Primary Cortical Dynamics for Visual Grouping

Zhaoping Li

Published in *Theoretical aspects of neural computation*,
K.M. Wong, I. King, D-Y Yeung (Eds), Springer-verlag. January, 1998

Abstract. The classical receptive fields of V1 neurons contain only *local* visual input features. The visual system must group separate *local* elements into meaningful *global* features to infer the visual objects in the scene. Local features can group into regions, as in texture segmentation; or into contours which may represent boundaries of underlying objects. I propose that the primary visual cortex (V1) contributes to both kinds of groupings with a single mechanism of cortical interactions/dynamics mediated by the horizontal connections, and that the dynamics enhance the saliencies of those features in the contours (compared with those in a noisy background) or near the region boundaries (compared with those away from the boundaries). Visual inputs specify the initial neural activity levels, and cortical dynamics modify the neural activities to achieve desired computations. Contours are thereby enhanced through dynamically integrating the mutual facilitation between contour segments, while region boundaries are manifested (and enhanced) in the dynamics because of the breakdown of translation invariance in image characteristics at the region boundaries. I will show analytically and empirically how global phenomena emerge from local features and finite range interactions, how saliency enhancement relates to the contour length and curvature, and how the neural interaction can be computationally designed for region segmentation and figure-ground segregation. The structure and behavior of the model are consistent with experimental observations.

1 Introduction

Visual inputs are first sampled as pixels. Subsequently, the images are processed by local transforms, such as the receptive fields in the primary visual cortex, to give local image features such as edge segments or bars. However, these local features are too small to represent global visual objects. The visual system must group local features into global and more meaningful ones for visual recognition and visual-motor tasks. One is to group local edge segments into global contours, and the other is to group local features into regions, as in texture segmentation. Global contours sometimes mark boundaries of regions. Other times, regions such as those defined by textures do not have definite or visible markings for boundaries, which we humans can nevertheless locate easily. In any case, a region and its boundary are complementary to each other. Knowing one can infer the other. It is desirable to have a single computational mechanism for detection or grouping of both contours and regions.

Both contour and region groupings are very important for visual segmentation, which is still a formidable problem in computer vision after more than two decades of research efforts (Kasturi and Jain 1991). The problem with contour grouping is that there are many candidate edge segments after the edge detection operation on an image, many of them are simply "noisey" contrast elements not belonging to any meaningful contour. The grouping algorithm has to discriminate between "signals" and "noises" using contextual information. Many computer vision algorithms on edge linking need user intervention, though more autonomous algorithms exist and they work under certain conditions (e.g., Shashua and Ullman 1988, see more references in Li 1997). For region grouping, all existing approaches require image feature extraction and/or region classification as a preprocessing stage to compute feature values or classification flags for every small area in an image. The regions are differentiated by these feature values to locate the boundaries (Haralick and Shapiro 1992). Such approaches have problems near the boundaries where features are indefinite. Furthermore, they can not segment two regions in Fig. 1 where the two regions would have the same feature or classification values. Segmentation outcomes from edge and region based approaches usually do not agree with each other, even though the two kinds of algorithms have been combined for better visual segmentation performance (Kasturi and Jain 1991). There is so far no *single* algorithm or mechanism that deals with both contour and region groupings.

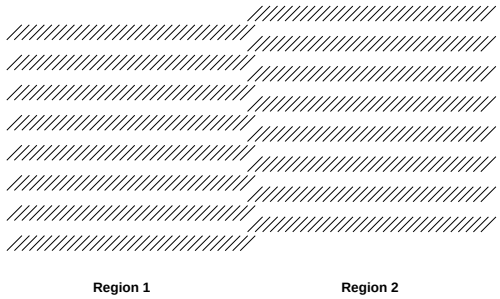

Fig. 1. The feature values in the two regions are the same. However one easily sees the boundary between two regions. Traditional approaches to segmentation using feature extraction and comparison will not be able to segment the regions in this example.

