LETTER Communicated by Ning Qian

Modeling Stereopsis via Markov Random Field

Yansheng Ming
ysming@nlpr.ia.ac.cn
Zhanyi Hu
huzy@nlpr.ia.ac.cn
National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing, 100190, P.R.C.

Markov random field (MRF) and belief propagation have given birth to stereo vision algorithms with top performance. This article explores their biological plausibility. First, an MRF model guided by physiological and psychophysical facts was designed. Typically an MRF-based stereo vision algorithm employs a likelihood function that reflects the local similarity of two regions and a potential function that models the continuity constraint. In our model, the likelihood function is constructed on the basis of the disparity energy model because complex cells are considered as front-end disparity encoders in the visual pathway. Our likelihood function is also relevant to several psychological findings. The potential function in our model is constrained by the psychological finding that the strength of the cooperative interaction minimizing relative disparity decreases as the separation between stimuli increases. Our model is tested on three kinds of stereo images. In simulations on images with repetitive patterns, we demonstrate that our model could account for the human depth percepts that were previously explained by the second-order mechanism. In simulations on random dot stereograms and natural scene images, we demonstrate that false matches introduced by the disparity energy model can be reliably removed using our model. A comparison with the coarse-to-fine model shows that our model is able to compute the absolute disparity of small objects with larger relative disparity. We also relate our model to several physiological findings. The hypothesized neurons of the model are selective for absolute disparity and have facilitative extra receptive field. There are plenty of such neurons in the visual cortex. In conclusion, we think that stereopsis can be implemented by neural networks resembling MRF.

1 Introduction

Stereopsis is a process that our visual system employs to estimate the distance of an object by measuring the binocular disparity, defined as the difference in positions of the object's images on two retinas. At the heart

of stereopsis lies the correspondence problem: How are images projected from the same 3D feature linked? This problem is not trivial because, aside from correct matches, there are also many false matches. The correspondence problem manifests itself in the primary visual cortex, the beginning of stereopsis in the visual pathway. Complex cells in the primary visual cortex are considered disparity detectors whose disparity selectivity can be predicted by the disparity energy model (Ohzawa, DeAngelis, & Freeman, 1990). However, previous experiments show that these disparity detectors respond not just to correct matches but also to false ones (Cumming & Parker, 1997, 2000). The ambiguous responses of complex cells cannot account for the unambiguous depth perception. Therefore, some mechanism must be employed to eliminate false matches and recover true disparity, which is equivalent to solving the correspondence problem.

In the computer vision field, advancements in the Markov random field (MRF) and associated inference algorithms like belief propagation (BP) have led to stereo vision algorithms with top performance (Sun, Zheng, & Shum, 2003; Yang, Wang, Yang, Stewenius, & Nister, 2009). These algorithms typically use a pixel as the basic matching unit. Tests on benchmark image sets have demonstrated that ambiguities of matching pixels are solved to a large extent.

This letter aims to explore such possibilities as whether MRF-based stereo algorithms are biologically plausible and if stereopsis can actually be implemented by neural networks that pass messages such as BP. As a response, a biological stereopsis model based on MRF is proposed in this study. There are two differences between our model and conventional computer stereo vision algorithms. First, local evidence of matching is computed from population responses of complex cells. In other words, the likelihood function is constructed on the basis of the disparity energy model. In section 2.2, we show that this likelihood function is related to some psychophysical findings. Second, we restrict our choice of potential functions by taking into account Petrov's psychological findings (Petrov, 2002). Some experiments show that our visual system tries to minimize relative disparities across the scene by altering the matching of individual features (Zhang, Edwards, & Schor, 2001; Goutcher & Mamassian, 2005). Petrov's experiment suggests that the strength of this interaction falls as the distance between the features increases. Potential functions are usually used in computer stereo vision algorithms in order to enforce the continuity constraint. We find that not every potential function could account for Petrov's finding. In section 2.3, a likelihood function that considers this finding is provided.

Different biological implementations of BP have been suggested in the literature (Rao, 2004, 2005; Ott & Stoop, 2006). Rao (2004, 2005) implemented BP in the log domain. Ott and Stoop (2006) took a more indirect approach. They proved that BP on a binary MRF could be implemented by the continuous Hopfield network, which was considered more biologically plausible. We adopt Rao's basic principle for two reasons. First, we have no desire to

restrict our model to a binary type. Second, Rao's implementation is more direct and closer to the mathematical form of BP. Because the topology of our network is different from that in his article, we add message neurons to represent messages. After all, details of BP implementation are not our main concern in this work in light of the lack of sufficiently informative physiological evidence in the literature. Instead, this letter relies on the basic idea of biological BP, which is to interpret neural dynamics as passing messages. In our model, these messages will excite nearby neurons, which have similar disparity preferences. Therefore, neurons in our model could be considered as having a facilitative extrareceptive field (ERF). The physiological relevance of this is discussed in section 4.

Our model is tested on three kinds of stereo images. First are images with repetitive patterns, such as sinusoidal gratings. This kind of image is popular to use in psychological experiments because there are many false matches in the central region. Experiments found in the literature showed that the disparity at the central region was controlled by the disparity at the ends as if our visual system tried to minimize relative disparities, thus conforming to the continuity constraint. In addition, the second-order mechanisms were hypothesized to encode disparity at the ends because the first-order mechanism, tuned to low spatial frequency, could not account for edge-based matching when the image contrast was low (McKee, Verghese, & Farell, 2004). However, in our model, the matching ambiguity of the central region is solved naturally by propagating messages from both ends. Our model is not affected by low contrast because complex cells tuned to low frequencies are not employed. The simulations show that our model could predict human depth percepts in many situations, and some mathematical analysis is carried out to give some quantitative explanation of experimental data.

Next, we test our model's ability to remove false targets on RDS and the natural scene images. RDS is famous for its abundance of false targets and absence of monocular cues. The simulations demonstrate that our model can solve the correspondence problem satisfactorily. Such simulations are necessary. Stereo vision algorithms in the computer vision field generally employ very discriminative likelihood functions. A good guess of the true disparity can be made by simply singling out the strongest peak of a likelihood function. In contrast, Figure 15 shows that our likelihood function is not ideal since it contains many spurious peaks with almost the same strengths as the true peak. It is possible that the Bayesian computation would fail because of these false peaks. Therefore, the simulation is the first validation that the disparity energy model can work well with MRF. This point could serve as the basis for our model's biological plausibility because removing false matching is a basic attribute of the biological stereo vision system. Furthermore, a comparison of our model and the coarse-to-fine model (Chen & Qian, 2004) reveals that our model can compute the absolute disparity of small objects with larger relative disparity. In summary, we

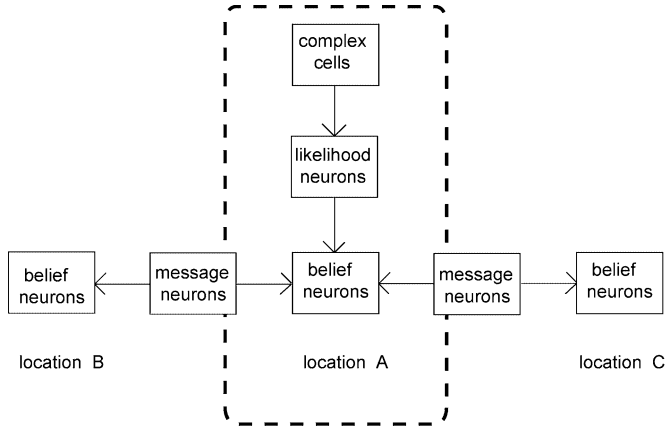

Figure 1: Diagram of the model proposed in this study. Arrows indicate neural connections. At each location, a group of likelihood neurons receives input from complex cells and sends output to belief neurons. Messages are sent between neighboring belief neurons. For example, belief neurons at location A communicate with neurons at nearby locations B and C through message neurons between them.

believe that stereopsis could be implemented by a neural network resembling MRF, and its biological implication deserves further investigation in related fields.

2 Model Descriptions

2.1 Overview. Our model has three layers. In the top layer, complex cells encode absolute disparity. At the bottom of each location, a group of neurons called belief neurons encode the visual system's confidence of disparities. In addition, message neurons are employed to implement BP. In the middle layer, likelihood neurons transform the response of complex cells into likelihood. The top layer is assumed to reside in V1, and the other two layers reside in the higher regions of visual cortex. A diagram of our model is shown in Figure 1.

At each location, belief neurons send messages to other nearby belief neurons through message neurons in a manner described by BP. The topology of the neural network at this layer is an important issue but is largely unknown. In this letter, we explore two kinds of first-order MRF, both shown in Figure 2, the line and the grid. In the first-order MRF, connections exist only between adjacent nodes. Although the neural system may employ more complicated circuits, experiments in section 3 show that this simplification can take us far. At the beginning of section 3, the strengths

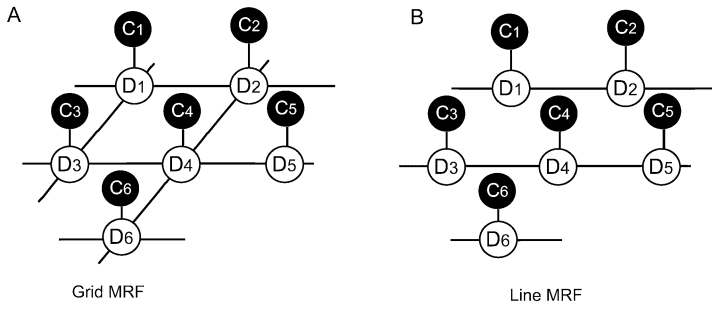

Figure 2: MRFs employed in our model. (A) The grid MRF. (B) The line MRF, which does not have z-axis connections. In both situations, node C_i represents the disparity likelihood at location i. In our model, the likelihood function was derived from the responses of complex cells of a single spatial frequency channel. Node D_i represents the visual system's confidence of disparities at location i.

and weaknesses of these topologies are discussed. However, the discussions in the rest of this letter apply to both topologies.

