LETTER Communicated by Hugh Wilson

A Coarse-to-Fine Disparity Energy Model with Both Phase-Shift and Position-Shift Receptive Field Mechanisms

Yuzhi Chen
chen@mail.cps.utexas.edu
Ning Qian
nq6@columbia.edu
Center for Neurobiology and Behavior and Department of Physiology and Cellular Biophysics, Columbia University, New York, NY 10032, U.S.A.

Numerous studies suggest that the visual system uses both phase- and position-shift receptive field (RF) mechanisms for the processing of binocular disparity. Although the difference between these two mechanisms has been analyzed before, previous work mainly focused on disparity tuning curves instead of population responses. However, tuning curve and population response can exhibit different characteristics, and it is the latter that determines disparity estimation. Here we demonstrate, in the framework of the disparity energy model, that for relatively small disparities, the population response generated by the phase-shift mechanism is more reliable than that generated by the position-shift mechanism. This is true over a wide range of parameters, including the RF orientation. Since the phase model has its own drawbacks of underestimating large stimulus disparity and covering only a restricted range of disparity at a given scale, we propose a coarse-to-fine algorithm for disparity computation with a hybrid of phase-shift and position-shift components. In this algorithm, disparity at each scale is always estimated by the phase-shift mechanism to take advantage of its higher reliability. Since the phase-based estimation is most accurate at the smallest scale when the disparity is correspondingly small, the algorithm iteratively reduces the input disparity from coarse to fine scales by introducing a constant position-shift component to all cells for a given location in order to offset the stimulus disparity at that location. The model also incorporates orientation pooling and spatial pooling to further enhance reliability. We have tested the algorithm on both synthetic and natural stereo images and found that it often performs better than a simple scale-averaging procedure.

1 Introduction

Position-shift and phase-shift (or phase-difference) models are two distinct mechanisms that have been proposed for describing reception field (RF) profiles and disparity sensitivity of V1 binocular cells (Bishop & Pettigrew,

1986; Ohzawa, DeAngelis, & Freeman, 1990; see Qian, 1997, for a review). The position-shift model assumes that the left and right RFs of a simple cell are always identical in shape but can be centered at different spatial locations, while according to the phase-shift model, the two RFs of a cell can have different shapes but are always tuned to the same spatial location. Recent studies indicate that both the phase- and position-shift models contribute to the representation of binocular disparity in the brain (Zhu & Qian, 1996; Ohzawa, DeAngelis, & Freeman, 1997; Anzai, Ohzawa, & Freeman, 1997, 1999a; Livingstone & Tsao, 1999; Prince, Cumming, & Parker, 2000) and that the phase-shift model appears to be more prominent although position shift may also play a significant role at high spatial frequencies (Ohzawa et al., 1997; Anzai et al., 1997, 1999a). These findings are consistent with earlier psychophysical observations that there is a correlation between the perceived disparity limit and spatial frequency as predicted by the phase model, but the disparity range at high spatial frequencies is larger than what a purely phase-shift model allows (Schor & Wood, 1983; Smallman & MacLeod, 1994).

These studies raise at least two questions: What are the relative strengths and weaknesses of the phase- and the position-shift mechanisms in disparity computation, and what is the advantage, if any, of having both mechanisms? We (Zhu & Qian, 1996; Qian & Zhu, 1997) and others (Smallman & MacLeod, 1994; Fleet, Wagner, & Heeger, 1996) have previously analyzed the similarities and the differences between the two RF models. However, that work mainly focused on the disparity tuning properties of a given cell. For the task of disparity estimation faced by the brain, it is the population responses of many cells to a given (unknown) disparity that is most relevant.[1] Tuning curves and population response curves have different shapes and properties in general; they are identical only under some special conditions (Teich & Qian, 2003). Qian and Zhu (1997) did compare the disparity maps computed from the population responses of the phase- and position-shift RF models, but the study did not examine the properties of the population responses in detail and was done under the condition where the difference between the two RF models is minimal (see section 2.1). In addition, previous studies have not explored how to properly combine the phase- and position-shift mechanisms in a stereo algorithm to achieve better results than either mechanism alone.

Another limitation of most previous studies on disparity computation is the exclusion of orientation tuning. Indeed, with a few exceptions (Mikaelian

[1] As we will detail in section 2, for a set of cells with a range of preferred disparity, a population response curve is obtained by plotting each cell's response to a fixed stimulus disparity against the cell's preferred disparity. The stimulus disparity can be estimated from either the peak or the mean of a population response curve. In this article, we use the peak location. In contrast, one cannot estimate stimulus disparity directly from a disparity tuning curve (Qian, 1997).

