A Hierarchical Network Model for Depth Segregation

Stella X. Yu*[1,3] Tai Sing Lee[2,3] Takeo Kanade[1,2]
Robotics Institute[1]
Department of Computer Science[2]
Carnegie Mellon University
Center for the Neural Basis of Cognition [3]
5000 Forbes Ave, Pittsburgh, PA 15213-3890
{*xingyu,tai,tk*}@*cs.cmu.edu*

Abstract

In this paper, we propose a Markov random field network model to segregate overlapping objects into depth layers using sparse local cues. A topology-dependent multiscale hierarchy is used to introduce rapid long range interaction. The operations within each level are identical across the hierarchy. The decision rules are encoded in clique potentials and their parameters are estimated using an optimization technique based on 2-object training samples. We find that this model generalizes successfully to 5-object test images, and provides a framework to reason about local and global interaction, as well as some of the recent neurophysiological findings in figure-ground segregation.

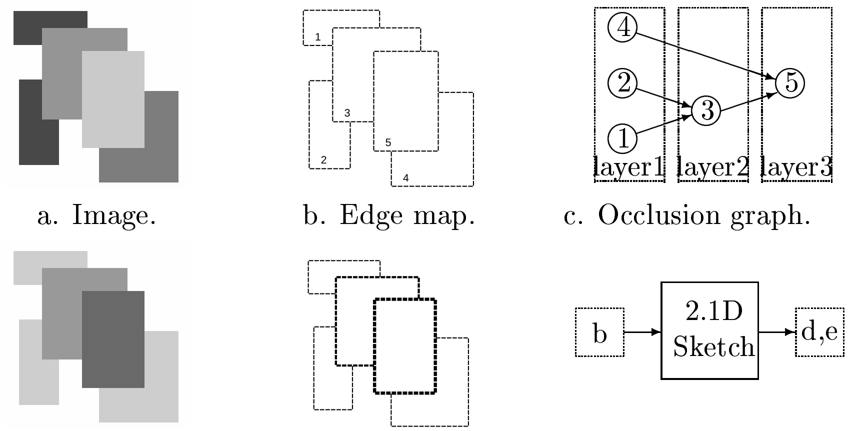
1 Introduction

In this work, we focus on the issue of global depth segregation based on sparse occlusion cues. The importance of local occlusion cues in determining global depth perception can be appreciated in our remarkable ability in inferring relative depth among objects in cartoon drawings (Fig. 1a). Occlusion cues such as T-junctions are relatively sparse in the image. Yet, they provide an important constraint for figure-ground segregation, which is an emergent global perceptual phenomenon. On the other hand, neurons in early visual cortex are locally connected to each other through horizontal axonal collaterals. These neurons can be considered as local decision units. In this work, we study how Markov random fields can be generalized to embed multiple explicit decision rules in the interaction between local units. This framework allows us to study the connection between local processes and global perceptual representation, the necessary representations and rules of decision units. Such an investigation should help us to better appreciate the computational link between some of the recent neurophysiological findings on figure-ground segregation [17, 9, 10] to perceptual behaviors.

[*] Corresponding author.

2 Methods

2.1 Problem formulation



a. Image. b. Edge map. c. Occlusion graph.

d. Pixel depth label. e. Edge depth label. f. Goal of our model.

Figure 1: Segregate depth into layers. Rectangular objects are numbered in b. Darker object surfaces/edges are in front of lighter object surfaces/edges in d/e. Given an edge map as input, our model produces two complementary depth maps as output.

For simplicity, we take an edge map (Fig. 1b) with complete and closed contours of rectangular shapes as input to our system. These shapes can overlap and occlude one another. The occluded part of an object is not visible. The system is to produce two complementary maps as output (Fig.1f): a pixel depth map (Fig. 1d) where a higher depth value is assigned to pixel depth units of a more proximal surface and a lower value to pixel units of a more distant surface; and an edge depth map in which the edge depth units at the border of a more proximal surface assume a higher value. The edge depth units assume the same depth value as the pixel depth units of the surface to which they belong to (Fig. 1e). These two representations are sufficient to specify the depth ordering sequence of objects in the scene.