Here I propose that the first step towards contour and region grouping is to enhance the saliencies of image elements on contours or near region boundaries against non-contour elements or elements away from the boundaries. In addition, I propose that a single mechanism, using the cortical interactions in the primary visual cortex, suffices for both kinds of saliency enhancement. This operation serves the most difficult task in contour and region grouping — to discriminate between contour and non-contour elements, or to locate the boundary elements. By using a same language — saliency — to distinguish both contours and region boundaries from background, it is feasible to have a single algorithm for both grouping purposes.

One may find it easier to accept saliency enhancement for contour elements than that for elements near region boundaries. In fact, it is only natural to enhance or mark the region boundaries for segmentation. This is because segmentation necessarily means boundary localization. A mere classification flag at every image area is sometimes not useful or necessary, as indicated by the counter example in Fig. (1), and at other times require an additional step to differentiate the classification values in order to segment.

A model of V1 is constructed to implement the proposal. The contextual influences beyond the classical receptive fields modify the neural activity levels initialized by external inputs to achieve desired visual computation. It will be shown analytically and empirically that contours are thereby enhanced through dynamical integration of the contextual facilitation along the contour, and that the enhancement increases with contour length and smoothness. On the other hand, region boundaries are manifested (and enhanced) in the dynamics (mediated by the translation invariant cortical neural connections) by the breakdown of translation invariance in image characteristics. Since translation symmtry breaking can be detected without feature classifications, our approach performs region segmentation without region classification. While boundary resions are problematic for traditional segmentation-by-classification approaches, they are high-lights in the present approach. The structure and behavior of the model are consistent with the experimental observations (Kapadia, Ito, Gilbert, and Westheimer 1995, Gallant, van Essen, Nothdruft 1994).

2 A V1 model of contour enhancement and region boundary enhancement

2.1 Model elements and structure

The selective saliency enhancing network is a model of V1. It consists of K neuron pairs at each spatial location i modeling a hypercolumn in V1 (Fig. 2). Each neuron has a receptive field center i and an optimal orientation $\theta = k\pi/K$ for $k = 1, 2, ...K$. A neuron pair consists of a connected excitatory neuron and inhibitory neuron which are denoted by indice $(i\theta)$ for their receptive field center and preferred orientation, and are referred to as an edge segment. An edge segment receives the visual input via the excitatory cell, whose output quantifies the saliency of the edge segment and projects to higher visual centers. The inhibitory cells are treated as interneurons. Such a local cortical circuit is modeled after the experimental observations (White 1989, Douglas and Martin 1990). Some edge elements excite each other via finite range excitatory-to-excitatory connections $J_{i\theta,j\theta'}$, while others inhibit each other via finite range disynaptic inhibition $W_{i\theta,j\theta'}$ which are excitatory-to-inhibitory connections. The system dynamics follow the equations of motion

