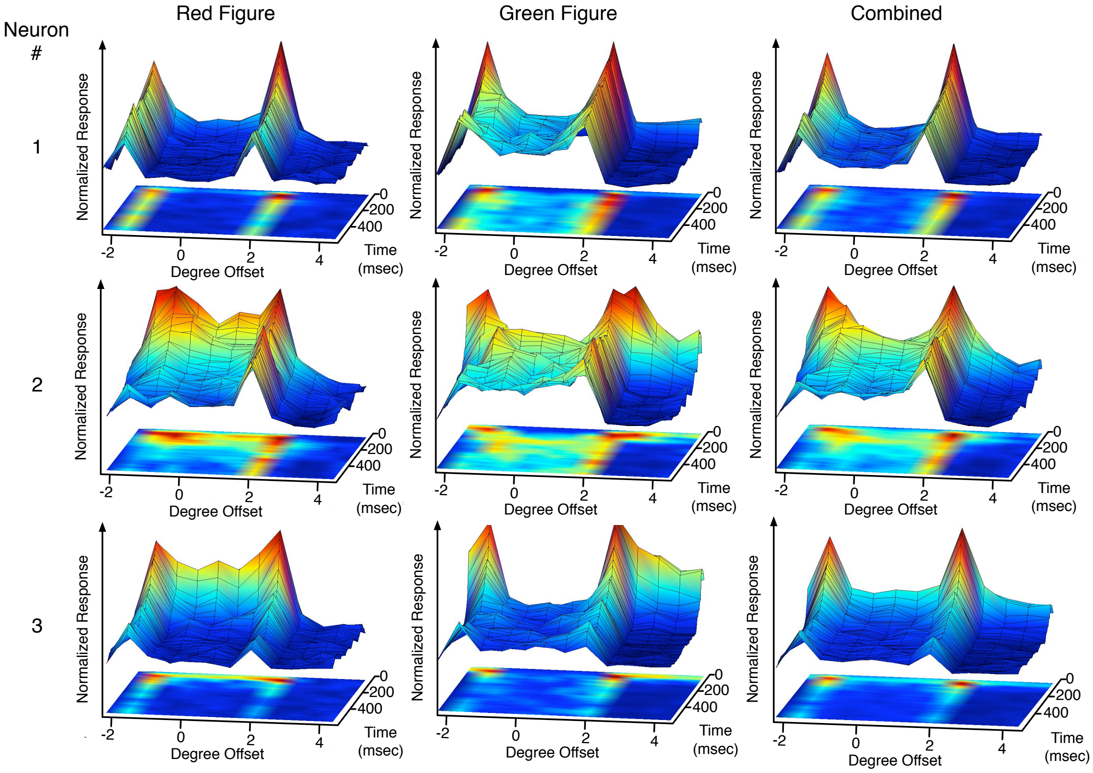


Figure 16: The temporal evolution of the responses of three different (typical) V1 neurons to different parts of the red figure in the green background (left column), and to different parts of the green figure in the red background (middle column). The right column sums the responses to the two stimuli at each spatial location to demonstrate the figure enhancement effect.

to the green figure than the response to the red background (middle column). Its response to the green background was weaker or about the same as the response to the red figure (left column). Therefore, the stronger response inside the green figure cannot be attributed to its preference to green color alone. The combined response (to red/green figure with green/red background) shows a moderate enhancement in the figure relative to the background, as well as a robust boundary response. Many cells exhibited this kind of behavior.

Neurons 2 (row 2) and 3 (row 3) preferred red color initially, as shown by their robust response to the onset of the red color in their receptive fields (left and middle columns). In fact, these neurons did not initially respond to the temporal onset of the green stimuli in their receptive fields. However, over time, the neurons started to respond more strongly inside the green figure to such an extent that the response in the green figure became stronger than the response to the red background (the color of which they initially preferred). In the case of the neuron 2, the later responses inside the figures were always stronger than the responses in the background, regardless of the color of the stimuli, with sharp discontinuities at the boundary. Its combined responses (the sum of the responses to the red and the green figures in contrast to the sum of the responses to the green and the red backgrounds) emphasized the enhancement responses inside the figure. Neuron 3 is similar to neuron 2 in exhibiting a dramatic reversal of the responses from initial preference for color to a later preference for figure (row 2, middle column). Its combined response emphasizes the boundary responses more than the enhancement of the figure.

These examples relate to prediction 1, since they show that some neurons responded more to the color region, others responded more to the boundaries, and a significant number of neurons, such as neuron 1, respond to both. In all these three neurons, the responses at the interior of the figure lag behind the responses at the figure border, particularly in the case of the green figure in a red background (middle column). This is manifested as a concave ‘wavefront’ in the responses to the green figure (middle column), which is consistent with the idea of information propagating inward from the border. This concave wavefront is not as obvious in the red figure case, perhaps because the initial chromatic change from grey (the initial screen color) to red within the receptive field provided a stronger bottom-up drive to the neurons as well. When this temporal transient in color is eliminated as in the experiment described in Figure 14, a progressive delay in the onset time of the interior response relative to the boundary response is observed.

