A Probabilistic Network Model for Integrating Visual Cues and Inferring Intermediate-level Representations

Kyungim Baek and Paul Sajda

Department of Biomedical Engineering

Columbia University

New York, NY 10027

kb2107,sajda@columbia.edu

Abstract

Psychophysical data have demonstrated that our visual system must integrate multiple, spatially local and non-local cues to construct the visual scene. In this paper we describe a probabilistic network model which integrates visual cues to infer intermediate-level visual representations. We demonstrate the network model for two example problems: inferring "direction of figure" (DOF) [15] and estimating perceived velocity. One can consider the assignment of DOF as essentially a problem in probabilistic inference, with DOF being a hidden variable, assigning "ownership" of an object's occluding boundary to a region which represents the "figure". The DOF is not directly observed but can potentially be inferred from local observations and "message passing". For example, our model combines contour convexity and similarity/proximity cues to form observations, with belief propagation (BP) used to integrate these observations with state probabilities to infer the DOF.

We extend the network model, integrating form and motion streams, to explain the coherence-based motion effects first demonstrated by McDermott et al. [11]. The extended model consists of two interacting network chains (streams), one for inferring DOF and the other for inferring scene motion. The local figure-ground relationships estimated in the DOF stream are subsequently used by the motion stream as evidence for surface occlusion, modulating the covariance of a Gaussian distribution used to model the velocity at apertures located at junction points. The distribution of scene motion ultimately is represented in velocity space as a mixture of these form-modulated Gaussians.

Simulation results show that the network's integration of cues can account for several examples of perceptual ambiguity in DOF, consistent with human perception. Also, the integration of form and motion representations qualitatively accounts for psychophysical results showing surface dependent motion coherence of oscillating edges [11]. We also show that the model naturally integrates top-down cues, leading to perceptual bias in interpreting ambiguous figures, such as Rubin's vase, as well as bias in the perceived coherence of object motion.

1. Introduction

A challenge in developing any biologically plausible model of intermediate or high-level visual processing is resolution of the *generalized aperture problem*–i.e neurons or populations of neurons sample only a limited extent of the visual field which can lead to "observations" that are ambiguous relative to the structure and motion of objects. The aperture problem is well-known in motion processing (e.g. [6, 10, 16]), but the problem is also evident in form processing, for example inferring figure-ground relationships. Ultimately intermediate and high-level visual processing must "infer the scene" by integrating spatially non-local cues.

Multiple modalities, or *streams*, also must be integrated to resolve perceptual ambiguities and/or maintain consistency with human perception. For example, form and motion streams can interact in sophisticated ways, with form cues having significant impact on the perception of coherent motion. McDermott, Weiss and Adelson [11] have shown, through psychophysical experiments, that perceived motion coherence can be affected by the strength of configural cues, such as border ownership, amodal completion, and depth segregation. Their results demonstrate that the configural cues are not purely local (i.e. not simply junction cues) and instead require spatially non-local integration across the scene – though not global consistency. Thus integration across space and stream can be seen to be intimately linked.

Weiss and colleagues have demonstrated that visual integration (and segmentation) processes are naturally defined within a Bayesian framework. Weiss has provided numerous examples of how Bayesian modeling can account for a variety of motion illusions [19], inference of figure-ground [18] and integration of form and motion cues for motion segmentation [17]. These impressive and broad ranging results suggest Bayesian inference as a possible mechanism underlying intermediate-level visual processing.

In this paper we begin by extending the ideas of Weiss [18] for computing direction-of-figure (DOF), a surface representation proposed by Sajda and Finkel [15]. One can consider the assignment of DOF as essentially a problem in probabilistic inference, with DOF being a hidden variable that is not directly observed but can potentially be inferred from local observations and some form of message passing, representing spatial integration. We develop a locally connected probabilistic network model for integrating both local and non-local spatial cues for

inferring DOF, and demonstrate how, through the probabilistic integration of cues, the model can account for several figure-ground ambiguities that are consistent with human perception. In addition, the model naturally allows for the integration of top-down cues, representing bias or prior beliefs, and therefore can also account for a variety of reversible ambiguous figures (e.g. Rubin's Vase). The model, though not biologically realistic (e.g. nodes in the network do not represent spiking neurons), demonstrates how a locally connected network with nodes which are restricted to local observations in the visual field can integrate multiple spatial cues for inferring surface representations.

Our model also begins to consider how form and motion streams, both confronted with the aperture problem, might interact to infer scene structure and attributes. In particular, we demonstrate Bayesian mechanisms for integration and inference which qualitatively account for the motion coherence results of McDermott *et al* [11]. We extend our model for computing DOF (form stream) and add a second stream dedicated to motion processing. The probabilistic representation of the DOF plays a critical role in that the degree of certainty of the DOF modulates the degree of certainty of the local motion. This degree of DOF certainty (i.e belief) ultimately changes the scene motion and the "perception" of the network.

2. Belief Propagation

Solving an inference problem often begins with representing the problem using some form of graphical structure. Examples of such graphical models are Bayesian (or belief) networks [14] and undirected graphs, also known as Markov networks. In a graphical model a node represents a random variable and links specify the dependency relationships between these variables [7]. The states of the random variables can be hidden in the sense that they are not directly observable, but it is assumed that they may have observations related to the state values. Graphical models allow for a compact representation of many classes of inference problems. Once the underlying graphical structure has been constructed, the goal is to infer the states of hidden variables from the available observations. *Belief Propagation* (BP) is an algorithm for solving inference problems based on local message passing. In this section we focus on undirected graphical models with pairwise potentials. It has been shown that most graphical

models can be converted into this general form [20].

Let x be a set of hidden variables and y a set of observed variables. The joint probability distribution of x given y is given by,

$$P(x_1, \ldots, x_n|y) = c \prod_{i,j} T_{ij}(x_i, x_j) \prod_i E_i(x_i, y_i) \tag{1}$$

where c is a normalizing constant, x_i represents the state of node i, $T_{ij}(x_i, x_j)$ captures the compatibility between neighboring nodes x_i and x_j, and $E_i(x_i, y_i)$ is the local interaction between the hidden and observed variables at location i. An approximate marginal probability of the joint probability (eq. (1)) at node x_i over all x_j other than x_i is called the local ***belief***, $b(x_i)$.