& Qian, 1997, 2000; Matthews, Meng, Xu, & Qian, 2003), most previous computational studies of the disparity energy model employed one-dimensional (1D) Gabor filters (Qian, 1994; Zhu & Qian, 1996; Fleet et al., 1996; Qian & Zhu, 1997; Tsai & Victor, 2003) instead of two-dimensional (2D), oriented filters found in the visual cortex.[2] Of course, a 1D algorithm can be readily extended to 2D by replacing 1D Gabor filters with 2D Gabor filters. However, without the additional refinements presented in this article, the performance of the resulting 2D algorithm is much worse than that of the corresponding 1D algorithm. The reason is that at depth boundaries, 2D filters tend to mix up regions of different disparities more than 1D filters do (Chen & Qian, unpublished observations).

In this article, we analyze the differences between the phase-shift and position-shift RF models in terms of both disparity tuning curves and population responses. We consider stimuli with and without orientations, and examine various offsets between the preferred orientation of the cells and the stimulus orientation. We find that although the two RF models generate disparity tuning curves of similar qualities, the phase-shift mechanism provides more reliable population responses than the position-shift mechanism does for relatively small stimulus disparity. Here, the reliability is measured by the standard deviation of the peak location of the tuning curve or population response curve when certain stimulus details unrelated to disparity (such as the lateral position of a bar or the dot distribution in a random dot pattern) are varied. Based on these and other considerations and our previous finding that the phase-shift-based disparity computation is accurate only when the stimulus disparity is considerably smaller than the RF sizes (Qian, 1994; Qian & Zhu, 1997), we propose an iterative algorithm that employs both the phase- and position-shift mechanisms in a specific way and demonstrate that the hybrid mechanism is more reliable and accurate than either mechanism alone. After incorporating pooling across spatial location, orientation, and spatial scale, we present a coarse-to-fine model for disparity computation and demonstrate its effectiveness on both synthetic and natural stereograms. We also compare this coarse-to-fine algorithm with a simple scale-averaging procedure applied to the population responses of the position-shift mechanism.

2 Analyses and Simulations

2.1 Phase-Shift vs. Position-Shift RF Models. The details of the disparity energy model used in this work have been described previously (Ohzawa et al., 1990; Qian, 1994). Briefly, according to quantitative physiological studies, spatial RFs of a binocular simple cell can be well described

[2] Some studies mentioned orientation pooling, but did not really address the issue due to the 1D filters used.

by 2D Gabor functions (Daugman, 1985; Jones & Palmer, 1987; Ohzawa et al., 1990). The position-shift and phase-shift RF models can be expressed as an overall positional difference between the left eye and right eye Gabor functions and as a phase difference between the sinusoidal modulations of the Gabor functions, respectively. We first consider a vertically oriented Gabor function centered at origin:

$$g(x, y, \phi) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \cos(\omega x + \phi), \tag{2.1}$$

where ω is the preferred spatial frequency, σ_x and σ_y determine the x and y RF dimensions, and ϕ is the phase parameter for the sinusoidal modulation. (Other preferred orientations and RF centers can be obtained by rotating and translating this function, respectively.) The phase-shift model (Ohzawa et al., 1990; DeAngelis, Ohzawa, & Freeman, 1991; Ohzawa, DeAngelis, & Freeman, 1996) posits that the left and right RFs of a simple cell are expressed as

$$g_l^{pha}(x, y) = g(x, y, \phi) \tag{2.2}$$

$$g_r^{pha}(x, y) = g(x, y, \phi + \Delta\phi). \tag{2.3}$$

Thus, the two RFs have the same position (both centered at origin), but there is a phase difference $\Delta\phi$ between their sinusoidal modulations. In contrast, the position-shift model assumes that the left and right RFs take the form

$$g_l^{pos}(x, y) = g(x, y, \phi) \tag{2.4}$$

$$g_r^{pos}(x, y) = g(x + d, y, \phi). \tag{2.5}$$

These two RFs have identical shape, but there is an overall shift d between their horizontal positions.