But perhaps, the most striking aspects of these plots are the uniform responses within the figure and the abrupt discontinuity at the border of the figures. This is similar to the findings for texture, and is a clear indication of a mechanism similar to nonlinear diffusion as prescribed by the computational models. However, it is not entirely clear at this stage whether the propagated signals are related to color perception, surface representation or perceptual saliency. Some neural models such as the BCFC model proposed by Grossberg and Mingola (1985) suggest that color information in V1 is only carried by center-surround color-opponent cells, which means that both luminance and color information are available only at the contrast border. This necessitates the propagation of color and luminance signals from the boundary to the interior during the inference of color for each region. Evidence from von der Heydt et al. (2003) seems to argue against this color-diffusion hypothesis. On the other hand, the propagation of neural activity from the border is also reminiscent of the border propagation of Horn’s surface inference from shading algorithm. Further experiments are needed to clarify these issues.

The nature of the enhancement signal

The evidence presented so far is broadly consistent with the nonlinear diffusion and boundary contraction predictions of the weak-membrane class of models. The main differences are the adaptation decay, and the enhancement within the figure that have been observed in neurophysiological studies.

Both of these two discrepancies can be understood in terms of the theory of hierarchical generative model for predictive coding (Mumford 1992, Rao and Ballard 1998). The rapid decay in response after the initial outburst in response to stimulus onset can be understood mechanistically in terms of synaptic adaptation in the forward connection (Chance et al. 1998), as surround inhibition, or as feedback inhibition. Alternatively, it can be understood in terms of V1 neurons losing interest on the input because the input stimulus is being explained or ‘predicted’ by the surrounding neurons or higher level neurons (Mumford 1992, Rao and Ballard 1998). The delayed enhancement in responses could reflect the differential predictions that are offered by the contextual surround. The stimulus features in a small region or compact figure are not as well predicted by the surround stimuli, thus they are considered more surprising and appear to be more salient, which can elicit stronger attention. Features in a larger region of similar features are better predicted by the surrounding context, and hence are less salient. The delay in the enhancement response simply reflects the amount of time required to integrate the signals over each region: the larger a region, the longer the delay. From this predictive coding perspective, the enhancement response can be viewed as a measure of surprise or saliency.

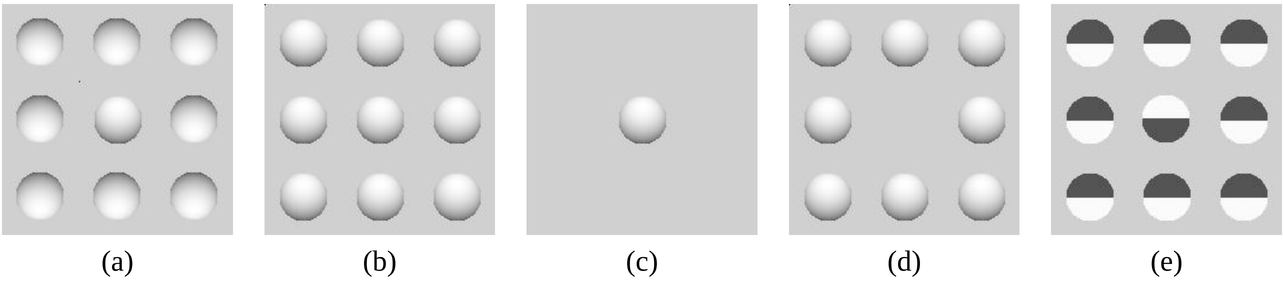


Figure 17: The basic stimuli conditions with LA (sphere with lighting from above) as the stimulus element presented to the receptive field of the neuron. The actual display contains many more stimulus elements repeated in the surround. (a) Oddball condition: RF stimulus is different from the surrounding stimulus elements. (b) Uniform condition: RF stimulus is the same as the surrounding stimulus elements. (c) Singleton condition. (d) Hole condition: RF not simulated, only the RF surround was stimulated. (e) An example of the 2D control. Oddball condition of the WA (white above) stimulus.

An alternative perspective, however, is also possible. The enhancement could be signaling a fitness measure (informally a “happiness factor”), which is proportional to the probability of how well a higher level description or model that is fitting the input data in V1. A compact figure fits the model of an object better because its smooth and compact boundary might fit the spatial and shape prior better. This explanation is more consistent with Tu et al.’s (2005) model in the second generation of segmentation theories in which a match of a model to the data can produce resonating happiness. This view is also compatible with the classical ideas of interactive activation and adaptive resonance (McClelland and Rumelhart 1981, Grossberg 1987). The delay in the enhancement is expected because matching data with top-down models would need to involve recurrent interaction with the extrastriate cortex, and this takes time. Such resonance could enhance the relevant part of the V1 representation (see also Deco and Lee 2004).

These two views are difficult to resolve at this stage, as they share many features in common, and it is likely both contain aspects of the truth. Both views involve recurrent bottom-up and top-down interaction, although the predictive coding theory includes both feedback and surround inhibition. Both views suggest that the response enhancement associated with a stimulus would be correlated related to the perceptual saliency of that stimulus.

It is possible that the computation of the ‘saliency’ of the texture and color figures can potentially be computed bottom-up using lateral inhibition. Lee et al. (2002) performed another experiment to establish that this enhancement does involve top-down processes and is quantitatively related to perceptual saliency of a region. They tested V1 and V2 neurons with a set of stimuli with different degrees of bottom-up contrast saliency and perceptual ‘pop-out’ saliency. Among them, a white-above (WA) stimulus (Figure 17e) has a high contrast and thus strong bottom-up saliency. Yet when surrounded by a set of white-below (WB) stimuli, the WA oddball is difficult to detect, thus with low perceptual pop-out saliency. On the other hand, a light-from-above (LA) stimulus (Figure 17a) has a lower stimulus contrast, but when surrounded by a set of light-from-below (LB) stimuli, the LA oddball easily pops out from the distractors.

