MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo No. 1613 August 1997
C.B.C.L. Memo No. 153

Visual segmentation without classification in a model of 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 significantly influence the neurons' activities in primary visual cortex [1, 2, 3, 4, 5]. We propose that such contextual influences are used to segment regions by detecting the breakdown of homogeneity or translation invariance in the input, thus computing *global* region boundaries using *local* interactions. This is implemented in a biologically based model of V1, and demonstrated in examples of texture segmentation and figure-ground segregation. By contrast with traditional approaches, segmentation occurs without classification or comparison of features within or between regions and is performed by exactly the same neural circuit responsible for the dual problem of the grouping and enhancement of contours.

This report describes research done at the Center for Biological and Computational Learning at Massachusetts Institute of Technology and the Computer Science Department of Hong Kong University of Science and Technology. The authors can be reached at M.I.T., Center for Biological and Computational Learning, 45 Carleton St., Cambridge MA 02142, USA. E-mail: zhaoping@ai.mit.edu

Recent experiments have pointed to the complexity of processing that occurs in V1[6, 7, 8, 9, 3]. Not only can this processing determine the gains and the classical tuning functions of cells,[6, 9, 10] but it also arranges for contextual influences on their activities from stimuli beyond their classical receptive fields (RFs)[1, 2, 3, 11, 4, 12, 13, 5]. The responses of cells depend on whether stimuli within and beyond the RFs share the same orientations[2, 4, 11, 5], and whether the stimuli within the RFs are part of different regions, such as figure or ground [12, 13]. Horizontal intra-cortical connections are suggested to mediate the contextual influences[7, 3]. While there have been substantial experimental interest and some modeling interest (e.g., [14]) in these contextual influences, computational understanding of their roles in visual processing is lagging far behind [1, 3].

We propose that the contextual influences in the primary visual cortex can serve the goal of visual grouping, i.e., inferring *global* visual objects such as contours and regions from the *local* features captured by the RFs. Local features can group into regions, as in texture segmentation; or into contours which may represent boundaries of underlying objects. We show how one form of global grouping, namely region segmentation, can emerge from a simple but biologically-based model of V1 which only involves finite-range cortical interactions.

It has always been assumed, implicitly or explicitly, that to segment one region from another, feature extraction and/or classification within a region and feature comparison between regions are required [15, 16, 17]. On the other hand, feature extraction or classification often require segmentation, thus creating a dilemma. In these traditional approaches, not only is feature classification problematic near the boundaries between regions, but also segmentation using feature comparison is tricky in cases such as figure (3D), where the two regions have the same texture feature value but are segmentable in natural vision. Therefore, feature extraction or classification is not always necessary nor sufficient for segmentation. In fact, even with distinguishable classification flags for all image areas in any two regions, segmentation is not completed until another processing step locates the boundary, perhaps by searching for where the classification flags change. Therefore, we propose that segmentation in its pre-attentive stage is segmentation without classification, i.e., segmentation without explicitly knowing the contents of the regions. This simplifies the segmentation process conceptually, making it feasible by low level processing in V1. This paper focuses on this pre-attentive segmentation. Additional processing is likely needed to improve the outcome based on pre-attentive segmentation, e.g., by filling in the contents of the regions.

The model focuses on simple texture segmentation, i.e., region grouping without color, motion, luminance, or stereo cues. A single texture region is defined by the homogeneity or translation invariance of the statistics of the input features that define it, no matter what features are involved or, for instance, whether or not they are textons[18]. If cortical interactions are translation invariant and do not induce spontaneous pattern formation (such as zebra stripes [19]) through the spontaneous breakdown of translation symmetry, then the cortical response to a homogenous region will itself be homogeneous. However, homogeneity is disrupted at the boundary of a region. Consequently, a neuron near the boundary and another far from the boundary experience different contextual influences, and thus exhibit different response levels. The location of the boundary can therefore be pinpointed by assessing where the contextual influences or neural response levels change. In the model, this breakdown in homogeneity gives relatively higher neural activities near the boundaries than away from them. This makes the boundaries relatively more salient, allowing them to pop out perceptually. Physiological experiments in V1 indeed show that activity levels are higher near texture boundaries[20].

