MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo No. 1640 April, 1998
C.B.C.L. Memo No. 163

Pre-attentive segmentation in the primary visual cortex

Zhaoping Li

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.

Abstract

Stimuli outside classical receptive fields have been shown to exert significant influence over the activities of neurons in primary visual cortex. We propose that contextual influences are used for pre-attentive visual segmentation, in a new framework called *segmentation without classification*. This means that segmentation of an image into regions occurs without classification of features within a region or comparison of features between regions. This segmentation framework is simpler than previous computational 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 tricky in conventional methods. The cortex 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. 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. This proposal is implemented in a biologically based model of V1, and demonstrated using examples of texture segmentation and figure-ground segregation. The model performs segmentation in exactly the same neural circuit that solves the dual problem of the enhancement of contours, as is suggested by experimental observations. Its behavior is compared with psychophysical and physiological data on segmentation, contour enhancement, and contextual influences. We discuss the implications of *segmentation without classification* and the predictions of our V1 model, and relate it to other phenomena such as asymmetry in visual search.

This report is the manuscript (except for a title change from "visual segmentation without classification in primary visual cortex") submitted to the journal *Neural Computation* on April 16th, 1998, as the revised version of the original manuscript submitted on November 24th, 1997 to the same journal. The authors can be reached at M.I.T., Center for Biological and Computational Learning, Cambridge MA 02139, USA. E-mail: zhaoping@ai.mit.edu

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, Miezin, and McGuinness, 1985, Knierim and Van Essen 1992, Gilbert, 1992, Kapadia, Ito, Gilbert, and Westheimer 1995, Sillito et al 1995, Lamme, 1995, Zipser, Lamme, and Schiller 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), suggesting that mechanisms within V1 itself are responsible (see 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). While the phenomena and the mechanisms of the contextual influences are studied experimentally and in some models (e.g., Somers Todorov, Siapas, and Sur 1995, Stemmler, Usher, and Niebur, 1995), insights into their computational roles have been limited to mainly contour or feature linking (Allman et al 1995, Gilbert, 1992, see more references in Li 1998).

We propose that contextual influences serve the goal of pre-attentive visual segmentation or grouping to infer *global* visual objects such as regions and contours from *local* RF features. Local features can group into regions, as in texture segmentation; or into contours which may represent boundaries of underlying objects. We show how region segmentation emerges from a simple but biologically-based model of V1 with only *local* cortical interactions between cells within a few RF sizes away from each other. Note that although the horizontal intra-cortical connections are termed as long range, they are still local with respect to the whole visual field since the axons reach only a few millimeters, or a few hypercolumns or receptive field sizes, away from the pre-synaptic cells. To attack the formidable problem of segmentation in such a low level visual area, we introduce a new computational framework — *segmentation without classification.*

2. The problem of visual segmentation

Visual segmentation is defined as locating the boundary between different image regions. For example, when regions are defined by their pixel luminance values, center-surround filters in the retina can locate boundaries between regions by comparing the classification (in this case, luminance) values between neighboring image areas. In general, regions are seldom classifiable by pixel luminances, and image filters are mainly to extract image features rather than to segment image regions(Haralick and Shapiro, 1992). For general region segmentation, previous computational approaches have always assumed, implicitly or explicitly, that segmentation requires (1) feature extraction or classification for every small image area, and, (2) comparisons of the classification flags (feature values) between neighboring image areas to locate the boundary as where the classification flags change (Haral-

ick and Shapiro, 1992, Bergen 1991, Bergen and Adelson, 1988, Malik and Perona 1990). This framework can be summarized as *segmentation by classification*. Over the years, many such segmentation algorithms have been developed both for computer vision(Haralick and Shapiro, 1992) and to model natural vision (Bergen and Adelson, 1988, Malik and Perona 1990). They are all forms of segmentation by classification, and differ chiefly as to how the region features are extracted and classified, whether it is by, e.g., 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). In such approaches, classification is problematic and ambiguous near boundaries between regions. This is because feature evaluations can only be performed in local areas of the image which are assumed to be sitting away from region boundaries, i.e., feature classification presumes some degree of region segmentation. A priori, the locations of the region boundaries are not known, and so the feature values at these places will be ambiguous. The probability that this ambiguity happens can be reduced by choosing smaller image areas for feature evaluation. However, this in turn gives less accurate feature values, especially in textured regions, and can lead to there being significant differences in the feature values even within a region, and thus to false region boundaries. One seems inevitably to face a fundamental dilemma — classification presumes segmentation, and segmentation presumes classification.

This dilemma can be dissolved by recognizing that segmentation does not presume classification. Natural vision can segment the two regions in Fig. 1 even though they have the same texture features (note that the plotted area is only a small part of an extended image). In this case, classification of the region features is neither sufficient, nor necessary, and segmentation is rather triggered by the sudden changes near the region boundary which is problematic in traditional approaches. In fact, even with distinguishable classification flags for all image areas in any two regions (such as the '|' and '—' in Fig. 3A), segmentation is not completed until another processing step locates the boundary, perhaps by searching for where the classification flags change. We propose that segmentation at its pre-attentive bare minimum is segmentation without classification, i.e., segmentation without explicitly knowing the feature contents of the regions. This simplifies the segmentation process conceptually, making it plausible that it can be performed by low level processings in V1. This paper focuses on pre-attentive segmentation. Additional processing is likely needed to improve the resulting segmentation, e.g., by refining the coarse boundaries detected at the pre-attentive stage and classifying the contents of the regions.

3. The principle and its implementation

The principle of *segmentation without classification* 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, or, for instance, whether they are colored 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 segmentation without classification, 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 color, motion, luminance, or stereo cues. To focus on the segmentation problem, the model includes mainly layer 2-3 orientation se-

lective 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 'x' patterns in Fig. 6D and the statistics of their spatial arrangements, or the estimated densities of bars of particular orientations, from which one can not recover the original input images. To reiterate, this model does not extract texture features in order to segment[1]. To avoid confusions, this 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 homogenous 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 msec after the initial cell responses (Nothdurft, 1994, Gallant, Van Essen, and Nothdurft 1995).

Fig. 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, θ) has RF center i and preferred orientation $\theta = k\pi/K$ for $k = 1, 2, ...K$, 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, and each model cell represents a collection of local cells of similar types. 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 β in the input image contribute to $I_{i\theta}$ by an amount $\hat{I}_{i\beta}\phi(\theta - \beta)$, where $\phi(\theta - \beta) = e^{-|\theta-\beta|/(\pi/8)}$ 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 $J_{i\theta,j\theta'}$ (resp. $W_{i\theta,j\theta'}$) mediate contextual influences via monosynaptic excitation (resp. disynaptic inhibition) from bar $j\theta'$ to $i\theta$ which have nearby but different RF centers, $i \neq j$, and similar orientation preferences, $\theta \sim \theta'$. The membrane potentials follow the equations:

$$\begin{aligned} \dot{x}_{i\theta} &= -\alpha_x x_{i\theta} - \sum_{\Delta\theta} \psi(\Delta\theta) g_y(y_{i,\theta+\Delta\theta}) + J_o g_x(x_{i\theta}) \\ &\quad + \sum_{j\neq i,\theta'} J_{i\theta,j\theta'} g_x(x_{j\theta'}) + I_{i\theta} + I_o \qquad (1) \\ \dot{y}_{i\theta} &= -\alpha_y y_{i\theta} + g_x(x_{i\theta}) \\ &\quad + \sum_{j\neq i,\theta'} W_{i\theta,j\theta'} g_x(x_{j\theta'}) + I_c \qquad (2) \end{aligned}$$

where $\alpha_x x_{i\theta}$ and $\alpha_y y_{i\theta}$ model the decay to resting potential, $g_x(x)$ and $g_y(y)$ are sigmoid-like functions modeling cells' firing rates in response to membrane potentials x and y, respectively,