MRF provides a good abstraction of neural circuits in Figure 1. The objective of BP is to find $\{d_i^*, 1 \le i \le M\}$ that maximizes the posterior probability function 2.1, given the responses of complex cells,

$$P(D|C) \propto \prod_{i=1}^{M} \phi_i\left(d_i, c_i\right) \prod_{j \in N(i)} \psi_{ij}\left(d_i, d_j\right), \tag{2.1}$$

where d_i denotes disparity at position i, M is the number of unobserved nodes, c_i represents the responses of complex cells encoding the disparity at position i, and $N(i)$ represents all the nodes at neighboring position i.

The likelihood functions $\phi_i\left(d_i, c_i\right)$ are represented by likelihood neurons that receive the responses of complex cells. Section 2.2 discusses the likelihood function in detail. The potential function $\psi_{ij}\left(d_i, d_j\right)$, which is responsible for continuity preservation, is selected as

$$\forall i,\ j \in N(i),\ \psi_{ij}(d_i, d_j) = \max\left(\exp\left(-\frac{(d_i - d_j)^2}{\sigma_d}\right), \eta\right). \tag{2.2}$$

Parameters σ_d and η control the strength of interaction. When they are large, the strength of interaction is weak, and the belief neurons at two locations tend to compute disparities independently. Two of the curves of this function are plotted in Figure 3. For the dashed curve, $\sigma_d = 100$ and $\eta = 0.5$. For the solid curve $\sigma_d = 5$ and $\eta = 0.2$, the continuity constraint is more tightly imposed in this latter case.

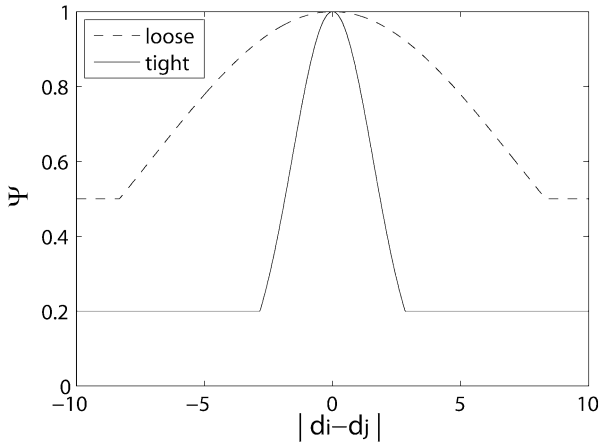

Figure 3: Plots of equation 2.2 under different sets of parameter settings. For the dashed curve ($\sigma_d = 100$ and $\eta = 0.5$), the constraint is relatively loose. For the solid curve ($\sigma_d = 5$ and $\eta = 0.2$), the constraint is much tighter.

In section 2.3, we demonstrate how this choice of potential function is related to the psychological findings. Here, we discuss the merits of representing the continuity constraint by MRF consisting of potential functions, such as equation 2.2. On the one hand, the continuity constraint has been proven to be a useful tool in eliminating false matches; on the other hand, it may smooth out real disparity boundaries if it is overemphasized. Therefore, the essence of applying this constraint is to seek a balance. We argue that MRF is suitable for this. Equation 2.2 penalizes any disparity discontinuities, thus favoring frontal-parallel planes. In this sense, it enforces the continuity constraint. However, the penalty is small if the disparity changes slightly, and it may not smooth out small disparity changes. Equation 2.2 has a lower bound η, preventing excessive penalty for large disparity changes in the scene. Due to this bound parameter, the penalty will not excessively increase when the disparity difference is beyond some threshold. With the proper choice of η, a sharp disparity change can be preserved if there is enough evidence to suggest its existence. In summary, our MRF-based model is able to accommodate true disparity discontinuities, although a penalty is assigned to every possible disparity discontinuity.

2.2 Converting Responses of Complex Cells to Likelihood. The response of a complex cell with position shift d and phase shift $\Delta\varphi$ can be expressed as

$$C(d, \Delta\varphi) = (L_1 + R_1)^2 + (L_2 + R_2)^2 \tag{2.3}$$

where

$$L_1 = \int_{-\infty}^{+\infty} I_l(a-\tau)\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\frac{\tau^2}{2\sigma_x^2}\right)\cos(\omega\tau+\varphi)d\tau \tag{2.4}$$

$$L_2 = \int_{-\infty}^{+\infty} I_l(a-\tau)\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\frac{\tau^2}{2\sigma_x^2}\right)\sin(\omega\tau+\varphi)d\tau \tag{2.5}$$

$$R_1 = \int_{-\infty}^{+\infty} I_r\,(b-\tau)\,\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\frac{\tau^2}{2\sigma_x^2}\right)\cos\left(\omega\tau+\varphi+\Delta\varphi\right)d\tau \tag{2.6}$$

$$R_2 = \int_{-\infty}^{+\infty} I_r(b-\tau)\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\frac{\tau^2}{2\sigma_x^2}\right)\sin(\omega\tau+\varphi+\Delta\varphi)d\tau, \tag{2.7}$$

where I_l and I_r are the left and right retinal images, respectively. For simplicity, we use the 1D Gabor function to model a neuron's receptive field (RF). The parameters a and b stand for the positions of the centers of RFs in the left and right retinal images, respectively. Moreover, $(a-b)$ equals the position shift d. The parameter φ is the phase of sinusoidal modulation, $\Delta\varphi$ the phase shift of the complex cell, σ_x the gaussian width, and ω the preferred spatial frequency.

We define $\tilde{L}$ and $\tilde{R}$ as the complex-valued monocular responses in equation 2.8, similar to that used by Fleet, Wagner, and Heeger (1996). Equations 2.3 and 2.8 lead to equation 2.9:

$$\tilde{L} = L_1 + L_2 i,\ \tilde{R} = R_1 + R_2 i \tag{2.8}$$

$$C(d,\Delta\varphi) = |\tilde{L}+\tilde{R}|^2. \tag{2.9}$$

Consider a complex cell with zero phase shift; L and R stand for its left and right complex-valued monocular responses, respectively:

$$C(d,0) = |L+R|^2. \tag{2.10}$$

Our mathematical basis for the likelihood is simple. If point a should match point b, L should equal R because the left and right RFs are identical with respect to the center positions, and so are the corresponding image patches covered by RFs. Because L and R are vectors, the differences in either magnitude or angle are supposed to decrease the plausibility of correspondence. Accordingly the likelihood ϕ is defined as

$$\phi = \max\left(\frac{|L+R|^2-|L-R|^2}{(|L|+|R|)^2},\varepsilon\right) = \max\left(\frac{4\,|L|\,|R|\cos\theta}{(|L|+|R|)^2},\varepsilon\right), \tag{2.11}$$

where θ is the angle between L and R and ε is the lower bound satisfying $0 < \varepsilon < 1$. As shown in the appendix,

$$|L - R|^2 = C(d, \pi) \tag{2.12}$$

$$(|L| + |R|)^2 = \max_{\Delta\varphi} C(d, \Delta\varphi). \tag{2.13}$$

Therefore, the likelihood ϕ of disparity d in equation 2.11 can be written as

$$\phi(d) = \max\left(\frac{C(d,0) - C(d,\pi)}{\max_{\Delta\varphi} C(d, \Delta\varphi)}, \varepsilon\right). \tag{2.14}$$

Equation 2.14, in fact, specifies how likelihood neurons transform the population responses of complex cells into likelihood functions. Note that we use the position shift to encode disparity and the phase shift to compute likelihood. Therefore, the detectable disparity range of our model corresponds to the range of position shift. In our simulation, the position shift ranges from -40 pixels to 40 pixels regardless of the neurons' preferred spatial frequency. In other words, our model is not bounded by size-disparity correlation (Prince & Eagle, 1999; Cumming & DeAngelis, 2001).

Note that the likelihood function can be directly obtained from the population response of the phase-shift model or position-shift model because curves of population responses generally have peaks near the ground truth disparity. In order to show that the likelihood function 2.14 is better than others constructed from population responses, we compare them in a simulation on RDS. The likelihood functions from population responses of the position-shift model (see equation 2.15) and the phase-shift model (see equation 2.16) are

$$\phi^{pos}(d) = \frac{C(d,0)}{\max_d C(d,0)} \tag{2.15}$$

$$\phi^{pha}(d) = \frac{C(0,\Delta\varphi)}{\max_{\Delta\varphi} C(0,\Delta\varphi)} = \frac{C(0,\omega d)}{\max_d C(0,\omega d)}. \tag{2.16}$$

Equation 2.16 makes use of the finding that the preferred disparity of a neuron with pure phase shift $\Delta\varphi$ is $\Delta\varphi \times \omega^{-1}$ (Qian, 1994). Note that the numerators in equations 2.15 and 2.16 are population responses and the denominators are normalization factors used to ensure that the maximal value of likelihood is 1. In this simulation, RFs are modeled as a 1D Gabor function, the gaussian width of which is fixed at 10 pixels. We constructed 1D RDS with disparities ranging from 0 to 10 pixels. For each disparity value, 500 RDS are generated. For each RDS, we compute $\phi(d_T)$, $\phi^{pos}(d_T)$, and $\phi^{pha}(d_T)$, where d_T stands for the true disparity of the RDS. The mean

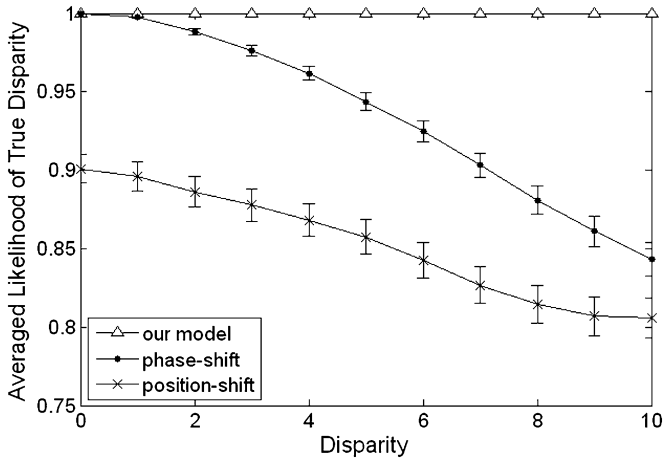

Figure 4: The averaged likelihood of the true disparity obtained from three likelihood functions. The lengths of error bars are twice the standard error of the mean. This figure show that the likelihood function used in our model is better than the two other likelihood functions computed directly from population responses, especially at large disparities. The figure also shows that the likelihood function from the population response of the phase-shift model is better than that of the position-shift model. The disparity of RDS varies from 0 to 10 pixels in steps of 1 pixel. For each disparity, results are computed from 500 RDS. Here the 1D Gabor RF is used, with a gaussian width of 10 pixels and a bandwidth of 1.14 octaves.

values and the standard error of the mean obtained from 500 simulations are then computed. The means are denoted by $\bar{\phi}(d_T)$, $\bar{\phi}^{pos}(d_T)$, and $\bar{\phi}^{pha}(d_T)$. We compare these entities because a good likelihood function should assign a high value to the ground truth disparity. The results are shown in Figure 4.