In this experiment, the RF stimulus (the center of each display in Figure 17) was presented to the center of the classical receptive field of the neuron, while the monkey performed a simple fixation task. Note that there were 6 types of stimuli tested in the actual experiment. Each of the stimulus elements (the target and distractors) was 1 degree in diameter while the receptive field ranged in size from 0.4-0.7 degrees. The center-to-center distance between the stimulus elements is 1.5 degree visual angles. The RF stimulus could be surrounded by identical stimulus elements (uniform condition) or the opposite stimulus elements (oddball condition). Can V1 neurons distinguish the difference in the surround stimuli between the oddball condition (Figure 17a) and the uniform condition (Figure 17b)?

We would expect that because of iso-orientation surround suppression, a vertically oriented neuron will respond more strongly to a vertical bar in the receptive field when the surround is populated by horizontal bars (oddball) than when the surround is populated by vertical bars (uniform). This has been observed by Knierim and Van Essen (1992) as well as other center-surround experiments based on sinewave gratings. However, the WA and WB stimuli, and likewise the LA and LB stimuli, would stimulate neurons of the same orientation. Since the iso-orientation suppression is not sensitive to the phase of the stimuli in V1, the amount of iso-orientation suppression from the surround will be the same for both the oddball and the uniform conditions. This is indeed the case for the WA and WB stimuli.

They found that, indeed, before the monkeys had learned to utilize the stimuli in some way (e.g. making a saccade to the oddball in the stimulus), V1 neurons were not sensitive to the difference in the surround stimuli between those two conditions, for both LA/LB and WA/WB stimuli. V2 neurons, on the other hand, responded more in the oddball condition than in the uniform condition for the LA/LB stimuli but this is not observed for the WA/WB stimuli. Ramachandran

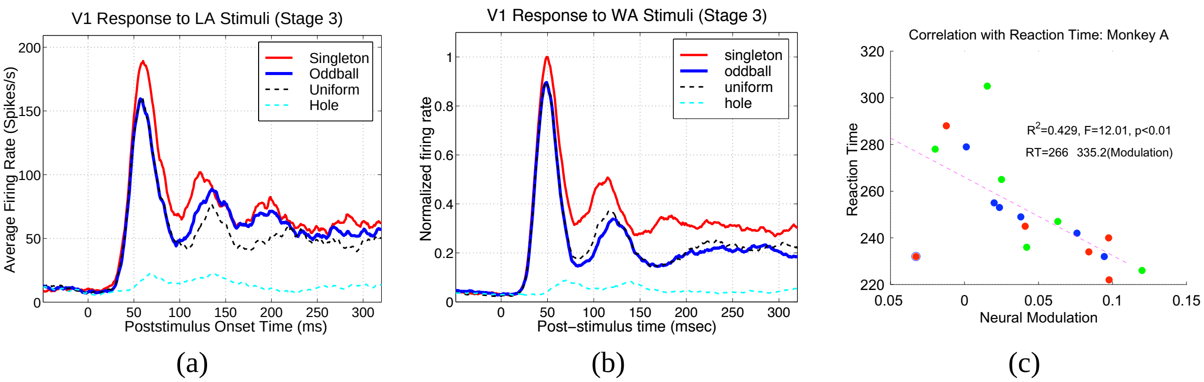


Figure 18: (a) The responses (PSTHs) of a neuron to the LA stimulus element in various contextual conditions. (b) The responses (PSTHs) of a neuron to the WA stimulus element in various contextual conditions. (c) The enhancement modulation ratio is found to be inversely correlated to the reaction time of the monkey in detecting the various stimuli with different degrees of saliency. The enhancement ratio is defined to be (A-B)/(A+B), where A is the response to the oddball condition and B is the response to the uniform condition. See Lee et al. 2002 for details.

(1988) pointed out that the LA/LB stimuli were more salient pop-out targets because they afford opposite 3D interpretation when a single lighting direction is assumed. For example, when the LA stimuli are considered convex, the LB stimuli in the same image will be considered concave (although it is also possible to perceive all the stimulus elements as convex spheres but with lighting coming from different directions). The WA/WB stimuli have stronger bottom-up contrast, and thus can drive V1 and V2 neurons more rigorously. Yet WA oddball does not jump out from the distractors because neither the WA or the WB stimuli offer 3D interpretations. Lee et al. (2002)'s initial negative finding of oddball enhancement effect in V1 but positive result in V2 might suggest that the 'predictive inhibition' mechanisms for such shape from shading stimuli may be based in V2.

Interestingly, after they trained the monkeys to make a saccade to the oddball in each stimulus (for both LA/LB and WA/WB), they found the V1 neurons started to respond better to the oddball condition than the uniform condition for the LA/LB stimuli but still not for the WA/WB stimuli, even when the monkeys were performing the same fixation task. Figure 18a shows that the singleton condition (i.e. no distractors) elicited the biggest response. The responses to the oddball and the uniform conditions are initially smaller than that for the singleton, perhaps due to surround suppression (if it is possible that lateral inhibition works at this scale). But at 100 ms, the response to the oddball condition became stronger than the response to the uniform condition, showing that the neuron was now sensitive to the difference in the surrounding stimuli between the two conditions. Observe that the latency of 100 ms is longer than the latency for the enhancement signals for the luminance/color figures (60 ms), and for the texture figures (80 ms) (see earlier figures). This longer delay probably reflects the greater complexity of the images being processed. In addition, the oddball (e.g. LA or LB) that is perceptually easier to detect is found to elicit a stronger enhancement responses than the oddball that are more difficult to detect (e.g. WA or WB). This suggests that the enhancement may reflect how salient a stimulus element is.