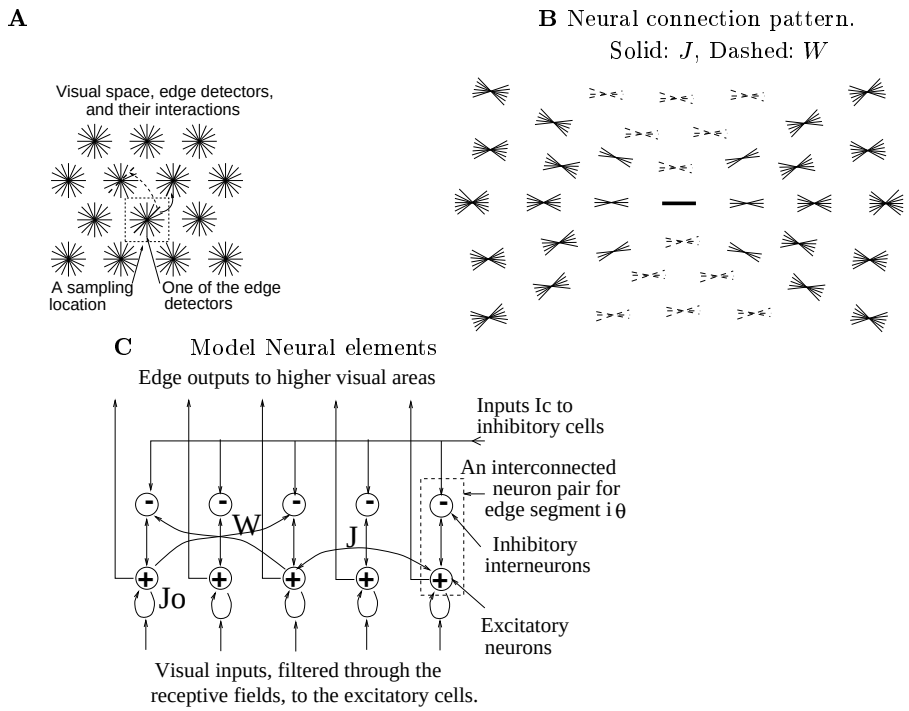

Fig. 2. A: Visual inputs are sampled in a discrete grid space by edge detectors modeling cells in V1. **B**: Edges interact with each other via monosynaptic excitation J (to thin solid edges) and disynaptic inhibition W (to dashed edges) within a finite distance (see also **C**). This is a coarse illustration of the neural connection pattern from the center (thick solid) edge to other edges in the neighborhood. All edges have the same connection pattern, suitably translated and rotated from this one. **C:** An edge segment is processed by an inter-connected pair of excitatory and inhibitory cells. The excitatory cells receive visual inputs and send outputs to higher centers.

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

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

where x and y are membrane potentials of the excitatory and inhibitory cells, $g_x(x)$ and $g_y(y)$ are cell output firing rates, $1/\alpha_x$ and $1/\alpha_y$ are the membrane time constants, $I_{i\theta}$ is the visual input, I_o and I_c are background or internally gen-

erated inputs, J_o is the self excitation connection, and $\psi(\Delta\theta)$ models the mutual inhibition spread within a hypercolumn. More details of the model parameters and structures can be found in (Li 1997).

The visual input pattern is $I_{i\theta}$, which is transformed by the neural interactions and dynamics to give an output pattern $g_x(x_{i\theta})$, the cell output activities from the excitatory cells. Usually $g_x(x_{i\theta}) \not\propto I_{i\theta}$. The relationship between the input I and output g_x patterns is determined by the network structure, in particular the neural connections J and W that mediate the contextual influences and induce the cortical dynamics. Therefore, J and W should be designed such that the network selectively enhances the saliencies $g_x(x_{i\theta})$ for image elements $(i\theta)$ within contours or near region boundaries.

2.2 Computational design of the cortical interactions for contextual influences

The neural connection structure is designed to satisfy the following conditions.

1. The connection strengths decreases with increasing distance between the edge segments, and becomes zero for large distances.
2. The connection structure has translation, rotation, and reflection invariance. This means the following. Let $i-j$ be the line connecting the centers of two edges $(i\theta)$ and $(j\theta')$, which form angles θ_1 and θ_2 with this connecting line. The connections $J_{i\theta,j\theta'}$ and $W_{i\theta,j\theta'}$ depend only on $|i-j|$, θ_1, and θ_2, and satisfy $J_{i\theta,j\theta'} = J_{j\theta',i\theta}$ and $W_{i\theta,j\theta'} = W_{j\theta',i\theta}$.
3. The connections should be such that the network gives stable and computationally desirable behavior: The network amplification (caused mainly by excitatory connections J) should be enough to give significant saliency enhancement to selected image elements, but not too much such as to unselectively give high saliencies to all elements in the image grid.
4. The mutual facilitation $J_{i\theta,j\theta'}$ between two edges $i\theta$ and $j\theta'$ is large if one can find a smooth or small curvature contour to connect $(i\theta)$ and $(j\theta')$, and generally decreases with increasing curvature of the contour.
5. The mutual inhibition $W_{i\theta,j\theta'}$ between two edges $i\theta$ and $j\theta'$ is strong when they are alternative choices in the route of a smooth contour, i.e., when they are close, have similar orientations, and displaced roughly in a direction perpendicular to their orientations.
6. The connection should be such that a translation invariant input pattern I (e.g., when $I_{i\theta}$ does not depend on i) will lead to a translation invariant output pattern g_x. This means, the system should not have spontaneous translation symmetry breaking (spontaneous pattern formation).
7. The balance between excitation J and inhibition W should be such that under a translation invariant input I, each visible element $i\theta$ receives an overall inhibition (at least under not too low an input strength) after combining the contextual influences from all the neighboring elements.