Figure (1) shows the elements of the model and their interactions. Based on experimental observations[8, 9], a cortical column is modelled by recurrently connected excitatory cells and inhibitory interneurons tuned to bars or edges. Quantities $x_{i\theta}$ and $y_{i\theta}$ are the membrane potentials of the excitatory and inhibitory cells having the RF center (or hypercolumn) location i and perferred orientation θ. The excitatory cell receives external visual input $I_{i\theta}$ to the cortical cell, which is the retinal image filtered through the RF. These edge or bar inputs to the model are merely image primitives, which are in principle like the image pixel primitives and are reversibly convertible from them. They are not to denote the texture feature values, e.g., the '+' or 'x' patterns and their spatial arrangements in the example of figure (3)C. Again, this model does not extract texture features in order to segment. The output from V1 is provided by the excitatory cells. Based on observations by Gilbert, Lund and their colleagues[7, 3], horizontal connections $J_{i\theta,j\theta'}$ and $W_{i\theta,j\theta'}$ link cells with different RF centers and similar orientation preferences to mediate contextual influences. The membrane potentials follow the equations:

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

where $\alpha_x x_{i\theta}$ and $\alpha_y y_{i\theta}$ model the decay to resting potentials, $g_x(x)$ and $g_y(y)$ are sigmoid-like functions modeling cells' firing rates $g_x(x)$ and $g_y(y)$ given membrane potentials x and y, respectively, $\psi(\Delta\theta)$ the inhibition spread within a hypercolumn, $J_o g_x(x_{i\theta})$ the self excitation, I_c and I_o are background inputs or inputs modeling the general and local activity normalization[21], and $J_{i\theta,j\theta'} g_x(x_{j\theta'})$ and $W_{i\theta,j\theta'} g_x(x_{j\theta'})$ model the contextual influences (see [22, 23]) for more details).

The activity levels of the neurons $g_x(x_{i\theta})$ are initially set by just the visual input $I_{i\theta}$. This input persists after its onset. The activities are then modified effectively within one membrane time constant by the cortical interaction that mediate the contextual influences. Mean field techniques and dynamic stability analysis are used to design the horizontal connections J and W to ensure that: (1) the system does not generate patterns spontaneously, i.e., the model gives spatially homogenous output for homogenous input images, (2) the region boundaries are relatively highlighted by modeling the physiologically observed iso-orientation suppression via the contextual influences (thereby making areas inside a region less salient), and (3) the same neural circuit performs contour enhancement (see [24] for more details).

The model was applied to a variety of textured inputs. Figure (2)A shows a sample input consisting of two regions, in which all the visible inputs $I_{i\theta}$ have the same strength. Figure (2)B,C shows the output of the model, indicating that the activities of the neurons at the boundary are significantly higher than others. Figure (2)D confirms that the boundary can be identified by thresholding the final activities.

Figure (3) shows other examples of input patterns and the thresholded outputs of the model. Note particularly in figures (3)A;B;C that the model copes well with textures defined by complex or stochastic patterns; from figure (3)D that it segments regions by detecting the breakdown of homogeneity even though the two regions have the same texture feature, a feat difficult in traditional approaches; in figure (3)E that both humans and the model have difficulty segmenting regions when the translation invariance is only broken very weakly; and in figure (3)H that when a region is very small, all parts of it belong to the boundary and it pops out from the background. Figure (3)F;G show other examples where regions differ by the orientations of the texture elements. Finally, figure (3)I confirms that exactly the same model, with the same elements and parameters, can also highlight contours against a noisy background. This can be seen as another example of a breakdown of translation invariance. Additional simulations confirm that the model also performs reasonably well on many other examples.

Our model to detect region boundaries is beyond and different from the early visual processing using center-surround filters or the like[25]. There, the filters are tuned to detect contrast in luminance, they can detect the edge primitives in a textured region, and their outputs can be used as inputs to our model. However, these filters can not detect feature changes from one region to another, e.g., figure (2)A, that are not apparent in average luminance changes. If one were to design a one stage filter to detect the feature changes between regions, the filter would be feature specific and many different kinds would be required to cover many possible region differences. The mechanism using cortical interactions in our model highlights conspicuous image locations or general feature changes from one region to another without specific tuning to any region features. While the early stage filters code image primitives[25], the mechanism in our model is aimed towards coding object surface primitives.