According to Chen and Qian (2004)'s simulation on RDS, the population response of the phase-shift model is more reliable than that of the position-shift model when the disparity is much smaller than the size of the RF; otherwise, both mechanisms become unreliable. Our simulation confirms their observations. For example, $\bar{\phi}^{pha}(d_T)$ is always larger than $\bar{\phi}^{pos}(d_T)$, and both curves drop as disparity increases in our simulations. In contrast, the likelihood function in our model always assigns the true disparity to the highest value, regardless of the disparity magnitude.

Equation 2.11 could account for some psychological findings. First, the factor $\frac{|L||R|}{(|L|+|R|)^2}$ ensures that the absolute contrast in two locations must be close in order to obtain a higher probability of correspondence. This property is shown in detail in Figure 5. Note that when $|L|$ varies in proportion to $|R|$, that is, $|L| = k|R|$, the factor $\frac{|L||R|}{(|L|+|R|)^2}$ and the likelihood ϕ are kept

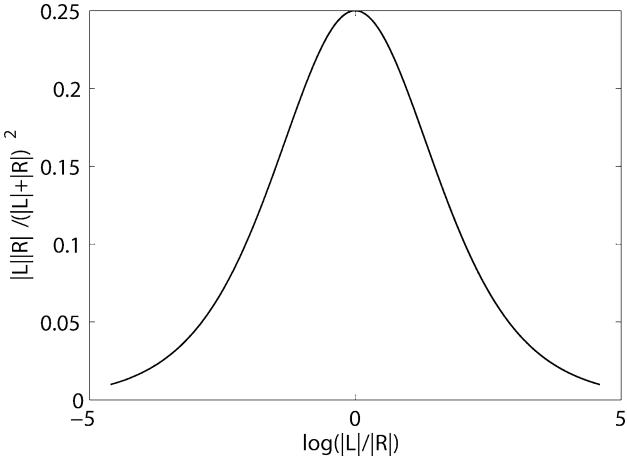

Figure 5: $\frac{|L||R|}{(|L|+|R|)^2}$ as a function of $\log(|L| \times |R|^{-1})$. The maxima is obtained when $|L| = |R|$.

unchanged as long as k does not change. This property could account for the psychological finding termed the contrast ratio constraint, which says that when the contrast of a feature in one eye increases, the contrast of the corresponding feature in the other eye must increase proportionally (Smallman & McKee, 1995).

Second, the likelihood is derived from neurons of a single frequency channel. This property enables our model to explain the depth perception of repetitive patterns without resorting to the second-order cue (detailed in section 3.1). Julesz (2006) designed some special RDS and demonstrated that stereo matching took place at different frequency channels in parallel. This fact can be accommodated in principle by our model because the disparity outputs from different frequency channels are not combined when they activate belief neurons, which in turn arouse depth perception. We thought that interactions of different scales take place at latter stages.

Tsang and Shi (2008) used a feature, derived from the population response (similar to ϕ in our model), to predict whether the true disparity was in the range of the preferred disparity of the phase-shift model. They found this feature worked better than other candidate features. However, they achieved good performance only when multiple features were combined. Similarly, we find that the likelihood function $\phi(d)$ in our case usually has more than one peak. In other words, false matches cannot be completely ruled out by simply considering $\phi(d)$. Therefore, message passing is essential in our model.

2.3 Modeling the Continuity Constraint. Many models of stereopsis rely on the continuity constraint to solve the correspondence problem (Marr

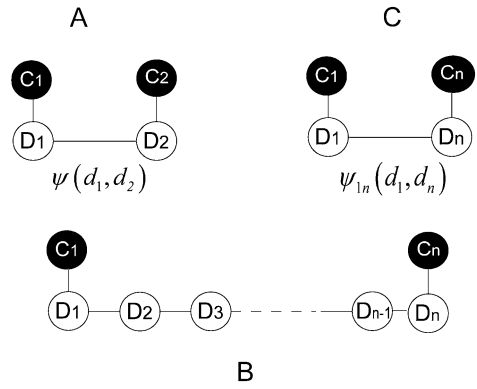

Figure 6: MRFs discussed in this section. (A) An MRF consisting of two adjacent nodes. (B) An MRF consisting of two remote nodes. (C) This MRF is equivalent to that in *B*, when equation 2.22 holds.

& Poggio, 1979; Prazdny, 1985). Although direct physiological mechanism has not been identified until now, its existence is supported by several psychological experiments. Petrov (2002) found that the strength of this constraint decreased when the separation between features increased. In our model, the interaction of adjacent nodes is reflected by the potential function in equation 2.2. We assume that the potential functions do not vary with position; therefore, we omit their subscripts and use ψ to stand for the potential functions between all adjacent nodes. First, we consider two adjacent nodes at location 1 and location 2 on a one-dimensional (line) MRF (see Figure 6A). BP will find the MAP configuration:

$$\begin{aligned}(d_1^*, d_2^*) &= \arg\max_{d_1,d_2} P(d_1, d_2|c_1, c_2) \\ &= \arg\max_{d_1,d_2} \phi_1(d_1, c_1)\phi_2(d_2, c_2)\psi(d_1, d_2). \end{aligned} \tag{2.17}$$

Now we consider two features at locations 1 and n, separated by a blank region, shown in Figure 6B. Similarly, their MAP disparities are determined as

$$\begin{aligned}(d_1^*, d_n^*) &= \arg\max_{d_1,d_n} \left(\max_{d_2 \ldots d_{n-1}} P(d_1, d_2 \ldots d_n|c_1, c_n) \right) \\ &= \arg\max_{d_1,d_n} \left(\phi_1(d_1, c_1)\, \phi_n(d_n, c_n) \max_{d_2 \ldots d_{n-1}} \prod_{i=1}^{n-1} \psi(d_i, d_{i+1}) \right). \end{aligned} \tag{2.18}$$

$\phi_2(d_2, c_2) \ldots \phi_{n-1}(d_{n-1}, c_{n-1})$ do not appear in equation 2.18 because we set $\phi_i(d_i, c_i) \equiv 1$, if the RFs cover only blank areas. Let $\psi_{1n}(d_1, d_n) =$

$\max_{d_2 \ldots d_{n-1}} \prod_{i=1}^{n-1} \psi(d_i, d_{i+1})$. We then have

$$(d_1^*, d_n^*) = \arg\max_{d_1, d_n} \phi_1(d_1, c_1)\phi_n(d_n, c_n)\psi_{1n}(d_1, d_n). \tag{2.19}$$

When equations 2.17 and 2.19 are compared, locations 1 and n can be considered as directly connected, and the strength of connection between two locations is reflected by the potential function $\psi_{1n}(d_1, d_n)$. The equivalent MRF is shown in Figure 6C. As we mentioned, the exact topology of the neural network enforcing the continuity constraint is still largely unknown, and belief neurons at location 1 may have a direct connection with belief neurons at location n rather than through the intermediate neurons as suggested in our model. However, this difference will not cause different depth percepts since the two features in Figure 6B have the same influence as those in Figure 6C. As shown in the appendix, the following inequality holds:

$$\forall d_1, d_n, n > 1, \quad \psi_{1n}(d_1, d_n) \geq \psi(d_1, d_n). \tag{2.20}$$

Equation 2.20 indicates that the penalty of discontinuity will not increase as the distance increases. In fact, the penalty will decrease under some mild conditions. For example, if we let $\psi(d_i, d_{i+1})$ take the form of equation 2.2 and additionally suppose that the disparity gradient $|d_i - d_{i+1}|$ and η are both small, we obtain

$$\begin{aligned} \forall i, \quad \psi(d_i, d_{i+1}) &= \max\left(\exp\left(-\frac{(d_i - d_{i+1})^2}{\sigma_d}\right), \eta\right) \\ &= \exp\left(-\frac{(d_i - d_{i+1})^2}{\sigma_d}\right). \end{aligned} \tag{2.21}$$

$$\begin{aligned} \psi_{1n}(d_1, d_n) &= \max_{d_2 \ldots d_{n-1}} \prod_{i=1}^{n-1} \psi(d_i, d_{i+1}) = \exp\left(-\frac{(d_1 - d_n)^2}{(n-1)\sigma_d}\right) \\ &= \psi(d_1, d_n)^{\frac{1}{n-1}}. \end{aligned} \tag{2.22}$$

Since the value of the potential function $\psi(d_1, d_n)$ is between 0 and 1, $\psi_{1n}(d_1, d_n)$ becomes flatter and flatter as n increases. In other words, the strength of the interaction between the two nodes decreases monotonically with an increase in distance between them.