In general, it is not possible to recover the exact depth ordering or overlap sequence in the scene since the solution is not unique. For example, there can be multiple choices when objects do not occlude each other directly (object 1 and 2 in Figure 1b) and when we cannot tell which object is occluding which (object 3 and 4 in Figure 1b). If we represent visible pairwise object occlusion relationships in a directed graph (Figure 1c), these two cases correspond to the existence of unconnected siblings of the same parent. Instead of recovering the overlap sequence, we can sort object depths into layers, ordered by occlusion. This problem is called the 2.1D sketch in [13]. If there is a directed cycle in the graph, then the depth cannot be segregated into layers. We define the depth assignment solution to be the set of smallest depth labels that satisfy all the visible occlusion relationships. For example, object 4 in 1c is on layer 1 rather than 2.

Long range influence can be mediated by local computation using Markov random fields

[4, 11]. Markov random field approach has been widely used in texture modeling [2], as well as in image segmentation [4, 6]. In this work, we suggest a broader view that clique potentials can be more general so that they can encode arbitrary local decision rules.

The depth segregation process modeled in MRF seeks to make correct depth labeling that corresponds to the most probable configuration or equivalently, to find the configuration of the minimum energy state. Prior knowledge in terms of a set of local rules is encoded in clique potential $V_c(h|g)$.

$$
\begin{aligned}
V_c(h|g) = {} & \textstyle\sum_{a=(i\circ j)\in c} \beta_1 \cdot \gamma(h_i = h_j) \cdot \chi(g_a = 0) && \text{rule 1}\\
+ {} & \textstyle\sum_{a=(i\circ j)\in c} \beta_2 \cdot \gamma(h_i \neq h_j) \cdot \chi(g_a = 1) && \text{rule 2}\\
+ {} & \textstyle\sum_{a=(i\circ j)\in c} \beta_3 \cdot \gamma\Big(h_a = \max(h_i, h_j)\Big) \cdot \chi(g_a = 1) && \text{rule 3}\\
+ {} & \textstyle\sum_{(a=i\circ j, b=k\circ l)\in c^l} \beta_4 \cdot \gamma(h_a = h_b) \cdot \chi(g_a = g_b = 1) && \text{rule 4}\\
+ {} & \textstyle\sum_{(a=i\circ j, b=k\circ l)\in c^l} \beta_5 \cdot \gamma\Big(\zeta(h_i - h_j) = \zeta(h_k - h_l)\Big) && \\
& \quad \cdot\chi(h_i \neq h_j, h_k \neq h_l) \cdot \chi(g_a = g_b = 1) && \text{rule 5}\\
+ {} & \textstyle\sum_{(a=i\circ k, b=j\circ k)\in c^c} \beta_6 \cdot \gamma(h_a = h_b) \cdot \chi(g_a = g_b = 1) && \text{rule 6}\\
+ {} & \textstyle\sum_{(a=i\circ k, b=j\circ k)\in c^c} \beta_7 \cdot \gamma\Big(\zeta(h_i - h_k) = \zeta(h_j - h_k)\Big) && \\
& \cdot\chi(h_a = h_b) \cdot \chi(g_a = g_b = 1) && \text{rule 7}\\
+ {} & \textstyle\sum_{(a=i\circ j, b=k\circ l, u=j\circ l, v=i\circ k)\in c^t} \beta_8 \cdot \Big(\gamma(h_a > h_u) + \gamma(h_b > h_u)\Big) && \\
& \cdot\chi\Big(\zeta(h_i - h_j) = 1 \cup \zeta(h_k - h_l) = 1\Big) \cdot \chi(g_a = g_b = g_u = 1 \cap g_v = 0) && \text{rule 8}\\
+ {} & \textstyle\sum_{(a=i\circ j, b=k\circ l, u=j\circ l, v=i\circ k)\in c^t} \beta_9 \cdot \Big(\gamma(h_i > h_j) + \gamma(h_k > h_l)\Big) && \\
& \cdot\chi\Big(\zeta(h_i - h_j) = 1 \cup \zeta(h_k - h_l) = 1\Big) \cdot \chi(g_a = g_b = g_u = 1 \cap g_v = 0) && \text{rule 9}\\
+ {} & \textstyle\sum_{(a=i\circ j, b=k\circ l, u=j\circ l, v=i\circ k)\in c^t} \beta_{10} \cdot \gamma(h_i = h_l) && \\
& \cdot\chi(g_a = g_u = 0 \cup g_b = g_v = 0) && \text{rule 10}
\end{aligned}
$$

