Network: Comput. Neural Syst. **10** (1999) 187–212. Printed in the UK PII: S0954-898X(99)96007-4

# Visual segmentation by contextual influences via intra-cortical interactions in the primary visual cortex

Zhaoping Li

Gatsby Computational Neuroscience Unit, 17 Queen Square, University College London, London WC1N 3AR, UK

Received 21 July 1998, in final form 4 January 1999

**Abstract.** Stimuli outside classical receptive fields have been shown to exert a significant influence over the activities of neurons in the primary visual cortex. We propose that contextual influences are used for pre-attentive visual segmentation. The difference between contextual influences near and far from region boundaries makes neural activities near region boundaries higher than elsewhere, making boundaries more salient for perceptual pop-out. The cortex thus computes *global* region boundaries by detecting the breakdown of homogeneity or translation invariance in the input, using *local* intra-cortical interactions mediated by the horizontal connections. This proposal is implemented in a biologically based model of V1, and demonstrated using examples of texture segmentation and figure–ground segregation. The model is also the first that performs texture or region segmentation in exactly the same neural circuit that solves the dual problem of the enhancement of contours, as is suggested by experimental observations. The computational framework in this model is simpler than previous approaches, making it implementable by V1 mechanisms, though higher-level visual mechanisms are needed to refine its output. However, it easily handles a class of segmentation problems that are known to be tricky. Its behaviour is compared with psycho-physical and physiological data on segmentation, contour enhancement, contextual influences and other phenomena such as asymmetry in visual search.

## 1. Introduction

In early stages of the visual system, individual neurons respond directly only to stimuli in their classical receptive fields (RFs) (Hubel and Wiesel 1962). These RFs sample the *local* contrast information in the input, but are too small to cover visual objects at a *global* scale. Recent experiments show that the responses of primary cortical (V1) cells are significantly influenced by stimuli nearby and beyond their classical RFs (Allman *et al* 1985, Knierim and van Essen 1992, Gilbert 1992, Kapadia *et al* 1995, Sillito *et al* 1995, Lamme 1995, Zipser *et al* 1996, Levitt and Lund 1997). These contextual influences are, in general, suppressive and depend on whether stimuli within and beyond the RFs share the same orientation (Allman *et al* 1985, Knierim and van Essen 1992, Sillito *et al* 1995, Levitt and Lund 1997). In particular, the response to an optimal bar in the RF is suppressed significantly by similarly oriented bars in the surround—iso-orientation suppression (Knierim and van Essen 1992). The suppression is reduced when the orientations of the surround bars are random or different from the bar in the RF (Knierim and van Essen 1992, Sillito *et al* 1995). However, if the surround bars are aligned with the optimal bar inside the RF to form a smooth contour, then suppression becomes facilitation (Kapadia *et al* 1995). The contextual influences are apparent within 10–20 ms after the cell's initial response (Knierim and van Essen 1992, Kapadia *et al* 1995),

0954-898X/99/020187+26$19.50  

suggesting that mechanisms within V1 itself are responsible (see the discussion later on the different time scales observed by Zipser *et al* (1996)). Horizontal intra-cortical connections linking cells with non-overlapping RFs and similar orientation preferences have been observed and hypothesized as the underlying neural substrate (Gilbert and Wiesel 1983, Rockland and Lund 1983, Gilbert 1992).

There have been some models on the underlying neural circuits (e.g. Somers *et al* 1995, Stemmler *et al* 1995) to explain the orientation- and contrast-dependent contextual influences observed physiologically. In terms of visual computation, the insight into the roles of the contextual influences is mainly limited to contour or feature linking (Allman *et al* 1985, Grossberg and Mingolla 1985, Gilbert 1992, see the discussions later and more references in Li 1998a). In a previous publication (Li 1998b), we studied how the contextual influences can indeed selectively enhance neural responses to segments of smooth contours against a noisy background. In this paper, we will show that the contextual influences serve a much more extensive computational goal of pre-attentive segmentation by enhancing neural responses to important image locations such as borders between texture regions and small figures against backgrounds, and that enhancing the contour segments is only a particular case of such a computation. The enhanced neural responses to these important image locations make them more salient, maybe even pop out, for further processing or to attract visual attention, thus serving pre-attentive segmentation. This is a computation on a *global* scale, such as on the texture regions and contours (which may represent boundaries of underlying objects) in an image, using *local* classical RF features and *local* intra-cortical interactions within a few RF sizes. Note that although the horizontal intra-cortical connections are termed long range in the literature, they are still local with respect to the whole visual field since the axons reach only a few millimetres (Gilbert 1992), or a few hypercolumns or receptive field sizes, away from the pre-synaptic cells. Although the primary visual cortex is a low-level visual area, we show below how it can play an important role in segmentation. In this paper, while we present the general framework of pre-attentive segmentation by V1 mechanisms using a model, the behaviour of the model is demonstrated mostly by examples of region segmentation and figure–ground segregation. Most details on the particular aspect of the computation—contour enhancement—can be found in a previous publication (Li 1998b).

## 2. The problem of visual segmentation

Visual segmentation is defined as locating the boundary between different image regions. There are two classes of approaches to segmentation—edge/contour based and region based (Kasturi and Jain 1991, Haralick and Shapiro 1992). For example, when regions are defined by their pixel luminance values, one can either use an edge-based approach to locate boundaries by finding (and linking) the edge or luminance contrast, which is large at the region boundaries, or use region-based approaches by comparing the luminance or the feature classification values between neighbouring image areas to find where features change. In general, regions are seldom classifiable by pixel luminances. Distracting luminance edges are often abundant within texture regions whose borders often do not correspond to any definite luminance edges or contour, making edge-based approaches difficult. Although most edge-based algorithms define edges as contrast in luminance, one may also use edges defined by contrasts in other features such as texture or motion; the resulting algorithm for edge or boundary finding requires classifying the region (e.g. texture or motion) features first (Kasturi and Jain 1991) and is then essentially a region-based approach as described

below. For region-based segmentation on general images, segmentation requires implicitly or explicitly:

(a) for every small image area, to extract and classify image feature (such as image statistics by pixel correlations, or the model parameters in the Markov random fields generating the image (Haralick and Shapiro 1992), or the outcomes from model neural filters (Bergen and Adelson 1988) or model neural interactions (Malik and Perona 1990));
(b) comparisons of the classification flags (feature values) between neighbouring image areas to locate the boundary as where the classification flags change.

In such approaches, classification is problematic and ambiguous near region boundaries where different features from different regions contribute to the feature estimate for an image area. One may also combine region- and edge-based approaches (Kasturi and Jain 1991). However, the outcomes of the two approaches often conflict with each other and combining them is seldom easy.


**Figure 1.** The two regions have the same feature values, and there are no vertical contrast edges at the vertical region border. Traditional approaches using region- or edge/contour-based approaches have difficulty in segmenting the regions.

Natural vision can easily segment the two regions in figure 1. However, this example is difficult for computer vision algorithms to segment whether one uses edge- or region-based algorithms. While edge-based segmentation algorithms would not identify any vertical luminance edges corresponding to the true boundary but many distracting edges within each region, region-based algorithms would find the same texture feature in both regions and thus no feature contrasts to locate the boundary. We propose that pre-attentive visual mechanisms in V1 locate the region boundaries by locating where homogeneities in inputs break down, and highlight such locations by higher neural responses. In principle, such a V1 algorithm corresponds to an edge/boundary-based approach in computer vision. However, it is more general than most boundary-based algorithms because the mechanism is not restricted to contrast in luminance or any particular feature such as texture, as long as input homogeneity is broken. Furthermore, as we will show below, the V1 mechanism locates the boundary without explicit feature classification and comparison between neighbouring image areas, thereby avoiding the difficulty forced by region-based approaches in the example of figure 1. In this sense, our proposed segmentation in V1 can be seen as *segmentation without classification*, in contrast with traditional approaches. This pre-attentive segmentation is in some sense the bare minimum component of segmentation, and is of low level and primitive such that not

all regions can be segmented from others, just as humans cannot pre-attentively segment, for example, an area of 'T's from an area of 'L's. This also means that higher visual areas are needed to improve and refine the segmentation outcomes from V1. However, the pre-attentive mechanism in V1 is of general purpose enough that it can handle contour enhancement (a special case of boundary-based segmentation), texture boundary location (highlight) and the pop-out of figures against a background within the same neural circuit, as will be shown below.

## 3. The principle and its model implementation

The basic computational principle is to detect region boundaries by detecting the breakdown of translation invariance in inputs. A single image region is assumed to be defined by the homogeneity or translation invariance of the statistics of the image features, no matter what the features are, for instance, whether they are coloured red or blue or whether or not the texture elements are textons (Julesz 1981). In general, this translation invariance should include cases such as the image of a surface slanted in depth, although the current implementation of the principle has not yet been generalized beyond images of fronto-parallel surfaces. Homogeneity is disrupted or broken at the boundary of a region. In pre-attentive segmentation, a mechanism signals the location of this disruption without explicitly extracting and comparing the features in image areas.

This principle is implemented in a model of V1. Without loss of generality, the model focuses on texture segmentation, i.e. segmentation without colour, motion, luminance or stereo cues. To focus on the segmentation problem, the model includes mainly layer 2–3 orientation-selective cells and ignores the mechanism by which their receptive fields are formed. Inputs to the model are images filtered by the edge- or bar-like local RFs of V1 cells. (The terms 'edge' and 'bar' will be used interchangeably.) The resulting bar inputs are merely image primitives, which are in principle like image pixel primitives and are reversibly convertible from them†. They are *not* texture feature values, such as the '+' or '$\times$' patterns in figure 5(E) and the statistics of their spatial arrangements, or the estimated densities of bars of particular orientations, from which one cannot recover the original input images. To avoid confusion, the rest of the paper uses the term 'edge' only for local luminance contrast, a boundary of a region is termed 'boundary' or 'border' which may or may not (especially for texture regions) correspond to any 'edges' in the image. The cells influence each other contextually via horizontal intra-cortical connections (Rockland and Lund 1983, Gilbert and Wiesel 1983, Gilbert 1992), transforming patterns of inputs to patterns of cell responses. If cortical interactions are translation invariant and do not induce spontaneous pattern formation (such as zebra stripes (Meinhardt 1982)) through the spontaneous breakdown of translation symmetry, then the cortical response to a homogeneous region will itself be homogeneous. However, if there is a region boundary, then two neurons, one near and another far from the boundary, will experience different contextual influences, and thus respond differently. In the model, the cortical interactions are designed (see below) such that the activities of neurons near the boundaries will be relatively higher. This makes the boundaries relatively more salient, allowing them to pop out perceptually for pre-attentive segmentation. Experiments in V1 indeed show that activity levels are robustly higher near simple texture boundaries only 10–15 ms after the initial cell responses (Nothdurft 1994, Gallant *et al* 1995).

† In practice, in the presence of noise, it is not possible to uniquely reconstruct the original pixel values in the input image from the 'edge' and 'bar' variables. For simplicity, the current implementation has not enforced this reversibility. However, the principle of no classification is adhered to by not explicitly comparing (whether by differentiation or other related techniques) the 'edge' and 'bar' values between image areas to find region boundaries.