[1] 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 manners) the 'edge' and 'bar' values between image areas to find region boundaries

$\psi(\Delta\theta)$ 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 modeling the general and local normalization of activities (Heeger, 1992) (see Li (1998) 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, Bauer, Jordan, Brosch, Kruse, Munk, and Reitboeck 1988), an intrinsic property of a population of recurrently connected excitatory and inhibitory cells (see Li (1998) for detailed parameters and dynamic analysis of the model). Temporal averages of $g_x(x_{i\theta})$ over several oscillation cycles (about 12 to 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, 1998), 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 (Fig. 2B), 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, Kandler, Fitzpatrick, and Katz 1995). In particular, J and W are chosen to satisfy the following three conditions (Li, 1997): (1) the system should not generate patterns spontaneously, i.e., homogenous input images give homogenous outputs, so that no illusory borders occur 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 is shown in Fig. 2B, and is thus a prediction of our model (see Appendix and Li (1998) for its derivation). Qualitatively, the connection pattern 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. 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), and their magnitudes decay with distance between RFs(Li, 1998).

Mean field techniques and dynamic stability analysis (shown in Appendix) are used to design the horizontal connections that ensure the 3 conditions above. Conditions (1) and (2) are strictly met only for (the particularly homogenous) 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 Fig. (6)), 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 types of homogenous 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 our 3 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 our model

quantitatively imprecise compared to physiological and psychophysical observations (see discussions later).

4. Performance of the model

The model was applied to a variety of input textures, as shown in examples in the figures. With two exceptions, the input values $\hat{I}_{i\theta}$ is the same for all visible bars in each example so that any difference in the outputs $g_x(x_{i\theta})$ of the bars are solely due to the effects of the intra-cortical interactions. The exceptions are the input taken from a photo (Fig. 10), and the input in Fig. (9D) which models an experiment on contour enhancement (Kapadia et al 1995). The difference in the outputs, which are interpreted as a difference in saliencies, are significant about one membrane time constant after the initial neural response (Li, 1998). 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 order 10 msec. The actual value $\hat{I}_{i\theta}$ used in all examples are 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 Figs. (7B) and the target bar in Fig. (9D) (Kapadia et al 1995, Kovacs and Julesz 1993). Intermediate levels are used for all visible bars in texture segmentation and figure-ground pop-out examples (Figs. (3, 4, 5, 6, 7A, and 8)). High input levels are used for all visible bars in Fig. (9A,B,C) and the contextual (background) bars in Fig. (9D) 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 photo image (Fig. (10)) is different for different $i\theta$ with $I_{i\theta} \leq 3.0$. The output saliency $g_x(x_{i\theta})$ ranges in $[0, 1]$. 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 $g_x(.)$ and $g_y(.)$, 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 difference between different examples are the differences in the model inputs $I_{i\theta}$ and possibly the different image grid structure (Manhattan or orthogonal grids) for better input sampling. All the model parameters needed to reproduce the results are listed in the Appendix of the reference (Li, 1998).

Fig. 3A shows a sample input containing two regions. Fig. 3B 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 Fig. (10). Fig. 3C plots the average saliency $S(c)$ of the bars in each column c in Fig. 3B, indicating that the most salient bars are indeed near the region boundary. Fig. 3D confirms that the boundary can be identified by thresholding the output activities using a threshold, denoted as, say, $thre = 0.5$ in Fig. 3D, the fraction of the highest output $\max_{i\theta}\{g_x(x_{i\theta})\}$ in the image. Note that V1 does not perform such thresholding, it is performed only for display purposes. Also, the value of the threshold is example dependent for better visualization. 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_{peak} be average saliency across the most salient grid column parallel and near the boundary, and $\bar{S}$ and σ_s be the mean and standard deviation in the saliencies of non-boundary locations, defined as being at least (say) 3 grid units away from the boundary. Define ($r \equiv S_{peak}/\bar{S}$, $d \equiv (S_{peak} - \bar{S})/\sigma_s$). A salient boundary should give large values for (r, d). One expects that at least one of r and d should be comfortably larger than 1 for the boundaries to be adequately salient. In Fig.

(3), $(r, d) = (4.5, 15.0)$. Notes 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 psychophysical observation that texture boundaries are stronger when the texture elements on one side of them are parallel to the boundaries (Walkson and Landy 1994). Fig (4A) shows an example with the same orientation contrast (90^o) at the boundary but different orientations of the texture bars. Here the saliency values distribute symmetrically across the boundary and the boundary strength is a little weaker. These model behaviors can be physiologically tested.

Fig. 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^o (Fig.(4C)) — 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 (Fig. 4B), this should lead to a bias in the estimation of the border location, as can be experimentally tested. This also suggests that outputs from pre-attentive segmentation need to be processed further by the visual system. The boundary strength also decreases if the orientations of the texture elements are somewhat random or the spacing between the elements increases (Fig. (5)). Boundary detection is difficult when orientation noise $> 30^o$ or when the spacing between bar elements is more than 4 or 5 grid points (or texture element sizes). These qualitative and quantitative results (on the cut off orientation contrast, orientation noise, and bar spacings) compare quite well with human performance on segmentation related tasks (Nothdurft 1985, 1991).

This model also copes well with textures defined by complex or stochastic patterns (Fig. (6)). In both Figs. 6A and 6B, the neighboring regions can be segmented even though they have the same bar primitives and densities. In particular, the two regions in Fig. 6A have exactly the same features, just like that in Fig. 1, and would be difficult to segment using traditional approaches.

When a region is very small, all parts of it belong to the boundary and it pops out from the background, as in Fig. 7A. In addition, Fig. 7B 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. Fig. (8) demonstrates such an example where a cross among bars pops out much more readily than a bar among crosses. Such asymmetry is quite natural in our framework — 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 fields (Knierim and van Essen 1992, Kapadia et al, 1995). In particular, Fig. (9A,B,C,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 colinear contextual bars that form a line. As experimentally observed (Knierim and van Essen 1992), suppression in the model is strongest when the surround bars are of the same orientation as the center bar, is weaker when the surround bars have random orientations, and is weakest when the surround bars are oriented orthogonally to the center bar. The relative degree of suppression is quantitatively comparable to that of the orientation contrast cells observed physiologically (Knierim and van Essen 1992). Similarly, Fig. (9D) closely simulates the enhance-

ment effect observed physiologically (Kapadia et al 1995) when bars in the surround are aligned with the central bar to form a line.

5. Summary and discussions

Summary of the results

This paper makes two main contributions. First, we propose a computational framework for pre-attentive segmentation — segmentation without classification. Second, we present a biologically based model of V1 which implements the framework using contextual influences, and we thereby demonstrate the feasibility of the framework.

Since it does not rely on classification, our segmentation framework is simpler than traditional methods, which explicitly or implicitly require classification. Consequently, not only can our framework be implemented using lower level visual mechanisms as in V1, but also, it avoids the dilemma which plagues the traditional computational approaches — segmentation presumes classification, classification presumes segmentation. A further consequence is that, our framework can easily handle some segmentation examples such as those in Fig. (6A), for which the two regions have the same classification values, that pose problems for the traditional computational approaches, but are easily segmentable by human pre-attentive vision.