Using these RF profiles, one can compute simple and complex cell responses based on the disparity energy model (Ohzawa et al., 1990; Qian, 1994). We focus on complex cell responses here because simple cell responses are much less reliable due to their stronger dependence on the Fourier phases of the stimuli (Ohzawa et al., 1990; Qian, 1994; Zhu & Qian, 1996; Qian & Zhu, 1997; Chen, Wang, & Qian, 2001). It can be shown (see the appendix) that the response of a complex cell (constructed from a quadrature pair of simple cells) to a stereo image patch of disparity D can be written as

$$r_q^{pha} \approx 4A^2 \cos^2\left(\frac{\omega D - \Delta\phi}{2}\right) + \frac{D}{\sigma_x} 4AB \cos\left(\frac{\omega D - \Delta\phi}{2}\right) \times \cos\left(\alpha - \beta + \frac{\omega D - \Delta\phi}{2}\right) \tag{2.6}$$

$$r_q^{pos} \approx 4A^2 \cos^2\left(\frac{\omega(D-d)}{2}\right) + \frac{D-d}{\sigma_x} 4AB \cos\left(\frac{\omega(D-d)}{2}\right) \times \cos\left(\alpha - \beta + \frac{\omega(D-d)}{2}\right) \tag{2.7}$$

for the phase- and position-shift RF models, respectively. Here A and α are the local Fourier amplitude and phase (evaluated at the RF's preferred frequency) of the stimulus patch filtered by the RF gaussian envelope; B and β are the similar amplitude and phase of the stimulus patch filtered by the first-order derivative of the RF gaussian envelope (see the appendix). The phases α and β are more dependent on the detailed luminance distribution of the stimulus than the amplitudes A and B. The two terms in each of equations 2.6 and 2.7 are, respectively, the zeroth and the first-order terms in D/σ_x or $(D-d)/\sigma_x$.

Before exploring the implications of the above expressions, we need to define disparity tuning curve and population response curve explicitly. For a given cell with fixed RF parameters, if we vary the stimulus disparity D and plot the response of the cell as a function of D, we obtain a disparity tuning curve. The preferred disparity of the cell is the stimulus disparity that generates the peak response in the tuning curve. For a fixed stimulus disparity D, we can consider a set of cells that prefer different disparities (e.g., have different $\Delta\phi$ or d parameters; see below) but are otherwise identical. If we plot the responses of these cells to the same D against their preferred disparities, we get a population response curve.

A complication is that a cell's preferred disparity depends not only on its intrinsic RF parameters but also on the stimulus to some degree (Poggio, Gonzalez, & Krause, 1988; Zhu & Qian, 1996; Chen et al., 2001). The abscissa of the population response curve, however, should not depend on any stimulus parameters as these parameters are assumed to be unknown during disparity computation. In other words, the abscissa should be only a function of the intrinsic RF parameters that uniquely label each cell in the population. We will therefore use an intrinsic parameter (or a combination of intrinsic parameters) as the abscissa that approximates the preferred disparity of each cell. To do so, note that in equations 2.6 and 2.7, the first term (zeroth order) is usually larger than the second term (first order). If we keep only the first term, as we did in some of our previous work (Qian, 1994; Zhu & Qian, 1996; Qian & Zhu, 1997), the preferred disparity of a cell is a function of its intrinsic RF parameters only, given by:

$$D_{pref}^{pha} \approx \frac{\Delta\phi}{\omega} \tag{2.8}$$

$$D_{pref}^{pos} \approx d \tag{2.9}$$

for the phase- and position-shift models, respectively. We will therefore use $\Delta\phi/\omega$ or d to label different cells along the abscissa of population response

curves. Also note that if we keep only the first terms, the stimulus disparity D can be estimated from the preferred disparity of the most active cell (denoted by $*$) of the population response curve according to

$$D_{est}^{pha} \approx \frac{\Delta\phi^*}{\omega} \tag{2.10}$$

$$D_{est}^{pos} \approx d^* \tag{2.11}$$

for the phase- and position-shift models, respectively (Qian, 1994; Zhu & Qian, 1996; Qian & Zhu, 1997). We will use the same equations for disparity estimation in this article as we did previously. However, the more accurate analyses done here (the second terms in equations 2.6 and 2.7) will allow us to determine the conditions under which equations 2.8 to 2.11 are good approximations, and will lead to a better hybrid model with both phase- and position-shift mechanisms.[3]