It has recently been argued that texture analysis is performed at a low level of visual processing[15], and indeed filter based models[16] and their non-linear extensions (e.g., [17]) capture well much of the phenomenology of psychophysical performance. However, all the previous models are based on the traditional approach of segmentation by feature classification/comparison, and thus share the problems associated with that approach. By performing segmentation without classification, our model differs from these in principle. Consequently, while our model employs only those low level visual operations that are consistent with experimental observations[7, 8, 9, 3], the model by Malik and Perona[17], for instance, uses complicated forms of cortical interactions such as winner-take-all operations and spatial derivatives for which there exists little experimental evidence. In addition, our model is the first to perform region segmentation and contour enhancement using exactly the same neural circuit. This is desirable since regions and their boundary contours are complementary to each other. Furthermore, in our framework, small regions naturally pop out, as in figure (3H), filling-in in a non-homogeneous region would be the perceptual consequence of the model's failing to highlight the non-homogeneity, and feature statistics in a region[26] are automatically accounted for for region segmentation.

The components of the model and its behavior are consistent with experimental evidence[7, 8, 9, 3, 20]. However, the model is obviously only an approximation to the true complexities of V1. For instance, all its elements are tuned to one scale, and exhibit none of the flexible adaptation that is pervasive in the real system. Therefore, the model sometimes finds it easier or more difficult to segment some regions than natural vision, for instance, not coping well with gradual changes in images caused by the tilt of a textured surface. Any given neural interaction will be more sensitive to some region differences than others. Hence, a more detailed model of the neural elements and the connection pattern would be required to capture exactly the psychophysical data on segmentation in natural pre-attentive vision. However, independent of such details, 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.

References

[1] J. Allman, F. Miezin, and E. McGuinness *Ann. Rev. Neurosci.* 8:407-30, (1985).

[2] J. J. Knierim and D. C. van Essen *J. Neurophysiol.* 67, 961-980. (1992)

[3] C. D. Gilbert *Neuron.* 9(1): 1-13. (1992)

[4] A. M. Sillito, K. L., Grieve, H. E. Jones, J. Cudeiro, and J. Davis *Nature* 378(6556):492-6. (1995)

[5] J. B. Levitt and J. S. Lund *Nature* 387 (6628): 73-6. (1997)

[6] A. M. Sillito, J. A. Kemp, J. A. Milson, and N. Beerardi. *Brain Research.* 194, 517-520.(1980)

[7] K.S. Rockland and J. S. Lund *J. Comp. Neurol.* 216, 303-318, (1983)

[8] E. L. White *Cortical circuits* (Birkhauser, Boston, 1989)

[9] R. J. Douglas and K. A. Martin in *Synaptic Organization of the Brain* G. M. Shepherd Ed. (Oxford University Press 1990) 3rd Edition,

[10] D. Ferster, S. Chung, and H. Wheat *Nature*, Vol. 380 p 249-252. (1996)

[11] M. K. Kapadia, M. Ito, C. D. Gilbert, and G. Westheimer *Neuron.* Oct; 15(4): 843-56. (1995)

[12] V. A. Lamme *Journal of Neuroscience* 15(2):1605-15. (1995)

[13] K. Zipser, V. A. Lamme, and P. H. Schiller *J. Neurosci.* 15, 16(22):7376-89. (1996)

[14] D. C. Somers, E. V. Todorov, A. G. Siapas, and M. Sur *A.I. Memo. NO. 1556,* (MIT. 1995)

[15] J. R. Bergen In *Vision and visual dysfunction* D. Regan Ed. Vo. 10B (Macmillan New York, 1991) p. 114-134.

[16] J.R. Bergen and E. H. Adelson. *Nature* 333:363-364, May (1988).

[17] J. Malik and P. Perona. *J. Opt. Soc. Am. A* 7(5):923-932, (1990)

[18] B. Julesz. *Nature* 290:91-97, (1981).

[19] H. Meinhardt, *Models of biological pattern formation* (Academic Press, London ; New York 1982.)