To establish this observation, Lee et al. (2002) compared the neural enhancement ratio against the speed and accuracy of the monkeys in detecting the oddball. They found that the enhancement ratio was inversely correlated to the reaction time of the monkeys (Figure 18c), and positively correlated with the accuracy (not shown) in the monkeys' ability in correctly locating the oddball. This finding confirms that the enhancement signal is correlated with the perceptual saliency of the target stimuli.

A simple interpretation is that the responses to the oddball were enhanced in the extrastriate cortex because the monkeys started to pay more attention to it. This attentional signal was then fed back to enhance the responses of the V1 neurons. Could this enhancement in V1 be a simple passive reflection of activities higher up, an epihenomenon that does not serve any functional purpose? First, the enhancement itself is not completely dependent on top-down attention because WA and WB stimuli fail to elicit the response even though the monkeys have been trained to detect them and should be paying as much attention to them as the LA and LB stimuli. The interaction between attention and the underlying processing of these stimuli is important. Second, Lee et al. (2002) argued from the high-resolution buffer perspective that V1 must provide the necessary segmentation boundary to constrain and contain the enhancement signals precisely. The participation of V1 is thus essential to produce a precise 'coloring' of the figure to highlight it for further processing. To prove this point, they showed that the enhancement response is limited to the oddball stimuli but not to the distractor stimuli (Lee et al. 2002).

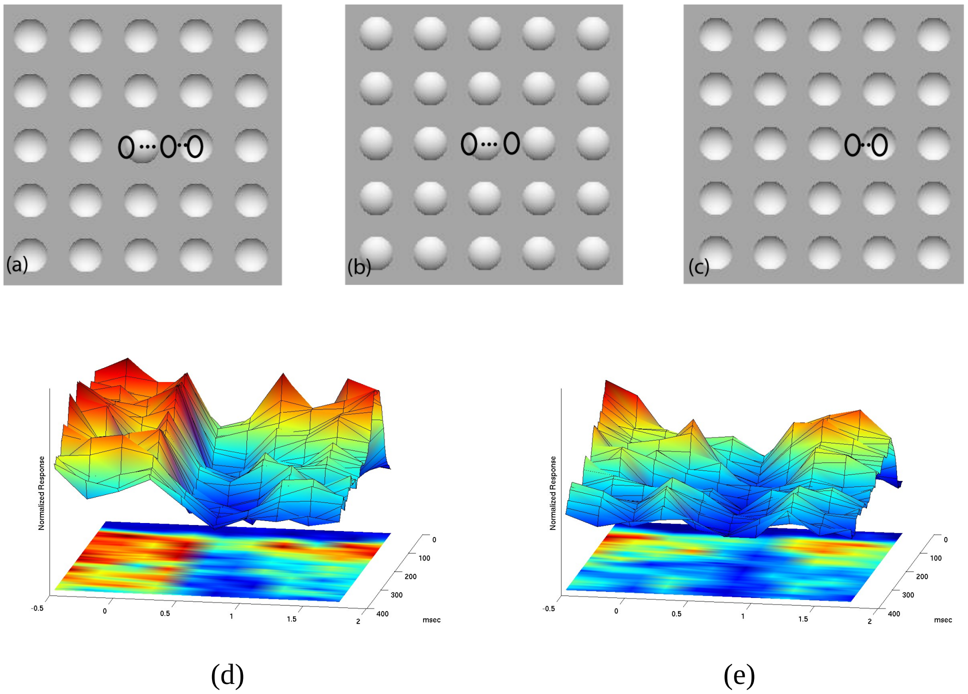


Figure 19: Spatial sampling of the shape from shading stimuli in oddball and uniform condition. (a) LA oddball condition: 10 positions sampled covering the entire oddball and part of the distractor. (b) LA uniform condition: 6 positions sampled covering the pattern of the oddball when it is surrounded by identical elements, for comparison with the response to the oddball in (a). (c) LB uniform condition: 4 positions sampled covering the pattern of the distractor when it is surrounded by identical elements, for comparison with the response to the distractor in (a). Second row shows a neuron's response to the (a) (left column) and (b) and (c) (right column). (d) The spatiotemporal response to the oddball condition (a) shows that the response was strong and the enhancement was localized inside the figure, while (e) the response to the distractor vanished over time with only a weak boundary effect.

A recent pilot study in Lee's laboratory applied the spatial sampling paradigm to examine the precise spatiotemporal responses of neurons to these stimuli. Ten positions were sampled in the LA oddball condition, including a position centered on the LA oddball and another centered on one of its immediate neighboring distractors (Figure 19a). Additionally, six positions were sampled over the LA stimulus in the uniform condition (there is no need to sample centered on a distractor, since it is the same as the target) (Figure 19b), and 4 positions were sampled over the shape of the distractor, also in the uniform condition (Figure 19. The idea is to compare the response to the oddball image against the uniform images while keeping the receptive field stimulus in each comparison constant. Figure 19d and e show the spatial activity profile of a cell in response to the oddball condition as well as the uniform conditions.

Several interesting phenomena can be observed. First, the neurons' responses inside the oddball were enhanced over time, with a sharp discontinuity at the boundary. By contrast, the responses to the distractors were initially strong, but decayed rapidly to very low levels. The later differential response between the oddball and the distractors presumably arose because the oddball was selected as the figure or target.