The BP algorithm iterates a local message computation and belief updates [20]. The message $M_{ij}(x_j)$ passed from a hidden node x_i to its neighboring hidden node x_j represents the probability distribution over the state of x_j. In each iteration, messages and beliefs are updated as follows:

$$M_{ij}(x_j) = c \int_{x_i} dx_i T_{ij}(x_i, x_j) E_i(x_i, y_i) \prod_{x_k \in N_i/x_j} M_{ki}(x_i) \tag{2}$$

$$b(x_i) = c E_i(x_i, y_i) \prod_{x_k \in N_i} M_{ki}(x_i) \tag{3}$$

where N_i/x_j denotes a set of neighboring nodes of x_i except x_j. M_{ij} is computed by combining all messages received by x_i from all neighbors except x_j in the previous iteration and marginalizing over all possible states of x_i (Figure 1). The current local belief is estimated by combining all incoming messages and the local observations.

It has been shown that, for singly-connected graphs, BP converges to exact marginal probabilities [20]. Although how it works for general graphs is not well understood, experimental results on some vision problems, such as motion analysis, also show that BP works well for graphs with loops [4].

3. Probabilistic Network Model for Inferring DOF

3.1 DOF Problem

Segmentation of a scene into foreground figures and background is a central problem in visual perception. Recent studies report evidence showing that shape selective cells in infero-

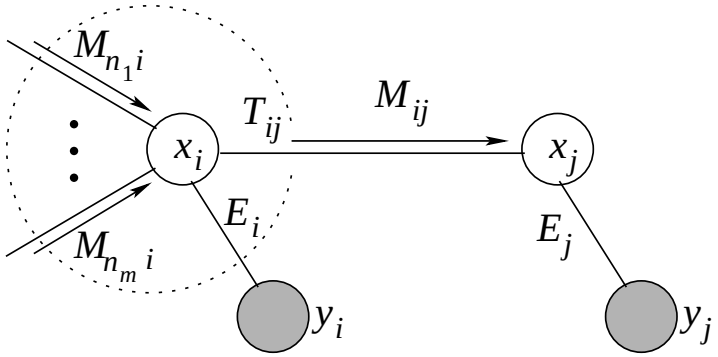

Figure 1: Illustration of local message passing from node x_i to node x_j. Open circles are hidden variables, while shaded circles represent observed variables. The local belief at node x_j is computed by combining the incoming messages from all its neighbors and the local interaction E_j.

temporal cortex and lateral occipital complex, known to be important for visual perception and recognition, respond selectively to figure-ground relationships [2, 9]. Although these results also indicate that figure-ground selectivity of cell responses is not entirely determined by the contours, psychological evidence of shape perception suggests that the figure-ground segregation starts at contours by assigning "ownership" of an object's occluding boundary to a region which represents the object's surface [12, 13]. The contour ownership, sometimes referred to as "direction of figure" [15], therefore specifies a figure-ground relationship directly for the occluding contours. The contour ownership is hypothesized to be assigned in early stages of visual processing. Recent neurophysiological findings reported by Zhou *et al.* support the hypothesis, showing that more than half of the neurons in extrastriate cortex (areas V2 and V4) are selective to contour ownership [21].

Compared to where the contour ownership is processed, less is known about how the ownership is resolved computationally. In previous studies, several computational models have been developed and applied to estimate DOF. In [15], Sajda and Finkel proposed a network-based neural computational model to represent object surfaces through contour ownership. In their model mechanisms, which utilize various local and global cues, cooperate to compute DOF. Simulation results show that the model can account for a variety of perceptual phenomena.

However, there are limitations in that their model does not generalize when one considers the convergence of cues provided through different stages – i.e. convergence of top-down, bottom-up and "horizontal" (within area) cues.

Recent work by Weiss has applied a probabilistic network model based on the BP algorithm to the DOF problem [18]. In his BP model, DOF computation is formulated as probabilistic inference where DOF is represented as a hidden variable that is not directly observable. The model infers DOF from local observations by propagating local estimates at different locations along the contour. The performance of the BP model is compared with three relaxation labeling algorithms, with results showing the BP model's superiority over other techniques in terms of both accuracy and convergence speed.

Weiss's model uses observations made using a single cue, which is local figure convexity. Although convexity is considered as having strong influence on perceiving figures compared to other cues such as contrast polarity or symmetry [8], using the convexity cue alone is insufficient in many cases, in particular those with perceptual ambiguity in figure-ground. In this section we describe our network model which infers DOF by combining figure convexity and similarity/proximity cues to form observations.

3.2 Graphical Model and Inference of DOF using BP

We formulate the DOF problem using an undirected chain, shown in Figure 2, similar to Weiss [18]. The hidden variable x_i represents a two dimensional binary DOF vector at location i along the boundary. The vector $x_i = (1, 0)^T$ specifies that the DOF is in the direction of local convexity, while $x_i = (0, 1)^T$ assigns the opposite direction to the DOF.

Since the graph is a chain, every node has two neighbors from which it receives messages. Furthermore, the hidden variables are discrete and have two possible states, with incoming messages and belief at x_j in Figure 2 computed as follows:

$$M_{ij}(x_j) = c\sum_{x_i} T_{ij}(x_i, x_j)E_i(x_i, y_i)M_{hi}(x_i) \quad (4)$$

$$M_{kj}(x_j) = c\sum_{x_k} T_{kj}(x_k, x_j)E_k(x_k, y_k)M_{lk}(x_k) \quad (5)$$

$$b(x_j) = cE_j(x_j, y_j)M_{ij}(x_j)M_{kj}(x_j) \quad (6)$$

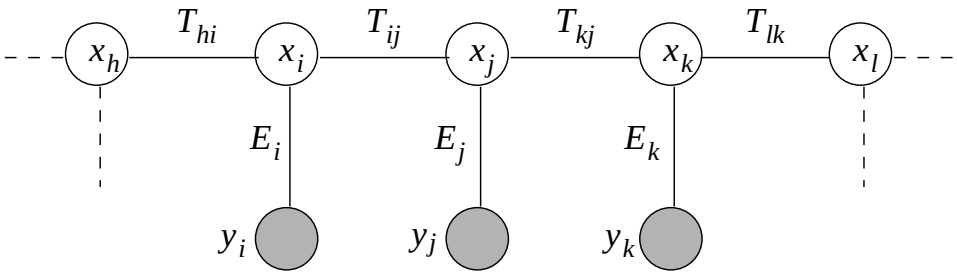

Figure 2: An undirected chain graph where each hidden node (open circle) has a corresponding observation (shaded circle).

where the sum is over two possible states of the hidden variables. The vector multiplication is performed element by element. Each element of $b(x_j)$ represents the degree of confidence for DOF for the corresponding direction at location j. When the algorithm converges the states of the hidden variables are determined by taking the direction having larger confidence.