Note that not every choice of ψ has the same property. Consider the following group of functions:

$$\left\{ \psi^\alpha(d_i, d_{i+1}) = \max\left(\exp\left(-\frac{|d_i - d_{i+1}|^\alpha}{\sigma_d}\right), \eta\right) \middle| \alpha = 1, 2, \ldots, n \right\}. \tag{2.23}$$

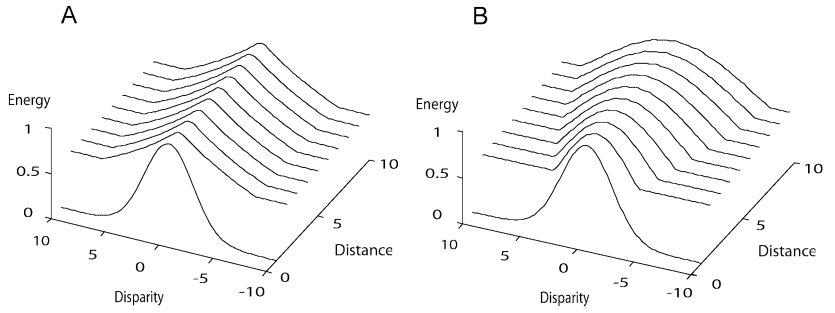

Figure 7: (A) The evolution of a message passing through 10 nodes connected by ψ^1. The initial message, a gaussian curve, is shown at distances 0. The message at distances 1 and 10 has the same shape. (B) The message in the case of ψ^2. The shape of the message becomes flatter and flatter as it passes through intermediate nodes. In other words, the power of the interaction decays as the distance increases.

Similar to the scenario where $\alpha = 2$ as in the previous discussion, we can obtain the following conclusions: if $\alpha > 2$, the strength will decay more rapidly, and if $\alpha = 1$, the strength of influence is a constant. Therefore, ψ^1 is not a biologically plausible choice, although ψ^1 and its variants are commonly used in computer stereo vision algorithms (Sun et al., 2003; Yang et al., 2009). With more psychophysical data available in the future, selection of the potential functions could be expected to be more tightly constrained. In the simulations in this letter, ψ^2 is used because we do not want the interactive strength to decay too rapidly.

We have discussed the strength of interactions in terms of potential functions. In our model, the maximum a posteriori (MAP) is obtained by passing messages. We can also analyze the problem from the perspective of passing messages. Here is a simulation. Node 1 influences node n by sending a message that is initially a gaussian. However, the message must go through intermediate $(n-1)$ nodes. Each time, the message is filtered by potential function in a way specified by update rule 2.29. We want to see how the message evolves with different ψ^α. Figure 7A shows that when ψ^1 is used, the message retains its strength. Figure 7B shows that when ψ^2 is used, the message decays. The parameters used in the simulation for both ψ^1 and ψ^2 are $\sigma_d = 20$ and $\eta = 0.5$. The gaussian width of the initial message is 10 pixels.

2.4 Biologically Plausible Message Passing. The max-product version of BP is as follows:

1. Initialize all messages.

$$m_{ji}(d_i) \equiv 1 \tag{2.24}$$

2. Update all the beliefs at each iteration.

$$b_i(d_i) \leftarrow k\phi_i(d_i) \prod_{j \in N(i)} m_{ji}(d_i), \tag{2.25}$$

where $\phi_i(d_i)$ in our model is shorthand for $\phi_i(d_i, c_i)$ and k is a normalization constant.

3. Update all messages at each iteration:

$$m_{ij}(d_j) \leftarrow k \max_{d_i} \psi_{ij}(d_i, d_j)\phi_i(d_i) \prod_{l \in N(i) \setminus j} m_{li}(d_i) \tag{2.26}$$

When all the iterations are complete, the MAP estimation $\{d_i^*, 1 \le i \le M\}$ can be obtained by finding each d_i^* that maximizes $b_i(d_i)$, provided there is only one disparity configuration assigned with the highest probability (Bishop, 2006),

$$\forall i, d_i^* = \arg\max_{d_i} b_i(d_i) \tag{2.27}$$

An alternative to this winner-take-all approach for disparity prediction is discussed in section 3.1.

Rao (2004, 2005) proposed performing Bayesian inferences in the log domain, inspired by the physiological evidence for the neural encoding of log probabilities (Carpenter & Williams, 1995). The log-space scheme is also used in our model. With the logarithm of update rules 2.25 and 2.26, the following equations can be obtained:

$$\log b_i(d_i) \leftarrow \log k + \log \phi_i(d_i) + \sum_{j \in N(i)} \log m_{ji}(d_i) \tag{2.28}$$

$$\log m_{ij}(d_j) \leftarrow \log k + \max_{d_i} \left(\log \psi_{ij}(d_i, d_j) + \log \phi_i(d_i) + \sum_{l \in N(i) \setminus j} \log m_{li}(d_i) \right). \tag{2.29}$$

Here two adaptations are made in our model. First, message neurons are introduced for representing messages in MRF although they are unnecessary for Rao's models with different topologies. Second, a specific kind of neural connection scheme is used for reducing the number of connections. Update rule 2.29 says that message neurons representing m_{ij} must receive input from all the message neurons sending information to location i except neurons representing m_{ji}. The number of input nodes in n-connected network is n, including $(n-1)$ message nodes and one likelihood node. The complexity of connections can be reduced if we substitute $\log b_i(d_i)$ for

$\log \phi_i(d_i) + \sum_{l \in N(i)} \log m_{li}(d_i)$. Accordingly, update rule 2.29 becomes

$$\log m_{ij}(d_j) \leftarrow \log k + \max_{d_i}(\log \psi_{ij}(d_i, d_j) + \log b_i(d_i) - \log m_{ji}(d_i)). \tag{2.30}$$

In this case, a message node receives input from only two nodes in n-connected network.

3 Simulations and Comparisons

Our model is tested on three types of stereo images. In section 3.1, we test our model using the line MRF on regularly spaced dots and sinusoidal gratings. These periodic stimuli are commonly used in psychophysical experiments. The depth perception of these stimuli is not likely to be explained by first-order models using low-spatial-frequency channels such as the coarse-to-fine model. Instead, the second-order mechanism was suggested previously (McKee et al., 2004). However, simulations show that our model could predict to a large degree the human depth percept for such stimuli.

In sections 3.2 and 3.3, our model is tested on RDS and natural scene images and compared with the coarse-to-fine model. The simulations show that our model is able to compute the absolute disparity of small objects with larger relative disparity.

We use the grid MRF on RDS and natural scene images because it allows signals to travel in two dimensions. Consequently, smoother disparity maps are obtained, although we find that ambiguities can be solved by line MRF to a large extent. However, the correctness of BP is guaranteed on the MRF, and not on grid MRF. In fact, BP fails to find the global maximum on grid MRF when tested with stimuli in section 3.1. This is one reason that line MRF is used in section 3.1. Another reason is that the stimuli in section 3.1 vary only horizontally, and messages traveling in vertical directions are duplicative. Therefore, the grid MRF will not have any advantage in this case.

3.1 Simulations on Periodical Stimuli. Mitchison and McKee (1987a, 1987b) used these kinds of stimuli to disclose the way our visual system solved ambiguity. The basic stereogram in their experiment includes rows of identical, regularly spaced dots. A row of dots is shown in Figure 8A. The left image and right images are identical except for the left-most dot in the left image and the right-most dot in the right image. These two dots are both displaced inward by s, a fraction of the interdot spacing denoted as L. Mitchison and McKee observed that when the duration of view was longer than 500 ms and interdot spacing was larger than 5 min of arc, a dot in the left eye must match a dot in the right eye (termed *discrete matching* in their work). Matching is ambiguous except for the dots at both ends. They also found that among all the possible matches for an internal dot in one eye, the

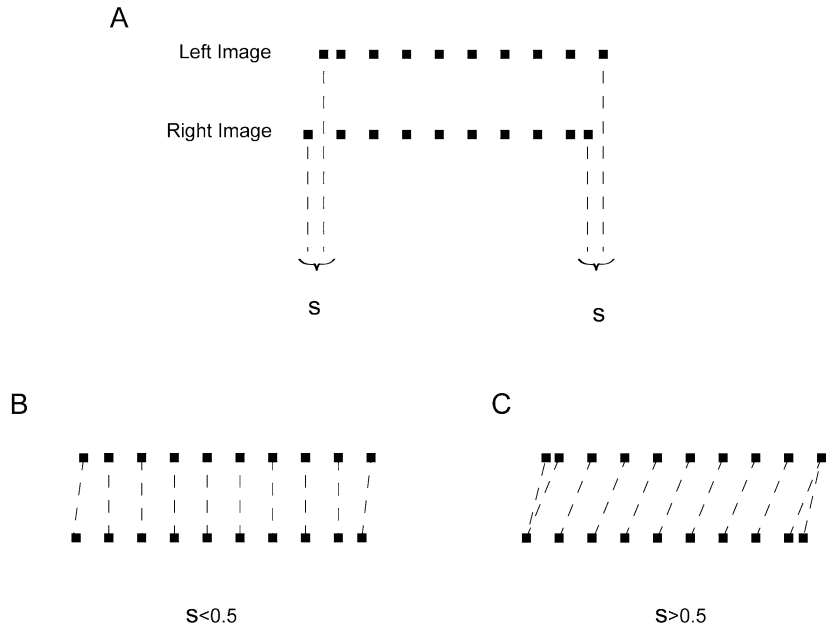

Figure 8: Ten-dot stimulus used in our simulation and two possible matching patterns. (A) The dots shown at the top are the left images, and those at the bottom are right images. The parameter s is the displacement of the left-most dot and the right-most dot, described as a fraction of interdot spacing. (B) The matching pattern when $s < 0.5$. The disparity of the internal dots is zero. (C) The matching pattern when $s > 0.5$. The disparity of the internal dots is the interdot spacing.

visual system chose the one whose disparity was the closest to the disparity of the ends. In other words, the matching of the internal dots was guided by the interpolated plane determined by dots at the ends. They called it the *nearest disparity rule*.