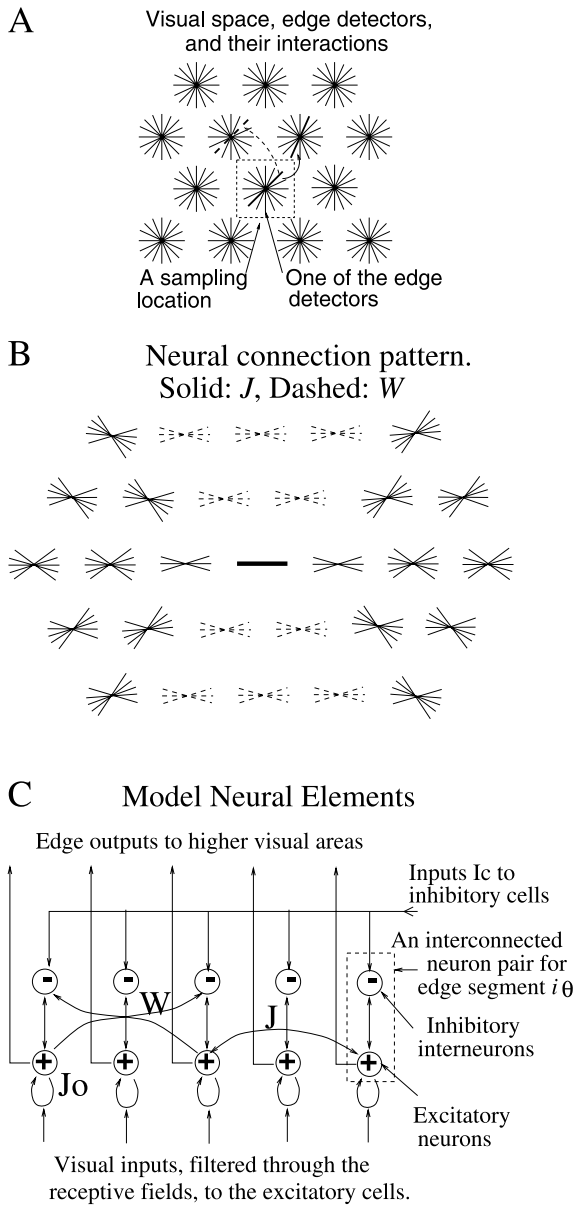


**Figure 2.** (A) Visual inputs are sampled in a discrete grid by edge/bar detectors, modelling RFs for V1 layer 2–3 cells. Each grid point has $K$ neuron pairs (see (C)), one per bar segment. All cells at a grid point share the same RF centre, but are tuned to different orientations spanning 180°, thus modelling a hypercolumn. A bar segment in one hypercolumn can interact with another in a different hypercolumn via monosynaptic excitation $J$ (the full arrow from one thick bar to another), or disynaptic inhibition $W$ (the broken arrow to a thick broken bar). See also (C). (B) A schematic of the neural connection pattern from the centre (thick full) bar to neighbouring bars within a finite distance (a few RF sizes). $J$'s contacts are shown by thin full bars. $W$'s are shown by thin broken bars. All bars have the same connection pattern, suitably translated and rotated from this one. (C) An input bar segment is associated with an interconnected pair of excitatory and inhibitory cells, each model cell models abstractly a local group of cells of the same type. The excitatory cell receives visual input and sends output $g_x(x_{i\theta})$ to higher centres. The inhibitory cell is an interneuron. The visual space has toroidal (wrap-around) boundary conditions.

Figure 2 shows the elements of the model and their interactions. At each location $i$ there is a model V1 hypercolumn composed of $K$ neuron pairs. Each pair $(i, \theta)$ has RF centre $i$ and preferred orientation $\theta = k\pi/K$ for $k = 1, 2, \ldots, K$, and is called (a neural representation of) an edge segment. Based on experimental data (White 1989, Douglas and Martin 1990), each edge segment consists of an excitatory and an inhibitory neuron that are connected with each other. Each model cell represents a collection of local cells of similar types, hence a 1:1 ratio of the number of model excitatory cells and inhibitory cells does not imply that there is a 1:1 ratio in the real cortex, for which the ratio is actually larger than 1. The excitatory cell receives the visual input; its output quantifies the response or salience of the edge segment and projects to higher visual areas. The inhibitory cells are treated as interneurons. An edge of input strength $\hat{I}_{i\beta}$ at $i$ with orientation $\beta$ in the input image contributes to $I_{i\theta}$ by an amount $\hat{I}_{i\beta}\phi(\theta - \beta)$, where $\phi(\theta - \beta) = \mathrm{e}^{-|\theta-\beta|/(\pi/8)}$ is the cell's orientation tuning curve. Based on observations by Gilbert, Lund and their colleagues (Gilbert and Wiesel 1983, Rockland and Lund 1983, Hirsch and Gilbert 1991), horizontal connections $J_{i\theta,j\theta'}$ (respectively, $W_{i\theta,j\theta'}$) mediate contextual influences via monosynaptic excitation (respectively, disynaptic inhibition) from bar $j\theta'$ to $i\theta$ which have nearby but different RF centres, $i \neq j$, and similar orientation preferences, $\theta \sim \theta'$. The membrane potentials follow the equations:

$$\dot{x}_{i\theta} = -\alpha_x x_{i\theta} - g_y(y_{i,\theta}) - \sum_{\Delta\theta\neq 0} \psi(\Delta\theta) g_y(y_{i,\theta+\Delta\theta}) + J_o g_x(x_{i\theta}) + \sum_{j\neq i,\theta'} J_{i\theta,j\theta'} g_x(x_{j\theta'}) + I_{i\theta} + I_o \tag{1}$$

$$\dot{y}_{i\theta} = -\alpha_y y_{i\theta} + g_x(x_{i\theta}) + \sum_{j\neq i,\theta'} W_{i\theta,j\theta'} g_x(x_{j\theta'}) + I_c \tag{2}$$

where $\alpha_x x_{i\theta}$ and $\alpha_y y_{i\theta}$ model the decay to the resting potential, $g_x(x)$ and $g_y(y)$ are sigmoid-like functions modelling cells' firing rates in response to membrane potentials $x$ and $y$, respectively, $\psi(\Delta\theta) \leqslant 1$ is the spread of inhibition within a hypercolumn, $J_o g_x(x_{i\theta})$ is self-excitation, $I_c$ and $I_o$ are background inputs, including noise and inputs modelling the general and local normalization of activities (Heeger 1992) (see the appendix for more details). Visual input $I_{i\theta}$ persists after onset, and initializes the activity levels $g_x(x_{i\theta})$. Equations (1) and (2) specify how the activities are then modified (effectively within one membrane time constant) by the contextual influences. Depending on the visual stimuli, the system often settles into an oscillatory state (Gray and Singer 1989, Eckhorn *et al* 1988), a common intrinsic property of a population of recurrently connected excitatory and inhibitory cells. Temporal averages of $g_x(x_{i\theta})$ over several oscillation cycles (about 12–24 membrane time constants) are used as the model's output. If the maxima over time of the responses of the cells were used instead as the model's output, the boundary effects shown in this paper would usually be stronger. That different regions occupy different oscillation phases could be exploited for segmentation (Li 1998b), although we do not do so here. The nature of the computation performed by the model is determined largely by the horizontal connections $J$ and $W$.

For view-point invariance, the connections are local, and translation and rotation invariant (figure 2(B)), i.e. every pyramidal cell has the same horizontal connection pattern in its egocentric reference frame. The synaptic weights are designed for the segmentation task, while staying consistent with experimental observations (Rockland and Lund 1983, Gilbert and Wiesel 1983, Hirsch and Gilbert 1991, Weliky *et al* 1995). In particular, $J$ and $W$ are chosen to satisfy the following three conditions (Li 1998a):

(1) the system should not generate patterns spontaneously, i.e. homogeneous input images should give homogeneous outputs, so that no illusory borders will be formed within a single region;

(2) neurons at region borders should give relatively higher responses; and
(3) the same neural circuit should perform contour enhancement.

Condition (3) is not only required by physiological facts (Knierim and van Essen 1992, Kapadia *et al* 1995), but is also desirable because regions and their boundary contours are complementary. The qualitative structure of the connection pattern satisfying the conditions resembles a 'bow tie': $J$ predominantly links cells with aligned RFs for contour enhancement, and $W$ predominantly links cells with non-aligned RFs for surround suppression (figure 2(B)). Both $J$ and $W$ link cells with similar orientation preferences, as observed experimentally (Rockland and Lund 1983, Gilbert and Wiesel 1983, Hirsch and Gilbert 1991, Weliky *et al* 1995). Since this qualitative connection pattern is derived from the three conditions given above, it is thus a prediction of our computational requirements. The connection strength between cells decays with distance between the RFs, and is zero between cells separated by long distances (see the appendix). Once the choice of the connection strengths is set by the three conditions, they are not varied in the application of our model to any visual input patterns. This is because, by our computational design, varying the connections beyond the bound of the three conditions will inevitably destroy the computational properties of the model. For instance, if the connections are changed such that condition (1) is no longer met, then the model will produce output highlights at image locations where there should be none, and these hallucinated highlights would compete with and even overwhelm the highlights generated by the actual region borders or pop-out targets at other locations in the image.

Mean field techniques and dynamic stability analysis (shown in the appendix) are used to design the horizontal connections that ensure the three conditions. Conditions (1) and (2) are strictly met only for (the particularly homogeneous) inputs $I_{i\theta}$ within a region that are independent of $i$, i.e. exactly the same inputs are received at each grid point. When a region receives more complex input texture patterns such as in stochastic or sparse texture regions (e.g. those in figure 5), conditions (1) and (2) are often met but not guaranteed. This is not necessarily a flaw in this model, since it is not clear whether conditions (1) and (2) can always be met for any type of homogeneous inputs within a region under the hardware constraints of the model or the cortex. This is consistent with the observations that sometimes a texture region does not pop out of a background pre-attentively in human vision (Bergen 1991). A range of quantitatively different connection patterns can meet the three restrictive conditions. Of course, this range depends on the particular structure and parameters of the model such as its receptive field sampling density. This makes the model quantitatively imprecise compared to physiological and psycho-physical observations (see discussions later). The quantitative model parameters used for horizontal connections and neural elements to reproduce all our results are also listed in the appendix.