where χ and γ denote two indicator functions, which map from {True, False} to $\{1, 0\}$ and $\{-1, 1\}$ respectively. $\gamma(\cdot) = 1 - 2\chi(\cdot)$. Let ζ denote the *sign* function, which takes on $-1, 0, 1$ for negative, zero and positive numbers respectively. The line site a between pixel i and j is denoted by $a = i \circ j$ and conversely, the set of pixels associated with the line is denoted by $a^\circ = (i, j)$, with i and j ordered from left to right or from top to bottom. In particular, $(i, j) \circ (i, j+1)$ and $(i, j) \circ (i+1, j)$ are abbreviated as $(i, j\circ)$ and $(i\circ, j)$ respectively.

c^l, c^c, c^t are the sets of cliques for aligned lines, corners and crosses respectively:

$$
\begin{aligned}
c^l &= \{(a, b) : a = (i\circ, j), b = (i\circ, j+1); a = (i, j\circ), b = (i+1, j\circ), a, b \in c\},\\
c^c &= \{(a, b) : a = (i\circ, j), b = (k, l\circ), |i - k| \leq 1, |j - l| \leq 1, a, b \in c\},\\
c^t &= \{(a, b, u, v) : (a, b) \in c^l, (u, v) \in c^l, \{a, b\} \cap \{u, v\} = \emptyset, a^\circ \cup b^\circ = u^\circ \cup v^\circ\}.
\end{aligned}
$$

The two indicator functions in the above, χ and γ, enable us to embed the conjunction of *if* conditionals into the clique potentials. Let us decode rule 1 as an example. Consider the line site a between pixel i and j. If the clause $(g_a = 0)$ is not true, i.e. there is an edge between the two pixels, then this first term is zero, no action will be taken; otherwise, if the clause $(h_i = h_j)$ is also true, i.e. the pixel depth values at the two sites are equal, then the term produces a reward of $-\beta_1$, lowering the energy. However, if it is not true, i.e. the depth values at the two pixel sites are different, then $V_c(h|g)$ gets β_1 on this term as a punishment, increasing the energy. Here we require all βs to be positive. Let's denote these two conditionals as A and B respectively, and summarize the 10 rules in Table 1.