For the stimulus show in Figure 8A, the nearest disparity rule predicts that when $s < 0.5$, the disparity of the internal dots is zero (see Figure 8B); when $s > 0.5$, the internal dots assume a disparity of interdot spacing L (see Figure 8C). This rule is supported by their experimental results (see Figure 9). A long-duration curve (dashed line) shows how the depth percept in Figure 8C becomes increasingly dominant over the depth percept with zero disparity (Figure 8B). As stated by Mitchison and McKee (1987a), the intermediate points in Figure 9 come from the average of the two end percepts.

This finding interests us because it seems that our visual system minimizes the relative disparities across the whole scene. Next, we explore

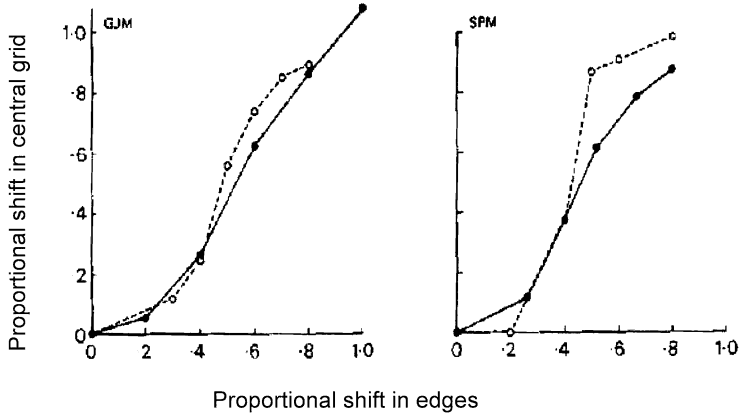

Figure 9: Experimental result of Mitchison and McKee (1987a). It shows how the perceived depth changes as a function of s (Figure 2 in their text). The dashed line is under the condition of long duration, when discrete matching happens. In their work, discrete matching means that a dot in one eye is paired with a dot in the other eye. The solid line is under the condition of short duration. The ordinate is the disparity shift as a fraction of interdot spacing. GJM and SPM are initials from Mitchison and McKee identifying two subjects in the experiment.

whether our model could produce similar results, and our results are quite encouraging. Here are our simulations.

Six pairs of 50×200 images are created with different values of s. Each image consists of a row of 10 black dots; Mitchison and McKee (1987a) used the same number of dots in a row for their experimental results in Figure 9. It is not necessary to use more than one row because when line MRF is applied, disparity computations for different rows are carried out independently and the disparity predictions are merely duplicates. Each dot is a 3×3 square. The interdot spacing L is 20 pixels. Therefore the disparity of the next-to-nearest matching is 20 pixels, and s increases from 0 to 1 in steps of 0.2.

The setting of our model is as follows. Line MRF is applied, and a final disparity estimation is obtained after passing the message 200 times. The BP is guaranteed to converge, and MAP is obtained because the line MRF does not have loops. RFs are modeled as 1D Gabor functions with $\sigma_x = 2$ pixels. The neurons' frequency bandwidth is also set to 1.14 octaves. The σ_d in equation 2.2 is set to 4. The position shift varies from -40 pixels to 40 pixels in steps of 1 pixel. The response of a likelihood neuron whose receptive field covers only the blank region is considered subthreshold. The lower bounds of equation 2.2 and equation 2.11 are $\eta = 0.01$ and $\varepsilon = 0.001$, respectively.

Table 1: Disparities (pixels) of the 10 Dots in the Left Image Computed by Our Model for Each Image Pair.

s	0	0.2	0.4	0.6	0.8	1
Dot 1	0	4	8	12	16	20
Dot 2–9	0	0	0	20	20	20
Dot 10	0	4	8	20	20	20

Notes: Dots are numbered from left to right. The interdot spacing is 20 pixels.

Our model's results in Table 1 are consistent with the nearest disparity rule proposed by Mitchison and McKee since the disparities of internal dots (dot 2– dot 9) are controlled by s in the same way. Unlike the matching pattern in Figure 8C, our model cannot link dot 10 simultaneously to the last two dots in the right image because in our model, a point in the left image has only one disparity value.

The deterministic prediction of our model is made by the winner-take-all approach shown in equation 2.27. In other words, our model assumes that the true disparity is encoded by the belief neuron with the maximal figuring rate. However, Mitchison and McKee's (1987a) results show that both depth percepts can occur when s is between 0 and 1, subject to different probabilities. The long-duration curve in Figure 9 can be considered as the probability of seeing the depth percept in Figure 8C because the intermediate points in Figure 9 give some averaged depth of the bistable percepts. The winner-take-all approach cannot fully account for the probabilistic nature of depth perception, and neither can the nearest disparity rule. In order to show that the probability of depth percepts is related to the population response of belief neurons, the winner-take-all approach needs to be relaxed, and belief neurons' responses should be interpreted in some probabilistic way. In this section, we assume that only the belief neurons that give rise to local maxima of the population response have a chance of arousing depth perception, and the probabilities of these depth percepts are computed from the softmax function whose arguments are the maxima of population responses.[1] As we can see in Figure 10, only disparities of 0 and L give rise to peaks of the population response curves. Therefore, only matching patterns in Figures 8B and 8C have nonzero probability in our model. According to our assumption, $p(0|s)$ and $p(L|s)$, which are the probabilities of the depth percepts in Figures 8B and 8C, should be computed

[1]Softmax function f is defined as $f(p_1, p_2, \ldots, p_n) = (q_1, q_2, \ldots, q_n)$ where

$$q_i = \frac{\exp(p_i)}{\sum_{j=1}^{n} \exp(p_j)}.$$

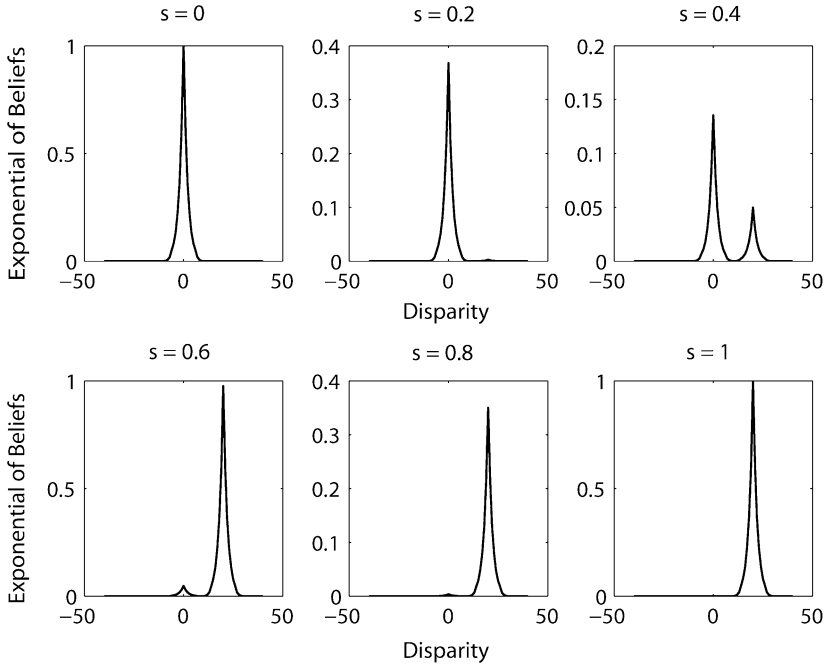

Figure 10: Exponentials of responses of belief neurons encoding the disparity of dot 5 at the center of the stimuli in Figure 8A. The population response curves possess only two local maxima at the disparities of 0 pixels and 20 pixels.

by the softmax function as

$$p(0|s) = \frac{\exp(bn(0|s))}{\exp(bn(0|s)) + \exp(bn(L|s))} = \text{sigmoid}(bn(0|s) - bn(L|s)) \tag{3.1}$$

$$p(L|s) = \frac{\exp(bn(L|s))}{\exp(bn(0|s)) + \exp(bn(L|s))} = \text{sigmoid}(bn(L|s) - bn(0|s)), \tag{3.2}$$

where $bn(D|s)$ denotes the response of the belief neuron encoding disparity D given s. The softmax function becomes the sigmoid function in the case of two percepts. Note that $p(0|s)$ and $p(L|s)$ are not the marginal probabilities of $P(D|C)$ in equation 2.1. The exponentials are necessary because belief neurons encode the logarithm of beliefs in equation 2.25. After the message

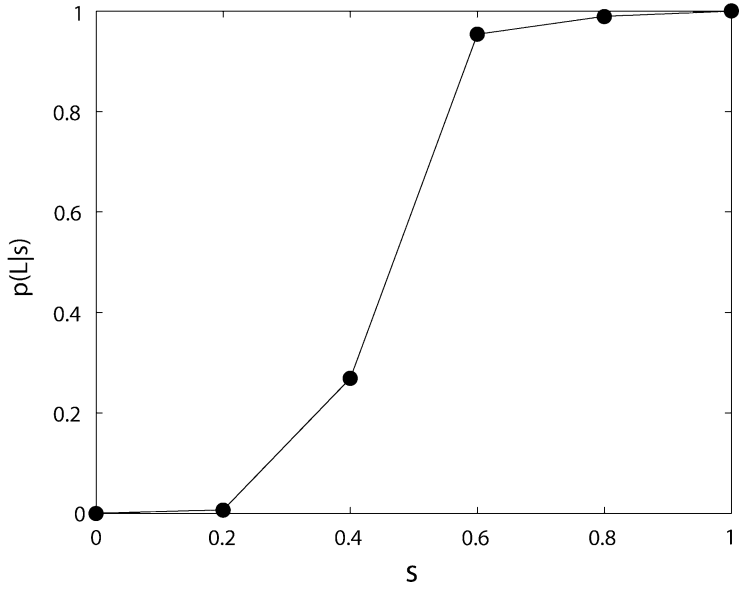

Figure 11: Probability of attaining the depth percept in Figure 8C, predicted by our model.

is passed 200 times, $p(L|s)$ is computed from the response of belief neurons encoding the disparity of dot 5 by equation 3.2. Although the plot of $p(L|s)$ shown in Figure 11 is not in exact alignment with the long-duration curves in Figure 9, the general trends are fairly consistent. Therefore, we thought the probabilistic aspect of the depth perception could be reflected in population response of belief neurons.