Since the computational framework is new, this is the first model of V1 that captures the effect of higher neural activities near region boundaries, as well as its natural consequence of pop-out of small figures against backgrounds and asymmetries in pop-out strengths between choices of figure and ground. The mechanism of the model is 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, our model is the first to use the same neural circuit for both the region boundary effect and contour enhancement — individual contours in a noisy or non-noisy background can also seen as examples of the breakdown of translation invariance in inputs. Putting these effects together, V1 is modeled 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 roles of contextual influences is pre-attentive segmentation.

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 non-linear extensions (e.g., Malik and Perona (1990)) capture well much of the phenomenology of psychophysical performance. However, all the previous models are in the traditional framework of segmentation by classification, and thus differ from our model in principle. For example, the texture segmentation model of Malik and Perona (1990) also employs neural-like (albeit much less realistic) 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 Fig. (1), and would not naturally capture figure pop-out, asymmetries between the figure and ground, or contour enhancement.

By locating the conspicuous image locations without specific tuning to (or classification of) any region features, our model is beyond early visual processing using center-surround filters 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 neighborhood, the neurons natu-

rally 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 in this framework as any feature to which the particular intra-cortical interactions are sensitive and discriminatory. Using orientation dependent interactions between neurons, our model agrees with previous ideas (Northdurft, 1994) that (texture) segmentation is primarily driven by orientation contrast. However the emergent network behavior is collective and accommodates characteristics of general regions beyond elementary orientations, as in Fig. 6. Furthermore, the psychophysical phenomena of filling-in (when one fails to notice a small blank region within a textured region) could be viewed in our framework as the instances when the network fails to highlight enough the non-homogeneity in inputs near the filled-in area.

Our pre-attentive segmentation without classification is quite primitive. It merely segments surface regions from each other, whether or not these regions belong to different visual objects. Furthermore, by not classifying, it does not characterize the region properties (such as by the 2+1/2 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, our model does not say whether a region is made of a transparent surface on top of another surface.

Our framework of segmentation without classification suggests that one should find experimental evidences of pre-attentive segmentation preceding and dissociated from visual classification/discrimination. Recent experimental evidence from V1 (Lamme, Zipser, and Spekreijse 1997, Zipser, private communication 1998) shows that the modulation of neural activities starts at the texture boundary and only later includes the figure surface, where the neural modulations take about 50 ms to develop after initial cell responses (Zipser et al 1996, Zipser, private communication, 1998). Some psychophysical evidences (Scialfa and Joffe 1995) suggest that information regarding (figure) target presence is available before information regarding feature values of the targets. V2 lesions in monkeys are shown to disrupt region content discrimination but not region border detection (Merigan, Mealey, and Maunsell, 1993). These results are consistent with our suggestion. Furthermore, neural modulation in V1, especially those in the figure surface (Zipser 1998, private communication), 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, these experimental evidences suggest the plausibility of the following computational framework. Pre-attentive segmentation without classification 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 to give top-down influence and refine the segmentation (perhaps to remove the bias in the estimation of the border location in the example of Fig. 4B), this latter process might be attentive and can be viewed as segmentation by classification; 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 studies 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 1998). However, we have not yet modeled how higher visual centers process the bottom up signals to generate the feedback.

The model's components and behavior 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 qualitative structure of the horizontal connections as in Fig. 2B, the dependence of the boundary highlights on the relative orientation between texture bars and texture borders (e.g., in Fig. 3B), and the biases in the estimated border location by the neural responses (e.g., Fig. 4B). Since the model is quite simplistic in the connection design, I expect that there will be significant differences between the model and physiological connections. For instance, two linked bars interact in the model either via monosynaptic excitation or disynaptic inhibition. In 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 1998 for analysis). Hence, the excitation (or inhibition) in our model could be interpreted as the abstraction of the predominance of excitation (or inhibition) between two linked bars. Currently, different sources of experimental data on the connection structure are not yet consistent with each other regarding the spatial and orientation dependence of excitation and inhibition (Fitzpatrick 1996, Cavanaugh, Bair, Movshon 1997, Kapadia, private communication 1998, Hirsch and Gilbert 1991, Polat, Mizobe, Pettet, Kasamatsu, Norcia 1998), partly due to different experimental conditions like input contrast levels or the nature of stimulus elements (e.g., bars or gratings). Our model 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 Figs. (4, 5, 9).

The modulations 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 illusion, 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 some recent psychophysical study (Kapadia, private communication, 1998) confirm such contextual dependent distortion in perceived orientation. V1 cells indeed display changes in orientation tunning under contextual influences (Gilbert and Wiesel 1990), although the magnitude and direction of the changes vary from cell to cell.

Comparison with other models

There are many other related models. Many cortical models are mainly concerned with contour linking, and the reference Li (1998) 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, Mingolla, and Ross, 1997). They proposed their '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 presumably 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 neighboring 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. Recently, Lee (1995) used

a Bayesian framework to infer the region features and boundary signals from initial image measurements using gabor filters. The feature and boundary values influence each other to update their values in iterative steps to decrease an energy functional derived from the Bayesian framework. Lee (1995) suggested hypothetically that a V1 circuit may implement this bayesian algorithm.

Our model contrasts to previous models as the only one that models the effect of region boundary highlights in V1. Hence, it is also the only one that models contour enhancement and region boundary highlights in the same neural circuit. 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). Although there are many simulation and computational models of V1, if they are not designed for it, V1 models are unlikely to perform region boundary highlights or contour enhancement. The reference Li (1998) discussed the difficulties in a recurrent network even for mere contour enhancement using only the elements and operations in V1. Our experience also shows that explicit design is necessary for a V1 contour enhancement model to additionally perform region boundary highlights (i.e., to meet conditions (1) and (2) in section 3).

Limitations and extensions of the model

Our 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 color, time, or stereo input dimensions. Also, the receptive field features used for our 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 "homogenous" 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 centers and the RFs from different hypercolumns barely overlap. This sampling arrangement is sparse compared with V1 sampling. Fig. (10) demonstrates the current model performance on a photo. The effects of single scale and sparse sampling (alising) are apparent in the model input image, which is more difficult than the photo image for human to segment. However, the most salient model outputs do include the vertical column borders as well as some of the more conspicuous horizontal streaks in the photo.

In addition to orientation and spatial location, neurons in V1 are tuned for motion direction/speed, disparity, ocularity, scale, and color (Hubel and Wiesel 1962, Livingstone and Hubel 1984). Our model should be extended to stereo, time, and color dimensions. The horizontal connections in the extended model will link edge segments with compatible selectivities to scale, color, ocular dominance, disparity, and motion directions as well as orientations, as suggested by experimental data (e.g., Gilbert 1992, Ts'o and Gilbert 1988). The model should also expand 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 help for better quantitative match between the model and human vision. For instance, Malik and Perona (1990) showed using psychophysical observations that the odd-symmetric receptive fields are not used for pre-attentive segmentation. The design of the horizontal connections between cells should respect these facts.

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 psychophysical 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 1991). It is desirable also 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 accordingly be modified. 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 — segmentation without classification. It introduces a simple and biological plausible model of V1 to implement the framework using mechanisms of contextual influences via intra-cortical interactions. Although the model is yet very primitive compared to the real cortex, our results show the feasibility of the underlying ideas, that region segmentation can occur without region classification, that breakdown of 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.