We can now consider the properties of disparity tuning curves and population response curves generated with the phase- and position-shift RF models. We are particularly interested in whether and how much these curves depend on the stimulus details other than disparity (such as the lateral position of a bar or the dot distribution in a random dot pattern). Obviously, such dependence is undesirable, as it will make disparity estimation unreliable. We first consider disparity tuning curves. According to the definitions in the appendix, α and β in equations 2.6 and 2.7 strongly depend on the detailed luminance profile of the stimulus. If the second terms in the equations can be neglected, then the first term will be scaled only by A. Consequently, the shape of the tuning curves will be largely independent of the stimulus details, and the peak locations will accurately follow equations 2.8 and 2.9 for the phase- and position-shift models, respectively. If the second terms are not very small, however, the tuning curves of a cell will change with stimulus details, and the peak locations may deviate significantly from equations 2.8 and 2.9. Since the stimulus disparity D has to vary over a wide range to generate a tuning curve, D/σ_x or $(D-d)/\sigma_x$ cannot always be small, and thus the second terms cannot always be neglected. Therefore, complex cell tuning curves based on either position-shift or phase-shift RF model must be somewhat unreliable, albeit much more reliable than simple-cell tuning curves (Qian, 1994; Zhu & Qian, 1996; Chen et al., 2001).

The situation is different for population responses, however. Here stimulus disparity D is fixed while the cell parameter $\Delta\phi$ or d varies over a wide range. For the phase-shift model, if D is fixed at a value much smaller than σ_x, then the second term is negligible over the entire range of $\Delta\phi$ because $\Delta\phi$

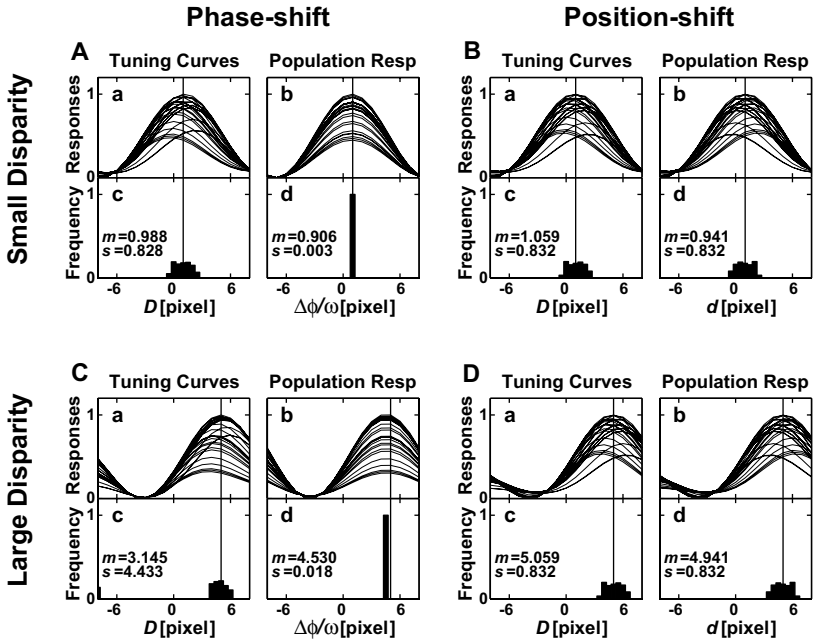
[3] Alternatively, one could solve D from equation 2.6 or 2.7, or use those equations as templates to estimate D. The resulting method, however, will involve complex procedures that are unlikely to be implemented by the visual cells.

appears only inside the bounded cosine functions. In this case, equation 2.10 is well satisfied at the peak location of the population response curve and can be used to estimate D accurately regardless of the other details of the stimulus. Therefore, population responses generated with the phase-shift RF model should be highly reliable when the stimulus disparity is small compared with the RF size. In contrast, population responses generated with the position-shift model can never be very reliable. The reason is that the cells' position-shift parameter d is present both inside and outside the cosine functions in equation 2.7. As d varies over a wide range, the second term cannot always be small for any fixed values of D. Intuitively, when d is significantly different from D, the image patches covered by the left and right RFs will be very different, and this difference will introduce large variability into the population response curve.

We therefore conclude that among the four cases of disparity tuning curve and population response curve generated with the phase- and position-shift RF mechanisms, only the population response curves from the phase-shift mechanism have highly reliable peak locations for small stimulus disparities. In all other cases, the results in general will vary with stimulus details unrelated to disparity. We have performed extensive computer simulations to confirm this conclusion. Two examples are shown in Figures 1 and 2 for bar and random dot stimuli, respectively. We did not include spatial pooling (Zhu & Qian, 1996; Qian & Zhu, 1997) in these simulations in order to see the difference between the two RF models more clearly. The difference will be reduced (but not vanish) when spatial or orientation pooling is introduced.