## 4. Performance of the model

Figures 3–10 show the results of applying the model to a variety of input patterns. With a few exceptions, the input value $\hat{I}_{i\theta}$ is the same for all visible bars in each example so that any differences in the responses $g_x(x_{i\theta})$ to the bars are solely due to the effects of the intra-cortical interactions. The exceptions are the input taken from a photograph (figure 10), an input in figure 9(C) to test the robustness of texture segmentation to random input strength variations, and the input in figure 8(D), which models an experiment on contour enhancement (Kapadia *et al* 1995). The differences in the outputs, which are interpreted as differences in saliencies, are significant about one membrane time constant after the initial neural response. This agrees with experimental observations (Knierim and van Essen 1992, Kapadia *et al* 1995, Gallant

*et al* 1995) if this time constant is assumed to be of the order of 10 ms. The actual value $\hat{I}_{i\theta}$ used in all examples is chosen to mimic the corresponding experimental conditions. In this model the dynamic range is $\hat{I}_{i\theta} = (1.0, 4.0)$ for an isolated bar to drive the excitatory neuron from threshold activation to saturation. Hence, we use $\hat{I}_{i\theta} = 1.2$, 2.0 and 3.5 for low-, intermediate- and high-contrast input conditions used in experiments. Low input levels are used to demonstrate contour enhancement—the visible bars in figure 6(C) and the target bar in figure 8(D) (Kapadia *et al* 1995, Polat and Sagi 1993, Field *et al* 1993, Kovacs and Julesz 1993). Intermediate or high levels are used for all visible bars in texture segmentation and figure–ground pop-out examples (figures 3–5, 6(A, B), 7 and 9). High input levels are used for all visible bars in figures 8(A–C) and the contextual (background) bars in figure 8(D) to model the high-contrast conditions used in physiological experiments that study contextual influence from textured and/or contour backgrounds (Knierim and van Essen 1992, Kapadia *et al* 1995). The input $I_{i\theta}$ from a photographic image (figure 10) is different for different $i\theta$ with $I_{i\theta} \leqslant 3.0$. The output saliency $g_x(x_{i\theta})$ ranges between 0 and 1. The widths of the bars in the figures are proportional to input or output strengths. The same model parameters (e.g. the dependence of the synaptic weights on distances and orientations, the thresholds and gains in the functions $g_x(\cdot)$ and $g_y(\cdot)$, and the level of input noises in $I_o$) are used for all the examples whether it is for the texture segmentation, contour enhancement, figure–ground segregation, or combinations of them. The only differences between different examples are the differences in the model inputs $I_{i\theta}$ and possibly the different image grid structure (Manhattan or hexagonal grids) for better input sampling. We visualize the most salient outputs in some figures by plotting the bars that induce response levels higher than some threshold. V1, of course, does not threshold its outputs, we use the threshold only for display purposes. It is possible that higher visual centres could attend to a selected portion of the input by such thresholds, with an increasing value of the threshold for decreasing areas of the attended input.

Figure 3(A) shows a sample input containing two regions. Figure 3(B) shows the model output. Note that the plotted region is only a small part of, and extends continuously to, a larger image. This is the case for all figures in this paper except figure 10. Figure 3(C) plots the saliency $S(c)$ averaged over the bars in each column $c$ in figure 3(B), indicating that the most salient bars are indeed near the region boundary. Figure 3(D) confirms that the boundary can be identified by thresholding the output activities using a threshold, denoted as, say, $thre = 0.5$ in figure 3(D), the fraction of the highest output $\max_{i\theta}\{g_x(x_{i\theta})\}$ in the image. To quantify the relative saliency of the boundary, define the net saliency at each grid point $i$ to be that of the most activated bar ($\max_\theta\{g_x(x_{i\theta})\}$), let $S_{\text{peak}}$ be the average saliency across the most salient grid column parallel and near the boundary, and $\bar{S}$ and $\sigma_s$ be the mean and standard deviation in the saliencies of all locations including the boundary. Define ($r \equiv S_{\text{peak}}/\bar{S}$, $z \equiv (S_{\text{peak}} - \bar{S})/\sigma_s$). A salient boundary should give large values $(r, z)$. One expects that at least one of $r$ and $z$ should be comfortably larger than 1 for the boundaries to be adequately salient. In figure 3, $(r, z) = (3.7, 4.0)$. Note that the vertical bars near the boundary are more salient than the horizontal ones. This is because the vertical bars run parallel to the boundary, and are therefore specially enhanced through the contour enhancement effect of the contextual influences. This is related to the psycho-physical observation that texture boundaries are stronger when the texture elements on one side of them are parallel to the boundaries (Wolfson and Landy 1994). Figure 4(A) shows an example with the same orientation contrast (90°) at the boundary as in figure 3, but for different orientations of the texture bars. Here the saliency values distribute symmetrically across the boundary and the boundary strength is a little weaker. Figures 3 and 4(A) together predict that the neural response near a texture border is tuned to the border orientation relative to the optimal orientation of the bar within the RF, given an orientation contrast at the border, and that the optimal orientation of the border is the same as that of the bar within the RF.

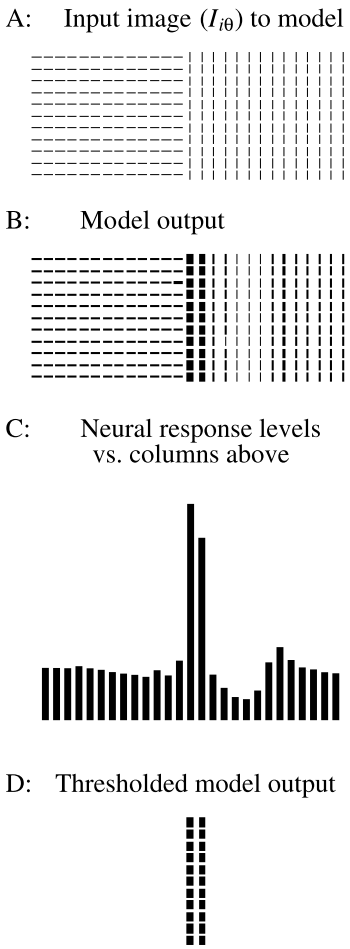


**Figure 3.** An example of the segmentation performance of the model. (A) Input $\hat{I}_{i\theta}$ consists of two regions; each visible bar has the same input strength. (B) Model output for (A), showing non-uniform output strengths (temporal averages of $g_x(x_{i\theta})$) for the edges. The input and output strengths are proportional to the bar widths. Because of the noise in the system, the saliencies of the bars in the same column are not exactly the same, this is also the case in other figures. (C) Output strengths (saliencies) averaged within each column versus lateral locations of the columns in (B), with the bar lengths proportional to the corresponding averaged output strengths. (D) The thresholded output from (B) for illustration, *thre* $= 0.5$. Boundary saliency measures $(r, z) = (3.7, 4.0)$.

Figure 4 shows examples using other orientations of the texture bars. The boundary strength decreases with decreasing orientation contrast at the region border. It is very weak when the orientation contrast is only 15° (figure 4(C))—here translation invariance in input is only weakly broken, making the boundary very difficult to detect pre-attentively. Note also that the most salient location in an image may not be exactly on the boundary (figure 4(C), see also figure 5(C)), this should lead to a bias in the estimation of the border location, and this can also be tested experimentally. This reinforces the point that outputs from pre-attentive segmentation need to be processed further by the visual system.

This model also copes well with textures defined by complex or stochastic patterns (figure 5). In figures 5(A–C) the neighbouring regions can be segmented even though they have the same bar primitives and densities. In particular, the two regions in figure 5(A) (or figure 5(C)) have exactly the same features, just like that in figure 1, and would be difficult to segment using traditional approaches. The model detects the true boundary in figure 5(D) even though there are orientation contrasts at many locations within each region. These examples work because the model is designed to detect where input homogeneity or translation invariance breaks down. In the example of figure 5(D), any particular vertical bar, within the right region far enough away from the border, has exactly the same contextual surround as any other vertical

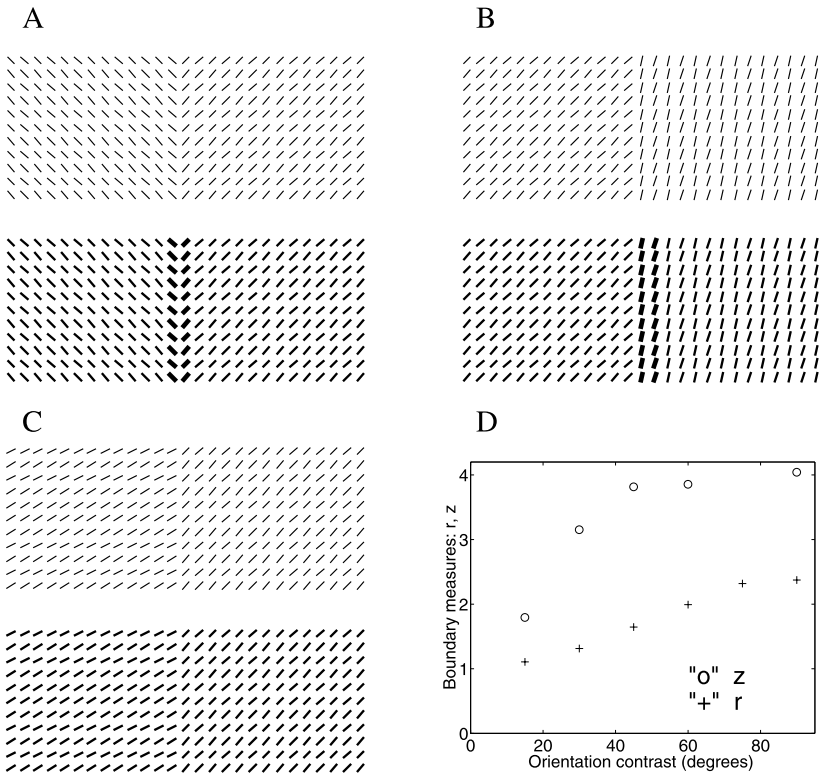


**Figure 4.** (A), (B), (C) Additional examples of model segmentation. Each is an input image as in figure 3(A) followed immediately below by the corresponding model outputs as in figure 3(B). In (A), (B), (C), respectively, the boundary measures are: $(r, z) = (1.4, 3.4)$, $(r, z) = (1.7, 3.7)$, $(r, z) = (1.03, 0.78)$. (D) The model segmentation performance, as measured by the boundary measure $r, z$ (indicated by '+' and '○', respectively), versus the orientation contrast at the texture border. Each data point is the average of all possible pairs of orientations of the two textures given an orientation contrast at the border. Again, each plotted region is only a small part of a larger extended image. The most salient column in (B) is, in fact, not exactly on the boundary, though the boundary column is only 6% less salient than its neighbour on the right, and $\sim 70\%$ more salient than areas away from the boundary. (C) contains two regions whose bar elements differ only slightly in orientation, the boundary bars are among the most salient ones, but only a very small fraction more salient than other bars (imperceptible in the line widths plotted in the output).

bar away from the border, i.e. they are all within the homogeneous or translation-invariant part of the region. Thus no one of such vertical bars will induce a higher response than any other, since they have the same direct input and the same contextual inputs. The same argument applies to the oblique bars or horizontal bars far away from the border in figure 5(D) as well as in figures 5(A–C). However, the bars at or near the border do not have the same contextual surround (i.e. contextual inputs) as those of the other bars, i.e. the homogeneity is truly broken, and thus they will induce a different response. By design, the border response will be higher. In other words, the model, with its translation-invariant horizontal connection pattern, only detects where input homogeneity breaks down, and the pattern complexity within a region does not matter as long as the region is homogeneous. The stochasticity in figure 5(E) implies a non-uniform response pattern even within each region. In that case, the border induces the