Appendix: Design analysis for horizontal connections

Connections J and W are designed to satisfy the 3 conditions listed in section 3. To illustrate, consider the example of a homogenous input

$$I_{i\theta} = \begin{cases} \bar{I}, & \text{when } \theta = \bar{\theta} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

of a bar oriented $\bar{\theta}$ 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 $\theta \neq \bar{\theta}$, and ignore all $(x_{i\theta}, y_{i\theta})$ with $\theta \neq \bar{\theta}$. Perturbations $(x'_i \equiv x_{i\bar{\theta}} - \bar{x}_{i\bar{\theta}}, y'_i \equiv y_{i\bar{\theta}} - \bar{y}_{i\bar{\theta}})$ around the mean field solution follow

$$\dot{Z} = AZ \tag{4}$$

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 $J_{i\bar{\theta},j\bar{\theta}}$ and $W_{i\bar{\theta},j\bar{\theta}}$ linking bar segments oriented all at $\bar{\theta}$. 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 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 which case the deviation from the mean field solution tends to be homogeneous although it will oscillate over time (Li, 1998). Iso-orientation suppression under supra-threshold input $\bar{I}$ is used to satisfy condition (2). This requires that every pyramidal cell $x_{i\bar{\theta}}$ in an iso-orientation surround should receive stronger overall disynaptic inhibition than monosynaptic excitation:

$$\sigma \sum_j W_{i\bar{\theta},j\bar{\theta}} > \sum_j J_{i\bar{\theta},j\bar{\theta}} \tag{5}$$

where $\sigma \equiv \psi(0) g'_y(\bar{y}_{i\bar{\theta}})$ 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 are of the form $I_{i\theta} \equiv \bar{I}_\theta$ where $\bar{I}_\theta$ is nonzero for two orientation indices θ. Condition (3) is ensured by strong enough monosynaptic excitation $\sum_{j\theta' \in \text{contour}} J_{i\theta,j\theta'}$ along any smooth contour extending from $i\theta$, and enough disynaptic inhibition between local, similarly oriented, and non-aligned bars to avoid enhancement of the noisy background (details in Li 1998), within the constraints of conditions (1) and (2).

References

[1] J. Allman, F. Miezin, and E. McGuinness (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.

[2] R. Eckhorn, 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**, 121-130.

[3] J. R. Bergen (1991) "Theories of visual texture perception". In *Vision and visual dysfunction* Ed. D. Regan Vo. 10B (Macmillan),pp. 114-134.

[4] J.R. Bergen and E. H. Adelson. (1988) Early vision and texture perception. *Nature* **333**, 363-364.

[5] Carpenter, G & Grossberg, S (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics and Image Processing,* **37**, 54-115.

[6] Cavanaugh, J.R., Bair, W., Movshon, J. A. "Orientation-selective setting of contrast gain by the surrounds of macaque striate cortex neurons" *Soc. Neuroscience Abstract* 227.2, 1997.

[7] Dayan, P, Hinton, GE, Neal, RM & Zemel, RS (1995). The Helmholtz machine. *Neural Computation,* **7**, 889-904.

[8] R. J. Douglas and K. A. Martin (1990) "Neocortex" in *Synaptic Organization of the Brain* ed. G. M. Shepherd. (Oxford University Press), 3rd Edition, pp389-438

[9] Field D.J., Hayes A., and Hess R.F. 1993. "Contour integration by the human visual system: evidence for a local 'associat ion field'" *Vision Res.* 33(2): 173-93, 1993

[10] Fitzpatrick D. "The functional organization of local circuits in visual cortex: Insights from the study of tree shrew striate cortex" *Cerebral Cortex*, V6, N3, 329-341. 1996b

[11] J. L. Gallant, D. C. van Essen, and H. C. Nothdurft (1995) "Two-dimensional and three-dimensional texture processing in visual cortex of the macaque monkey" In *Early vision and beyond* eds. T. Papathomas, Chubb C, Gorea A., and Kowler E. (MIT press), pp 89-98.

[12] Geman S. and Geman D. (1984) " stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images" *IEEE trans PAMI* **6** 721-741.

[13] C. D. Gilbert (1992) "Horizontal integration and cortical dynamics" *Neuron.* **9**(1), 1-13.

[14] Gilbert C.D. and Wiesel T.N. (1983) "Clustered intrinsic connections in cat visual cortex." *J Neurosci.* **3**(5), 1116-33.

[15] Gilbert C. D. and Wiesel T. N. " The influence of contextual stimuli on the orientation selectivity of cells i n primary visual cortex of the cat." *Vision Res* 30(11): 1689-701. 1990

[16] C. M. Gray and W. Singer (1989) "Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex" *Proc. Natl. Acad. Sci. USA* **86**, 1698-1702.

[17] Grenander, U (1976-1981). *Lectures in Pattern Theory I, II and III: Pattern Analysis, Pattern Synthesis and Regular Structures.* Berlin: Springer-Verlag., Berlin, 1976-1981).

[18] Grossberg S. and Mingolla E. (1985) "Neural dynamics of perceptual grouping: textures, boundaries, and emergent segmentations" *Percept Psychophys.* **38** (2), 141-71.

[19] Grossberg S. Mingolla E., Ross W. (1997) "Visual brain and visual perception: how does the cortex do perceptual grouping?" *TINS* vo. 20. p106-111.

[20] Haralick R. M. Shapiro L. G. (1992) *Computer and robot vision* Vol 1. Addison-Weslley Publishing.

[21] D. J. Heeger (1992) "Normalization of cell responses in cat striate cortex" *Visual Neurosci.* **9**, 181-197.

[22] Hirsch J. A. and Gilbert C. D. (1991) ' Synaptic physiology of horizontal connections in the cat's visual cortex." *J. Neurosci.* **11**(6): 1800-9.

[23] Hubel D. H. and Wiesel T. N. (1962) "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex." *J. Physiol.* **160**, 106-154.

[24] B. Julesz. (1962) Visual pattern discrimination *IRE Transactions on Information theory IT-8* 84-92.

[25] B. Julesz. (1981) Textons, the elements of texture perception and their interactions. *Nature* **290**, 91-97.

[26] Karni A, Sagi D. "Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity. *Proc. Natl. Acad. Sci. USA* 88 (11): 4977, (1991).

[27] M. K. Kapadia, M. Ito, C. D. Gilbert, and G. Westheimer (1995) "Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in V1 of alert monkeys." *Neuron.* **15**(4), 843-56.

[28] Kovacs I. and Julesz B. " A closed curve is much more than an incomplete one: effect of closure in figure-ground segmentation." *Proc Natl Acad Sci USA.* 15; 90(16): 7495-7 1993

[29] J. J. Knierim and D. C. van Essen (1992) " Neuronal responses to static texture patterns ion area V1 of the alert macaque monkeys." *J. Neurophysiol.* **67**, 961-980.

[30] Livingstone M. S. and Hubel, D. H. 1984. "Anatomy and physiology of a color system in the primate visual cortex." *J. Neurosci.* Vol. 4, No.1. 309-356. 1984

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

[32] Lamme V. A. F., Zipser K. and Spekreijse H. "Figure-ground signals in V1 depend on consciousness and feedback from extrastriate areas" *Soc. Neuroscience Abstract* 603.1, 1997.

[33] Lee Tai Sing "A bayesian framework for understanding texture segmentation in the primary visual cortex." *Vision Research* Vol. 35, p 2643-57, (1995).