3.3 Combining Multiple Cues

Perceptual studies have shown that various cues are used for figure-ground discrimination. Examples of such cues are figure convexity, similarity/proximity, contour closure, contour continuation, symmetry, T-junctions, and binocular disparity [8, 13, 15]. Although it has been suggested that some cues have stronger influence than others (e.g. convexity prevails over symmetry [8]), generally a single cue is not capable of explaining all aspects of the perceptual phenomena. In this work, we combine local figure convexity and similarity/proximity cues and show that it can further extend the ability of inferring DOF of the BP model.

The convexity at point i is determined by the local angle of the contour at the location. Let a_i be the local angle. The local interaction between hidden variable x_i and observed variable $y_{i,cvx} = (\theta, 2\pi - \theta)$ is defined as,

$$E_{i,cvx}(x_i, y_{i,cvx}) = (exp(-\theta), exp(-(2\pi - \theta))) \tag{7}$$

where $\theta = min(a_i, 2\pi - a_i)$ (Figure 3). This formulation prefers smaller angles.

To compute the similarity/proximity cue, we first look for points having similar local tangent angle that lie in a direction orthogonal to the contour at a given location. Once those points

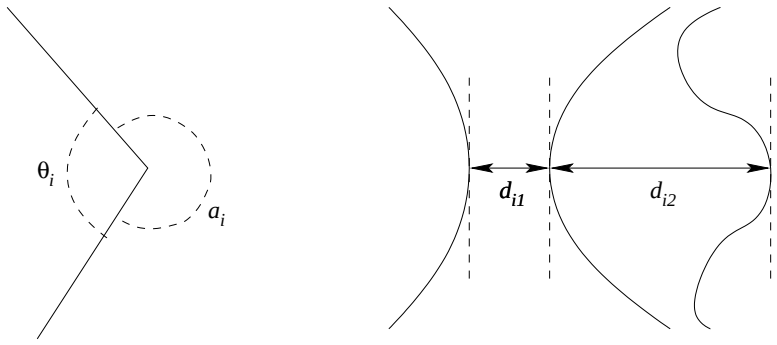

Figure 3: *Left*: The local angle a_i at point i is greater than π, so the angle θ in the direction of convexity is $2\pi - a_i$. *Right*: The distances from point i to the closest points having similar local tangent angle. Those points should lie in the direction orthogonal to the contour at point i.

are found, which side of the local convexity direction the points reside is determined. Let $y_{i,sim} = (d_{i1}, d_{i2})$, where d_{i1} and d_{i2} are the distances to the closest similar points found in each direction (Figure 3). Then, the local interaction is again defined using an exponential, i.e.

$$E_{i,sim}(x_i, y_{i,sim}) = (exp(-d_{i1}), exp(-d_{i2})) \tag{8}$$

Thus, the influence of the similarity/proximity cue decreases as the distance increases.

The two local interactions defined in the equation (7) and equation (8) are combined based on a *weak fusion model* [3]. The weak fusion model is a simple scheme which suggests that a property of the environment is estimated separately by each independent cue, and then combined in a weighted linear fashion to compute the overall effect. Following weak fusion, the total local interaction E_i is computed by weighted average of the interactions made by two separate observations $y_{i,cvx}$ and $y_{i,sim}$:

$$E_i = w_{cvx}E_{i,cvx} + w_{sim}E_{i,sim} \tag{9}$$

4. Probabilistic Network Model for Velocity Estimation

Based on the data from various psychophysical experiments indicating that the strength of form cues can significantly affects the perception of motion coherence, we add to the previous

network model a separate, structurally identical but functionally different processing stream. As a result, the proposed network model consists of two streams: the form stream inferring DOF and the motion stream performing velocity estimation (Figure 4). In current model the two streams interact unidirectionally so the influence flows from form stream to motion stream only. Both DOF and velocity are inferred from local observations using belief propagation.

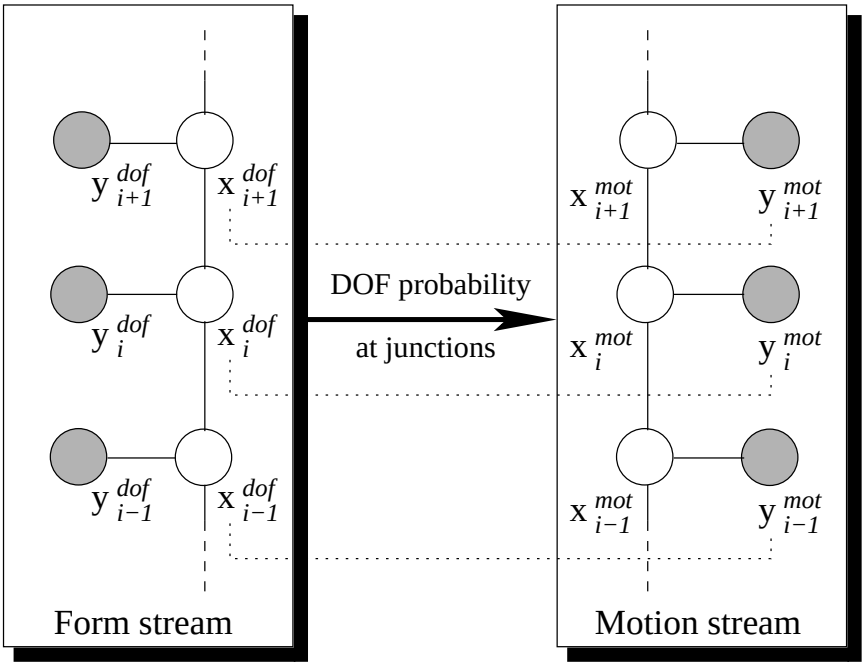

Figure 4: Bayesian network model for form and motion streams. The DOF probability computed in the form stream affects the motion stream by modulating the covariances of distributions used to model the velocity at apertures located at junction points. In the present model, the interaction between form and motion streams is unidirectional.