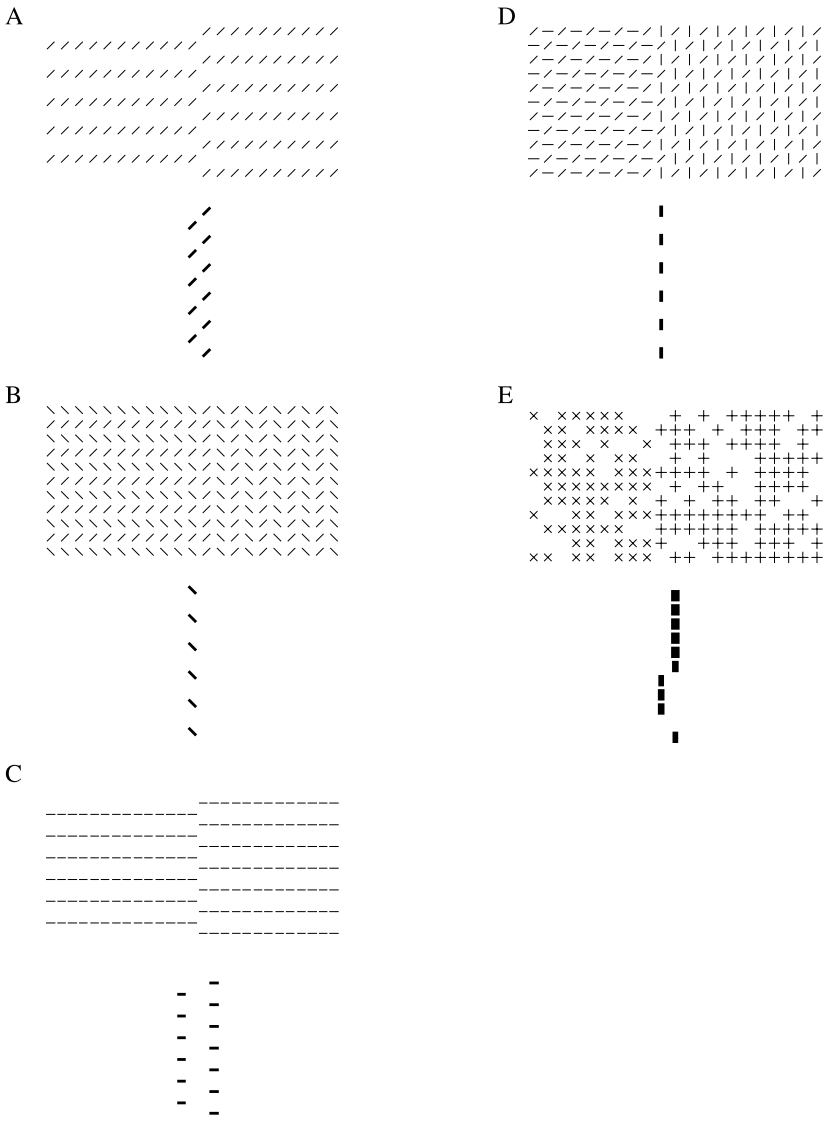


**Figure 5.** (A), (B), (C), (D) Model performance on regions with complex texture elements, and (E) regions with stochastic texture elements. Each plot is the model input ($\hat{I}_{i\theta}$) followed immediately below by the output ($g_x(x_{i\theta})$) highlights. For (A), (B), (C), (D), (E), respectively, the boundary measures are $(r, z) = (1.13, 3.5), (1.1, 1.5), (1.06, 2.6), (1.4, 2.9), (2.3, 3.3)$, the thresholds to generate the output highlights are *thre* $= 0.91, 0.9, 0.94, 0.85, 0.56$. In (C), even though the boundary saliency is only a fraction higher than others, this fraction is significant since it is $z = 2.6$ deviations away from the mean saliency.

highest response because it is where homogeneity breaks most strongly (see figure 9 for more examples of input randomness). The vertical array of highlights in figure 5(C) may correspond to the physiologically observed responses of V1 to illusory contours, i.e. the perceptual vertical contour without its corresponding vertical contour in the image (Grosof *et al* 1993). Note, however, that the most salient output in figure 5(C) is shifted a little away from the region

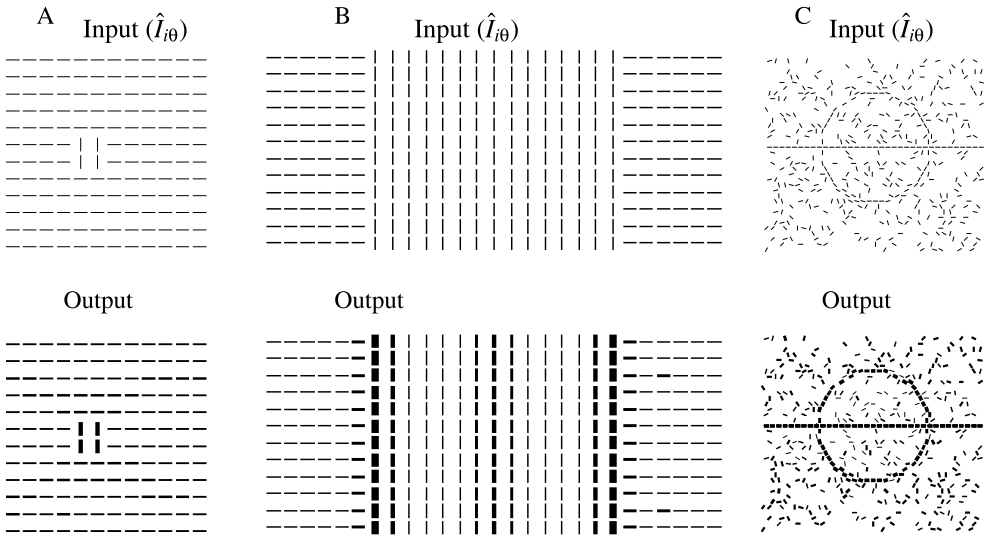


**Figure 6.** Model behaviour for other types of inputs. (A) A small region pops out since all parts of it belong to the boundary. The figure saliency is 0.336, which is 2.42 times the average ground saliency. (B) When the figure has a small and finite width in a figure–ground input, the figure borders have the highest saliencies, and the centre of the figure sometimes also shows a secondary saliency peak, as seen in the experiments by Lee *et al* (1998). The border saliency averages 0.56, the figure axis saliency is $\sim 0.28$, and the average background saliency is $\sim 0.13$. (C) Exactly the same model circuit (and parameters) performs contour enhancement. The input strength is $\hat{I}_{i\theta} = 1.2$. The contour segments' saliencies are $0.42 \pm 0.03$, and the background elements' saliencies are $0.18 \pm 0.08$.

border. This shift arises in the model as follows. Each horizontal bar segment receives collinear excitation from its neighbours. The most effective collinear bars are to its left and right in the same line. Less effective ones are in the neighbouring lines and displaced from it somewhat obliquely, but still with sufficient collinearity. Each bar also receives iso-orientation suppression from neighbours exactly or approximately above and below it in the neighbouring lines. The bars near the border have fewer collinear excitatory neighbours in the same line, but also fewer iso-orientation, suppressive, neighbours in the neighbouring lines. These two conflicting changes in contextual influences at the border not only reduce the summed strength of the border highlight, but also cause the highlight to be shifted away from the border. In fact, this example does not work well in the model if we use inputs of lower contrast. It may relate to the findings and arguments that V1 does not respond to illusory contours as well as V2 (von der Heydt *et al* 1984), and that the responses of the V1 neurons may indeed be border insensitive in this input pattern—further physiological investigations on this matter would be useful. It may also be a fault of the model, which, for simplicity, has not included end-stopped cells which are observed physiologically.

When a region is very small, all parts of it belong to the boundary and it pops out from the background, as in figure 6(A). When the figure region is a bit larger, but still of finite size, the centre of the figure sometimes shows a saliency peak, which is smaller than those of the figure borders. This is observed in the physiological experiments of Lee *et al* (1998), who argued that such peaks form a medial axis representation of objects, tracing out their skeletons (Blum 1973). These secondary peaks are a consequence of the *finite-size effect* of a region, and do not contradict the computational requirement for homogeneous output strength within a

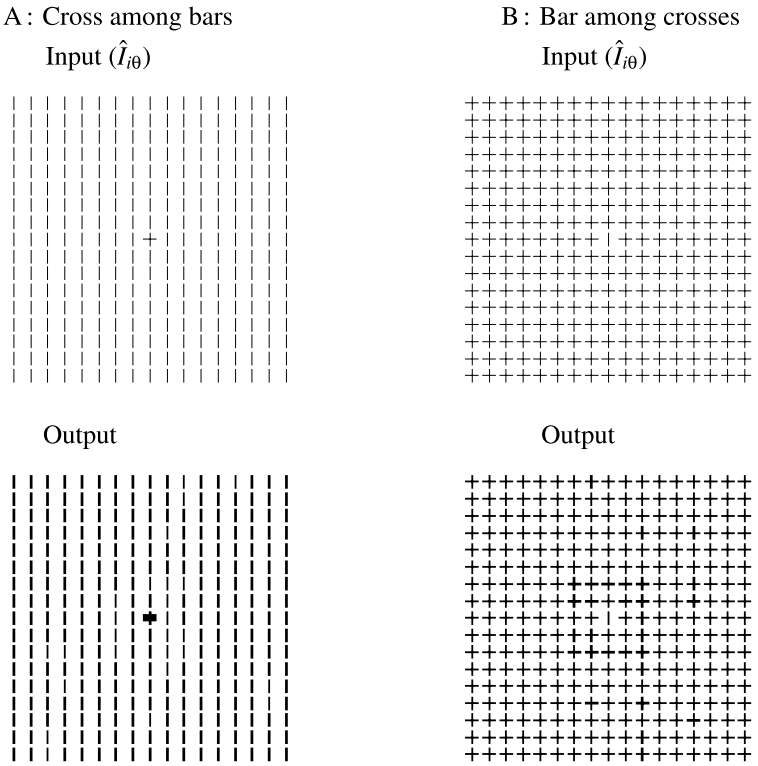


**Figure 7.** Asymmetry in pop-out strength. (A) The cross is 3.4 times as salient (measured as the saliency of the horizontal bar in the cross) as the average background. (B) The area near the central vertical bar is the most salient part in the image, and is no more than 1.2 times as salient as the average background. The target bar itself is actually a bit less salient than the average background.

homogeneous and *infinitely large* input region. The area at and near the border is by definition not homogeneous, and the effect of this spills over into the region by a small distance, of the order of the longest horizontal connection length. Often this causes a ripple effect near the border—the saliency is highest near the border, it then undergoes quickly decaying oscillations before reaching a homogeneous level into the region. Hence, the medial axis effect is most apparent when the size of the figure is about twice the wavelength of the ripple wave near the border. Indeed, Lee *et al* (1998) observed such an effect to appear only for certain figure sizes. However, in contrast to their proposal that this effect is caused by feedback from higher visual centres, our model suggests that V1 mechanisms alone could be mainly responsible. Figure 6(C) confirms that exactly the same model, with the same elements and parameters, can also highlight contours against a noisy background—another example of a breakdown of translation invariance.

Our model also accounts for the asymmetry in pop-out strength observed in psychophysics (Treisman and Gormican 1988), i.e. item A pops out among item B more easily than vice versa. Figure 7 demonstrates such an example where a cross among bars pops out much more readily than a bar among crosses. Other typical examples of asymmetry observed in psychophysics can also be simulated in this model (Li 1999a). Such asymmetry is quite natural given the basic idea underlying the model—the nature of breakdown of translation invariance in the input is quite different depending on which one is the figure or background.

The model replicates the results of physiological experiments on contextual influences from beyond the classical receptive field (Knierim and van Essen 1992, Kapadia *et al* 1995). In particular, figures 8(A–D) demonstrate that the response of a neuron to a bar of preferred orientation in its receptive field is suppressed by a textured surround, but enhanced by collinear contextual bars that form a line. As observed experimentally (Knierim and van Essen 1992), suppression in the model is strongest when the surround bars are of the same orientation as the centre bar, is weaker when the surround bars have random orientations and is weakest when the surround bars are oriented orthogonally to the centre bar. The relative degree of suppression is quantitatively comparable to that of the orientation contrast cells observed physiologically (Knierim and van Essen 1992). Similarly, figure 8(D) closely simulates the enhancement effect observed physiologically (Kapadia *et al* 1995) when bars in the surround are aligned with the central bar to form a line.