[34] J. B. Levitt and J. S. Lund (1997) Contrast dependence of contextual effects in primate visual cortex. *Nature* **387**(6628), 73-6.

[35] Z. Li (1998) "A neural model of contour integration in the primary visual cortex" in *Neural Computation* in press. Also, see Z. Li, (1997) "A neural model of visual contour integration" in *Advances in neural information processing systems 9*, eds. M. C. Mozer, M. Jordan, and T. Petsche Eds (MIT press, 1997) pp. 69-75.

[36] Z. Li (1997) "Primary cortical dynamics for visual grouping" in *Theoretical aspects of neural computation* Eds. Wong, K.Y.M, King, I, and D-Y Yeung, Springer-Verlag, 1998.

[37] J. Malik and P. Perona. (1990) Preattentive texture discrimination with early vision mechanisms. *J. Opt. Soc. Am. A* **7**(5),923-932.

[38] D. Marr (1982) *Vision* A computational investigation into the human representation and processing of visual information (Freeman).

[39] H. Meinhardt, (1982) *Models of biological pattern formation* (Academic Press).

[40] Merigan W. H. Mealey T.A., Maunsell J. H. (1993) "Visual effects of lesions of cortical area V2 in macaques" *J. Neurosci* **13** 3180-91.

[41] Nothdurft H. C. "Sensitivity for structure gradient in texture discrimination tasks" *Vis. Res.* Vol. 25, p 1957-68, (1985).

[42] Nothdurft H. C. "Texture segmentation and pop-out from orientation contrast". *Vis. Res.* Vol. 31, p 1073-78, (1991).

[43] H. Nothdurft (1994) "Common properties of visual segmentation" in *Higher-order processing in the visual system* eds. Bock G. R., and Goode J. A. (Wiley & Sons), pp245-268

[44] Polat U. Mizobe K. Pettet M. Kasamatsu T. Norcia A. "Collinear stimuli regulate visual responses depending on cell's contrast threshold" *Nature* vol. 391, p. 580-583, 1998.

[45] K.S. Rockland and J. S. Lund (1983) " Intrinsic Laminar lattice connections in primate visual cortex" *J. Comp. Neurol.* **216**, 303-318

[46] Scialfa C. T. and Joffe K. M. "Preferential processing of target features in texture segmentation. " *Percept Psychophys* 57(8). p1201-8. (1995).

[47] A. M. Sillito, K. L., Grieve, H. E. Jones, J. Cudeiro, and J. Davis (1992) " Visual cortical mechanisms detecting focal orientation discontinuities" *Nature* **378**(6556), 492-6.

[48] Sireteanu R. and Rieth C "Texture segregation in infants and children" *Behav. Brain Res.* 49. p. 133-139 (1992).

[49] D. C. Somers, E. V. Todorov, A. G. Siapas, and M. Sur (1995) "Vector-based integration of local and long-range information in visual cortex" *A.I. Memo. NO. 1556,* MIT.

[50] Stemmler M, Usher M, Niebur E (1995) "lateral interactions in primary visual cortex: a model bridging physiology and psychophysics" *Science* **269**, 1877-80.

[51] Treisman A. and Gormican S. (1988) "Feature analysis in early vision: evidence for search asymmetries" *Psychological Rev.* **95**, 15-48.

[52] Ts'o D. and Gilbert C. "The organization of chromatic and spatial interactions i n the primate striate cortex" *J. Neurosci.* 8: 1712-27. 1988

[53] Weliky M., Kandler K., Fitzpatrick D. and Katz L. C. (199 5) " Patterns of excitation and inhibition evoked by horizontal connections in visual cortex share a common relationship to orientation columns" *Neurons* **15**, 541-552.

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

[55] E. L. White (1989) *Cortical circuits* (Birkhauser).

[56] Wolfson S. and Landy M. S. "Discrimination of orientation-defined texture edges" *Vis. Res.* Vol. 35, Nl. 20, p 2863-2877, 1995.

[57] K. Zipser, V. A. Lamme, and P. H. Schiller (1996) " Contextual modulation in primary visual cortex." *J. Neurosci.* **16** (22), 7376-89.

Acknowledgement: 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, 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.

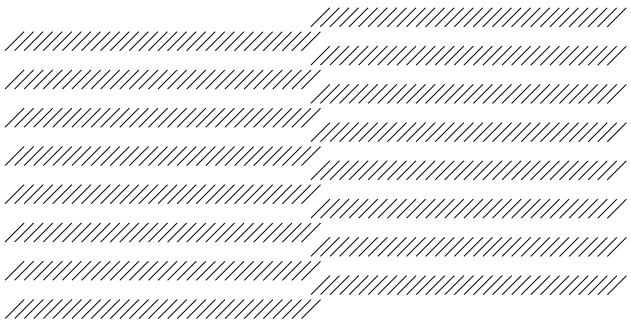

Figure 1: The two regions have the same feature values. Traditional approaches to segmentation using feature extraction and comparison have difficulty in segmenting the regions.

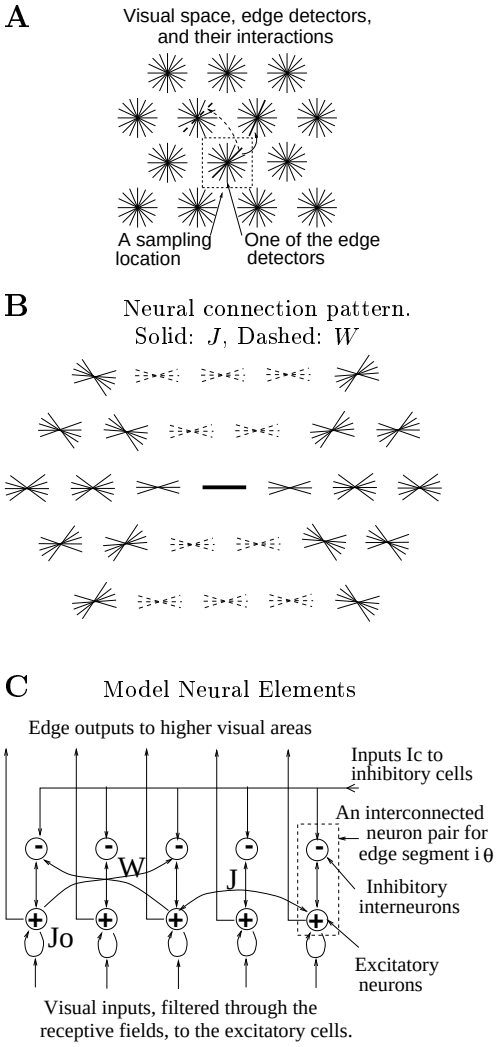

Figure 2: **A:** Visual inputs are sampled in a discrete grid by edge/bar detectors, modeling RFs for V1 layer 2-3 cells. Each grid point has K neuron pairs (see **C**), one per bar segment. All cells at a grid point share the same RF center, but are tuned to different orientations spanning 180^o, thus modeling a hypercolumn. A bar segment in one hypercolumn can interact with another in a different hypercolumn via monosynaptic excitation J (the solid arrow from one thick bar to another), or disynaptic inhibition W (the dashed arrow to a thick dashed bar). See also **C**. **B**: A schematic of the neural connection pattern from the center (thick solid) bar to neighboring bars within a finite distance (a few RF sizes). J's contacts are shown by thin solid bars. W's are shown by thin dashed 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 centers. The inhibitory cell is an interneuron. The visual space has toroidal (wrap-around) boundary conditions.