[20] J. L. Gallant, D. C. van Essen, and H. C. Nothdurft In *Linking psychophysics, neurophysiology and computational vision.* T. Papathomas, A. Gorea Eds. (MIT press, Cambridge MA 1994.)

[21] D. J. Heeger *Visual Neurosci.* 9, 181-197. (1992)

[22] Z. Li *Neural Computation* in press.

[23] Z. Li, in *Advances in neural information processing systems 9*, M. C. Mozer, M. Jordan, and T. Petsche Eds (MIT press, Cambridge, 1997)

[24] Z. Li in *Theoretical aspects of neural computation* (Springer-Verlap, Hong Kong, 1997) in press.

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

[26] B. Julesz. *IRE Transactions on Information theory IT-8* p. 84-92, (1962).

[27] C. M. Gray and W. Singer *Proc. Natl. Acad. Sci. USA* 86: 1698-1702. (1989)

[28] R. Eckhorn et al. *Biol. Cybern.* 60:121-130, (1988).

Acknowledgement: I thank Peter Dayan for many helpful discussions, he and John Hertz for their careful reading and helpful comments on the paper. This work is supported by the Hong Kong Research Grant Council and the Center for Biological and Computational Learning at MIT.

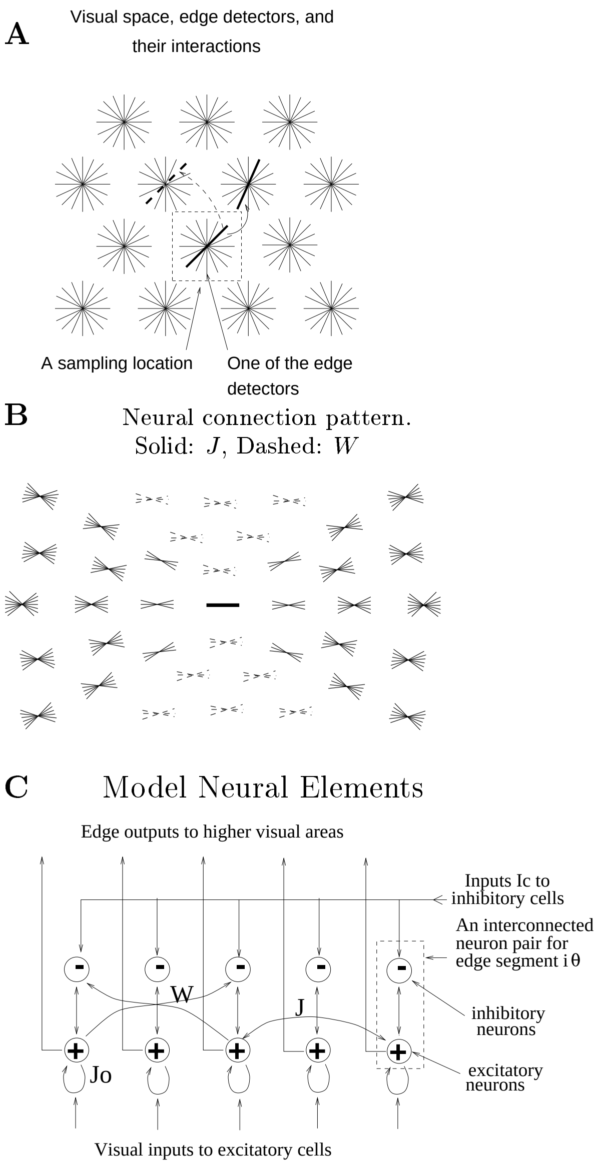

Figure 1: **A:** Visual inputs are sampled in a discrete grid by edge/bar detectors, referred to as edge or edge segments, modeling RFs in V1. Each grid point has K neuron pairs (see **C**), one per edge segment. All cells at a grid point share the same RF center, but are tuned to different orientations spanning 180^o, thus modeling a hypercolumn. An edge 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) edge to neighboring edges within a finite distance. J's contacts are shown by thin solid edges. W's are shown by thin dashed edges. All edges have the same connection pattern, suitably translated and rotated from this one. **C:** An input edge 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. Activity levels $g_x(x_{i\theta})$ often oscillate over time [27, 28], which is an intrinsic property of a population of recurrently connected excitatory and inhibitory cells. Temporal averages over multiple time constants after input onset are taken as the model output. The region dependence of the phases of the oscillations in this model could be exploited for segmentation[22], although it is beyond this paper. The visual space has toroidal (wrap-around) boundary conditions.