In realistic situations, there are many sources of input randomnesses or noise (other than neural noise), arising either from the image sampling stage, i.e. the transformation from the grey-scale image to the filtered and sampled inputs to the pyramidal cells, or randomness in the input scenes themselves. The overall performance of our pre-attentive visual system depends on both the image sampling stage and the intra-cortical interaction stage. Since our model in its current implementation has a much sparser sampling than that in human vision, its problem of sampling noise (e.g. aliasing) is much more serious. A better visual system, such as the human visual system, should have a many-fold over-complete and much denser sampling, i.e. the sizes or the tuning widths of the receptive fields in space, orientation and spatial frequency, etc are much larger than the distance (in space, orientation and spatial frequency, etc) between two nearby sampling nodes. With such dense sampling, there will be little aliasing (in space, orientation, scale) and thus the sampling noise can be negligible compared to the noise in the visual scene itself. So far we have omitted the image sampling stage in the applications of our model by directly applying the accurate inputs $I_{i\theta}$ to the pyramidal cells. In this way, we can demonstrate the power of the intra-cortical interaction by isolating its effects, without having to build an almost perfect and many-fold over-complete sampling stage and thus many more pyramidal cells in the model, whose computer simulation would overwhelm my currently available computing power. By introducing noise and variations in input $I_{i\theta}$ to the pyramidal cells, figure 9 shows that the texture boundary or the pop-out strength decreases if the orientations and/or the input strengths of the texture elements are somewhat random or the spacing between the elements increases. Boundary detection or figure pop-out in the model is difficult when orientation noise $>30^\circ$ or when the spacing between bar elements is more than four or five grid points (or texture element sizes). The qualitative and quantitative performances of the model on the cut-off orientation contrast, orientation noise and bar spacings compare reasonably well with human performance on segmentation related tasks (Nothdurft 1985, 1991, Li 1999b). This means that, if the input randomness is caused by the unsatisfactory sampling and filtering stage (the transform from grey-scale images to $I_{i\theta}$) in the model, then the degradation in the model performance is comparable with that of the human performance with equivalently degraded input scenes. The graceful degradation of the model's segmentation performance with the noise in input strengths, orientations and positions of the image elements (figure 5(E)) generally applies to all kinds of input patterns, including the ones in this paper. In fact, the temporal behaviour of the model, e.g. the neural response synchrony between elements of the same texture patch in figure 9 (although this is not used for segmentation purposes in this paper), changes only very mildly or negligibly when the added randomness is not enough to destroy the segmentation performance. Figure 10 demonstrates the current model performance on a photograph. The effects of single-scale sampling and the noise arising from the sparse sampling (aliasing) are apparent in the model input image,

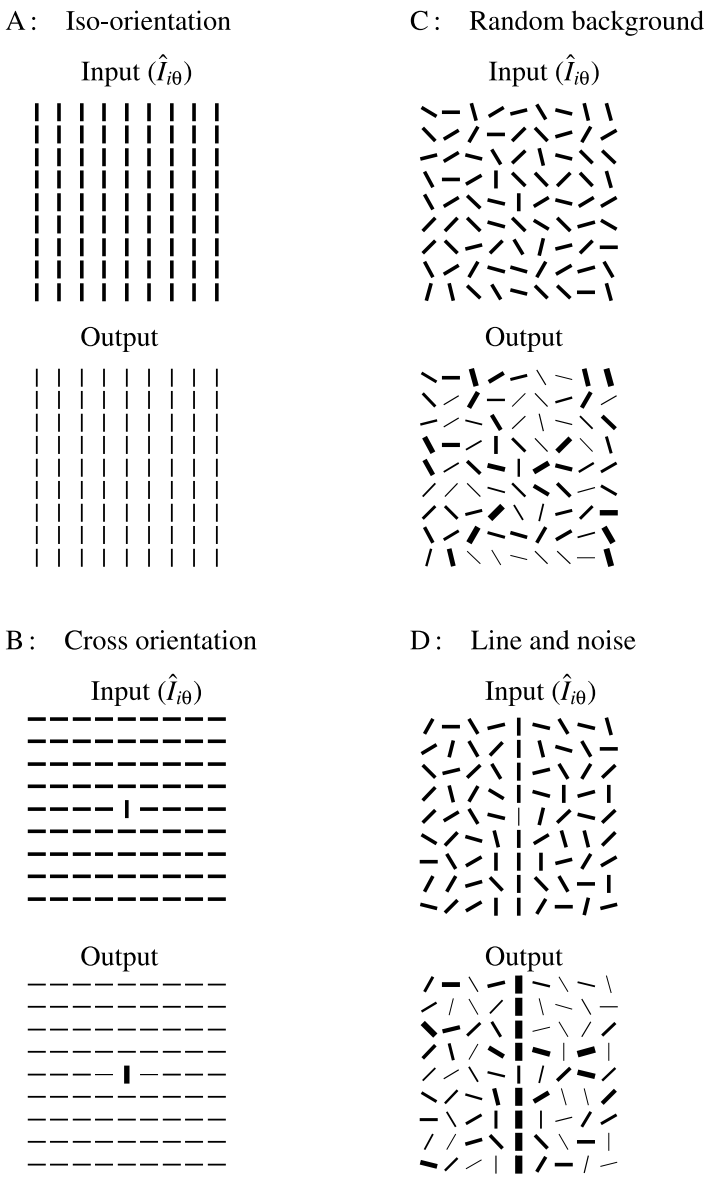


**Figure 8.** Model behaviour under inputs resembling those in physiological experiments. The input stimuli are composed of a vertical (target) bar at the centre surrounded by various contextual stimuli. All the visible bars have a high-contrast input $\hat{I}_{i\theta} = 3.5$, except for the target bar in (D) where $\hat{I}_{i\theta} = 1.2$ is near threshold. (A), (B), (C) simulate the experiments of Knierim and van Essen (1992) where a stimulus bar is surrounded by contextual textures of bars oriented parallel, orthogonal or randomly to it, respectively. The saliencies of the (centre) target bars in (A), (B), (C) are, respectively, 0.23, 0.74, and 0.41 (averaged over different random surrounds). An isolated bar of the same input strength would have a saliency 0.98. (D) simulates the experiment by Kapadia *et al* (1995) where a low-contrast (centre) target bar is aligned with some high-contrast contextual bars to form a line in a background of randomly oriented high-contrast bars. The target bar saliency is 0.39, about twice as salient as an isolated bar at the same (low) input strength, and roughly as salient as a typical (high input strength) background bar. Contour enhancement also holds in (D) when all bars have high input values, simulating the psychophysics experiment by Field *et al* (1993).

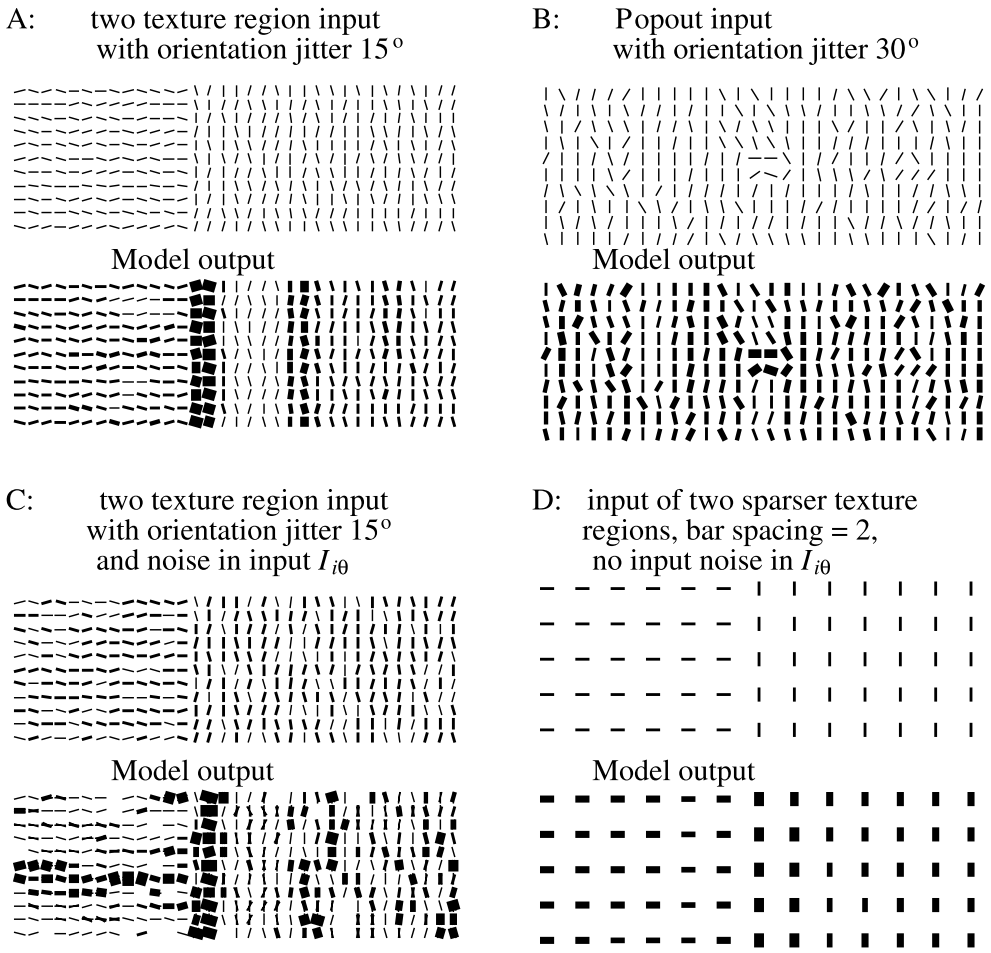


**Figure 9.** Four examples of model inputs ($\hat{I}_{i\theta}$) and outputs ($g_x(x_{i\theta})$) to show how performance changes with noises in orientation, noises in the input strength and the spacings between the texture elements. (A), (B) Stimulus patterns are made by adding orientation noises to the horizontally or vertically oriented bars of the two texture regions or figure–ground pop-out inputs, respectively. The orientations of the bars are randomly jittered from the average orientation by up to 15° in (A) and up to 30° in (B). (C) The input is the same as in (A) except that there are additional random variations in the strength of the input to the bars. The input strength $\hat{I}_{i\theta}$ to visible bars is randomly and uniformly distributed in the range [1, 3] rather than a fixed input value $\hat{I}_{i\theta} = 2.0$ as in (A). (D) Model input ($\hat{I}_{i\theta}$) and output ($g_x(x_{i\theta})$) highlights for two texture regions made of bars oriented horizontally and vertically. The spacing between neighbouring bars is two grid spacings. The average response to the vertical bars at the boundary is about 57% higher than the responses to the average background bars. In (B) the saliencies of the four target bars are $0.333 \pm 0.02$, the background saliency has a mean and standard deviation of $0.159 \pm 0.05$. The boundary measures for (A), (C), (D), respectively, are $(r, z) = (4.0, 3.4)$, $(3.5, 2.2)$ and $(1.5, 3.3)$. The input noise makes the saliency values quite non-uniform near the boundary for (A), (C) making the boundary measures $(r, z)$ less meaningful.

which is more difficult than the photographic image for humans to segment. However, the most salient model outputs given its unsatisfactorily sampled inputs do include the vertical column borders as well as some of the more conspicuous horizontal streaks in the photograph.