According to Mitchison and McKee (1987a), two depth percepts in Figures 8B and 8C have almost equal probability of occurring when $s = 0.5$. In the following, we give some mathematical analysis to account for this finding. According to equations 3.1 and 3.2,

$$p(L|s) = p(0|s) \Leftrightarrow \frac{\exp(bn(L|s))}{\exp(bn(0|s))} = 1. \tag{3.3}$$

The posterior probability function in equation 2.1 takes the disparity at each pixel as a variable, and it is difficult to give an analytical analysis for such a function of so many variables. As shown in section 2.3, this function can be approximated by a function depending only on the disparities of the dots in Figure 8. An approximation of potential functions can accordingly be obtained only by considering the disparities of adjacent dots as

$$\frac{\exp(bn(L|s))}{\exp(bn(0|s))} = \frac{\prod_{i=1}^{10} \phi_i(d_i, c_i) \prod_{i=1}^{9} \psi_{i,i+1}(d_i, d_{i+1})}{\prod_{i=1}^{10} \phi_i(d'_i, c_i) \prod_{i=1}^{9} \psi_{i,i+1}(d'_i, d'_{i+1})}, \tag{3.4}$$

where $d_1 = d_{10} = Ls,\ d_2 = d_3 \ldots = d_9 = L,\quad d'_1 = d'_{10} = Ls, d'_2 = d'_3 \ldots = d'_9 = 0$, and $\psi_{i,i+1}$ is the potential function of adjacent dots. Because every dot must match another dot, all the ϕ_is cancel out. Then equation 3.4 becomes

$$\frac{\exp(bn(L|s))}{\exp(bn(0|s))} = \frac{\psi_{1,2}(Ls, L) \times \psi_{2,3}(L, L) \times \ldots \psi_{8,9}(L, L) \times \psi_{9,10}(L, Ls)}{\psi_{1,2}(Ls, 0) \times \psi_{2,3}(0, 0) \times \ldots \psi_{8,9}(0, 0) \times \psi_{9,10}(0, Ls)}. \tag{3.5}$$

By assuming $\psi_{i,j}(d_i, d_j) = g(|d_i - d_j|)$ where $g(x)$ is a function that decreases monotonically when x increases and satisfies $g(0) = 1$ (the whole group of functions specified by equation 2.23 can be written in this form), equation 3.5 becomes

$$\frac{\exp(bn(L|s))}{\exp(bn(0|s))} = \frac{g(|Ls - L|)^2}{g(|Ls|)^2}. \tag{3.6}$$

Since $\frac{g(|Ls-L|)^2}{g(|Ls|)^2} = 1 \Rightarrow |Ls - L| = |Ls| \Rightarrow s = 0.5$, the depth percept will change around this point.

We further tested our model on sinusoidal gratings, another popular periodical stimulus. McKee et al. (2004) observed that the disparity of gratings with moderate or high frequencies was determined by the disparity of edges. However, low-frequency gratings were usually seen in the fixation plane regardless of the edge disparity.

In the simulation, we first made sinusoidal gratings with a frequency of 0.1 cycle per pixel. The width of the grating is 180 pixels. There are 18 cycles in the window, the same as in the gratings of moderate spatial frequency in McKee et al.'s (2004) experiment. Five image pairs with different edge disparities are used. The largest edge disparity is one period of grating. The complex cells in our model are tuned to the frequency of the gratings. Other parameters of our model are the same as in Figure 11. For each image pair, our model's disparity estimation is uniform throughout the scene. The simulation result in Figure 12 shows that the predicted disparity of the whole grating is always the disparity of edges.

Second, we apply our model to gratings of higher spatial frequency (0.2 cycle/pixel) and lower frequency (0.033 cycle/pixel). The maximal disparities are 5 and 30 pixels for gratings of higher spatial frequency and lower frequency, respectively. The numbers of cycles in the window are the same as those in McKee et al.'s (2004) experiment. In both cases, the disparity of edges determines the disparity of the central region (see Figure 12). Thus, the predictions of our model agree with human depth percepts on moderate- and high-frequency gratings but disagree on low-frequency gratings.

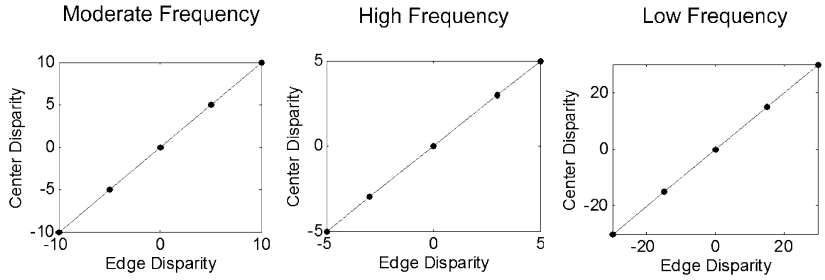

Figure 12: The disparity of the central region estimated by our model as a function of edge disparity. From left to right, the spatial frequencies of the sinusoidal gratings are 0.1 cycle per pixel, 0.2 cycle per pixel, and 0.033 cycle per pixel. The maximal disparity is one period of grating. The disparity of the central region always equals the disparity of the edge regardless of the spatial frequency. The complex cells in our model are always tuned to the frequency of the gratings.

The reason that our model can predict the depth percept aroused by sinusoidal gratings of moderate and high frequency is that messages from both edges disambiguate the responses of belief neurons at the central region. The difference between our model's prediction and McKee et al.'s (2004) result at low spatial frequencies implies that the continuity constraint is not the unique constraint imposed in stereopsis. Our visual system also has a tendency to match a feature with its nearest neighbor in the other eye, so that the absolute disparity is minimized (Mallot & Bideau, 1990). This constraint would penalize any nonzero disparities and could somehow be more influential in matching low-frequency gratings. One reason for this could be that the edge disparity of low-frequency gratings in McKee et al.'s experiment is larger than those of moderate- and high-frequency gratings. Hence, the edge disparity was penalized more severely.

3.2 Simulations on RDS. Seven pairs of 128×128 pixels random dot stereograms are created with a dot density of 50% and a dot size of 1 pixel. The central region has 30×30 pixels, and the disparity of the background is 0 pixels. The disparity of the central region is from 4 to 16 pixels in steps of 2 pixels.

The parameter setting of the coarse-to-fine model is as follows. The RFs of complex cells are represented by a 1D Gabor filter because for the stimuli used in this section, the 1D Gabor filter usually leads to faster and more reliable disparity computations compared to the 2D Gabor filter. The horizontal gaussian width σ_x of seven scales follows a geometric series with a ratio of $\sqrt{2}$. The σ_x of the largest scale is 32 pixels, which is larger than the largest disparity in all image pairs. In our simulations, the output is not

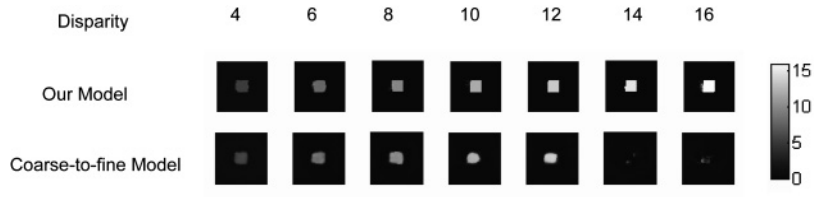

Figure 13: The disparity maps computed from seven pairs of RDS with increasing disparity of the central region. The first row shows the outputs of our model, which are not affected by the disparity increase. However, the coarse-to-fine model fails to detect the central square when the disparity is larger than 12 pixels.

sensitive to this parameter. The σ_x of the smallest scale is 2 pixels. At all scales, the neurons' frequency bandwidth is set to 1.14 octaves by fixing the product $\omega\sigma_x = \pi$. Spatial pooling is applied at each scale using gaussian filters with the same standard deviation σ_x.

The parameter setting of our model is as follows. The RFs of complex cells are also represented by 1D Gabor filters. The $\sigma_x = 2$ pixels are the same as the smallest scale in the coarse-to-fine model. The neurons' frequency bandwidth is also set to 1.14 octaves and $\sigma_d = 4$. The position shifts range from -40 pixels to 40 pixels in steps of 1 pixel. The lower bounds $\varepsilon = 0.001$ and $\eta = 0.01$. The final disparity estimation is obtained after the message is passed 150 times. The disparity estimation converged in all simulations.

The results are shown in Figure 13. Our model shows valid disparity estimation of the central region for all image pairs. The coarse-to-fine model, on the contrary, fails to detect the disparity of the central region when it is larger than 12 pixels. This could be explained by when the disparity of the central region is large, neurons with large receptive fields are required by the coarse-to-fine model. The 30×30 central region becomes quite small compared to the area covered by the receptive field. In fact, the receptive field would cover much more background than the central region. Since the complex cell computes averaged disparity in its receptive field as its disparity estimation, the estimation of a coarse scale is very close to the disparity of the background. When the relative disparity between the central region and the background is large enough, the subsequently finer scales will fail due to their limited disparity detection range. However, our model does not have this problem since it does not use neurons with large RF.

The above explanation implies that when the size of the RDS's central region is large enough, the large disparities could be recovered by the coarse-to-fine model. To test this, seven pairs of 128×128 RDS whose central regions have 64×64 pixels are used for the coarse-to-fine model. The largest disparity is still 16 pixels. All the parameters are unchanged. The

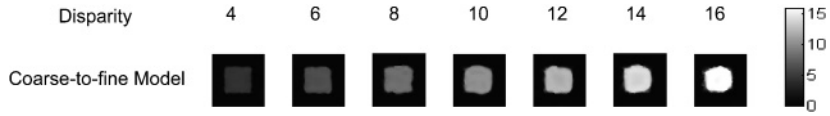

Figure 14: The disparity maps computed from seven pairs of RDS with a larger central region. The coarse-to-fine model correctly computes the disparity of all central regions.