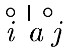
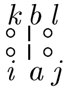
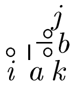
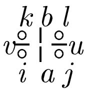
Clique c	Condition A	Pattern B	Score C	Rule #	Meaning
	$g_a = 0$	$h_i = h_j$	β_1	1	Depth continues in surface.
	$g_a = 1$	$h_i \neq h_j$	β_2	2	Depth breaks at edges.
	$g_a = 1$	$h_i \neq h_j$	β_3	3	Edges belong to surface in front.
	$g_a = g_b = 1$	$h_a = h_b$	β_4	4	Depth continues along contour.
	$g_a = g_b = 1$ $h_i \neq h_j$ $h_k \neq h_l$	$\zeta(h_i - h_j)$ $=$ $\zeta(h_k - h_l)$	β_5	5	Depth polarity continues along contour.
	$g_a = g_b = 1$	$h_a = h_b$	β_6	6	Depth continues around corners.
	$g_a = g_b = 1$ $h_a = h_b$	$\zeta(h_i - h_k)$ $=$ $\zeta(h_j - h_k)$	β_7	7	Depth polarity continues around corners.
	$g_a = g_b = 1$ $g_u = 1, g_v = 0$ $\zeta(h_i - h_j) = 1$ or $\zeta(h_k - h_l) = 1$	$h_a > h_u$	β_8	8	Depth breaks on edges at T-junctions.
		$h_b > h_u$	β_8		
		$h_i > h_j$	β_9	9	Depth breaks in surface at T-junctions.
		$h_k > h_l$	β_9		
	$g_a = g_u = 0$ or $g_b = g_v = 0$	$h_i = h_l$	β_{10}	10	Depth continues in surface.

Table 1: Encoding rules in clique potentials. Each of these β terms encodes a logic rule, which in general reads like this: if current clique configuration does not satisfy condition A, it gets a score of 0; otherwise, if condition A is satisfied, pattern B is expected; if B is also satisfied, then it gets a negative score $-C$; otherwise it gets a positive score C. a, b, u and v are labels for line sites while i, j, k, l are labels for pixel sites in the cliques.

From neural modeling perspective, the units in the network are not neurons with linearly weighted inputs and sigmoidal activation functions, but are capable of performing complicated logical computations individually. Recent findings and models in cellular neurophysiology [8, 12, 1] suggest neurons are capable of computations more sophisticated than previously assumed. The relative importance of the weights βs in the depth segregation can be estimated using a variety of methods. Here, we developed a method called *learning from rehearsal* to derive the constraints on the parameters of the network, and used linear programming to solve for the parameters.

The MRF model thus described suffers from being myopic [7] in local computation and sluggish at propagating constraints between widely separated processing elements [15]. This problem can be overcome by embedding the MRF in a hierarchy using multigrid techniques. We will give a more complete description of the parameter estimation scheme and the multiscale computation scheme in the full paper.

3 Results

Learning on a small set of training images containing *two objects* singles out a unique set of values for β, where $\beta = [18, 9, 97, 23.3, 3.2, 86.7, 3.35, 16.5, 42.5, 137, 20.8]$. With this set of parameters, the model produces reasonable results for a set of test images that the system has never experienced before.

Figure 2 shows how the system responds to a test image with *five overlapping rectangles* in the scene. The system generalizes very well in its response to this new input configuration. A sequence of 8 snap shots are taken at different time points during the evolution of the system. Snap shot 1 shows the system detecting T-junctions and starting propagating its initial result one level up the hierarchy. Snap shot 2 shows the information has propagated to the third level, and propagation of depth information within surface is now evident at the second level. Snap shots 3 and 4 show the information has propagated to the fourth and fifth levels respectively. Snap shot 5 shows the information starts to propagate down the hierarchy, introducing rapid filling-in of surface depth and depth segregation in snap shot 6. Snap shots 7 and 8 show the completion of surface/contour depth interpolation and segregation. All these are completed very rapidly in two iterations up and down the hierarchy.

We think this HMRF model for depth segregation might provide a plausible computational framework for reasoning about and understanding the basic computational constraints and neural mechanisms underlying local and global integration and figure-ground segregation in the brain. This work provides us with several insights to some psychological and neurophysiological phenomena and a few interesting experimental predictions.

First, brightness has been observed to propagate in from the border in the psychophysical experiment by Paradiso and Nakayama [14]. Such phenomenon has been postulated to be mediated by horizontal connections at the level of V1, for example in Grossberg and Mingolla's model [5]. Here, we show that a hierarchical framework can speed up the depth assignment process considerably, and yet also show a diffusion-like process of depth assignment propagating in from the border. In fact, traversing up and down the hierarchy twice is sufficient to complete the computation. This suggests the psychological phenomenon observed and the depth segregation computation could also be mediated by the feedback from V2 and V4, which are known to have receptive fields two and four times larger than those of V1 respectively.