4.1 Interaction between Form and Motion Streams and Velocity Estimation

Due to the aperture problem, a single local measurement along the contour cannot determine the object's motion, and therefore motion integration across space is required. There are also several features, such as corners, line terminators, and junctions, that may provide unambigu-

ous local motion and dominate the motion integration. In this work, we focus on the influence of a form cue (DOF) on local motion at junctions, and subsequently on the motion integration.

The degree of certainty in DOF at junctions inferred in the form stream can be used to distinguish between *intrinsic terminators*, that are due to the natural line ending of an object, and *extrinsic terminators*, that are not created by the end of a line itself but rather via occlusion by another surface. Intrinsic terminators provide an unambiguous signal for the true velocity of the line, while extrinsic terminators provide a locally ambiguous signal which must be suppressed for accurate motion computation [12]. DOF is an explicit surface representation, and therefore the degree of certainty (i.e belief) in DOF at junctions can be used to represent the strength of the evidence for surface occlusion, which determines the terminator type.

The hidden node in the motion stream represents a velocity of the corresponding location along the contour. We assume that both the pairwise compatibility $T_{ij}(x_i^{mot}, x_j^{mot})$ and the local interaction $E_i(x_i^{mot}, y_i^{mot})$ that models the velocity likelihood at apertures are Gaussian. T_{ij} is set manually and E_i is defined by a mean of normal velocity at point i and a local covariance matrix Cov_i (Figure 13). Before the BP algorithm starts, the variance at junction points is modulated by a function of DOF belief $b(x_i^{dof})$ as follows:

$$Cov_i = e^{\alpha\{b(x_i^{dof})-0.5\}} \cdot Cov_i \tag{10}$$

Initially, the covariance matrices of hidden variables are set to represent infinite uncertainty, and mean vectors are set to zero. When the BP algorithm converges, the motion integration is performed by computing the mixture of Gaussians:

$$p(v) = \sum_i p(v|i)p(i) \tag{11}$$

where $p(v|i)$ is the probability of velocity v from the resulting Gaussian of hidden variable x_i^{mot}, and $p(i)$'s are the mixing coefficients.

5. Simulation Results

Throughout the experiments described in this section, we set the pairwise compatibility as:

$$T_{ij}(x_i, x_j) = \begin{cases} 0.995 & \text{if } x_i = x_j \\ 0.005 & \text{if } x_i \neq x_j \end{cases} \tag{12}$$

Thus, T_{ij} reflects strong preference for neighboring points to have the same DOF or similar velocity. Also, all results are obtained using the same set of weight values for w_{cvx} and w_{sim}. The number of iterations in the BP algorithm was set to the number of nodes to make sure that information from each node has a chance to propagate to all other nodes.

5.1 Inferring DOF

The right figure in Figure 5 shows the predicted DOF by the probabilistic network model for an arbitrary shaped figure. Along with it, the initial local estimate of DOF is shown on the left. The color of the DOF indicator represents the confidence, with increasing confidence as one moves from blue to red. The initial estimate on the left clearly shows the preference for local convex direction with stronger confidence around locations of high local curvature. Although the network model quickly converges to the correct DOF, the confidence levels continue updating, resulting in a very high confidence – greater than 95% – in every location as shown in Figure 5 (Right).

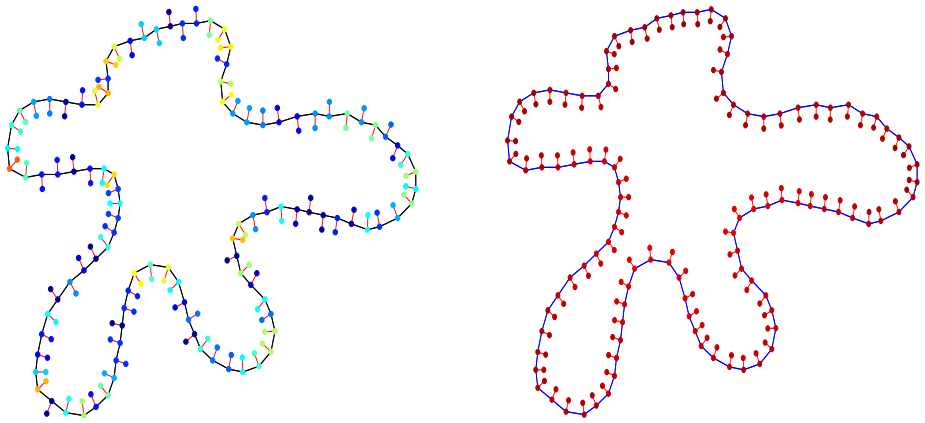

Figure 5: *Left*: Initial local estimate of DOF along the contour. *Right*: DOF prediction after convergence. The confidence is very high ($> 95\%$) in every location on the contour. Degree of confidence increases as the color changes from blue to red.

5.1.1 Ambiguous Figure-Ground

There are several classic examples in which discriminating figure from background is not immediately clear. The first two square spiral figures shown in Figure 6 are such examples [15]. Discrimination of the figure's surface in the first spiral is difficult , with difficulty increasing for regions close to the center of the spiral[1]. On the other hand, discriminating the figure in the second spiral appears to be straightforward. Immediately, we tend to perceive the thin strip in the center as figure, however this is incorrect. In this case, the width of the spiral increases as it winds around toward the center, generating an incorrect perception of figure-ground. Unlike the first two figures, correct figure-ground discrimination can be correctly made almost instantly for the third spiral.