Figure 1 shows simulated disparity tuning curves (of a given complex cell in response to a range of stimulus disparities) and population response curves (of an array of model complex cells to a given stimulus disparity) for both phase- and position-shift RF mechanisms. The orientation of both the RFs and the stimuli (bars) was vertical in these simulations. To measure reliability in each case, we simulated 1000 disparity tuning curves or 1000 population response curves by randomly varying the lateral bar position while keeping all the other parameters constant (see the figure caption for details). For each bar position, a given stimulus disparity was introduced by shifting the left and right eyes' images in opposite directions by half the disparity value. For clarity, only 30 tuning curves and 30 population curves are shown in panels a and b, respectively, but the peak location histograms in panels c and d are compiled from all 1000 simulations. The numbers inside each histogram panel are the mean (m) and standard deviation (s) of the distribution. The vertical lines in the tuning curve panels indicate the cell's preferred disparity as determined by its parameters according to equation 2.8 (for phase shift) or equation 2.9 (for position shift). The vertical lines in the population response panels indicate the stimulus disparity. For tuning curves, *small* or *large disparity* refers to the cell's preferred disparity (1 or 5 pixels). For population response curves, *small* or *large disparity* refers to the stimulus disparity (also 1 or 5 pixels).

It should be clear from the figure that the disparity tuning curves (panels a and c) are not very reliable in all cases: the peak location distributions all show significant spread, indicating some dependence of the preferred disparity on the lateral bar position. For the population responses, the same unreliability holds for the position-shift model (panels b and d in Figures 1B and 1D). In contrast, the population response curves of the phase-shift model to small stimulus disparity (panels b and d in Figure 1A) are both reliable (the peaks from 1000 simulations are well aligned) and accurate (the peak location agrees with the actual stimulus disparity). These results are consistent with our analysis above. The population response of the phase-shift model to large stimulus disparity (panels b and d in case Figure 1C) is also reliable. As we will show in Figure 2, this is not generally true, but happens to be so for the bar stimuli because the α and β parameters in equation 2.6 approximately cancel each other (see the appendix) and the two terms of the equation can be combined. Note, however, that in this case, the peak location is not accurate as it underestimates the stimulus disparity. This underestimation is due to a zero-disparity bias of the phase-shift model demonstrated previously (Qian & Zhu, 1997), and it grows with stimulus disparity size. If equation 2.6 is expanded up to the second-order term, then the zero disparity bias of the phase-shift model will become apparent (results not shown). Also note that the curves in Figure 1 usually have side peaks (Qian, 1994; Zhu & Qian, 1996) outside the range plotted here.



We also repeated the above simulations with random dot stereograms, and the results are shown in Figure 2 with the same format as Figure 1. The disparity tuning curves and population response curves were simulated with 1000 sets of random dot stereograms that all contain the same disparity values but different dot patterns. These curves are more variable than those in Figure 1, as reflected by the larger standard deviations of all the histograms, due to the stochastic nature of random dots. Nevertheless, Figure 2 clearly shows that the population response of the phase-shift model to small disparity (panels b and d in Figure 2A) is much more reliable and accurate than all the other cases, consistent with the results for bars in Figure 1 and with the prediction of our analysis. Note that in the large disparity case, both phase-shift and position-shift mechanisms show similarly unreliable population responses. This explains why Qian and Zhu (1997) found that the two RF mechanisms generate similar results; that study was done under the large disparity condition ($D/\sigma_x \geq 0.5$).

Figure 1: *Facing page*. Disparity tuning and population response curves of model complex cells with the phase- and position-shift RFs in response to bar stereograms. Four cases are considered: (A) small disparity with the phase-shift RFs; (B) small disparity with the position-shift RFs; (C) large disparity with the phase-shift RFs; and (D) large disparity with the position-shift RFs. For the tuning curves of a given cell to a range of stimulus disparities, *small* or *large disparity* refers to the cell's preferred disparity of 1 or 5 pixels (indicated by the vertical lines). For population response curves from an array of cells to a given stimulus disparity, *small* or *large disparities* refers to a stimulus disparity of 1 or 5 pixels (also indicated by the vertical lines). We use $\Delta\phi/\omega$ and d as approximate measures of the cells' preferred disparities in the population response plots (see equations 2.8 and 2.9). To test the reliability of tuning curve and population response, 1000 curves were obtained for each case by randomly varying the bar's lateral position in the cells' RF with subpixel interpolation. For the tuning curves, the bar's disparity varied from -8 to 8 pixels in a step of 1 pixel. For the population response curves, a group of model complex cells whose preferred disparities varied from -8 to 8 pixels in a step of 1 pixel were used. For clarity, only 30 tuning or population response curves are shown in panels a or b, and they are normalized by the strongest response. The peak location distribution histogram in panels c or d was compiled from all 1000 curves. Each peak location was computed by a parabolic fit of three points around the peak. The numbers in the histograms represent the mean peak location m and the standard deviation s, respectively. Parameters: In each stereogram, the bar's lateral position was confined within ± 8 pixels from the center of RF. The width and height of bar were 8 and 97 pixels, respectively. Both RFs and bars were oriented vertically. The RF parameters of model complex cells were $\omega/2\pi = 1/16$ cycle per pixel, $\sigma_x = 8$ pixels, $\sigma_y = 16$ pixels. The RFs were computed in a 2D region of 49×97 pixels. The bin size in each histogram was 0.5 pixel.