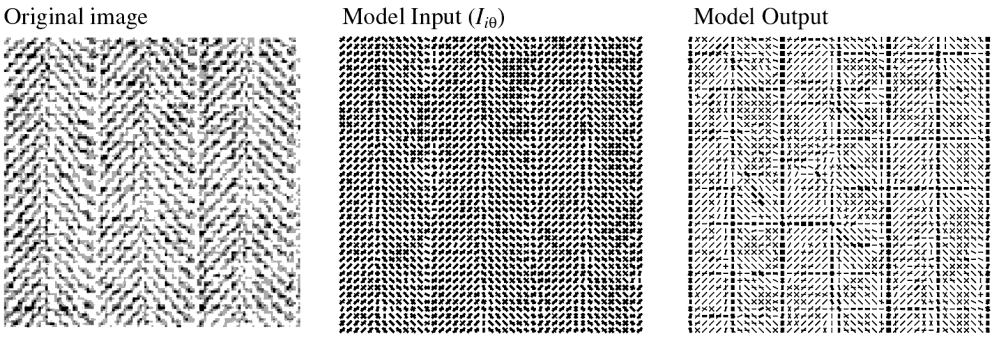


**Figure 10.** Model behaviour on a photographic image composed of 126 × 129 pixels. The input to the model is modelled as $I_{i\theta} = (e^2 + o^2)^{1/4}$, where $e$ and $o$ are the outputs from the even and odd Gabor-like filters at grid sampling points $i$ with preferred orientation $\theta$, the power $\frac{1}{4}$ coarsely models some degree of contrast gain control. The sampling points $i$ are spaced three pixels apart from the neighbours vertically or horizontally, the Gabor filters are of the form $\exp(-(4y^2 + x^2)/26)\cos(1.4y)$ and $\exp(-(4y^2 + x^2)/26)\sin(1.4y)$, where $x$ and $y$ are measured in pixels. At each grid point, bars of almost all $K = 12$ orientations have non-zero input values $I_{i\theta}$. For display clarity, no more than two strongest input or output orientations are plotted at each grid point in model input and output above. The second orientation bar is plotted only if input or output values at the grid point are not uni-modal, and the second strongest modal is at least 30% in strength of the strongest one. The strongest $I_{i\theta} = 3.0$ in the whole input. The more salient locations in the model output include some vertical borders of the columns in the input texture, as well as horizontal streaks, which are often also conspicuous in the original image. Note that this photograph is sampled against a blank background on the left and right, hence the left and right sides of the photograph area are also highlighted.

## 5. Summary and discussions

### *5.1. Summary of the results*

In this paper, we first proposed that contextual influences in V1 serve pre-attentive segmentation by highlighting the important image locations. Secondly, we validated this by presenting a biologically based model, which implements segmentation. The model can easily handle some segmentation examples such as those in figure 5(A) (in which the two regions have the same texture features and the region border does not have a definite edge signal) that pose problems for traditional approaches, but are easily segmentable by human pre-attentive vision.

This is the first model of V1 that captures the effect that neural activities are higher near region boundaries, as well as the natural consequence of this on pop-out of small figures against backgrounds and asymmetries in pop-out strengths between alternative choices of figure and ground. Underlying the model are the *local* intra-cortical interactions that modify individual neural activities depending on the contextual visual stimuli, thus detecting the region boundaries by detecting the breakdown of translation invariance in inputs. Furthermore, the model uses the same neural circuit for both the region boundary effect and contour enhancement—individual contours in a noisy or non-noisy background can also be seen as examples of the breakdown of translation invariance in inputs. Putting these effects together, V1 is modelled as a saliency network that highlights the conspicuous image areas in inputs. These conspicuous areas include region boundaries, and smooth contours or small figures against backgrounds, thus serving the purpose of pre-attentive segmentation. This V1 model, with its intra-cortical interactions designed for pre-attentive segmentation, successfully explains the contextual influences beyond the classical receptive fields observed in physiological

experiments (Knierim and van Essen 1992, Kapadia *et al* 1995). Hence, we suggest that one of the main roles for contextual influences is pre-attentive segmentation.

### *5.2. Relation to other studies*

It has recently been argued that texture analysis is performed at low levels of visual processing (Bergen 1991)—indeed, filter-based models (Bergen and Adelson 1988) and their nonlinear extensions (e.g. Malik and Perona 1990) capture well much of the phenomenology of psycho-physical performance. However, these previous segmentation models use a region-based approach, i.e. locating the border by classifying the region textures first, and thus differ from our model in principle. For example, the texture segmentation model of Malik and Perona (1990) also employs neural-like interactions in a parallel network. However, their interactions are designed to *classify* or extract region features. Consequently, the model requires a subsequent feature comparison operation (by spatial differentiation) in order to segment. It would thus have difficulties in cases like figure 1, and would not naturally capture figure pop-out, asymmetries between the figure and ground, or contour enhancement.

The task of locating conspicuous image locations without specific tuning to (or classification of) any region features is significantly more sophisticated computationally than standard early visual processing using centre–surround filters (which would not be able to detect boundaries between two different texture regions with the same mean luminance) or the like (Marr 1982). While the early stage filters code image primitives (Marr 1982), our mechanism should help in object surface representation. Since they collect contextual influences over a whole neighbourhood, the neurons naturally account for the statistical nature of the local image characteristics that define regions. This agrees with Julesz's conjecture of segmentation by image statistics (Julesz 1962) without any restriction to being sensitive only to the first- and second-order image statistics. Julesz's concept of textons (Julesz 1981) could be viewed within this framework as any feature to which the particular intra-cortical interactions are sensitive and discriminatory. Using orientation-dependent interactions between neurons, the model agrees with previous ideas (Northdurft 1994) that (texture) segmentation is primarily driven by orientation contrast. However, the emergent network behaviour is collective and accommodates characteristics of general regions beyond elementary orientations, as in figure 5. Furthermore, the psycho-physical phenomena of filling-in (when one fails to notice a small blank region within a textured region) could be viewed as the instances when the network fails to sufficiently highlight the non-homogeneity in inputs near the filled-in area.

Our basic framework is quite primitive. It merely segments surface regions from each other, whether or not these regions belong to different visual objects. Furthermore, by segmenting under a boundary/border-based approach without classifying region features, it does not characterize the region properties (such as by the (2 + 1)/two-dimensional surface representations (Marr 1982)) more than what is already implicitly present in the raw image pixels or the cell responses in V1. Hence, for example, the model output does not indicate whether a region is made of a transparent surface on top of another surface.

Key support for the model comes from experimental evidence that pre-attentive segmentation precedes and is dissociated from visual classification/discrimination of surfaces and regions. Recent experimental evidence from V1 (Lamme *et al* 1997, Zipser 1998) shows that the modulation of neural activities starts at texture boundaries and only later includes figure surfaces, where the neural modulations take about 50 ms to develop after the initial cell responses (Zipser *et al* 1996, Zipser 1998). Some psycho-physical evidence (Scialfa and Joffe 1995) suggests that information regarding (figure) target presence is available before information regarding feature values of the targets. Also consistent with the model is that V2

lesions in monkeys are shown to disrupt region content discrimination but not region border detection (Merigan *et al* 1993). Furthermore, neural modulation in V1, especially at figure surfaces (Zipser 1998), is strongly reduced or abolished by anaesthesia or lesions in higher visual areas (Lamme *et al* 1997), while experiments by Gallant *et al* (1995) show that activity modulation at texture boundaries is present even under anaesthesia. Taken together, this experimental evidence suggests the plausibility of the following computational framework. Pre-attentive segmentation (e.g. border highlights) in V1 precedes region classification; region classification after pre-attentive segmentation is initialized in higher visual areas; the classification is then fed back to V1 in the form of top-down influences, which can refine the segmentation (perhaps to remove the bias in the estimation of the border location in the example of figures 4(B) and 5(C)), this latter process might be attentive; the bottom-up and top-down loop can be iterated to improve both classification and segmentation. Top-down and bottom-up streams of processing have been studied by many others (e.g. Grenander 1976, Carpenter and Grossberg 1987, Ullman 1994, Dayan *et al* 1995). Our model is of the first step in the bottom-up stream, which initializes the iterative loop. The neural circuit in our model can easily accommodate top-down feedback signals, which, in addition to the V1 mechanisms, selectively enhance or suppress the neural activities in V1 (see examples in Li 1998b). However, we have not yet modelled how higher visual centres might process the bottom-up signals to generate the feedback.

The model's components and behaviour are based on and consistent with experimental evidence (Rockland and Lund 1983, White 1989, Douglas and Martin 1990, Gilbert 1992, Nothdurft 1994, Gallant *et al* 1995). The experimentally testable predictions of the model include the tuning of the neural response near texture border to the relative orientation between texture bars and texture borders (e.g. in figures 3(B) and 4(A)), the qualitative structure of the horizontal connections as in figure 2(B), and the biases in the estimated border location by the neural responses (e.g. figures 3(B) and 4(B)). Since the model is quite simplistic in the design of the connections, it is likely that there will be significant differences between the model and the anatomical and physiological connections. For instance, the model connections link cells up to several RF sizes or several hypercolumns away from each other. The connection can indeed be as long in cats (Gilbert and Wiesel 1983), but shorter in monkeys (Rockland and Lund 1983). The connection length in the model depends on the particular implementation of the model, it could be shortened without destroying the computational capability of the model when the sampling or RF density of the model is increased to resemble that of the real cortex. Also, two linked pyramidal cells interact in the model either via monosynaptic excitation or disynaptic inhibition. In the real cortex, two linked cells could often interact via both excitation and inhibition, making the overall strength of excitation or inhibition input contrast dependent (e.g. Hirsch and Gilbert 1991, see Li 1998b for analysis). Hence, the excitation (or inhibition) in the model could be interpreted as the abstraction of the predominance of excitation (or inhibition) between two linked bars. There is presently no agreement in the experimental data as to the spatial and orientation dependence of excitation and inhibition (Fitzpatrick 1996, Cavanaugh *et al* 1997, Kapadia 1998, Hirsch and Gilbert 1991, Polat *et al* 1998), partly due to different experimental conditions such as input contrast levels or the nature of stimulus elements (e.g. bars or gratings). The model's performance is also quantitatively dependent on input strength. One should bear this fact in mind when viewing the comparisons between the model and experimental data in figures 4, 9 and 8.

The modulation of neural activity by cortical interactions should have perceptual consequences other than contour/region boundary enhancement and figure pop-out. For instance, the preferred orientation of the cells can shift depending on contextual bars. Under population coding, this will lead to tilt illusions, i.e. the change in perceived orientation of the

target bar. The perceived orientation of the target bar could shift away or towards the orientation of the contextual bars, depending on the spatial arrangement (and the orientations) of the contextual bars. This is in contrast to the usual notion that the orientation of the target bar tends to shift away from those of the contextual bars. Both our model and a recent psycho-physical study (Kapadia 1998) confirm such context-dependent distortion in perceived orientation. V1 cells do display changes in orientation tuning under contextual influences (Gilbert and Wiesel 1990), although the magnitude and direction of the changes vary from cell to cell.