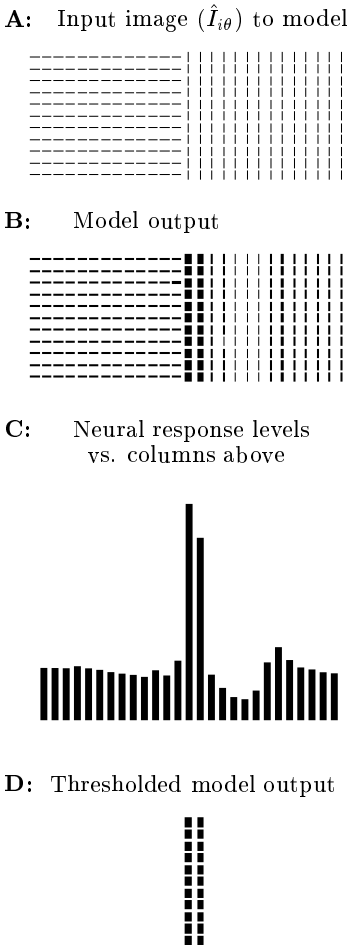

Figure 3: An example of the segmentation performance of the model. **A**: Input $\hat{I}_{i\theta}$ consists of two regions; each visible bar has the same input strength. **B**: Model output for **A**, showing non-uniform output strengths (temporal averages of $g_x(x_{i\theta})$) for the edges. The input and output strengths are proportional to the bar widths. **C**: Average output strengths (saliencies) vs. lateral locations of the columns in **B**, with the bar lengths proportional to the corresponding edge output strengthes. **D:** The thresholded output from **B** for illustration, $thre = 0.5$. Boundary saliency measures $(r, d) = (4.5, 15.0)$.

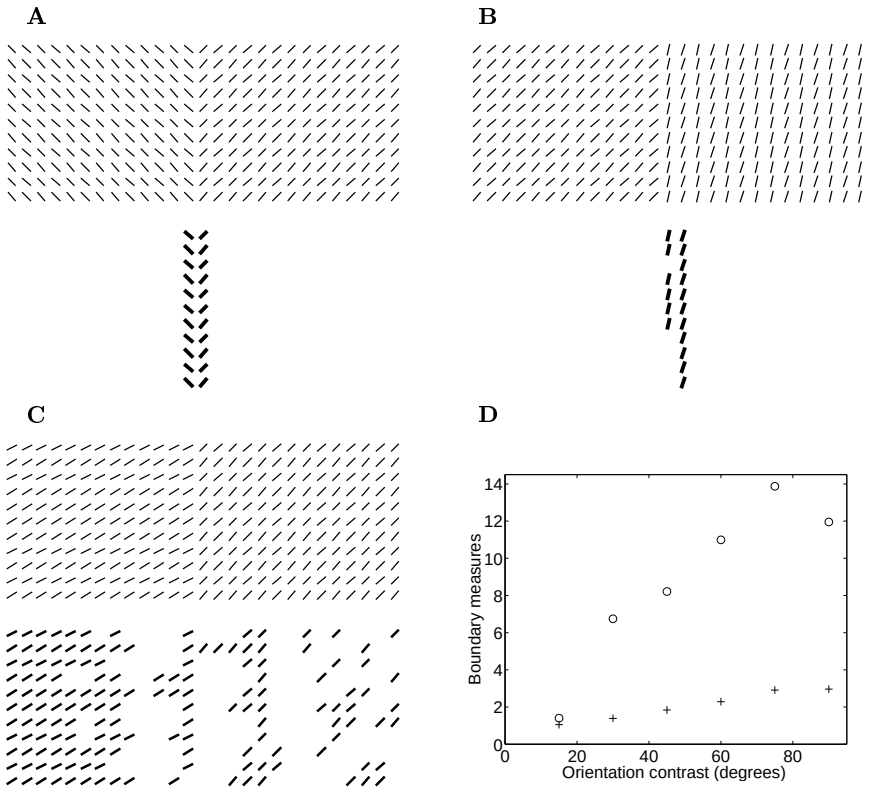

Figure 4: **A, B, C:** Additional examples of model segmentation. Each is an input image as in Fig. 3A followed immediately below by the corresponding thresholded (strongest) model outputs as in Fig. 3D. In **A, B, C** respectively, the boundary measures are: $(r, d) = (1.4, 9.0)$, $(r, d) = (1.77, 12.2)$, $(r, d) = (1.05, 1.24)$, and $thre = 0.77, 0.902, 0.8775$ to obtain the output highlights. **D:** Plots of boundary strengths (r,d) (symbols '+' and 'o' respectively) vs. orientation contrast at boundaries. A data point for each given orientation contrast is the average of 2 or 3 examples of different texture bar orientations. Again, each plotted region is only a small part of a larger extended image. Note that the most salient column in **B** is not exactly on the boundary, though the boundary column (on its left) is only 6% less salient numerically, and $\sim 70\%$ more salient than areas away from the boundary. Also, **C** contains two regions whose bar elements differ only slightly in orientation, giving a perceptually weak vertical boundary in the middle. Because of the noise in the system, the saliencies of the bars in the same column in **A, B, C** are not exactly the same, this is also the case in other figures.