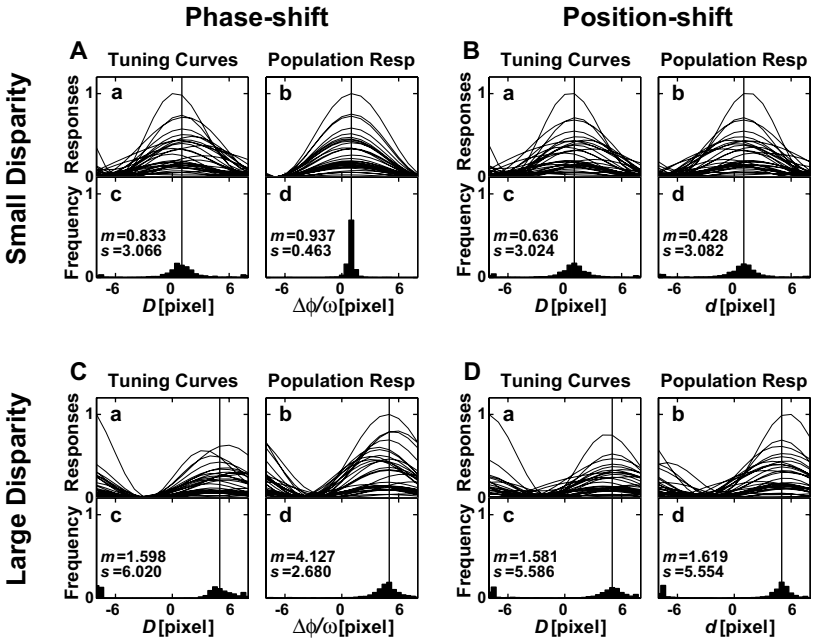


Figure 2: Disparity tuning and population response curves of model complex cells with the phase- and position-shift RFs in response to random dot stereograms. The size of the random dot stereograms is 49 × 97 pixels. The dot size and density are 2 × 2 pixels and 50%, respectively. All RF parameters and the presentation format are identical to those in Figure 1.

2.2 An Iterative Hybrid Algorithm for Disparity Estimation . We have shown above that for the phase-shift model and small stimulus disparity (relative to the cells' RF size), the peak location of the population response curve provides both reliable and accurate estimation of the stimulus disparities, while the position-shift model is always less reliable in comparison. However, the phase-shift model has its own limitations. First, for cells with preferred horizontal spatial frequency ω, the range of disparity they can detect is confined between $-\pi/\omega$ and π/ω (Qian, 1994) due to the periodicity of the population response as a function of $\Delta\phi$, which in turn is due to the periodicity of the Gabor RFs as a function of ϕ. Any disparity beyond this range cannot be correctly detected. Second, for stimulus disparity within the range, the reliability of the population responses gets worse as the disparity approaches the limits of the range (compare Figure 2Cd with Figure 2Ad). The accuracy also decreases with increasing disparity magnitude because the underestimation caused by the zero disparity bias increases.