### *5.3. Comparison with other models*

There are many other related models. Many cortical models are mainly concerned with contour linking, and Li (1998b) has a detailed citation of these models and comparisons with our model. For instance, Grossberg and his colleagues have developed models of visual cortex over many years (Grossberg and Mingolla 1985, Grossberg *et al* 1997). They proposed a 'boundary contour system' as a model of intra-cortical and inter-areal neural interactions in V1 and V2 and feedback from V2 to V1. The model aims to capture illusory contours, which link line segments and line endings, and the authors claim that such linking affects segmentation. Other models are more concerned with regions, namely, to classify region features and then to segment regions by comparing the classifications. To obtain texture region features, Malik and Perona (1990) use local intra-cortical inhibition. Geman and Geman built a model based on Markov random fields to restore images, in which neighbouring image features influence each other statistically (Geman and Geman 1984). Such local interactions improve the outcomes from the prior and preliminary feature classifications to drive segmentation.

Our model contrasts with previous models by modelling the effect of region boundary highlights in V1, and by using the same neural circuit as used for contour enhancement. Equally, its instantiation in V1 means that our model does not perform operations such as the classification and smoothing of region features and the sharpening of boundaries as carried out in some other models (e.g. Lee 1995, Malik and Perona 1990). There are many other models of visual processing (e.g. Grossberg and Mingolla 1985, Zucker *et al* 1989, Yen and Finkel 1997) that use a bow-tie-shaped interaction pattern that is qualitatively similar to ours (figure 2(B)). This, however, does not mean that those models would necessarily perform the same computations as our model. Using dynamic systems theory, a recent work (Li and Dayan 1999) has shown that if the units in the network interact with each other directly and reciprocally (or symmetrically in the terminology of recurrent neural networks), as in the models of Grossberg, Zucker and co-workers, the network cannot produce the desired contour enhancement and (texture) boundary highlighting even with the same qualitative pattern of interactions. In particular, the three design conditions (in section 3 or the appendix) for the synaptic weights are not likely to be satisfied simultaneously in those networks without significantly compromising the selective enhancement of the contours and texture borders. In other words, if those networks enhance the contours or texture boundaries sufficiently, then they are likely also to hallucinate illusory boundary or contour highlights even when there is none in input. Apparently, the biological fact that the principal (pyramidal) cells inhibit each other indirectly via inhibitory interneurons offers significant dynamic advantages over the seemingly simpler but less plausible networks with direct and reciprocal inhibition.

### *5.4. Limitations and extensions of the model*

The model is still very primitive compared to the true complexity of V1. We have yet to include multiscale sampling or the over-complete input sampling strategy adopted by V1, or to

include colour, time or stereo input dimensions. Also, the receptive field features used for the bar/edges should be determined more precisely. The details of the intra-cortical circuits within and between hypercolumns should also be better determined to match biological vision.

Multiscale sampling is needed not only because images contain multiscale features, but also to model V1 responses to images from flat surfaces slanted in depth—such a region should also be seen as 'homogeneous' or 'translation invariant' by V1, such that it has uniform saliency. Merely replicating and scaling the current model to multiple scales is not sufficient for this purpose. The computation requires interactions between different scales. We also have yet to find a better sampling distribution even within a single scale. Currently, the model neurons within the same hypercolumn have exactly the same RF centres and the RFs from different hypercolumns do not overlap. This sampling arrangement is much sparser than the V1 sampling.

In addition to orientation and spatial location, neurons in V1 are tuned for motion direction/speed, disparity, ocularity, scale and colour (Hubel and Wiesel 1962, Livingstone and Hubel 1984). Our model should be extended to these dimensions. The horizontal connections in the extended model will link edge segments with compatible selectivities to all of these facets as well as orientation, as suggested by experimental data (e.g. Li and Li 1994, Gilbert 1992, Ts'o and Gilbert 1988). The model should also be expanded to include details such as on and off cells, cells of different RF phases, non-orientation selective cells, end stopped cells and more cell layers. These details should lead to better quantitative match between the model and human vision.

Any given neural interaction will be more sensitive to some region differences than others. Therefore, the model sometimes finds it easier or more difficult to segment some regions than natural vision. Physiological and psycho-physical measurements of the boundary effect for different types of textures can help to constrain the connection patterns in an improved model. Experiments also suggest that the connections may be learnable or plastic (Karni and Sagi 1991, Sireteanu and Rieth 1992, Polat and Sagi 1994). It is also desirable to study the learning algorithms to develop the connections.

We currently model saliency at each location quite coarsely by the activity of the most salient bar. It is mainly an experimental question as to how to best determine the saliency, and the model should be modified accordingly. This is particularly the case once the model includes multiple scales, non-orientation selective cells and other visual input dimensions. The activities from different channels should somehow be combined to determine the saliency at each location of the visual field.

In summary, this paper proposes a computational framework for pre-attentive segmentation in the primary visual cortex. It introduces a simple and biologically plausible model of V1 to implement the framework using mechanisms of contextual influences via intra-cortical interactions. Although the model is as yet very primitive compared to the real cortex, our results show the feasibility of the underlying ideas, that breakdown of input translation invariance can be used to segment regions, that region segmentation and contour detection can be addressed by the same mechanism, and that low-level processing in V1 together with *local* contextual interactions can contribute significantly to visual computations at *global* scales.

## Acknowledgments

I thank Peter Dayan for many helpful discussions and conversations over the duration of this research, he, John Hertz, Geoffrey Hinton and John Hopfield for their careful readings and helpful comments on various versions of the paper, the two anonymous reviewers for their comments and many other colleagues for their questions and comments during and after my

seminars, and the Center for Biological and Computational Learning at MIT for hosting my visit. This work is supported in part by the Hong Kong Research Grant Council and the Gatsby Foundation.

## Appendix

### *A.1. Design analysis for horizontal connections*

Connections $J$ and $W$ are designed to satisfy the three conditions listed in section 3. To illustrate the design for conditions (1) and (2), consider the example of a homogeneous input

$$I_{i\theta} = \begin{cases} I_o > 0 & \text{when } \theta = \theta_o \\ 0 & \text{otherwise} \end{cases} \tag{A1}$$

of a bar oriented $\theta_o$ at every sampling point. By symmetry, a mean field solution $(\bar{x}_{i\theta}, \bar{y}_{i\theta})$ is also independent of spatial location $i$. For simplicity assume $\bar{x}_{i\theta} = 0$ for $\theta \neq \theta_o$, and ignore all $(x_{i\theta}, y_{i\theta})$ with $\theta \neq \theta_o$. To study whether this mean field solution $(\bar{x}_{i\theta_o}, \bar{y}_{i\theta_o})$ is stable, look at the perturbations $(x'_i \equiv x_{i\theta_o} - \bar{x}_{i\theta_o}, y'_i \equiv y_{i\theta_o} - \bar{y}_{i\theta_o})$ around it. It follows that

$$\dot{Z} = AZ \tag{A2}$$

where $Z = (x'^T, y'^T)^T$. Matrix $A$ results from expanding equations (1) and (2) around the mean field solution, it contains the horizontal connections $J_{i\theta_o, j\theta_o}$ and $W_{i\theta_o, j\theta_o}$ linking bar segments oriented all at $\theta_o$. Translation invariance in $J$ and $W$ implies that every eigenvector of $A$ is a cosine wave in space for both $x'$ and $y'$. To ensure condition (1), either every eigenvalue of $A$ should be negative so that the mean field solution is stable and no perturbation from the homogeneous mean field solution is self-sustaining, or the eigenvalue with largest positive real part should correspond to the zero-frequency cosine wave in space. In the latter case, the deviation or perturbation from the mean field solution tends to be homogeneous, and thus does not spontaneously form non-homogeneous spatial patterns, although it will oscillate over time (Li 1998b). Iso-orientation suppression under supra-threshold input $I_o$ is used to satisfy condition (2). This requires that every pyramidal cell $x_{i\theta_o}$ in an iso-orientation surround should receive stronger overall disynaptic inhibition than monosynaptic excitation:

$$\sigma \sum_j W_{i\theta_o, j\theta_o} > \sum_j J_{i\theta_o, j\theta_o} \tag{A3}$$

where $\sigma \equiv \psi(0) g'_y(\bar{y}_{i\theta_o})$ comes from the inhibitory interneurons. The excitatory cells near a region boundary lack a complete iso-orientation surround, they are less suppressed and so exhibit stronger responses, meeting condition (2). We tested conditions (1) and (2) in simulations using these simple and other general input configurations including the cases when input within a region is of the form $I_{i\theta} \equiv \tilde{I}_\theta$, where $\tilde{I}_\theta$ is non-zero for two orientation indices $\theta$. The design to ensure condition (3) that the system should properly enhance isolated smooth contours has been shown in Li (1998b) which, to avoid an over-long manuscript, did not discuss conditions (1) and (2). Briefly, condition (3) is ensured by strong enough but limited strength monosynaptic excitation $\sum_{j\theta' \in \text{contour}} J_{i\theta, j\theta'}$ along any smooth contour extending from $i\theta$—to give enough enhancement without exciting any bar/edge elements beyond the end of non-closed contours—and enough disynaptic inhibition between local, similarly oriented and non-aligned bars—to avoid enhancement of the noisy background.

*A.2. A complete list of model parameters*

One can use the following parameters to reproduce all results in the paper, using equations (1) and (2):

$\alpha_x = \alpha_y = 1 \qquad K = 12$

$$g_x(x) = \begin{cases} 0 & \text{if} \quad x < T_x \\ (x - T_x) & \text{if} \quad T_x \leqslant x \leqslant T_x + 1 \\ 1 & \text{if} \quad x > T_x + 1 \end{cases}$$

$$g_y(y) = \begin{cases} 0 & \text{if} \quad y < 0 \\ g_1 y & \text{if} \quad 0 \leqslant y \leqslant L_y \\ g_1 L_y + g_2(y - L_y) & \text{if} \quad 0 < L_y \leqslant y \end{cases}$$

$T_x = 1 \qquad L_y = 1.2 \qquad g_1 = 0.21 \qquad g_2 = 2.5$

$$\psi(\theta) = \begin{cases} 1 & \text{when} \quad \theta = 0 \\ 0.8 & \text{when} \quad |\theta| = \pi/K = 15^\circ \\ 0.7 & \text{when} \quad |\theta| = 2\pi/K = 30^\circ \\ 0 & \text{otherwise} \end{cases}$$

$I_c = 1.0 + I_{\text{noise}}$

$I_o = 0.85 + I_{\text{normalization}} + I_{\text{noise}}$

$$I_{\text{normalization}}(i\theta) = -2.0\left[\frac{\sum_{j \in S_i} \sum_{\theta'} g_x(x_{j\theta'})}{\sum_{j \in S_i} 1}\right]^2$$

$S_i =$ all $j$ such that $|i - j| \leqslant 2$ grid distances

$I_{\text{noise}} =$ zero mean, random, mean temporal width 0.1, mean amplitude 0.1

$J_o = 0.8$

$$J_{i\theta, j\theta'} = \begin{cases} 0.126\mathrm{e}^{-(\beta/d)^2 - 2(\beta/d)^7 - d^2/90} & \text{if } 0 < d \leqslant 10.0 \text{ and } \beta < \pi/2.69 \\ & \text{or } 0 < d \leqslant 10.0 \text{ and } \beta < \pi/1.1 \\ & \text{and } |\theta_1| < \pi/5.9 \text{ and } |\theta_2| < \pi/5.9 \\ 0 & \text{otherwise} \end{cases}$$

$$W_{i\theta, j\theta'} = \begin{cases} 0 & \text{if } d = 0 \text{ or } d \geqslant 10 \text{ or } \beta < \pi/1.1 \\ & \text{or } |\Delta\theta| \geqslant \pi/3 \text{ or } |\theta_1| < \pi/11.999 \\ 0.14(1 - \mathrm{e}^{-0.4(\beta/d)^{1.5}})\mathrm{e}^{-(\Delta\theta/(\pi/4))^{1.5}} & \text{otherwise.} \end{cases}$$

The parameters $d$, $\beta$ and $\Delta\theta$ in the expressions for $J_{i\theta, j\theta'}$ and $W_{i\theta, j\theta'}$ are determined as follows. Let $|i - j| = d$, and denote the angles between the edge elements and the line $i - j$ by $\theta_1$ and $\theta_2$, where $|\theta_1| \leqslant |\theta_2| \leqslant \pi/2$ and $\theta_{1,2}$ are positive or negative depending on whether the edges rotate clockwise or counter-clockwise towards the connecting line $i - j$ in no more than a $\pi/2$ angle. Denote $\beta = 2|\theta_1| + 2\sin(|\theta_1 + \theta_2|)$, $\Delta\theta = \theta - \theta'$ with $|\theta - \theta'| \leqslant \pi/2$. These same parameters were listed in Li (1998b).