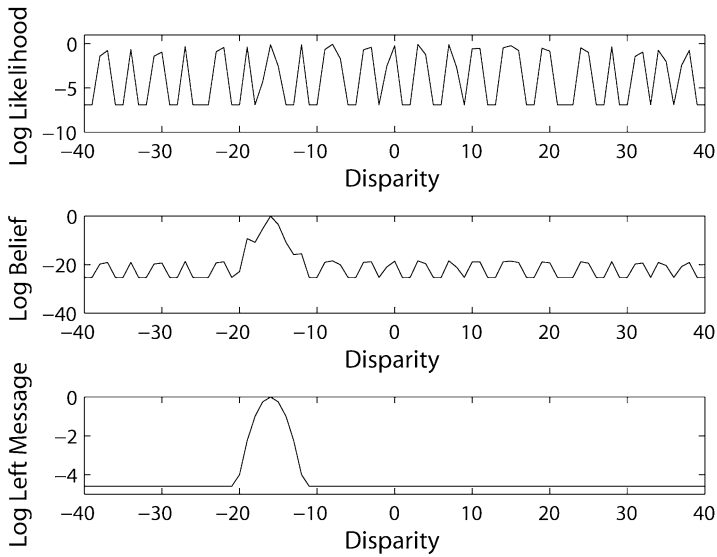

Figure 15: Simulated activity of neurons encoding the disparity of the central region of the seventh image pair. The first row shows the likelihood derived from the response of complex cells. Many spurious peaks appeared. The second row shows the response of belief neurons. The last row shows messages coming from the left side of the location. We can see that the strength of every false match diminishes by pooling messages from every direction.

result in Figure 14 shows that the coarse-to-fine model indeed succeeds in this case.

A key ingredient in our model is that false matches can be rejected by integrating disparity information at different locations. Figure 15 shows that there are many false peaks in the responses of likelihood neurons because the disparity detection range (80 pixels) is much larger than a period of neurons' preferred frequency channel (4 pixels). However, messages that carry the integrated disparity information are indeed free of false peaks.

Our model consumes much more time than the coarse-to-fine model. While the coarse-to-fine algorithm requires only about 4 seconds for 1 RDS, our simulation program in Matlab on a Core 2 1.86 GHz PC requires

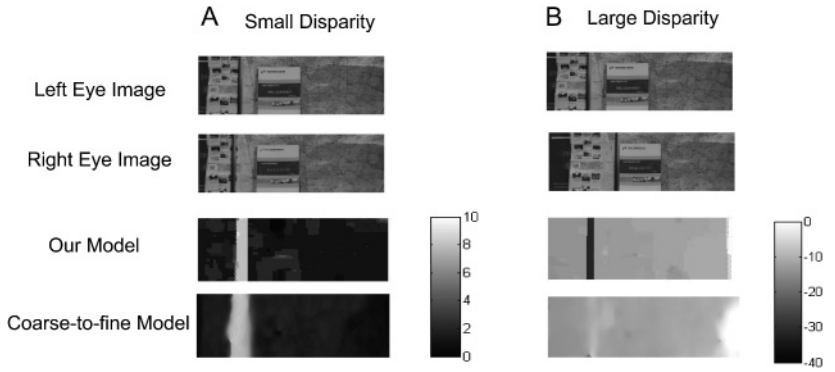

Figure 16: Computed disparity maps using the two models. The pencil is the focus of the simulation. (A) A stereo images pair with small disparity and the estimated disparity maps. Note that both the models successfully detected the disparity of the pencil. (B) When the disparity of the pencil is much larger than the background, the coarse-to-fine model did not detect it, but our model did.

about 40 seconds for a single iteration of BP. However, as the disparity computation in our model at each location can be done by neurons in parallel, the run-time difference of the two models might not be so significant for the neural system.

3.3 Simulations on Natural Scene Images. Our model's larger detection range for relative disparity discussed in section 3.2 can also be demonstrated with natural scene images. Two pairs of stereo images and the output of two models are shown in Figure 16. In Figure 16B, the disparity of the pencil is about -32 pixels, and that of the background is about -12 pixels. The coarse-to-fine model fails to detect the disparity of the pencil. In contrast, the pencil is prominent in the disparity map computed by our model. In Figure 16A, the disparity of the pencil is about 8 pixels, and the background disparity is about 2 pixels. Both models successfully detect the pencil in this case.

The parameter setting of the coarse-to-fine model is as follows. The RFs are modeled as 2D Gabor functions. The σ_x of the largest scale is 32 pixels for the small disparity image pair (see Figure 16A), and the σ_x of the largest scale is 64 pixels for the large disparity image pair (see Figure 16B). Spatial pooling is applied using 2D gaussian filters with standard deviations equal to σ_x in each scale. Orientation pooling is also applied over five orientations ranging from 30 to 150 degrees in steps of 30 degrees. Other parameters are the same as in section 3.2.

The parameter setting of our model is as follows: $\sigma_d = 2$, $\sigma_x = 2$. The total number of iterations is 300. The disparity map computation has converged. Other parameters are the same as in section 3.2.

It should be noted that the performance of both the coarse-to-fine model and our model must be to some extent affected by the parameter setting. However, since the parameters of both our model and those of the coarse-to-fine model are kept unchanged when disparity changes from a small one to a large one, we thought the comparative results reported in this work are not exclusively due to the parameter settings, and they would, to some extent, reflect some inherent characteristics of the two models.

4 Physiological Relevance

At the top layer of our model, disparity is encoded by complex cells that are widely accepted as disparity encoders (Hubel & Wiesel, 1962; Bishop & Pettigrew, 1986; Ohzawa et al., 1990; Prince, Cumming, & Parker, 2002). The remaining part of our model is largely based on conjectures. We make three key conjectures:

1. For every location of the image, there are neurons called belief neurons whose responses arouse depth perception in a probabilistic way.
2. There exist recurrent connections among belief neurons, and these connections as a whole can be modeled as an MRF.
3. The MAP of the MRF is computed by BP. Neural dynamics between belief neurons can be interpreted as messages of BP.

Currently, there is no direct physiological evidence for or against the existence of MRF, and a complex neural network like this is hard to identify and locate. However, we propose that our model is closely related to these physiological records focusing on the disparity selectivity of a neuron and the contextual effect.

First, the belief and message neurons in our model are selective for absolute disparity. In both ventral and dorsal visual pathways, physiologists have found plenty of neurons that are selective for absolute disparity (Parker, 2007). Since there are so many neurons in V1 representing absolute disparities, it seems illogical that such neurons in higher areas merely represent absolute disparities. They are likely to accomplish some more complicated tasks, such as those demonstrated in our model.

Second, a belief neuron is most active if the disparities in its RF and a larger neighborhood equal its preferred disparity. Then, it can be considered to have a facilitative extra receptive field. Bakin, Nakayama, and Gilbert (2000) have found neurons in V2 whose responses are determined by stimulus disparity in both the RF and the context. It is interesting to note that 62% of neurons in Bakin et al.'s experiment showed the same selectivity for disparities in the RF and the context. These neurons have the same type of ERF as the belief neurons in our model.

Finally, a large part of the connections in our model is between adjacent neurons with similar disparity selectivity. To save space in the brain, these neurons are better off grouped together, like neurons in V1 with the same

orientation selectivity. DeAngelis and Newsome (1999) did find such organization in MT. In addition, in orientation columns, clusters of neurons with similar ERF property were found (Yao & Li, 2002). This architecture is also constructive for implementing our model because all the hypothesized neurons have the same kind of ERF.

5 Discussion

Julesz (2006) made a distinction between local and global stereopsis. *Local stereopsis* is used for the mechanism that compares local image patches in two eyes, and *global stereopsis*, which is responsible for the removal of false targets, is considered qualitatively different from local stereopsis. By now, the knowledge of local stereopsis has been tremendously advanced by the disparity energy model, but the neural mechanism for the global stereopsis remains largely a mystery. This letter explores the possibility that neural circuits for global stereopsis could be modeled by a neural network simulating MRF. Our proposed model has several merits. First, our model relies only on message passing to remove false matches. In section 3.1, false matches introduced by repetitive patterns are removed without resorting to second-order disparity detectors. In simulation on RDS and natural scene images, the false matches caused by the bandpass RF of complex cells are removed without resorting to complex cells tuned to low spatial frequency. Second, sections 3.2 and 3.3 show that our model can have a larger detection range of relative disparity for small objects, compared to the coarse-to-fine model. Third, we show in section 2.3 that the potential functions can be chosen to match the signatures of continuity constraint discovered in Petrov's experiment (Petrov, 2002). Besides these attributes, our model does not need any neurons whose disparity selectivity is totally different from the neurons recorded in physiological experiments. In fact, we thought the belief neurons could appear as neurons with facilitative ERF and are selective for the absolute disparity.

Our model shares some features with several existing models of the global stereopsis. To our knowledge, the continuity constraint was first used in Marr and Poggio (1979)'s cooperative algorithm. This algorithm is not considered to be physiologically plausible because neurons' RFs are much larger than a dot, which is the basic unit for matching in the cooperative algorithm (Qian & Zhu, 1997). Unlike Marr and Poggio's algorithm, our model uses complex cells as the basic matching units. The RFs of complex cells used in our simulations cover many dots. Read (2002a, 2002b) built a Bayesian model of stereopsis. A Bayesian prior that embodied a preference for small absolute disparity was integrated in her model. In contrast, our model prefers small relative disparities. A more sophisticated Bayesian model was proposed by Tsang and Shi (2008).

In models such as Fleet et al.'s (1996) model and the coarse-to-fine model, neither coarse scale nor fine scale is dispensable. In contrast, our model

computes disparity from neural responses at a single scale, demonstrating that disparity information at different scales is quite redundant. Although our results question the necessity of coarse scales in removing false targets, the possibility of a coarse-to-fine scheme is not excluded. The advantage of coarse scales could be their promptness (Menz & Freeman, 2003). In addition to the physiological study cited, there are also psychophysical studies on the coarse-to-fine process in stereo vision and vergence eye movement.