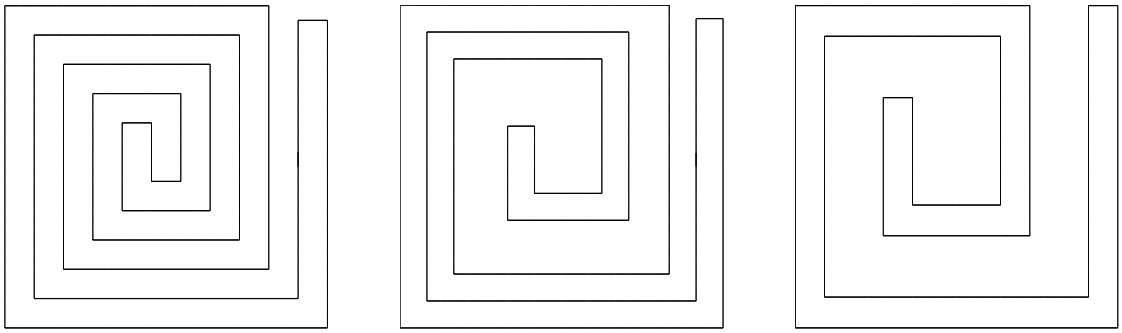

Figure 6: *Left*: Spiral figure in which figure-ground discrimination is ambiguous unless serially tracing the contour. *Middle*: Increasing spiral width as it winds around toward the center generates incorrect figure-ground percept. *Right*: Figure in which the correct percept of figure-ground can be made immediately. All figures are reproduced from [15].

As one might expect, a BP model using convexity cues alone predicts DOF in the same way for all three cases as shown in Figure 7 (top). Convexity cues conflict only on the boundaries of the outermost and innermost strips, so the confidence of DOF is very high, except on the two sides where the cue conflicts. As a result, the pattern of DOF prediction is the same regardless of the spiral type (top row in Figure 7). As discussed in [15], similarity/proximity cue can

[1]Note that all DOF relationships can be determined if one traces the contours of the spiral and "propagates" high confidence DOF assignments to low confidence regions. However in this discussion we ignore such "tracing", though such information could be integrated as another form of prior cue.

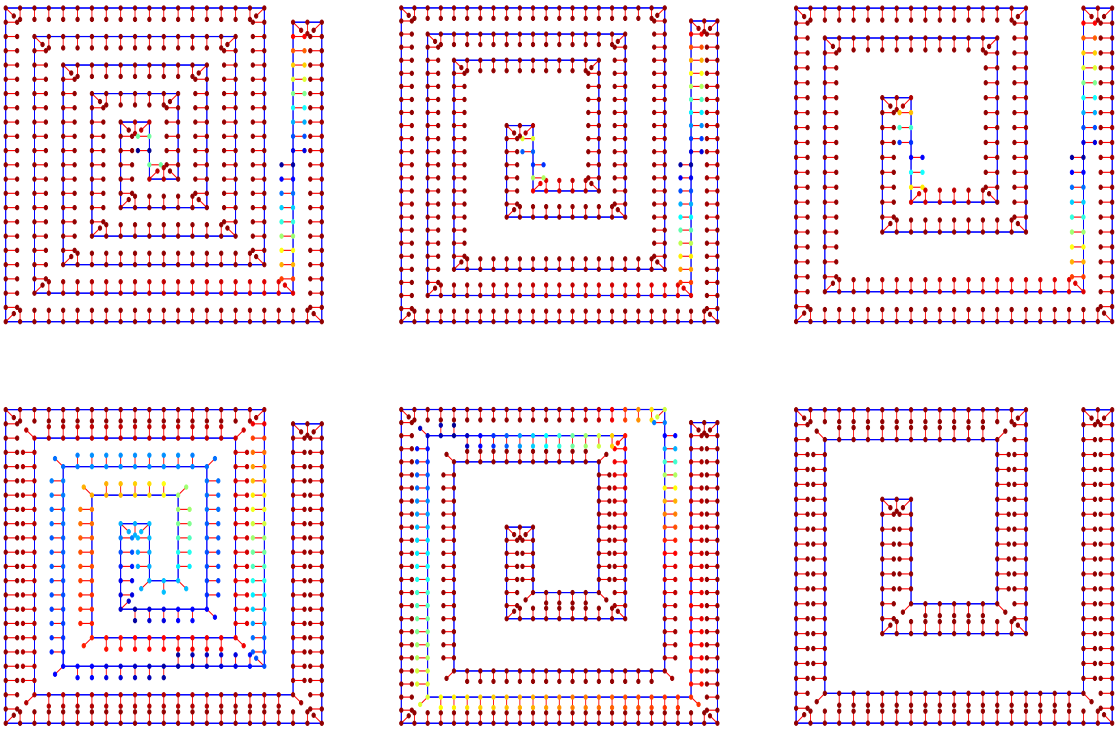

Figure 7: DOF estimation results for the three spiral images in Figure 6. *Top row*: DOF estimate using local convexity cue alone. *Bottom row*: DOF estimate by the model using both convexity and similarity/proximity cues consistent with human perception. Note the low certainty (blue) for the central part of the first spiral, indicating ambiguity of DOF, consistent with human perception. Also note the high certainty (red) for the central part of the second spiral, an incorrect percept which is consistent with human interpretation.

be considered as a main source of the ambiguous and inconsistent interpretation. Indeed, the network model using both convexity and similarity/proximity cues predicts DOF very close to human perception. DOF figures shown in the bottom row of Figure 7 illustrate increasing ambiguity and/or incorrect interpretation for the center region of the first two spirals, and perfect figure segmentation for the last spiral. Although they are not shown here, results show that it takes longer for the network model to converge for the first two spirals, with more oscillations in the DOF assignment, as compared to the third example. The results are slightly different, especially near the periphery of the spirals, from those obtained by the neural computational model described in [15]. We believe that this is because the current network model does not exploit observations from closure cues.

5.1.2 Perceptual Shift Induced by Prior Information

Certain figures are perceptually ambiguous in that figure and ground can shift or even oscillate. One of the most famous figures that demonstrates this perceptual ambiguity is Rubin's vase, shown in Figure 8. In this figure, one can perceive either faces or a vase (never both simultaneously), and whenever the perceived object is shifted the contour ownership is also changed accordingly.

Figure 8: Rubin's vase (from [1]).

One can bias the interpretation of the ambiguous figures by providing prior information. For example, prior cues might emerge from recognition of distinct visual features (e.g. nose, eyes, chin). In our probabilistic network model, prior information can be considered as another

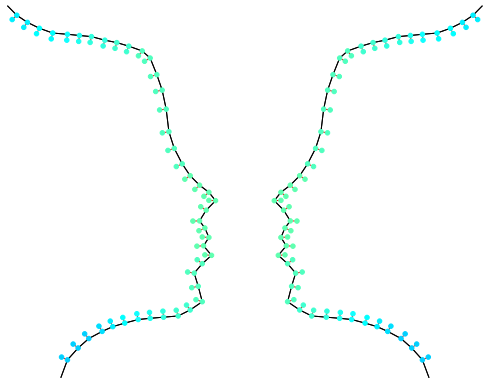

Figure 9: DOF prediction for Rubin's vase without using prior information.