Condition (1) requires local interaction for global grouping behavior. Feromagnetism is another example in nature of global behavior with local interac-

tions. Condition (2) ensures view point independence of the desired computation. In addition, the translation symmetry in interaction is required to detect the translation symmetry breaking at the boundary between two image regions. Condition (3) ensures that the model output is under the input control in a computationally desirable way. For instance, when the input contains a contour of finite length among a noisy background, the network should not extend the contour to infinite length, nor should it leave the contour unenhanced against the noisy background. Conditions (4) and (5) are for the contour enhancement, relative to the background. Note that these two conditions imply that two edges of similar orientations are more likely to interact with each other (see Fig. (2), whether it is mutual excitation or inhibition. Condition (6) ensures that the system does not find any region boundaries when it should not. Spontaneous pattern formation under translation invariant interactions are not uncommon in nature. Zebra stripe formation is one such example. Hence the model interaction should be within the subset of all translation invariant interactions that avoids spontaneous pattern formation. Condition (7) ensures that when translation symmetry is broken by an input image, the region boundary area has higher, rather than lower, saliency values than areas away from the boundaries. This is because by conditions (4) and (5), similar image elements interact with each other more strongly than non-similar elements. The image elements near region boundaries are surrounded by fewer similar neighbors, and consequently receive less overall inhibition. In the example of a region composed of many parallel lines, combining conditions (4), (5) and (7) leads to the following: a line segment in the middle of the region receives less contour enhancement excitation than the overall iso-orientation suppression from its flanking neighbors in nearby parallel lines.

2.3 Contour integration

In addition to the external visual input $I_{i\theta}$, an edge element $i\theta$ within a contour also receives excitation $\Delta I \equiv \sum_{j\theta' \in \text{contour}, j\theta' \neq i\theta} J_{i\theta,j\theta'} g_x(x_{j\theta'})$ from other contour elements $j\theta'$. To analyse contour enhancement, consider for simplicity an edge segment in a long enough curve whose curvature is changing slowly enough, and assume that there is no inhibition between contour elements. Then it can be shown (Li, 1997) that the response ratio between a curve segment and an isolated segment is $(g'_y(\bar{y}) + 1 - J_o)/(g'_y(\bar{y}) + 1 - J_o - \sum_{j\theta' \in \text{contour}, j\theta' \neq i\theta} J_{i\theta,j\theta'})$, where $\bar{y}$ is the average response of the inhibitory interneurons and g' is the derivative of g. Therefore, the degree of contour enhancement increases with $\sum_{j\theta' \in \text{contour}, j\theta' \neq i\theta} J_{i\theta,j\theta'}$, the integration of mutual excitation connection within a contour. Since such an integration is computationally designed to increase with contour smoothness and length, one can then relate the degree of enhancement with these contour characteristics (see Li 1997 for more examples and detailed analysis). In the computational design for $J_{i\theta,j\theta'}$, the scale of J should be chosen such that the integration of facilitation $J_{i\theta,j\theta'}$ along a contour is enough for significant saliency enhancement within a contour, even to fill in the gaps in a

contour, but not enough to excite segments beyond the ends of a contour. Fig. (3) demonstrates the performance of contour enhancement against noise.