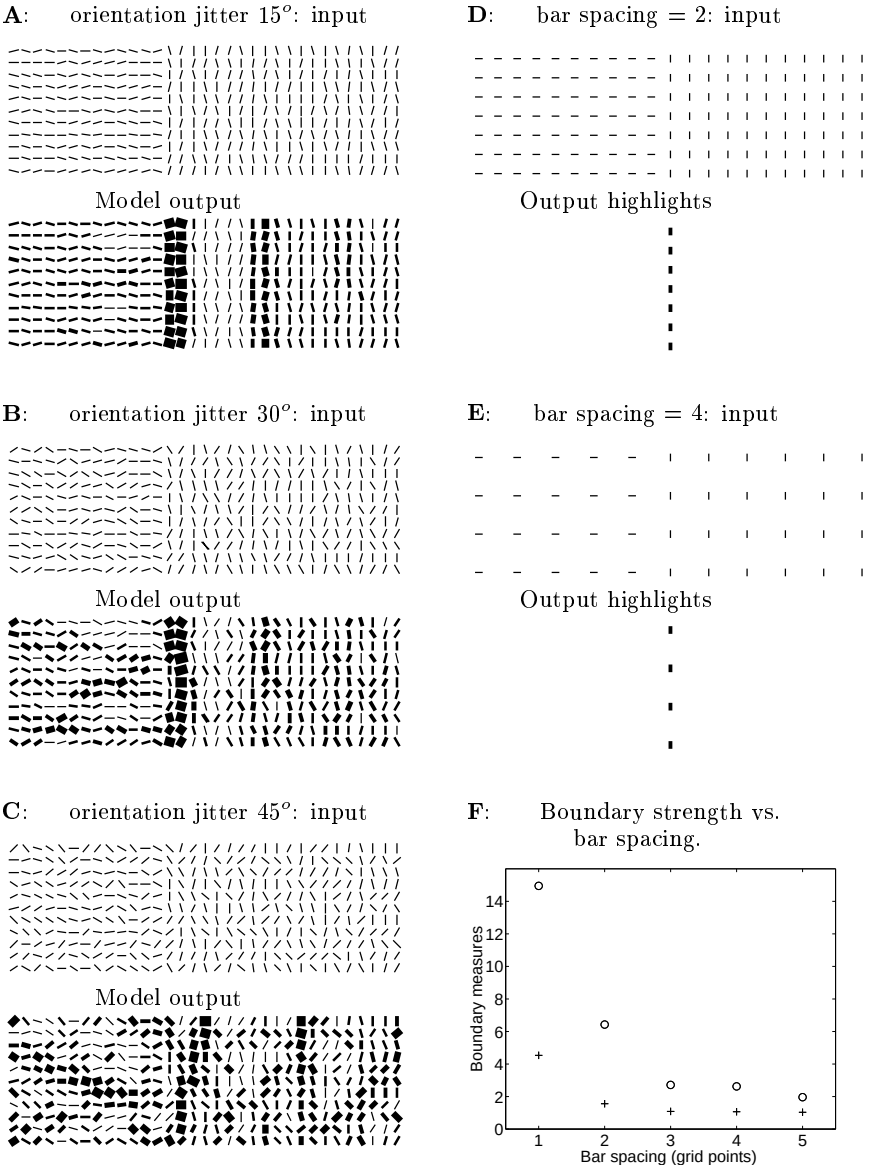

Figure 5: The boundary strength changes with orientation noise and the spacings between the bars in the textures. **A, B, C:** Model inputs ($\hat{I}_{i\theta}$) and outputs ($g_x(x_{i\theta})$) for two texture regions made of bars oriented, on average, respectively, horizontally and vertically. Each bar's orientation is randomly jittered from the average orientation by up to 15^o, 30^o, and 45^o, respectively. The orientation noise makes the saliency values quite non-uniform near the boundary, making the boundary measures (r, d) less meaningful. Boundary detection is difficult or impossible with orientation jitter $> 30^o$. **D, E:** Model inputs ($\hat{I}_{i\theta}$) and output ($g_x(x_{i\theta})$) highlights for two texture regions made of bars oriented horizontally and vertically. The spacing between neighboring bars are 2 and 4, respectively, grid spacings. **F:** Plots of boundary strengthes (r, d) (symbol '+' and 'o' respectively) vs. bar spacings for stimuli like **D, E**. To obtain output highlights in **D, E** respectively, $thre = 0.92, 0.95$. Note that although the boundary saliency is only a fraction higher than the non-boundary saliency as bar spacing increases, the boundary is still the most salient output when the region features are not noisy. The line widths for model outputs are plotted with one scale for **A, B, C** and another for **D, E**.

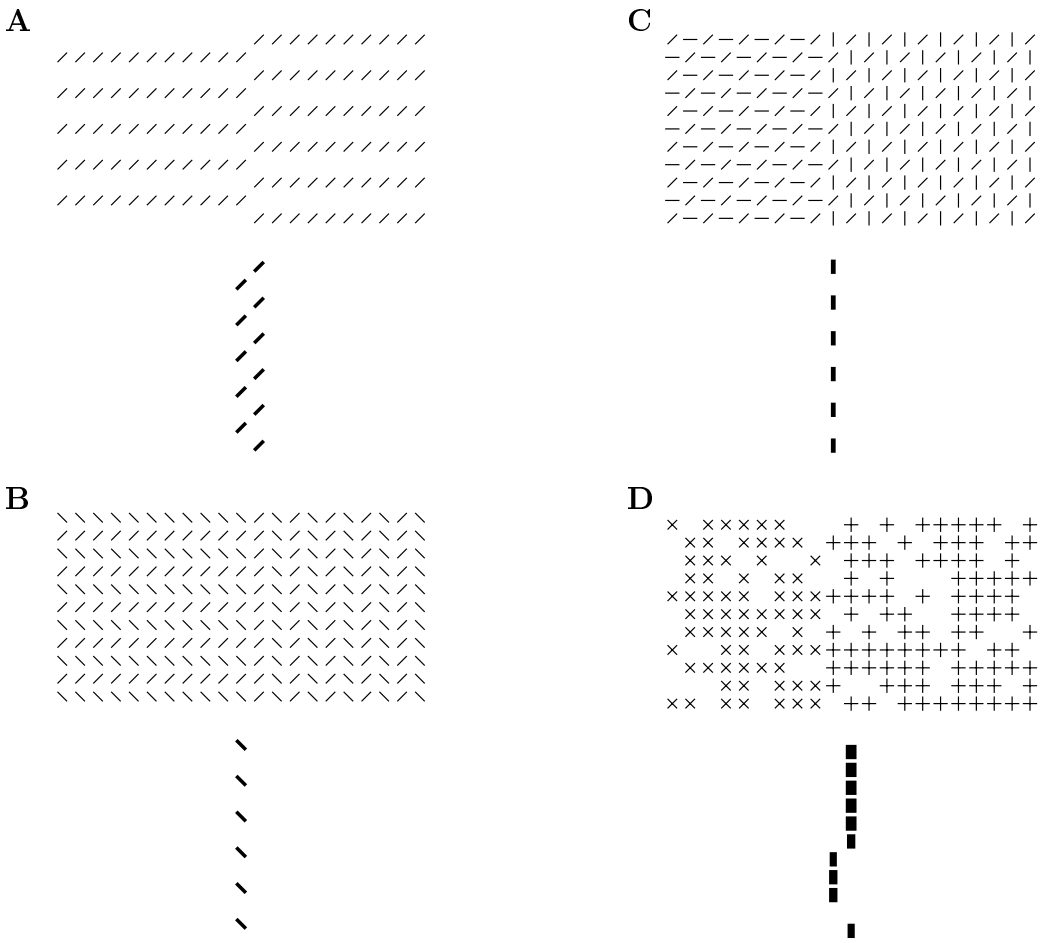

Figure 6: **A, B, C:** Model performance on regions with complex texture elements, and **D:** regions with stochastic texture elements. Each plot is the model input ($\hat{I}_{i\theta}$) followed immediately below by the output ($g_x(x_{i\theta})$) highlights. For **A, B, C, D** respectively, the boundary measures r and d are $(r, d) = (1.14, 6.3)$, $(r, d) = (1.1, 2.0)$, $(r, d) = (1.5, 4.5)$, and $(r, d) = (2.56, 5.6)$, the threshold to generate the output highlights are $thre = 0.91, 0.9, 0.85, 0.56$.

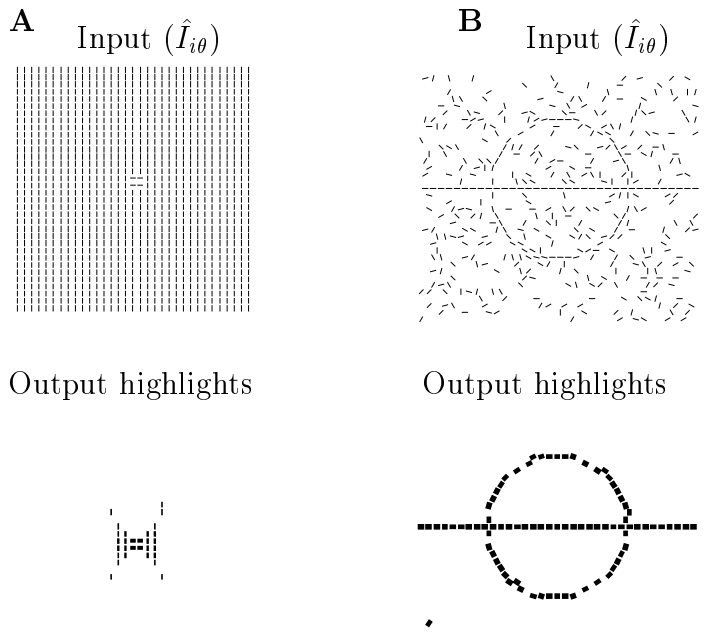

Figure 7: Model behavior for other types of inputs. **A:** A small region pops out since all parts of it belong to the boundary. The figure saliency is 0.336, which is 2.42 times of the average ground saliency. **B:** Exactly the same model circuit (and parameters) performs contour enhancement. The input strength is $\hat{I}_{i\theta} = 1.2$. The contour segments' saliencies are 0.42 ± 0.03, and the background elements' saliencies are 0.18 ± 0.08. To obtain the output highlights in **A, B** respectively, $thre = 0.46, 0.73$.