Since there is ample evidence indicating that both the phase- and position-shift mechanisms are involved in disparity processing (Schor & Wood, 1983; Smallman & MacLeod, 1994; Zhu & Qian, 1996; Anzai et al., 1997, 1999a; Livingstone & Tsao, 1999; Prince et al., 2000), it is natural to consider whether a hybrid of the two RF models could provide a better solution. Similar to the derivations of equations 2.6 and 2.7, the response of a complex cell with both position shift d and phase shift $\Delta\phi$ between the two eyes can be written approximately as

$$\begin{aligned} r_q^{hyb} \approx\ & 4A^2 \cos^2\left(\frac{\omega(D-d)-\Delta\phi}{2}\right) \\ & + \frac{(D-d)}{\sigma_x} 4AB \cos\left(\frac{\omega(D-d)-\Delta\phi}{2}\right) \\ & \times \cos\left(\alpha-\beta+\frac{\omega(D-d)-\Delta\phi}{2}\right) \end{aligned} \tag{2.12}$$

If we only consider the first term in equation 2.12, the preferred disparity of the cell is given by

$$D_{pref}^{hyb} \approx \frac{\Delta\phi}{\omega} + d. \tag{2.13}$$

Since $(D-d)$ appears outside the cosine functions in the second term of equation 2.12 just as in equation 2.7, whenever the position-shift parameter d is varied over a range for a population response curve, the computed disparity will not be reliable. Therefore, one should always rely on the phase difference $\Delta\phi$ for disparity computation, and d should be kept a constant close to D for all cells used in a given population response curve. The best scenario occurs if d happens to be equal to D since the second term of equation 2.12 will vanish, and the residual disparity $(D-d)$ for the phase mechanism to estimate will be zero, and this is when the phase model is most accurate.

These considerations lead us to the following iterative algorithm with a hybrid of both phase- and position-shift RFs. For each image location, we start (iteration 0) with a population of cells all with $d=0$ and with $\Delta\phi$ covering the full range from $-\pi$ to π, and apply the phase mechanism to get an estimation D_0 of the stimulus disparity D as we did before (Qian, 1994). Next, we use a set of cells all with the fixed $d=D_0$ and the full range of $\Delta\phi$, and apply the phase mechanism again to get a new estimate D_1 (iteration 1). Since the original stimulus disparity D has been offset by the constant position shift $d=D_0$ of all the cells, D_1 is a measure of the residual disparity $D-d=D-D_0$. Thus, the estimated stimulus disparity D_{est} from the first iteration is D_0+D_1. This process can be repeated such that at the nth iteration, cells will all have the same position-shift $d=D_0+D_1+\cdots+D_{n-1}$ and the newly computed D_n from the phase mechanism can be added to d to form

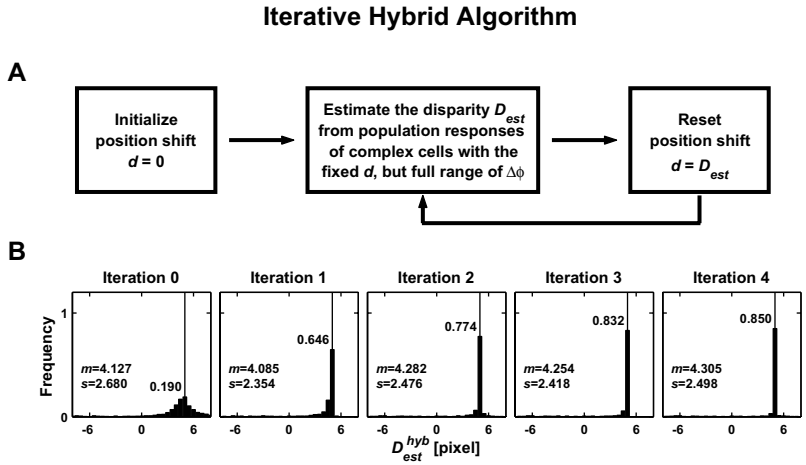


Figure 3: Iterative hybrid algorithm. (A) Flowchart of the algorithm. For a given stimulus, first initialize the position shift d to zero, and compute the population responses of a set of complex cells all with the same d but the full range of the phase shift $\Delta\phi$ from $-\pi$ to π. The peak location of population responses is extracted as the estimated disparity D_{est} according to equation 2.13. Then reset the position shift $d = D_{est}$, and repeat the above procedure until a stable disparity value is obtained. (B) The performance histograms for the first five iterations of the algorithm, compiled from the same 1000 random dot stereograms used in Figure 2Cd. The format of the histograms is the same as that of Figure 2Cd except that here, the maximum value of each distribution is shown above the peak in each panel. All parameters but d are the same as those used in Figure 2Cd. To simply the simulations, we generated in advance 17 populations of cells with fixed d's from -8 to 8 pixels in steps of 1 pixel. At each iteration, we selected the population whose fixed d was closest to the estimated disparity D_{est} in the previous iteration.

the current estimation of disparity $D_{est} = D_0 + D_1 + \cdots + D_n$. This algorithm is shown schematically in Figure 3A. Note that at each iteration, the (residual) disparity is always computed with the phase mechanism. The position-shift parameter d is simply introduced to offset the stimulus disparity D so that the phase mechanism will operate on a progressively smaller (residual) disparity and will therefore become more and more reliable and accurate.