In addition, the discontinuity in response at the boundary of the distractors was weak and poorly localized. Similar weak discontinuity responses at the boundaries occur for both types of stimuli in the uniform condition. This suggests that the process of target selection and the process of segmentation might be tightly coupled: segmentation constrains the target enhancement, but segmentation itself also depends on target selection.

These oddball enhancement effects require several mechanisms. There is an initial stage where all the objects (the target and distractors) are detected. Next is a selection stage where the target is selected. In the case of oddball, this can be mediated by surround competition. Finally there is an enhancement of the response within the target and a strengthening of the discontinuity at its boundary. The oddball enhancement might arise from the same principle of the 'figure enhancement' observed in texture, luminance and color figures. When there is only one figure in the scene, the figure is salient and will be selected almost automatically as a target. When there are multiple figures in the scene, the less predicted one (i.e. the oddball) will pop out as the preferred target. As with the figure enhancement phenomena, this oddball enhancement or target selection phenomenon goes beyond the scope of the first class of segmentation models. It is partially consistent with theories where the recognition of objects and their segmentation are integrated (Tu et al. 2005). In this theory, bottom-up processing activates higher-level object models, which feed back to explain the early representation. As before, such ideas can be understood either in terms of the predictive coding theory or in terms of interactive activation or resonance theory. Whether the excitement in the neurons is a sign of surprise or ‘happiness’ remains to be elucidated.

Summary and Discussion

In this chapter, we have described how image segmentation can be formulated in terms of obtaining an efficient encoding of the image. We introduced a class of computational models consistent with this viewpoint and made predictions from these models. We described several experiments consistent with these conjectures which strongly suggest that neurons in V1 are involved in image segmentation. At the very least, we showed that activity in V1 is significantly more complex than the standard models of V1 as a filter-bank, or as way to efficiently encode intensity. Here is a summary of the evidence that has been discussed.

1. Evidence of region and boundary representations: Neurons in V1 responded to the interior and the boundaries of regions, even when there were no image features (oriented or otherwise) inside their classical receptive fields (Figure 14). While some cells responded solely to boundaries and others responded strongly inside regions, many cells responded to both. This suggests that the boundary and surface representations might be more distributed than the simple dual representations in boundary cells and region cells that some earlier computational models envisioned.

2. Evidence of interaction between the surface and boundary processes: While the initial spatial response profile of a neuron tends to be smooth across boundaries, the later spatial response profile of the neuron always exhibited a sharp discontinuity between regions (Figure 16). The gradual sharpening of the boundary responses coincides with the development of abrupt discontinuity in responses across regions, which suggests that the two effects might be coupled (Figure 10). However, the boundary sharpening seems to continue to progress (200 to 300 ms) even after the response discontinuity has developed (100 ms) (Figure 11).

3. Evidence of nonlinear diffusion in regional representation: There was a delay between the responses at the center of the figure versus the response close to the boundary (Figure 16). The delay is progressively larger as the size of the figure or the distance away from the border increases, suggesting that the signal is propagated (diffuses) from the border to the interior surface of the figure (Figure 14). The abrupt discontinuity in response between regions suggests diffusion is blocked by the region boundary (hence making it nonlinear).

4. Evidence of gradual sharpening of the boundary representation: Gradual sharpening of the boundary response was observed for both impulse edges (i.e. boundaries defined by lines) as well as for step edges (Figure 10). This boundary sharpening may result from the continuation mechanism as p decreases.

5. Evidence of model selection or predictive coding between different visual areas: The enhancement of responses inside the figure is not predicted by most of the segmentation models. The experimental evidence from the oddball detection experiment (Figure 19) suggested that feedback from other cortical areas is likely involved, and that top-down feedback in turn can facilitate segmentation, while segmentation helps to confine the enhancement process to precise spatial locations. Empirical evidence suggests the enhancement is proportional to perceptual saliency of the target. However, there are multiple interpretations on the nature of the enhancement effect. The predictive coding perspective suggests the enhancement is a measure of surprise, while the model selection perspective suggests the enhancement is a measure of happiness (or fitness or resonance) which emerges from the match between the selected top-down model and the V1 representation.

We propose that, in addition to furnishing a wavelet/filter bank for efficient detection and representation of image details, V1 performs image segmentation, and represents boundary locations and region properties in the activity of its neurons. This is consistent with Lee and Mumford’s earlier proposal (Lee et al. 1998, Lee and Mumford 2003) that V1 can serve as a high-resolution buffer to support *all* visual reasoning and interpretations which require high spatial precision and fine details available explicitly only in V1. We are not arguing that V1 by itself is sufficient for robust scene segmentation. In our view, early visual inference such as segmentation cannot be robust or complete without the interaction with global context and higher order visual reasoning (e.g. object recognition). Such interaction can be mediated by the feedforward/feedback loops in the visual hierarchy (Lee et al. 1998). Recent works in computer vision makes it clear that segmentation can be enhanced by object recognition (Borenstein and Ullman 2001, Yu and Shi 2004, Tu et al. 2004). Processing required to perform segmentation may be partially performed in higher-level areas using top-down feedback to V1. In this scenario, V1 performs an initial processing of the image and excites higher level models and processing in other visual areas, which then in turn feed back to V1 to refine the representation.