## References


Allman J, Miezin F and McGuinness E 1985 Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local–global comparisons in visual neurons *Ann. Rev. Neurosci.* **8** 407–30

Bergen J R 1991 Theories of visual texture perception *Vision and Visual Dysfunction* vol 10B, ed D Regan (New York: Macmillan) pp 114–34

Bergen J R and Adelson E H 1988 Early vision and texture perception *Nature* **333** 363–4

Blum H 1973 Biological shape and visual science *J. Theor. Biol.* **38** 205–87

Carpenter G and Grossberg S 1987 A massively parallel architecture for a self-organizing neural pattern recognition machine *Comput. Vis. Graph. Image Process.* **37** 54–115

Cavanaugh J R, Bair W and Movshon J A 1997 Orientation-selective setting of contrast gain by the surrounds of macaque striate cortex neurons *Soc. Neurosci. Abstract* **227** 2

Dayan P, Hinton G E, Neal R M and Zemel RS 1995 The Helmholtz machine *Neural Comput.* **7** 889–904

Douglas R J and Martin K A 1990 Neocortex *Synaptic Organization of the Brain* ed G M Shepherd (Oxford: Oxford University Press) 3rd edn, pp 389–438

Eckhorn R, Bauer R, Jordan W, Brosch M, Kruse W, Munk M and Reitboeck H J 1988 Coherent oscillations: a mechanism of feature linking in the visual cortex? Multiple electrode and correlation analysis in the cat *Biol. Cybern.* **60** 121–30

Field D J, Hayes A and Hess R F 1993 Contour integration by the human visual system: evidence for a local 'association field' *Vis. Res.* **33** 173–93

Fitzpatrick D 1996 The functional organization of local circuits in visual cortex: insights from the study of tree shrew striate cortex *Cerebral Cortex* **6** 329–41

Gallant J L, van Essen D C and Nothdurft H C 1995 Two-dimensional and three-dimensional texture processing in visual cortex of the macaque monkey *Early Vision and Beyond* ed T Papathomas, C Chubb, A Gorea and E Kowler (Cambridge, MA: MIT Press) pp 89–98

Geman S and Geman D 1984 Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images *IEEE Trans. PAMI* **6** 721–41

Gilbert C D 1992 Horizontal integration and cortical dynamics *Neuron* **9** 1–13

Gilbert C D and Wiesel T N 1983 Clustered intrinsic connections in cat visual cortex *J. Neurosci.* **3** 1116–33

——1990 The influence of contextual stimuli on the orientation selectivity of cells in primary visual cortex of the cat *Vis. Res.* **30** 1689–701

Gray C M and Singer W 1989 Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex *Proc. Natl Acad. Sci., USA* **86** 1698–702

Grenander U 1976–1981 *Lectures in Pattern Theory I, II and III: Pattern Analysis, Pattern Synthesis and Regular Structures* (Berlin: Springer)

Grosof D H, Shapley R M and Hawken M J 1993 Macaque V1 neurons can signal 'illusory' contours *Nature* **365** 550–2

Grossberg S and Mingolla E 1985 Neural dynamics of perceptual grouping: textures, boundaries, and emergent segmentations *Percept. Psychophys.* **38** 141–71

Grossberg S, Mingolla E and Ross W 1997 Visual brain and visual perception: how does the cortex do perceptual grouping? *Trends Neurosci.* **20** 106–11

Haralick R M and Shapiro L G 1992 *Computer and Robot Vision* vol 1 (Reading, MA: Addison-Wesley)

Heeger D J 1992 Normalization of cell responses in cat striate cortex *Visual Neurosci.* **9** 181–97

Hirsch J A and Gilbert C D 1991 Synaptic physiology of horizontal connections in the cat's visual cortex *J. Neurosci.* **11** 1800–9

Hubel D H and Wiesel T N 1962 Receptive fields, binocular interaction and functional architecture in the cat's visual cortex *J. Physiol.* **160** 106–54

Julesz B 1962 Visual pattern discrimination *IRE Trans. Inform. Theory* **8** 84–92

——1981 Textons, the elements of texture perception and their interactions *Nature* **290** 91–7

Kapadia 1998 Private communication

Kapadia M K, Ito M, Gilbert C D and Westheimer G 1995 Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in V1 of alert monkeys *Neuron* **15** 843–56

Karni A and Sagi D 1991 Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity *Proc. Natl Acad. Sci., USA* **88** 4977

Kasturi R and Jain R C (eds) 1991 *Computer Vision, Principles* (Piscataway, NJ: IEEE Press)

Knierim J J and van Essen D C1992 Neuronal responses to static texture patterns ion area V1 of the alert macaque monkeys *J. Neurophysiol.* **67** 961–80

Kovacs I and Julesz B 1993 A closed curve is much more than an incomplete one: effect of closure in figure–ground

segmentation *Proc. Natl Acad. Sci., USA* **90** 7495–7
Lamme V A 1995 The neurophysiology of figure–ground segregation in primary visual cortex *J. Neurosci.* **15** 1605–15
Lamme V A F, Zipser K and Spekreijse H 1997 Figure–ground signals in V1 depend on consciousness and feedback from extra-striate areas *Soc. Neurosci. Abs.* **603** 1
Lee T S 1995 A Bayesian framework for understanding texture segmentation in the primary visual cortex *Vis. Res.* **35** 2643–57
Lee T S, Mumford D, Romero R and Lamme V A F 1998 The role of the primary visual cortex in higher level vision *Vis. Res.* **38** 2429–54
Levitt J B and Lund J S 1997 Contrast dependence of contextual effects in primate visual cortex *Nature* **387** 73–6
Li Z 1998a Primary cortical dynamics for visual grouping *Theoretical Aspects of Neural Computation* ed K Y M Wong, I King and D-Y Yeung (Berlin: Springer)
——1998b A neural model of contour integration in the primary visual cortex *Neural Comput.* **10** 903–40
——1999a A V1 model of pop out and asymmetry in visual search *Advances in Neural Information Processing Systems 11* ed M S Kearns, S A Solla and D A Cohn (Cambridge, MA: MIT Press) in press
——1999b Pre-attentive segmentation in primary visual cortex *Spatial Vis.* in press
Li Z and Dayan P 1999 Computational differences between asymmetric and symmetric networks *Network: Comput. Neural Syst.* **10** 59–77
Li C Y and Li W 1994 Extensive integration field beyond the classical receptive field of cat's striate cortical neurons—classification and tuning properties *Vis. Res.* **34** 2337–55
Livingstone M S and Hubel D H 1984 Anatomy and physiology of a color system in the primate visual cortex *J. Neurosci.* **4** 309–56
Malik J and Perona P 1990 Preattentive texture discrimination with early vision mechanisms *J. Opt. Soc. Am.* **7** 923–32
Marr D 1982 A computational investigation into the human representation and processing of visual information *Vision* (San Francisco, CA: Freeman)
Meinhardt H 1982 *Models of Biological Pattern Formation* (New York: Academic)
Merigan W H, Mealey T A and Maunsell J H 1993 Visual effects of lesions of cortical area V2 in macaques *J. Neurosci.* **13** 3180–91
Nothdurft H C 1985 Sensitivity for structure gradient in texture discrimination tasks *Vis. Res.* **25** 1957–68
——1991 Texture segmentation and pop-out from orientation contrast *Vis. Res.* **31** 1073–78
——1994 Common properties of visual segmentation *Higher-Order Processing in the Visual System* ed G R Bock and J A Goode (New York: Wiley) pp 245–68
Polat U, Mizobe K, Pettet M, Kasamatsu T and Norcia A 1998 Collinear stimuli regulate visual responses depending on cell's contrast threshold *Nature* **391** 580–3
Polat U and Sagi D 1993 Lateral interactions between spatial channels: suppression and facilitation revealed by lateral masking experiments *Vis. Res.* **33** 993–9
——1994 Spatial interactions in human vision: from near to far via experience-dependent cascades of connections *Proc. Natl Acad. Sci., USA* **91** 1206–9
Rockland K S and Lund J S 1983 Intrinsic laminar lattice connections in primate visual cortex *J. Comput. Neurol.* **216** 303–18
Scialfa C T and Joffe K M 1995 Preferential processing of target features in texture segmentation *Percept. Psychophys.* **57** 1201–8
Sillito A M, Grieve K L, Jones H E, Cudeiro J and Davis J 1995 Visual cortical mechanisms detecting focal orientation discontinuities *Nature* **378** 492–6
Sireteanu R and Rieth C 1992 Texture segregation in infants and children *Behav. Brain Res.* **49** 133–9
Somers D C, Todorov E V, Siapas A G and Sur M 1995 Vector-based integration of local and long-range information in visual cortex *AI Memo.* vol 1556 (Cambridge, MA: MIT Press)
Stemmler M, Usher M and Niebur E 1995 Lateral interactions in primary visual cortex: a model bridging physiology and psychophysics *Science* **269** 1877–80
Treisman A and Gormican S 1988 Feature analysis in early vision: evidence for search asymmetries *Psychol. Rev.* **95** 15–48
Ts'o D and Gilbert C 1988 The organization of chromatic and spatial interactions in the primate striate cortex *J. Neurosci.* **8** 1712–27
Ullman S 1994 Sequence seeking and counterstreams: a model for bidirectional information flow in the cortex *Large-Scale Theories of the Cortex* ed C Koch and J Davis (Cambridge, MA: MIT Press) pp 257–70
Weliky M, Kandler K, Fitzpatrick D and Katz L C 1995 Patterns of excitation and inhibition evoked by horizontal connections in visual cortex share a common relationship to orientation columns *Neurons* **15** 541–52
von der Heydt R, Peterhans E and Baumgartner G 1984 Illusory contours and cortical neuron responses *Science* **224** 1260–2

White E L 1989 *Cortical Circuits* (Boston, MA: Birkhäuser)
Wolfson S and Landy M S 1995 Discrimination of orientation-defined texture edges *Vis. Res.* **35** 2863–77
Yen S-C and Finkel L H 1997 Salient contour extraction by temporal binding in a cortically-based network *Advances in Neural Information Processing Systems 9* ed M C Moser, M I Jordan and T Petsche (Cambridge, MA: MIT Press)
Zipser K 1998 Private communication
Zipser K, Lamme V A and Schiller P H 1996 Contextual modulation in primary visual cortex *J. Neurosci.* **16** 7376–89
Zucker S W, Dobbins A and Iverson L 1989 Two stages of curve detection suggest two styles of visual computation *Neural Comput.* **1** 68–81