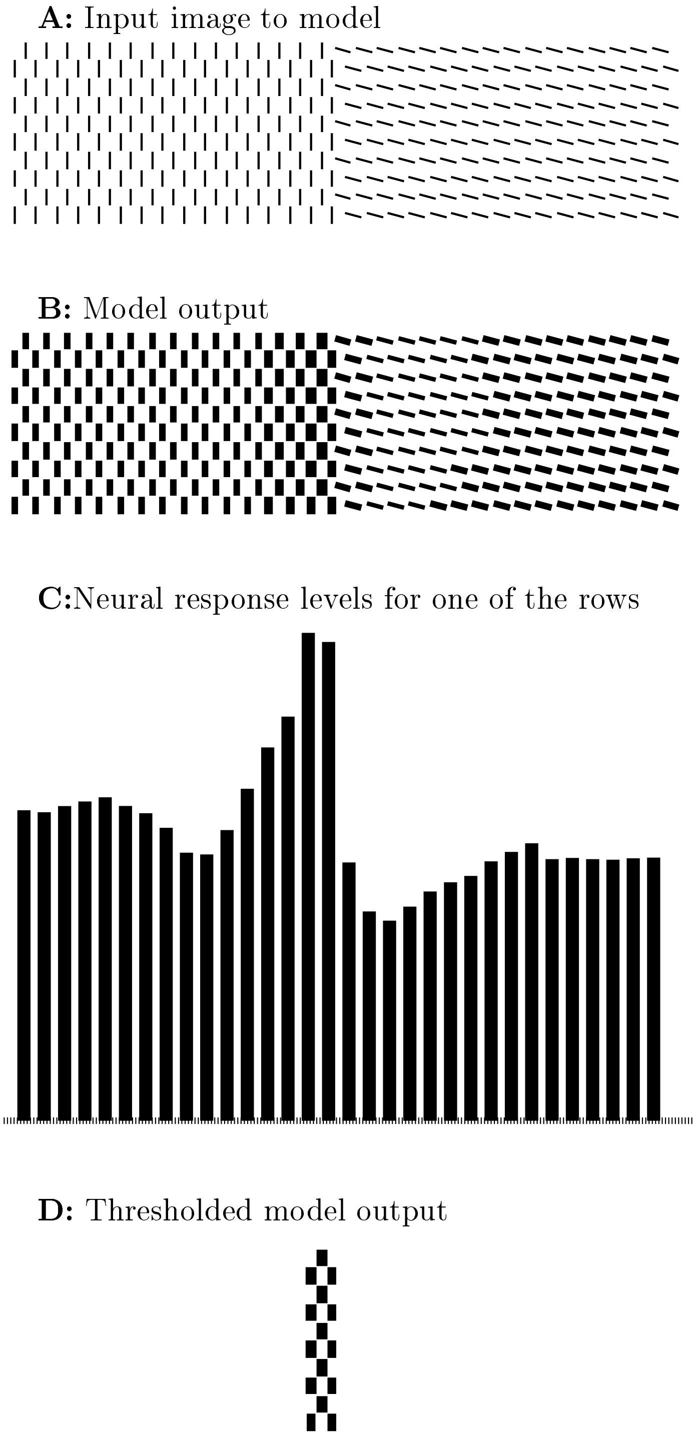

Figure 2: **A**: Input $I_{i\theta}$ of two regions; each visible edge has the same input strength. **B**: Model output for **A**, showing non-uniform output strengths (temporal averages of $g_x(x_{i\theta})$) for the edges. The input and output edge strengths are proportional to the edge thicknesses shown. **C**: Output strengths (saliencies) vs. lateral locations of the edges for a row like the bottom row in **B**, with the bar lengths proportional to the corresponding edge output strengthes. **D:** The thresholded output from **B** for illustration. Each plotted region shown here is actually a small part of, and extends continuously to, a larger image. The same format is used in other figures in this paper.