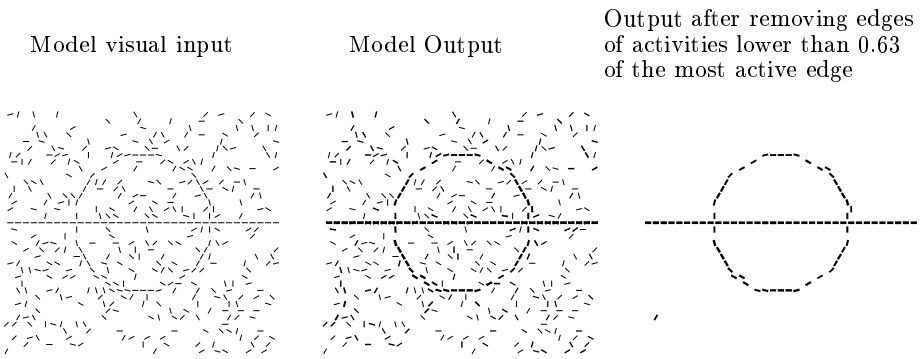

Fig. 3. Contour enhancement and noise reduction. The input and output edge strengths are denoted proportionately by the thicknesses of the edges. The same format applies to other figures in this paper. The model outputs are the temporal averages of $g_x(x)$. All visible edges have the same strength in the input, and are differentially enhanced or suppressed at the output. On average, the line and circle segments are roughly 2.5 times as salient as the "noise" segments. For demonstration, we display the outputs after thresholding out the weaker edges (right). Note that the apparent gaps between the circle segments are caused by the lack of sampling points there in the particular sampling grid arrangement. No gaps actually exist and hence no filling-in is needed.

2.4 Region boundary enhancement

In order to have well controlled region boundary enhancement, the synaptic connections J and W are examined to see whether conditions (6) and (7) in section 2.2 are satisfied. We check this for translation invariant inputs $I_{i\theta} = \tilde{I}$ for all i and any given θ, and $I_{i\theta'} = 0$ for $\theta' \neq \theta$. Find the mean field solution $\bar{x}$ and $\bar{y}$ by setting to zero the right hand sides of the equations (1) and (2), and setting $x_{i\theta} = \bar{x}$, $x_{i\theta'} = 0$ for $\theta' \neq \theta$, $y_{i\theta} = \bar{y}$. Condition (6) is satisfied if this mean field solution is stable, or if it is unstable, whether the dominant mode in the deviation from the mean field solution is also translation invariant. Stability and dominant mode analysis are studied by a perturbation analysis around the mean field solution, using the linear expansion or small amplitude approximation. Condition (7) is satisfied for that input if $\sum_{j\neq i} J_{i\theta,j\theta} < \sum_{j\neq i} W_{i\theta,j\theta} g_y'(\bar{y})\psi(0)$ for a reasonable range of $\bar{y}$. This inequality is derived by noting that edges excite each other directly by mono-synaptic, excitatory-to-excitatory connections J, and inhibit each other indirectly by disynaptic, excitatory-to-inhibitory connections W. Conditions (6) and (7) are further confirmed by simulations. After the conditions (6) and (7) are met for all choices of θ for those translation invariant input images, we hope, and check by some simulation examples, that the same conditions are also met for arbitrary translation invariant images.