Julesz (2006) suggested that the fusion of sparse lines and dots was fundamentally different from the fusion of complex image pairs such as RDS. This is indeed the case with our model. If the responses of likelihood neurons are not ambiguous, message passing does not need to be activated, and the belief neurons could simply copy signals from the likelihood neurons.

As mentioned in section 3.2, our model is very time-consuming. Therefore, we intend to find biologically plausible ways to speed up message passing. Furthermore, a large part of our model is based on the assumptions listed in section 4. We hope that with the advances in neurobiology, our model could find more concrete support from related findings in the field in the future.

Appendix: Derivation of Equations 2.12, 2.13, and 2.20

Substituting equations 2.4 and 2.5 into equation 2.8, we have

$$\begin{aligned}\tilde{L} &= \int_{-\infty}^{+\infty} I_l(a-\tau)\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\frac{\tau^2}{2\sigma_x^2}\right)(\cos(\omega\tau+\varphi)+i\sin(\omega\tau+\varphi))d\tau \\ &= \int_{-\infty}^{+\infty} I_l(a-\tau)\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\frac{\tau^2}{2\sigma_x^2}\right)\exp(i(\omega\tau+\varphi))d\tau. \qquad \text{(A.1)}\end{aligned}$$

Similarly,

$$\begin{aligned}\tilde{R} &= \int_{-\infty}^{+\infty} I_r(a+d-\tau)\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\frac{\tau^2}{2\sigma_x^2}\right)\exp(i(\omega\tau+\varphi+\Delta\varphi))d\tau \\ &= \exp(i\Delta\varphi)\int_{-\infty}^{+\infty} I_r(a+d-\tau)\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\frac{\tau^2}{2\sigma_x^2}\right)\exp(i(\omega\tau+\varphi))d\tau. \qquad \text{(A.2)}\end{aligned}$$

L and R in equation 2.10 are the complex valued monocular responses of a neuron with zero phase shifts. By setting $\Delta\varphi = 0$ in equation A.2, we

have

$$\tilde{L} = L \tag{A.3}$$

$$\tilde{R} = R \cdot \exp(i\Delta\varphi). \tag{A.4}$$

Substituting these two equations into equation 2.9, we have

$$C(d, \Delta\varphi) = |L + R \cdot \exp(i\Delta\varphi)|^2, \tag{A.5}$$

and equation A.6 is just equation 2.12:

$$C(d, \pi) = |L + R \cdot \exp(i\pi)|^2 = |L - R|^2 \tag{A.6}$$

Let $\Delta\varphi'$ denote angle of $L \cdot \bar{R}$, where $\bar{R}$ is the complex conjugate of R. If we let $\Delta\varphi = \Delta\varphi'$, vectors $\tilde{L}$ and $\tilde{R}$ will have the same direction. Consequently, $C(d, \Delta\varphi)$ shown in equation A.5 will reach its maximum. Therefore, we obtain equation 2.13:

$$\max_{\Delta\varphi} C(d, \Delta\varphi) = C(d, \Delta\varphi') = (|L| + |R|)^2. \tag{A.7}$$

The derivation of equation 2.20 is

$$\begin{aligned}
\psi_{1n}(d_1, d_n) &= \max_{d_2 \ldots d_{n-1}} \prod_{i=1}^{n-1} \psi(d_i, d_{i+1}) \\
&\geq \prod_{i=1}^{n-1} \psi(d_i, d_{i+1}) \Big| d_2, d_3 \ldots d_{n-1} = d_1 \\
&= \psi(d_1, d_n).
\end{aligned} \tag{A.8}$$

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant 60820012. We thank Lianqing Yu for his contributions to the simulation programs. We are also grateful to Eric K. C. Tsang for his inspiring comments.

References

Bakin, J. S., Nakayama, K., & Gilbert, C. D. (2000). Visual responses in monkey areas V1 and V2 to three-dimensional surface configurations. *Journal of Neuroscience*, *20*(21), 8188–8198.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.
Bishop, P. O., & Pettigrew, J. D. (1986). Neural mechanisms of binocular vision. *Vision Research*, *26*(9), 1587–1600.
Carpenter, R. H. S., & Williams, M. L. L. (1995). Neural computation of log likelihood in control of saccadic eye-movements. *Nature*, *377*(6544), 59–62.
Chen, Y. H., & Qian, N. (2004). A coarse-to-fine disparity energy model with both phase-shift and position-shift receptive field mechanisms. *Neural Computation*, *16*(8), 1545–1577.
Cumming, B. G., & DeAngelis, G. C. (2001). The physiology of stereopsis. *Annual Review of Neuroscience*, *24*, 203–238.
Cumming, B. G., & Parker, A. J. (1997). Responses of primary visual cortical neurons to binocular disparity without depth perception. *Nature*, *389*(6648), 280–283.
Cumming, B. G., & Parker, A. J. (2000). Local disparity not perceived depth is signaled by binocular neurons in cortical area V1 of the macaque. *J. Neurosci.*, *20*(12), 4758–4767.
DeAngelis, G. C., & Newsome, W. T. (1999). Organization of disparity-selective neurons in macaque area MT. *Journal of Neuroscience*, *19*(4), 1398–1415.
Fleet, D. J., Wagner, H., & Heeger, D. J. (1996). Neural encoding of binocular disparity: Energy models, position shifts and phase shifts. *Vision Research*, *36*(12), 1839–1857.
Goutcher, R., & Mamassian, P. (2005). Selective biasing of stereo correspondence in an ambiguous stereogram. *Vision Research*, *45*(4), 469–483.
Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in cat's visual cortex. *Journal of Physiology-London*, *160*(1), 106–154.
Julesz, B. (2006). *Foundations of cyclopean perception*. Cambridge, MA: MIT Press.
Mallot, H. A., & Bideau, H. (1990). Binocular vergence influences the assignment of stereo correspondences. *Vision Research*, *30*(10), 1521–1523.
Marr, D., & Poggio, T. (1979). Computational theory of human stereo vision. *Proceedings of the Royal Society of London Series B—Biological Sciences*, *204*(1156), 301–328.
McKee, S. P., Verghese, P., & Farell, B. (2004). What is the depth of a sinusoidal grating? *Journal of Vision*, *4*(7), 524–538.
Menz, M. D., & Freeman, R. D. (2003). Stereoscopic depth processing in the visual cortex: A coarse-to-fine mechanism. *Nature Neuroscience*, *6*(1), 59–65.
Mitchison, G. J., & McKee, S. P. (1987a). The resolution of ambiguous stereoscopic matches by interpolation. *Vision Research*, *27*(2), 285–294.
Mitchison, G. J., & McKee, S. P. (1987b). Interpolation and the detection of fine-structure in stereoscopic matching. *Vision Research*, *27*(2), 295–302.
Ohzawa, I., DeAngelis, G. C., & Freeman, R. D. (1990). Stereoscopic depth discrimination in the visual-cortex—neurons ideally suited as disparity detectors. *Science*, *249*(4972), 1037–1041.
Ott, T., & Stoop, R. (2006). The neurodynamics of belief propagation on binary Markov random fields. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems*, *19* (pp. 1057–1064). Cambridge, MA: MIT Press.
Parker, A. J. (2007). Binocular depth perception and the cerebral cortex. *Nature Reviews Neuroscience*, *8*(5), 379–391.

Petrov, Y. (2002). Disparity capture by flanking stimuli: A measure for the cooperative mechanism of stereopsis. *Vision Research*, *42*(7), 809–813.

Prazdny, K. (1985). Detection of binocular disparities. *Biological Cybernetics*, *52*(2), 93–99.

Prince, S. J. D., & Eagle, R. A. (1999). Size-disparity correlation in human binocular depth perception. *Proceedings of the Royal Society of London Series B—Biological Sciences*, *266*(1426), 1361–1365.

Prince, S. J. D., Cumming, B. G., & Parker, A. J. (2002). Range and mechanism of encoding of horizontal disparity in macaque V1. *Journal of Neurophysiology*, *87*(1), 209–221.

Qian, N. (1994). Computing stereo disparity and motion with known binocular cell properties. *Neural Computation*, *6*(3), 390–404.

Qian, N., & Zhu, Y. D. (1997). Physiological computation of binocular disparity. *Vision Research*, *37*(13), 1811–1827.

Rao, R. P. N. (2004). Bayesian computation in recurrent neural circuits. *Neural Computation*, *16*(1), 1–38.

Rao, R. P. N. (2005). Bayesian inference and attentional modulation in the visual cortex. *Neuroreport*, *16*(16), 1843–1848.

Read, J. C. A. (2002a). A Bayesian model of stereopsis depth and motion direction discrimination. *Biological Cybernetics*, *86*(2), 117–136.

Read, J. C. A. (2002b). A Bayesian approach to the stereo correspondence problem. *Neural Computation*, *14*(6), 1371–1392.

Smallman, H. S., & McKee, S. P. (1995). A contrast ratio constraint on stereo matching. *Proceedings of the Royal Society of London Series B—Biological Sciences*, *260*(1359), 265–271.

Sun, J., Zheng, N. N., & Shum, H. Y. (2003). Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(7), 787–800.

Tsang, E. K. C., & Shi, B. E. (2008). Normalization enables robust validation of disparity estimates from neural populations. *Neural Computation*, *20*(10), 2464–2490.

Yang, Q. X., Wang, L., Yang, R. G., Stewenius, H., & Nister, D. (2009). Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(3), 492–504.

Yao, H. S., & Li, C. Y. (2002). Clustered organization of neurons with similar extra-receptive field properties in the primary visual cortex. *Neuron*, *35*(3), 547–553.

Zhang, Z., Edwards, M., & Schor, C. M. (2001). Spatial interactions minimize relative disparity between adjacent surfaces. *Vision Research*, *41*(23), 2995–3007.

Received April 21, 2009; accepted February 11, 2010.