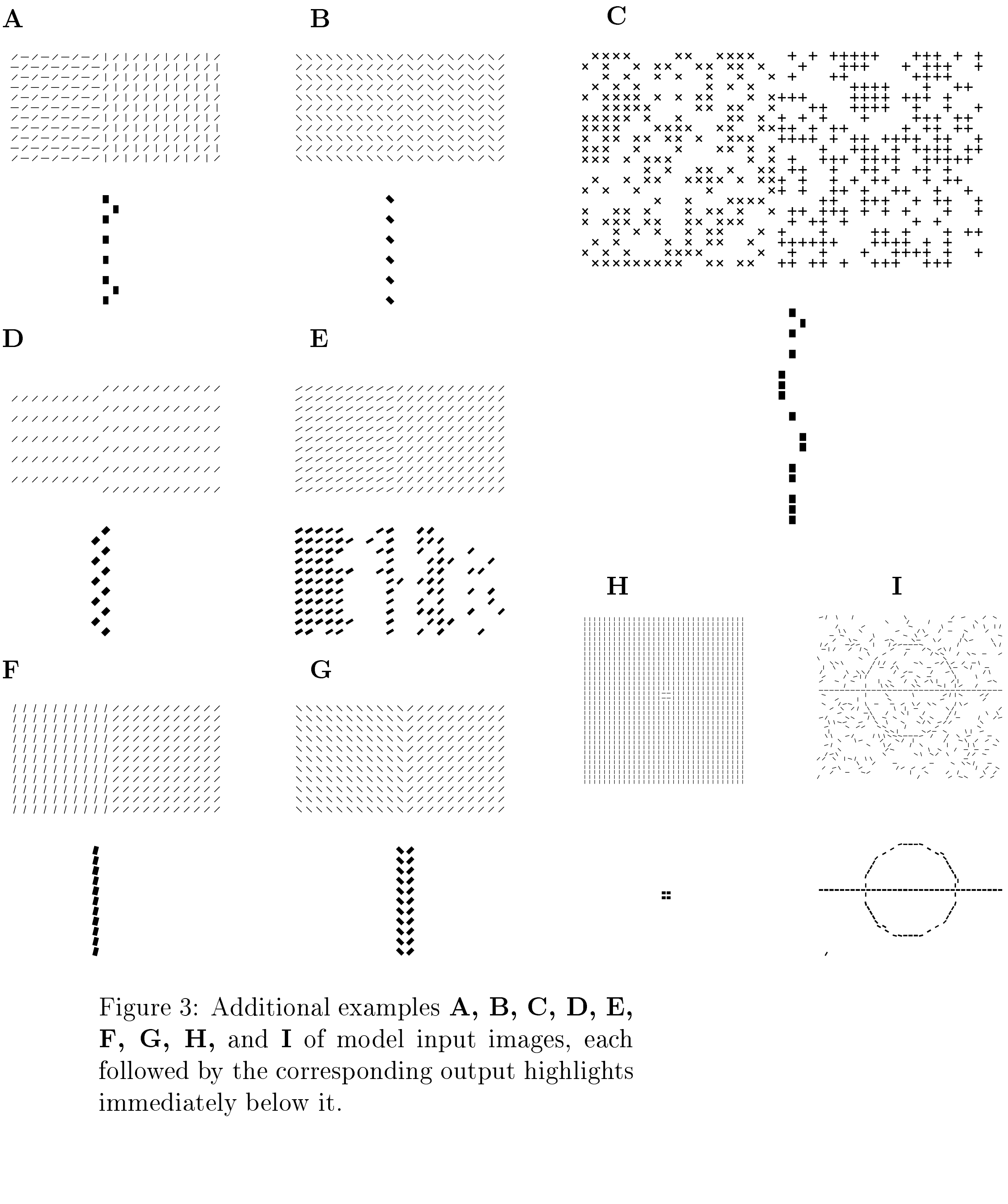

Figure 3: Additional examples **A, B, C, D, E, F, G, H,** and **I** of model input images, each followed by the corresponding output highlights immediately below it.