Direct evidence in support of V1 performing higher order process such as figure-ground segregation (border-ownership) and 3D surface encoding are either weak or unconvincing at this stage (Zipser et al. 1996, MacEvoy et al. 1998). Current compelling evidence seem to suggest that the representation of border-ownership (Zhou et al. 2000) and the representation of surface (Bakin et al. 2001, Thomas et al. 2002) might start at V2. Many color and brightness illusion that are tied to surface perception have also been mainly observed in V2 (Hung et al. 2001, Huang et al. 2002). V1's sensitivity to shape from shading information as demonstrated in Lee et al. (2002) is probably originated from feedback from V2. It is possible that the granularity of surface geometry representation is much coarser than the granularity for representing region and image properties, hence it is logical to factorize the functional representations into two separate areas. If surface inference and segmentation are to be integrated together, V1 and V2 have to work closely together.

How regions, boundaries and their models can be encoded flexibly as a whole in V1 or in the extrastriate areas remains an open question. An interesting but controversial hypothesis is that cells belonging to the same region or same contour can synchronize together, exhibiting a higher degree functional connectivity. This is related to von der Malsburgh (1981)'s binding-by-synchrony theory. The experimental evidence in support of this idea is mixed (Shadlen and Movshon 1999; Singer and Gray 1995). Emerging evidence suggests that synchrony due to similarity in bottom-up input could potentially serve as a mechanism for Gestalt grouping (Samonds and Bonds 2005), which might be related to the mechanisms for the affinity-based model (Shi and Malik 2000, Yu and Shi 2003, Sharon et al. 2001, Tolliver and Miller 2005) and Geman et al.'s (2002) compositional system. Further experiments are needed to clarify the connection between the phenomena of neuronal synchrony and figure enhancement, and their role in the encoding of the regions.

The computational models described in this chapter have provided important insights into the computational constraints and algorithms of segmentation. The biological predictions based on the weak-membrane class of models (Koch et al. 1987, Lee 1995) have motivated much of the experimental research discussed in this chapter. These first generation theories are, at best, first order approximations to true theories of segmentation since their performance on segmenting natural images is still limited. But the increasing successes of the next generation of theories, when evaluated on datasets with ground truth, suggests that the computational vision models for scene segmentation might be on the right track, and should be taken seriously in our investigation of visual systems. Insights from computer vision might prove to be instrumental in guiding our study on how V1 interprets images, extracting and representing abstract information rather than merely coding the raw input images. Theoretical framework will guide us where to look, and what to analyze. However, extra caution must be taken to guard against over-interpreting results to fit the theory in any theory-driven neurophysiological research.

The experimental results described in this chapter support some of the predictions derived from the computational models developed from computer vision. It is reassuring that the discovery of new phenomena such as region enhancement was parallel to the development of new computational algorithms in the computer vision community such as integration of top-down object recognition and bottom-up segmentation in various probabilistic inference frameworks. While current evidence, based on relatively simple image stimuli and experimental paradigm, cannot distinguish whether the enhancement within the figure is a sign of surprise (predictive coding) or happiness (model fitting and resonance), its correlation with perceptual saliency, and its various image-dependent properties provide credence to the hypothesis that image segmentation is a major computational task being performed in V1. Segmentation of the scene is a process of inference that produces a simpler and more compact description of the scene based on regions and boundaries, and their associated models. It can thus be considered as a form of efficient coding that goes beyond raw image coding.

Acknowledgement:

This paper benefits from the helpful discussion with David Mumford, David Tolliver, Gary Miller, Dan Kersten, Zili Liu, Jason Samonds, Tom Stepleton, Matthew Smith and HongJing Lu. We thank Cindy Yang, Ryan Kelly, Lei Lu, My Nguyen, Xiaogang Yan and Brian Potetz for technical assistance in the described experiments.

Tai Sing Lee is supported by NSF IIS-0413211 and NIMH MH64445 and Alan L. Yuille is supported by NSF 0413214.

References

Ambrosio L, Tortorelli VM. (1990). On the approximation of free discontinuity problems. Preprints di Matermatica. 86: Pisa, Italy: Scuola Normale Superiore.

Angelucci A, Levitt JB, Walton EJ, Hupe JM, Bullier J, Lund JS. (2002). Circuits for local and global signal integration in primary visual cortex. J Neurosci. 22:8633-8646.

Atick JJ, Redlich AN (1992) What does the retina know about natural scenes? Neural Comp., 4, 196–210.

Bakin JS, Nakayama K, Gilbert CD. (2000) Visual responses in monkey areas V1 and V2 to three-dimensional surface configurations. J Neurosci. 20:8188-8198.

Belhumeur, P (1996) A Bayesian Approach to Binocular Stereopsis. International Journal of Computer Vision, 19(3): 237-260.

Bell AJ, Sejnowski TJ. (1997) The ”independent components” of natural scenes are edge filters. Vision Res. 37(23):3327-38.

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

Borenstein E, Ullman S. (2001) Class specific top down-segmentation. Proceedings of the European Conference on Computer Vision, 110-122.

Canny J. (1986) A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intel B. 207:187-217.

Carandini M, Demb JB, Mante V, Tolhurst DJ, Dan Y, Olshausen BA, Gallant JL, Rust NC. (2005) Do we know what the early visual system does? J Neurosci. 25(46):10577-97.

Chance FS, Nelson SB, Abbott LF (1998) Synaptic depression and the temporal response characteristics of V1 cells. J Neurosci. 18(12):4785-99.

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

Daugman JG (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. J. Opt. Soc. Amer. 2(7): 1160-1169.