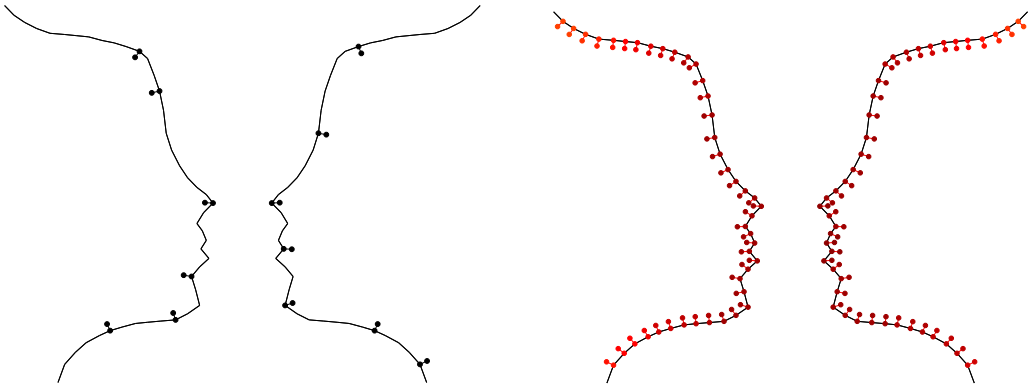

Figure 10: *Left*: Prior cue for face features. *Right*: DOF estimate of the network model integrating prior information. Figure shows the bias of DOF toward faces, reflecting strong influence of the prior cue.

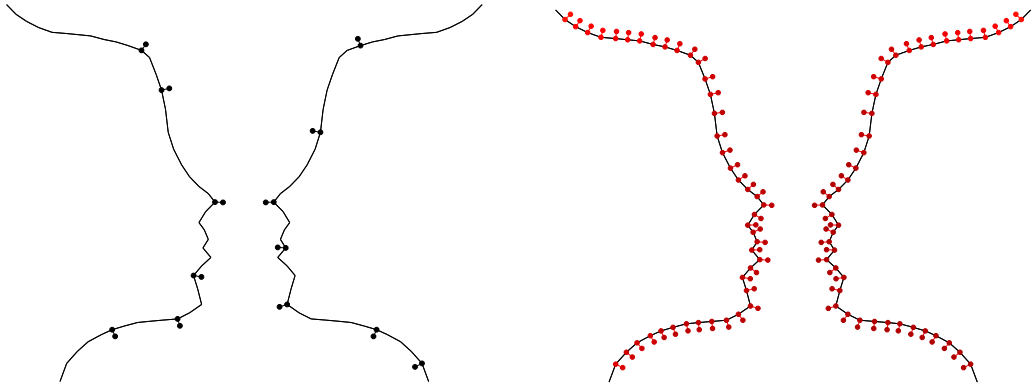

Figure 11: *Left*: Prior cues for vase features. *Right*: DOF estimate of the network model integrating prior information. Figure shows the bias of DOF toward vase, reflecting strong influence of the prior cue.

cue for DOF and therefore combined in the same way as described in section 3.3. In this case, the network model integrates bottom-up (convexity), horizontal (similarity/proximity), and top-down (prior information) cues using a simple combination mechanism.

Figure 9 shows the DOF estimation for the Rubin's vase image without using prior information. The network assigns the contour to faces mainly because of the influence from the convexity cue, however the confidence level is low throughout the entire contour. Prior information is added by explicitly specifying figure direction at several locations on the contour (top figures in Figure 10 and Figure 11). The results of integrating the prior cue are shown in the bottom of Figures 10 and 11. Even if a small weight is assigned for the prior cue[2], it strongly biases the DOF.

5.2 Velocity Estimation

Figure 12 shows the two stimuli used for these experiments. They are 90 degree rotated versions of the diamond stimuli described in McDermott *et al.*[3] [11], but the perceptual effect is basically the same. The motion of moving line segments is identical between the two stimuli. A pair of vertical line segments moves together sinusoidally in a horizontal direction while

[2]The ratio of w_{prior} over $(w_{cvx} + w_{sim})$ is $1 : 8$.

[3]See http://koffka.mit.edu/~kanile/master.html for flash demos.

the two horizontal lines move in a vertical direction. The vertical and horizontal motions are 90 degrees out of phase. The difference between the two stimuli is the shape of the occluders, which alters the perceived motion. A single coherent rotation is perceived for stimulus B, while we are more likely to see two separate motions of the line segments for stimulus A [11].

We argue that the belief for DOF represents the strength of surface occlusion which determines the property of line terminators at junctions and as a result, changes the motion perception. For stimulus A, the belief for DOF at junction points is weaker than that for stimulus B. Therefore, the junctions in stimulus A are considered more likely to be intrinsic terminators whose motion is less ambiguous (left figure in Figure 13). Conversely, for stimulus B, the belief for DOF for the occluding surfaces at junction points is very strong so the ambiguity of the local motion increases. Therefore, the sharp peak shown in the left figure of Figure 13 flattens out.

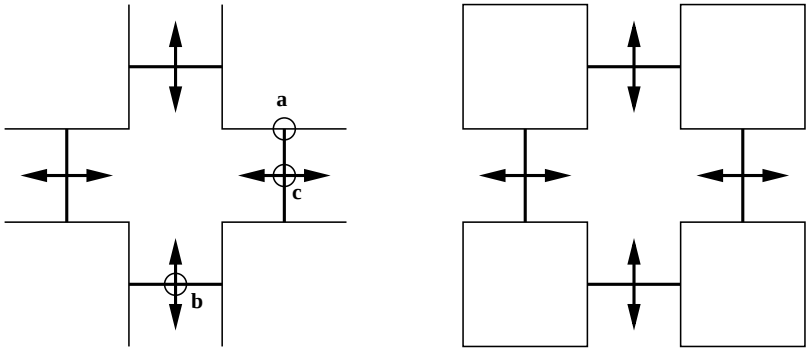

Figure 12: Stimuli generated from four line segments that move sinusoidally, 90 degrees out of phase, in vertical and horizontal directions. The line segments are presented with L-shaped occluders (stimulus A) and closed square occluders (stimulus B). The presence of the occluding surface alters the motion perception. Shown in stimulus A are three local apertures (a, b, c), which would have identical information content in stimulus B.