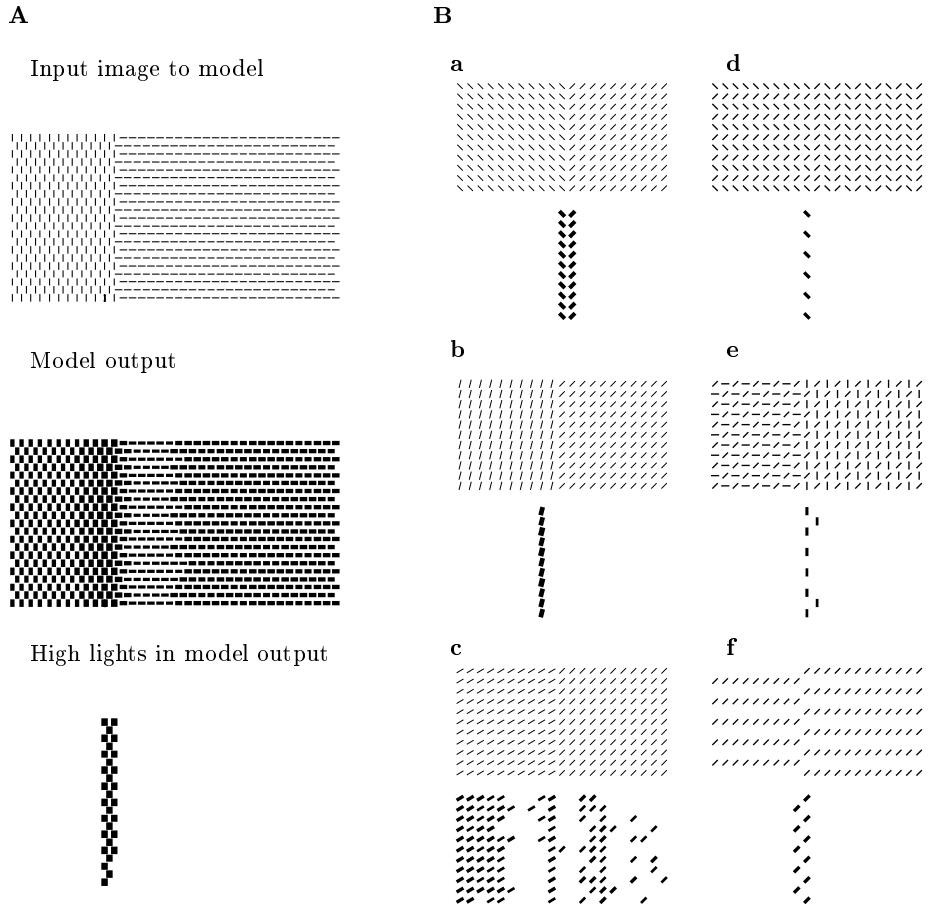

Fig. 4. **A:** An example of region boundary enhancement. Note that the output activities $g_x(x_{i\theta})$ are higher near the boundaries even though each visible edge has the same input strength. **B:** Six additional examples (**a, b, c, d, e** and **f**) of model input images, each followed by the corresponding output high lights immediately below it. Note that both humans and the network find it difficult to segment two regions in the example **c**. Traditional segmentation techniques can not segment **f** (cf. Fig. (1)).

Fig. (4) shows the model performance for some examples on region boundary enhancement. Note that the plots in Fig. (4) and later figures include only small portions of the input and output images in the model for illustration purpose. So the boundaries of the plots should not be taken as the boundaries of the texture regions (and hence they are not high-lighted at the model outputs). This model can also enhance boundaries for regions defined by stochastic image elements (Fig. (5)). The pop-out phenomena can also be accounted for — when a region is very small, all parts of the region belong to the boundary. The small region is

thus enhanced as a whole and pops out from the background (Fig. (5c).

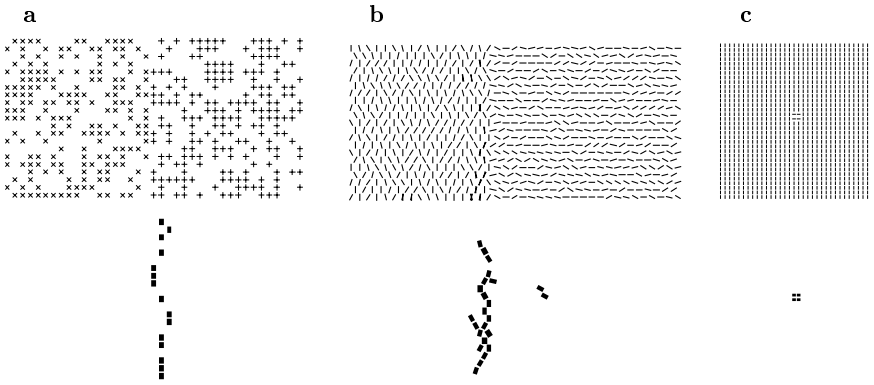

Fig. 5. Examples of segmentation of stochastic texture regions **a** and **b**, and popout **c**. Visual inputs are at the top row, followed by the respective output high-lights below.

The principle and mechanism of region boundary enhancement can be intuitively understood by the following analogy with physical systems. One may take an edge element as an atom, a composite pattern primitive such as a "+" in Fig. (5**a**) as a molecule, interactions between edges or composite pattern primitives as interactions between atoms and molecules, regular texture regions as lattice structured materials, and stochastic regions as non-lattice liquids such as glass. Usually, two blocks of different materials somehow joined together are likely to break near their junctions, because molecular interactions are not translation invariant near the junctions and they are manifested by stronger molecular vibrations there. Analogously, neural activities near the region boundaries are likely to be the high lights, relatively enhanced by the underlying neural interactions.