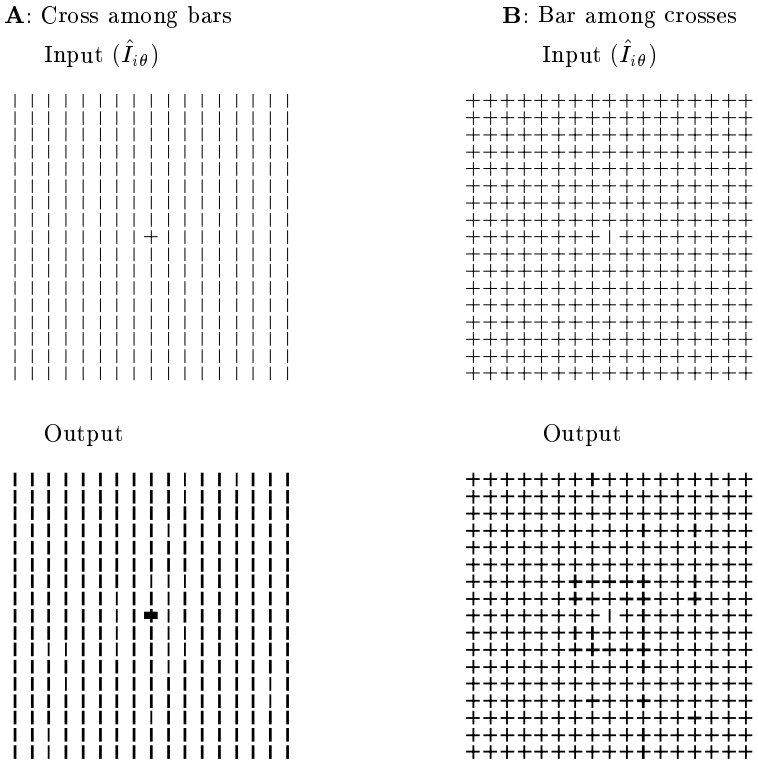

Figure 8: Asymmetry in pop-out strength. **A**: The cross is 3.4 times as salient (measured as the saliency of the horizontal bar in the cross) as the average background. **B**: The area near the central vertical bar is the most salient part in the image, and is no more than 1.2 as salient as the average background. The target bar itself is actually a bit less salient than the average background.

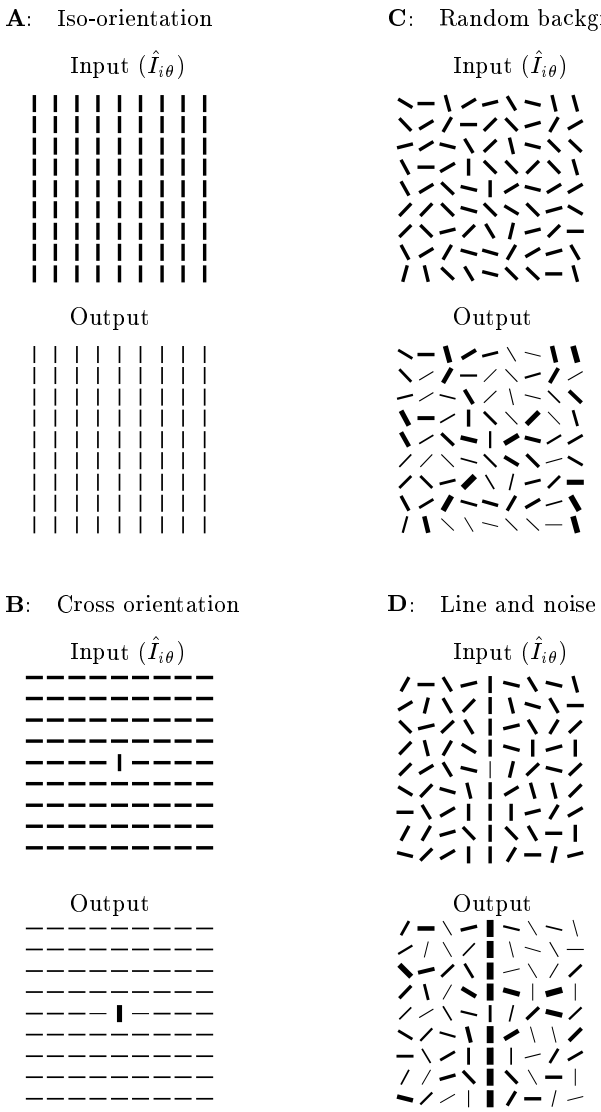

Figure 9: Model behavior under inputs resembling those in physiological experiments. The input stimuli are composed of a vertical (target) bar at the center surrounded by various contextual stimuli. All the visible bars have high contrast input $\hat{I}_{i\theta} = 3.5$ except for the target bar in **D** 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 (center) 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 (center) target bar is aligned with some high contrast contextual bars to from 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, Hayes, and Hess 1993).

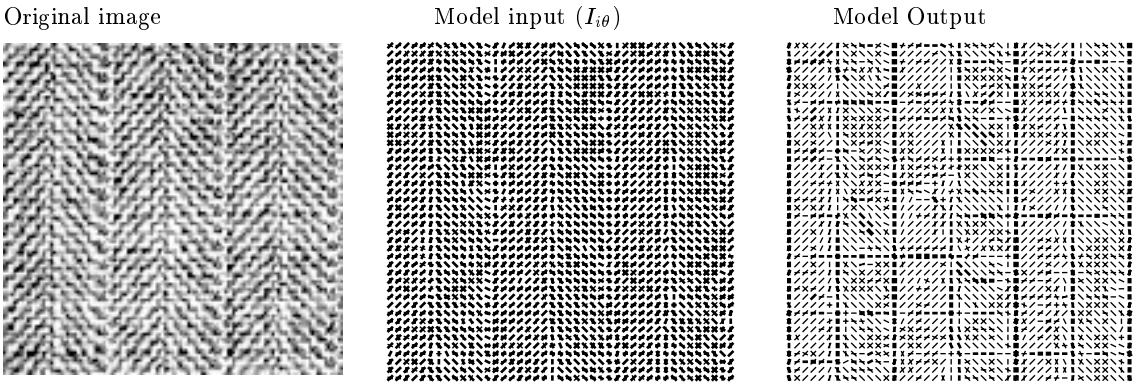

Figure 10: Model behavior on a photo image. The input to the model is modeled as $I_{i\theta} = (e^2 + o^2)^{1/4}$, where e and o are the outputs from the even and odd gabor-like filters at grid sampling point i with preferred orientation θ, the power 1/4 coarsely models some degree of contrast gain control. At each grid point, bars of almost all orientations have nonzero input values $I_{i\theta}$. For display clarity, no more than 2 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 is 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 photo is sampled against a blank background on the left and right, hence the left and right sides of the photo area are also highlighted.