Second, while Paradiso and Nakayama's experiment suggests diffusion in the brightness domain, the similarity in dynamics between the brightness diffusion and our depth assignment process suggest depth segregation and assignment might be the underlying process that carries the brightness diffusion along. By the same reasoning, one would expect other surface cues such as color, texture, and stereo disparity should also be accompanying, if not following, the depth assignment process. It will indeed be interesting to examine experimentally whether the propagation of surface cues follow the depth assignment process or occur simultaneously. The fact that Dobbins et al. [3] found that a significant number of V1, V2 and V4 cells sensitive to distance even in monocular viewing conditions suggest depth assignment might intertwined with many early visual processes.

Finally, the hierarchy presented is not simply a multiscale network in that when the information travels up, the topological relationships between different objects are taken into consideration in such a way that the same relaxation procedure can be applied at each level. For example, edges of overlapping shapes are kept, whereas the edges of two nearby shapes appearing side by side would disappear at a coarser resolution. This operation can be achieved by taking the algebraic sum of depth polarities. In order to accomplish this in the network, depth polarity of edges needs to be computed and represented explicitly. This might provide a computational rationale for the existence of the depth-polarity sensitive cells von der Heydt and his colleagues found in V1, V2 and V4 [17, 16]. In addition, the model makes explicit prediction on how adjacent edges of different depth polarities would be processed in higher visual cortical areas.

References

[1] Larry F. Abbott. Integrating with action potentials. *Neuron*, 26:3–4, 2000.

[2] Haluk Derin and Howard Elliott. Modeling and segmentation of noisy and textured images using Gibbs random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):39–55, 1987.

[3] Allan C. Dobbins, Richard M. Jeo, Jozsef Fiser, and John M. Allman. Distance modulation of neural activity in the visual cortex. *Science*, 281:552–5, 1998.

[4] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–41, 1984.

[5] Stephen Grossberg and Ennio Mingolla. Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading. *Psychological Review*, 92:173–211, 1985.

[6] Tsai Hong Hong, K. A. Narayanan, Shmuel Peleg, Azriel Rosenfeld, and Teresa Silberberg. Image smoothing and segmentation by multiresolution pixel linking: further experiments and extensions. *IEEE Transactions on Systems, Man, and Cybernetics*, 12:611–22, 1982.

[7] Tsai Hong Hong and Azriel Rosenfeld. Compact region extraction using weighted pixel linking in a pyramid. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(2):222–9, 1984.

[8] Christof Koch. Computation and the single neuron. *Nature*, 385:207–210, 1997.

[9] Victor Lamme. The neurophysiology of figure-ground segregation in primary visual cortex. *Jounral of neuroscience*, 10:649–69, 1995.

[10] Tai Sing Lee, David Mumford, Rick Romero, and Victor Lamme. The role of primary visual cortex in higher level vision. *Vision Research*, 38:2429–54, 1998.

[11] Stan Z. Li. *Markov random field modeling in computer vision.* Springer-Verlag, 1995.

[12] Henry Markram, Joachim Lubke, Michael Frotscher, and Bert Sakmann. Regulartion of synaptic efficacy by coincidence of postsynpatic APS and EPSPs. *Science*, 275:213–215, 1997.

[13] Mark Nitzberg. *Depth from Overlap*. PhD thesis, The Division of Applied Sciences, Harvard University, 1991.

[14] Michael A. Paradiso and Ken Nakayama. Brightness perception and filling-in. *Vision Research*, 31(7/8):1221–36, 1991.

[15] Demetri Terzopoulos. Image analysis using multigrid relaxation methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(2):129–39, 1986.