3 Summary and Discussion

This paper proposes that groupings of local visual features into global contours and regions can be carried out in the first stage by local, finite range, neural interactions to enhance the saliencies of image elements within contour or near the region boundaries. This proposal is implemented in a model of V1 composed of edge/bar detectors and horizontal connections mediating contextual influences. The structure of the horizontal connections is computationally designed for the requirement of contour and region boundary enhancement. The model is studied analytically and empirically to understand that contours are enhanced by integrating the mutual facilitation between contour segments, while region boundaries are detected by the breakdown of translation invariance in the image characteristics near the boundaries. The performance of the model is demonstrated by examples.

The main contributions of this work are the following. First, it is the first of the kind to deal with both contour and region groupings with a single mechanism. Regions and their boundary contours are complementary to each other. It is computationally desirable to handle both groupings by a single mechanism. Second, this work introduces an entirely new approach to region segmentation — region segmentation without region classification. It avoids some problems and the ad hoc flavor in the traditional approach to region segmentation of the last two decades, and is computationally simpler. Third, compared to many other models (see references in Li 1997), this model achieves contour enhancement using only known V1 neural elements and interactions without requiring higher visual centers or biologically non-plausible operations. (It has been difficult to model contour enhancement using only V1 elements largely because of the dynamic stability problems in a recurrent neural network.) It thus answers the question of whether the difficult problem of global contour integration could be first attempted in a lower visual stage such as V1. In fact, even though the structure of the neural connections in this model is designed by the computational requirements of contour and region grouping, this structure is consistent with the structure observed physiologically (e.g., Gilbert 1992). The behavior of this model in both the contour and region boundary enhancement is also consistent with experimental data (Kapadia et al. 1995, Gallant et al 1994). These facts further strengthen the plausibility of our proposal.

This model has many weaknesses. First of all, it has not yet implemented multiscale samplings and interactions in visual space. Consequently the model can not for instance enhance fine detailed contours or to detect and segment regions of very small sizes. Also, the model neural circuit, in particular, the structure of the horizontal connections, must be different from the reality at least quantitatively. Therefore, the model behaves differently, at least quantitatively, from human performance. For instance, the model sometimes find it easier or more difficult to segment some regions than humans do. However, I believe that these and many others are mainly the weaknesses of this particular, still primitive, model implementation of V1. The principle of contour enhancement by integrating mutual facilitation and region segmentation by detecting the breakdown of translation invariance in inputs should still hold. More informative experimental data on the V1 structure should help to build a better V1 model to implement the above principles for visual grouping.

References

1. Douglas R.J. and Martin K. A. "Neocortex" in *Synaptic Organizat ion of the Brain* 3rd Edition, Ed. G. M. Shepherd, Oxford University Press 1990.
2. Gallant J. L., van Essen D. C. and Nothdurft H. C. "Two-dimensional and three-dimensional texture processing in visual cortex of the macaque monkey" In: Papathomas T., Gorea A. (Eds.) *Linking psychophysics, neurophysiology and computational vision.* MIT press, Cambridge MA 1994.
3. Gilbert C.D. "Horizontal integration and cortical dynamics" *Neuron.* 9(1): 1-13. 1992

4. Haralick R. M. Shapiro L. G. *Computer and robot vision* Vol 1. Addison-Weslley Publishing, 1992
5. Kapadia M.K., Ito M., Gilbert C.D. and Westheimer G. "Improvement in visual sensitivity by changes in local context: parallel studies in human ob servers and in V1 of alert monkeys." *Neuron.* Oct; 15(4): 843-56. 1995
6. Kasturi, R., and Jain R. C., Eds. *Computer vision, Principles* IEEE Computer Society Press, 1991.
7. Li, Zhaoping "A neural model of contour integration in the primary visual cortex" Technical report HKUST-CS97-05, Hong Kong University of Science and Technology. Submitted for publication.
8. Shashua A. and Ullman S. "Structural Saliency" *Proceedings of the International Conference on Computer Vision.* Tempa, Florida, 482-488. 1988
9. White E. L. *Cortical circuits* 46-82, Birkhauser, Boston, 1989

This article was processed using the LaTeX macro package with LLNCS style