Deco G, Lee TS. (2004) The role of early visual cortex in visual integration: a neural model of recurrent interaction. European J Neurosci. 20:1089-1100.

Geiger D, Yuille AL (1991) A common framework for image segmentation, International Journal of Computer Vision, 6(3) 227-243.

Geman S, Geman D. (1984) Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. IEEE Trans on Patt Anal Mach Intel. 6:721-741.

Geman S, Potter D, Chi Z. (2002) Composition systems. Quarterly of Applied Mathematics, LX, 707-736.

Gregory, RL. (1970) The Intelligent Eye. London: Weidenfeld and Nicolson.

Grosof DH, Shapley RM, Hawken MJ. (1993) Macaque V1 neurons can signal ‘illusory’ contours. Nature. 365:550-552.

Grossberg S, Mingolla E. (1985) Neural dynamics of perceptual grouping: textures, boundaries, and emergent segmentations. Perception & psychophysics. 38:141-171.

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

Horn BKP. (1986) Robot Vision, MIT Press, Cambridge, MA & McGraw-Hill, New York, NY.

Huang X, MacEvoy SP, Paradiso MA. (2002) Perception of brightness and brightness illusions in the macaque monkey. J Neurosci. 22:9618-25.

Hubel DH, Wiesel TN. (1978) Functional architecture of macaque monkey visual cortex. Proc Royal Soc B (London). 198:1-59.

Hung CP, Ramsden BM, Chen LM, Roe AW. (2001) Building surfaces from borders in Areas 17 and 18 of the cat. Vision Res. 41:1389-1407.

Hupe JM, James AC, Payne BR, Lomber SG, Girard P, Bullier J. (1998) Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. Nature. 394:784-787.

Jones JP, Palmer LA. (1987). An Evaluation of the Two-dimensional Gabor filter model of simple receptive fields in the cat striate cortex. J. Neurophysiology 58, 1233-1258.

Kanizsa G. (1979) Organization in Vision. Praeger, New York.

Kapadia MK, Westheimer G, Gilbert CD. (2000) Spatial distribution of contextual interactions in primary visual cortex and in visual perception. J Neurophysiol. 84:2048-2062.

Knierim JJ, Van Essen DC. (1992) Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. J Neurophysiol. 67:961-980.

Koch C, Marroquin J, Yuille AL (1986). Analog “neuronal” networks in early vision. Proc. Natl. Acad. Sci, U.S.A. 83: 4263-4267.

Lamme VAF. (1995) The neurophysiology of figure-ground segregation in primary visual cortex. J Neurosci. 15:1605-1615.

Lamme, VAF; Zipser, K; Spekreijse, H (1998) Figure-ground activity in primary visual cortex is suppressed by anesthesia Proceedings of the national academy of sciences, USA, 95 (6): 3263-3268.

Lamme, VAF; Zipser, K; Spekreijse, H Figure-ground signals in V1 depend on extrastriate feedback Investigative ophthalmology & visual science, 38 (4) (Part 2): 4490.

Leclerc YG. (1989) Constructing Simple Stable Descriptions for Image Partitioning. International Journal of Computer Vision, 3(1): 73-102.

Lee TS, Mumford D, Yuille A. (1992) Texture segmentation by minimizing vector-valued energy functionals: the coupled-membrane model. Lecture Notes in Computer Science. 588:165-173.

Lee TS. (1995) A Bayesian framework for understanding texture segmentation in the primary visual cortex. Vision Research. 35:2643-2657.

Lee TS. (1996) Image representation using 2D Gabor wavelets. IEEE Trans Patt Anal Mach Intel. 18:959-971.

Lee TS, Mumford D, Romero R, Lamme VAF. (1998) The role of the primary visual cortex in higher level vision. Vision Research. 38:2429-2454.

Lee TS, Nguyen M. (2001) Dynamics of subjective contour formation in the early visual cortex. Proc Nat Acad Sci. 98:1907-1911.

Lee TS, Yang C, Romero R, Mumford D. (2002) Neural activity in early visual cortex reflects perceptual saliency determined by stimulus attributes and experience. Nature Neurosci. 5:589-597.

Lee TS, Mumford D. (2003) Hierarchical Bayesian inference in the visual cortex. J Opt Soc Amer A. 20:1434-1448.

Lewicki, MS, Olshausen, BA (1999) Probabilistic framework for the adaptation and comparison of image codes. J. Opt. So. Am. A 16(7): 1587-1601.

Li CY, Li W. (1994) Extensive integration field beyond the classical receptive field of cat’s striate cortical neuron–classification and tuning properties. Vision Res. 34:2577-2598.

Marcus DS, Van Essen DC. (2002) Scene segmentation and attention in primate cortical areas V1 and V2. J Neurophysiol. 88:2648-2658.

MacEvoy SP, Kim W, Paradiso MA. (1998) Integration of surface information in primary visual cortex. Nature Neurosci 1:616-620.

Maffei L, Fiorentini A. (1976) The unresponsive regions of visual cortical receptive fields. Vision Res 16:1131-1139.

Martin D, Fowlkes C, Tai D and Malik J. A Database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In proceedings International Conference of Computer Vision. Vol 2, pp 416-424. 2001.

Marr D. (1982). Vision. New Jersey, WH Freeman & Company.

Marr D, Hildreth E. (1980) Computational theory of edge detection. Proc. Royal Society, London B 207, 187-217.

McClelland JL, Rumelhart DE (1981). An interactive activation model of context effects in letter perception. Part I: An account of basic findings. Psychological review, 88: 375-407.