[16] R von der Heydt, H Zhou, and HS Friedman. Representation of stereoscopic edges in monkey visual cortex. *Vision Research*, 40(15):1955–67, 2000.

[17] H Zhou, HS Friedman, and R von der Heydt. Coding of border ownership in monkey visual cortex. *Journal of Neuroscience*, 20(17):6594–611, 2000.

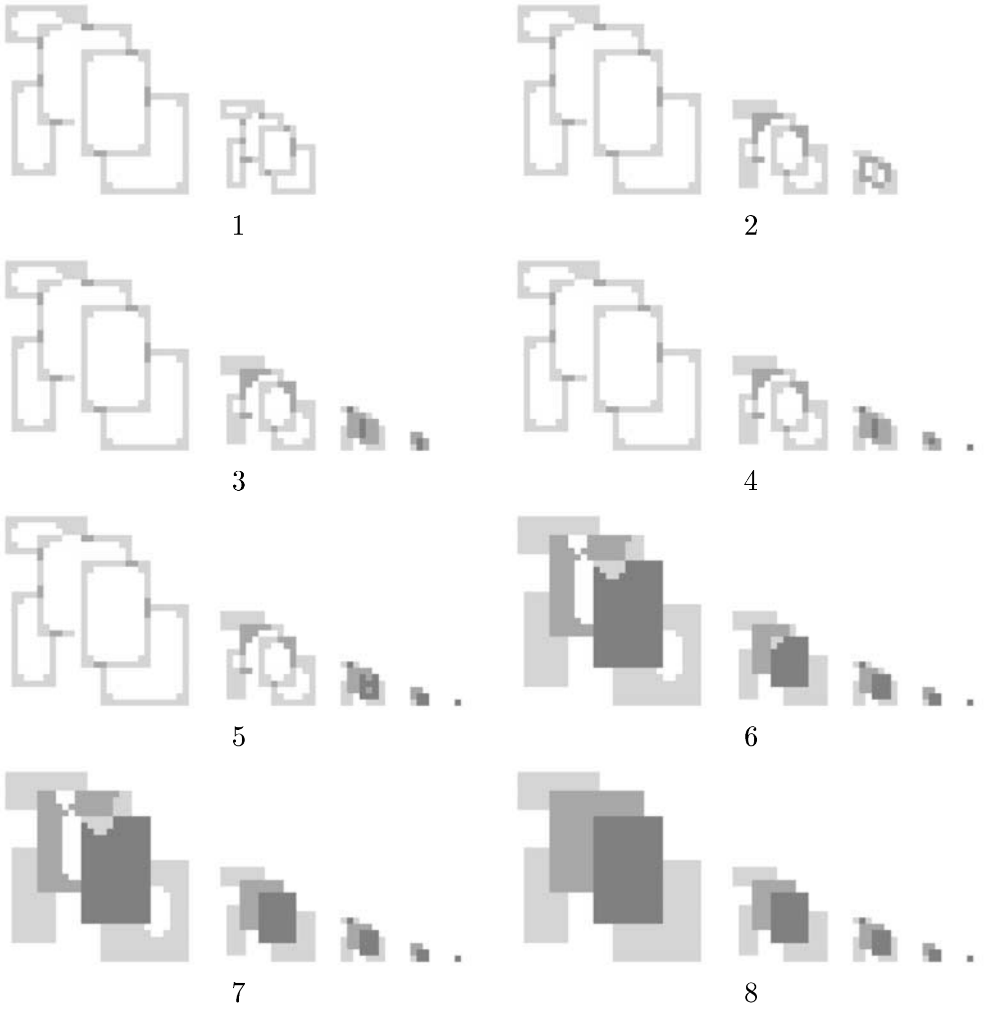


Figure 2: Dynamics of the HMRF's response to a 5-object test image. The parameters are learned on a few 2-object images. Shown here are a number of snaps shots taken at different time points during the depth segregation computation. The hierarchy is traversed twice till its complete convergence to the correct labeling.