Figure 14 shows the resulting velocity estimated by the network model for stimulus A (top row) and stimulus B (bottom row). Since the estimated motions along the line segments that move simultaneously are almost identical after convergence, the first two columns show only the motions at locations **b** and **c** (shown in Figure 12) in the velocity space. In the third column,

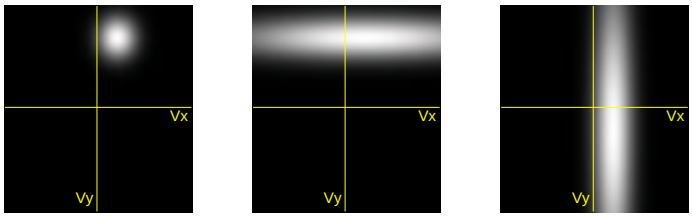

Figure 13: Velocity likelihood for three apertures shown in Figure 12. *Left*: Likelihood at junction point **a**. *Middle*: Likelihood at location **b** on a line segment moving vertically. *Right*: Likelihood at location **c** on a line segment moving horizontally.

which displays the integrated motion, we clearly see the bimodal distribution for stimulus A, while a single peak is formed at the intersection of two distributions for stimulus B. This implies that we perceive two separate motions for stimulus A, and a single coherent motion for stimulus B. Figure 15 illustrates the resulting velocity estimates for six successive frames sampled from a period of the sinusoidal motion. The maximum a posterior estimate for each frame follows a circular trajectory for stimulus B, which is consistent with the perception of rotation.

A prior cue shown in Figure 16 can be added for inferring DOF in stimulus A. Psychophysically, this simulates priming subjects or using stereo to add depth information so that subjects are biased to see the occlusion. Adding priors to the form stream in the model strengthens the belief for DOF in the indicated direction, and consequently more coherent motion would be expected. Figure 16 shows the results with weak and strong weights on the prior cue. Compared to the top row of Figure 14, the solution becomes more unimodal and, with a strong prior it produces a single peak at the intersection similar to stimulus B.

6. Conclusion

We have presented a probabilistic network model for integrating multiple cues, including both spatially local and non-local cues as well as different modalities such as form and motion streams. The "perception" of the network model was demonstrated for two problems involving intermediate-level surface representations: inferring direction of figure and velocity estimation.

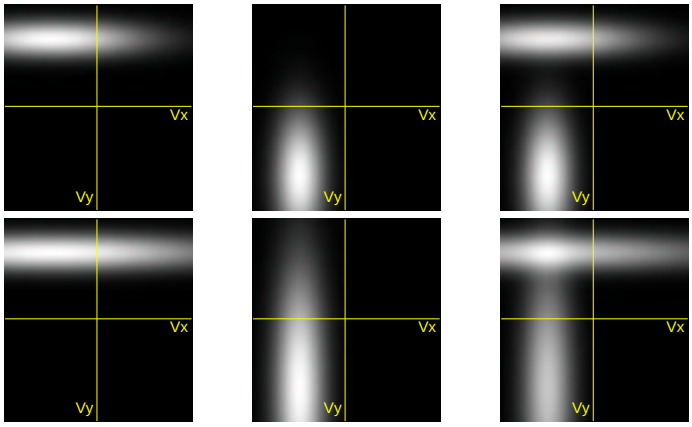

Figure 14: Velocity estimation results for stimulus A (*top*) and stimulus B (*bottom*). *Left*: Estimation at location **b**. *Middle*: Estimation at location **c**. *Right*: Velocity computed by combining the two estimation in the first two columns using mixture of Gaussian.

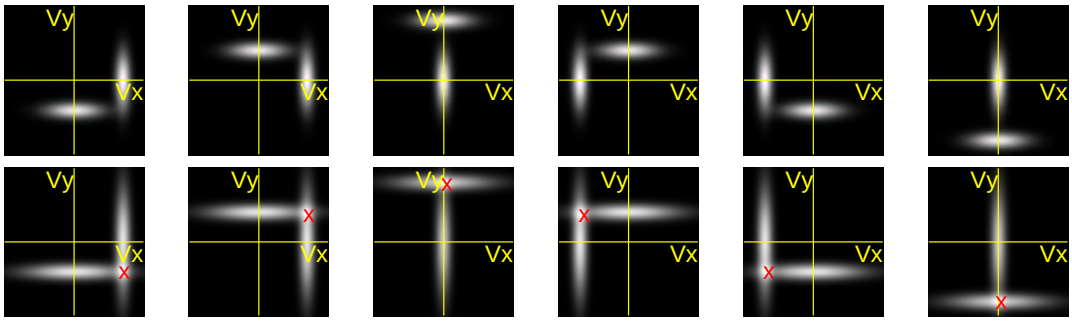

Figure 15: A sequence of resulting velocity estimation for six successive frames sampled from a period of sinusoidal motion with a regular interval. Top row shows two separate motions oscillating in the direction normal to the line segment for stimulus A. The bottom sequence shows a coherent motion forming a circular trajectory.