Miller EK, Desimone R. (1994) Parallel neuronal mechanisms for short-term memory. Science. 28: 263(5146): 520-522.

Mumford D, Shah J. (1989) Optimal approximations by piecewise smooth functions and associated variational problems. Comm. on Pure and Appl. Math., XLII:577 - 685.

Mumford D. (1992) On the computational architecture of the neocortex: II. The role of cortico-cortical loops. Biological Cybernetics. 66: 241-251.

Nitzberg M, Mumford D, Shiota T. (1993) Filtering, segmentation and depth. Springer-Verlag New York, Inc.

Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381: 607-609.

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

Ramachandran VS (1988) Perception of shape from shading. Nature 331: 163-166.

Rissanen J. (1987) Minimum Description Length Principle in Encyclopedia of Statistical Sciences, J. Wiley, New York, NY, 5: 523-527.

Rossi AF, Desimone R, Ungerleider LG. (2001) Contextual modulation in Primary Visual Cortex of Macaques. J Neurosci. 21:1698-1709.

Samonds JM, Bonds AB. (2005) Gamma oscillation maintains stimulus structure-dependent synchronization in cat visual cortex. J Neurophysiol. 93:223-236.

Shadlen MN, Movshon JA. (1999). Synchrony unbound: a critical evaluation of the temporal binding hypothesis. Neuron. 24(1):67-77, 111-25.

Sharon E, Brandt A, Basri R, (2001) Segmentation and Boundary Detection Using Multiscale Intensity Measurements, Proceedings IEEE Conference on Computer Vision and Pattern Recognition, I:469-476, Kauai, Hawaii, 2001.

Shannon, CE (1948) A mathematical theory of communication. Bell System Technical Journal, 27: 379-423 and 623-656, July and October, 1948.

Sheth BR, Sharma J, Rao SC, Sur M. (1996) Orientation maps of subjective contours in visual cortex. Science. 274:2110-2115.

Shi J, Malik J (2000) Normalized Cuts and Image Segmentation. IEEE Trans of pattern analysis and machine intelligence. 22: 8, 888-905.

Simoncelli EP (2003) Vision and the statistics of the visual environment. Curr Opin Neurobiol. 13(2):144-149. Review.

Singer W, Gray CM (1995) Visual feature integration and the temporal correlation hypothesis. Annu Rev Neurosci. 18:555-586

Thomas OM, Cumming BG, Parker AJ. (2002) A specialization for relative disparity in V2. Nature Neurosci. 5:472-478.

Tso DY, Gilbert CD, Wiesel TN (1988) Relationships between horizontal interactions and functional architecture in cat striate cortex as revealed by cross correlation analysis. Journal of neuroscience. 6:1160-1170.

Tolliver D, Miller GL (2005) Graph partitioning by spectral rounding: applications in image segmentation and clustering. Submitted to CVPR.

Tu ZW, Chen XR, Yuille AL, Zhu SC (2005) Image parsing: unifying segmentation, detection and recognition International Journal of Computer Vision, 63(2): 113-140.

Tu Z, Zhu SC (2002) Image segmentation by data-driven Markov chain Monte Carlo. IEEE Transactions on pattern analysis and machine intelligence 24(5): 657-673.

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

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

von der Heydt R, Friedman HS, Zhou H (2003) Searching for the neural mechanisms of color filling-in. in Filling-in: From Perceptual Completion to Cortical Reorganization. Ed. Pessoa L, De Weerd P. Oxford University Press, Oxford. 106-127

von der Malsburg C. (1981) The correlation theory of brain function. Internal Report. Goettingen, Germany: Max-Planck Institute for Biophysical Chemistry.

Williams LR, Jacobs DW (1997) Stochastic Completion Fields: A neural model of illusory contour shape and salience. Neural Computation 9(4): 837-858.

Winkler, G. (1995). Image Analysis, Random Fields and Dynamic Monte Carlo Methods. Springer, Berlin.

Yan, XG, Lee TS (2000), Informatics of spike trains in neuronal ensemble, Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, WC2000, 5978-65226, 1-6.

Young RA (1985) The Gaussian derivative theory of spatial vision: analysis of cortical cell receptive field line-weighting profiles. General Motors Research Technical Report, GMR-4920.

Yu, SX, Shi J. (2003) Multiclass Spectral Clustering. Proceedings of the ninth international conference of computer vision: 313-319.

Yu SX, Shi J. (2004) Segmentation given partial grouping constraints. IEEE Trans Patt Anal Machine Intel. 26:173-183.

Yuille AL, Grzywacz NM (1998) A theoretical framework for visual motion. In High-level motion processing, Ed. T. Watanabe, A Bradford Book, The MIT Press, 1998

Zhou H, Friedman HS, von der Heydt R. (2000) Coding of border ownership in monkey visual cortex. J Neurosci. 20:6594-6611.

Zhu SC, Lee TS, 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.

Zhu S, Mumford D (1997) Prior learning and Gibbs reaction diffusion. IEEE Transaction of machine intelligence and pattern recognition, 19(11):1236–1250.

Zhu SC, Yuille AL (1996) Region Competition: Unifying Snake/ balloon, Region Growing and Bayes/MDL/Energy for multi-band Image Segmentation. IEEE Trans.on Pattern Analysis and Machine Intelligence, 18(9): 884-900.

Zipser K, Lamme VAF, Schiller PH. (1996) Contextual modulation in primary visual cortex. J Neurosci. 16:7376-7389.