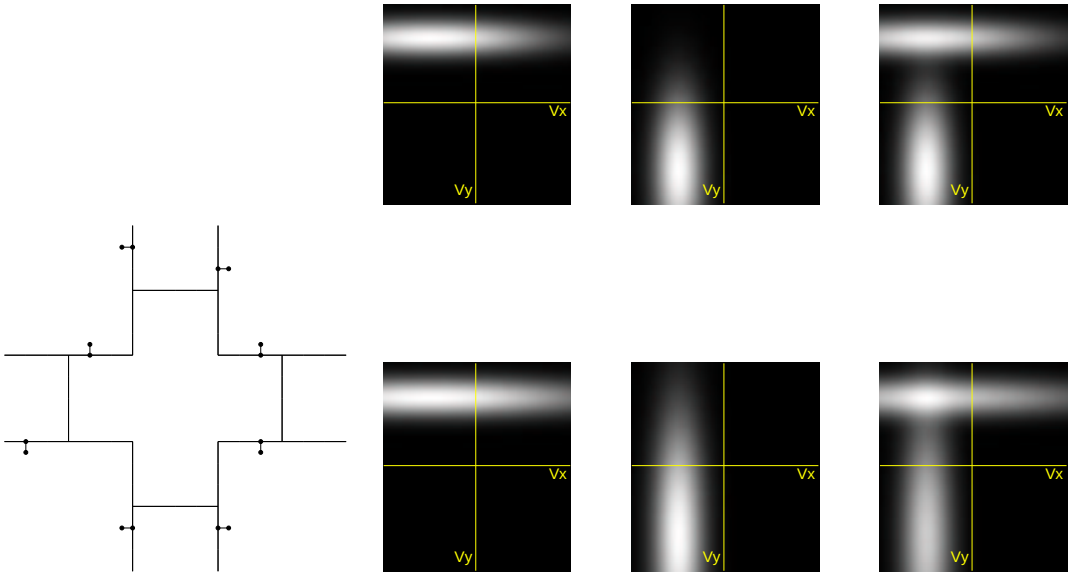

Figure 16: *Left*: Prior cue for stimulus A used for inferring DOF. *Right*: Velocity estimation results with weak (*top*) and strong (*bottom*) weights on prior cues.

The inference scheme of the model is based on a local belief propagation algorithm. Simulation results show that the integration of multiple cues extends the network model's ability to correctly infer DOF, allowing it to account for certain cases of perceptual ambiguity, consistent with human perception. Results for velocity estimation also demonstrate qualitative agreement with psychophysical motion coherence experiments.

Our implementation for integrating form and motion streams is not meant to represent a complete model, instead it suggests that spatial cues from multiple streams might be integrated using a common framework to "construct" a scene. Others have attempted to develop more biologically realistic models of form and motion integration [5], with mechanisms that are perhaps more "neural". However these models are difficult to analyze, particularly in terms of understanding the general role/effect of "uncertainty" in the observations and reversal or ambiguity in perception. In so much that we do not claim that ours is as a biologically realistic model, it is worth noting that it is constructed as a network model, with no requirements for global connectivity or consistency. An interesting question to explore is whether the human visual system possesses the necessary neural machinery for locally constructing distributions of observations and integrating these via local "message passing".

Acknowledgments

This work was supported by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-1-0625, and a grant from the National Imagery and Mapping Agency, NMA201-02-C0012.

References

[1] T. J. Andrews, D. Schluppeck, D. Homfray, P. Matthews, and C. Blakemore, "Activity in the fusiform gyrus predicts conscious perception of Rubin's vase-face illusion," *NeuroImage*, Vol. 17, pp. 890-901, 2002.

[2] G. Baylis and J. Driver, "Shape-coding in IT cells generalizes over contrast and mirror reversal, but not figure-ground reversal," *Nature Neuroscience*, Vol. 4, pp. 937-942, 2001.

[3] J. J. Clark and A. L. Yuille, *Data fusion for sensory information processing systems*, Kluwer Academic, Boston, 1990.

[4] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *International Journal of Computer Vision*, Vol. 40, No. 1, pp. 25-47, 2000.

[5] S. Grossberg, E. Mingolla and L. Viswanathan, "Neural dynamics of motion integration and segmentation within and across apertures," *Vision Research*, Vol 41, pp 2521-2553,, 2001.

[6] B.K.P. Horn and B.G. Schunck, "Determining optical flow," *Artificial Intelligence*, Vol. 17, pp 185-203, 1981.

[7] M. Jordan and C. Bishop, *An introduction to graphical models*, MIT press, 2003 (in press).

[8] G. Kanizsa, *Organization in vision*, Praeger, New York, 1979.

[9] Z. Kourtzi and N. Kanwisher, "Representation of perceived object shape by the human lateral occipital complex," *Science*, Vol. 293, pp. 1506-1509, 2001.

[10] D. Marr and S. Ullman, “Directional selectivity and its use in early visual processing,” *Proc. R. Soc. Lond. B Biol. Sci.*, Vol 211, pp 151-180, 1981.

[11] J. McDermott, Y. Weiss, and E. H. Adelson, “Beyond junctions: Nonlocal form constraints on motion interpretation,” *Perception*, Vol. 30, pp. 905-923, 2001.

[12] K. Nakayama, S. Shimojo, and G. H. Silverman, “Stereoscopic depth: its relation to image segmentation, grouping, and the recognition of occluded objects,” *Perception*, Vol. 18, pp. 55-68, 1989.

[13] K. Nakayama, “Binocular visual surface perception,” *Proc. Natl. Acad. Sci. USA*, Vol. 93, pp. 634-639, 1996.

[14] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann, 1988.

[15] P. Sajda and L.H. Finkel, “Intermediate-level visual representations and the construction of surface perception,” *Journal of Cognitive Neuroscience*, Vol. 7, No. 2, pp. 267-291, 1995.

[16] H. Wallach, “Ueber visuell whargenommene bewegungrichtung,” *Psychol. Forsch*, Vol. 20,pp 325-380, 1935.

[17] Y. Weiss and E.H. Adelson, “Perceptually organized EM: A framework for motion segmentation that combines information about form and motion,” *MIT Media Lab. Perceptual Computing Section TR*, No.315, 1994.

[18] Y. Weiss, “Interpreting images by propagating Bayesian beliefs,” in M. C. Mozer, M. I. Jordan, and T. Petsche (Eds.), *Advances in Neural Information Processing Systems*, Vol. 9, pp. 908-915, 1997.

[19] Y. Weiss, E.P. Simmoncelli and E.H. Adelson, “Motion illusions as optimal percepts,” *Nature Neuroscience*, Vol 5, pp 598-604, 2002.

[20] J. S. Yedidia, W. T. Freeman, and Y. Weiss, “Understanding belief propagation and its generalizations,” in G. Lakemeyer and B. Nebel (Eds.), *Exploring Artificial Intelligence in the New Millennium*, pp. 239-269, 2003.

[21] H. Zhou, H. S. Friedman, and R. von der Heydt, “Coding of border ownership in monkey visual cortex,” *Journal of Neuroscience*, Vol. 20, No. 17, pp. 6594-6611, 2000.