A demonstration of this algorithm is presented in Figure 3B. We applied the method to the same 1000 random dot stereograms (all with $D = 5$ pixels) as in Figure 2C. Figure 3B shows the distribution histograms of estimated disparity up to the fourth iterations; little change occurs afterward. Since d equals 0 in the 0th iteration (far left panel), its histogram is identical to that in Figure 2Cd, with a broad distribution of estimated disparity around

the true disparity (vertical line). The situation quickly improved with a few iterations. The fraction of estimations falling within a bin of 0.5 pixel width around the true disparity increased from 19% to 85% in four iterations. This result suggests that the iterative hybrid algorithm converges quickly and that the estimation is much more accurate and reliable than a pure phase- or position-shift mechanism alone.

It should be noted in Figure 3B that at each iteration, the estimated disparity always has some very small probability of being far away from the true stimulus disparity (note the negative disparity tail barely visible in the histograms). A closer examination reveals that if an estimation is far from the true disparity at the start (e.g., a wrong sign), subsequent iterations usually cannot correct the error. The reason is that if the estimated disparity is very wrong, the fixed d introduced in the next step may make the residual disparity larger than the original disparity and the phase mechanism will become less reliable and harder to recover from the error. We will show that one way to reduce the occurrence of such runaway behavior is to pool across space, orientation, and scale.

2.3 Pooling. Pooling information from different sources can often improve performance. We (Zhu & Qian, 1996; Qian & Zhu, 1997) have previously demonstrated that spatial pooling of quadrature pair responses within a small neighborhood can significantly improve the quality of computed disparity maps. Here we focus on orientation pooling and spatial frequency (i.e., scale) pooling.

2.3.1 Orientation Pooling. Before considering orientation pooling, we first examine how the disparity population responses of model complex cells depend on their preferred orientations. To generate RFs with orientation θ (measured from the positive horizontal axis), one can rotate the corresponding vertically oriented RF (equation 2.1) by $\theta - 90^\circ$ with respect to the RF center (positive angle means counterclockwise rotation). For the phase-shift RF model, the left and right RF centers are the same. Therefore, the response of a complex cell with preferred orientation θ to any stimulus is equal to the response of the corresponding vertically oriented complex cell to the stimulus rotated by an angle of $90^\circ - \theta$. If the original stimulus has a horizontal disparity D, the rotated stimulus then has components $D \sin\theta$ and $D \cos\theta$ orthogonal and parallel to the RF orientation, respectively. Since the RF profile along the preferred orientation changes much more slowly than that along the orthogonal axis, the parallel component $D \cos\theta$ can be ignored, and the complex cell response with the phase-shift mechanism and preferred orientation θ can be obtained approximately by replacing D in equation 2.6 by $D \sin\theta$:

$$r_q^{pha}(\theta) \approx 4A'^2 \cos^2\left(\frac{\omega D \sin\theta - \Delta\phi}{2}\right)$$

$$+ \frac{D\sin\theta}{\sigma_\perp} 4A'B' \cos\left(\frac{\omega D\sin\theta - \Delta\phi}{2}\right)$$
$$\times \cos\left(\alpha' - \beta' + \frac{\omega D\sin\theta - \Delta\phi}{2}\right), \tag{2.14}$$

where A', B', α', and β' are similar to A, B, α and β in equations 2.6 and 2.7 except that they refer to the stimulus rotated an angle of $90^\circ - \theta$. Here, $\sigma_\perp$ represents the gaussian width of the RF in the direction perpendicular to the cell's preferred axis; it equals σ_x for vertically oriented RFs.

For the position-shift mechanism, the left and right RF centers are not the same in general. Therefore, the rotational equivalence mentioned above has to be applied to the left and right RF responses separately before binocular combination. The final result is that the complex cell response with the position-shift mechanism and preferred orientation θ can be obtained approximately by replacing $D - d$ in equation 2.7 by $(D - d)\sin\theta$:

$$r_q^{pos}(\theta) \approx 4A'^2 \cos^2\left(\frac{\omega(D-d)\sin\theta}{2}\right)$$
$$+ \frac{(D-d)\sin\theta}{\sigma_\perp} 4A'B' \cos\left(\frac{\omega(D-d)\sin\theta}{2}\right)$$
$$\times \cos\left(\alpha' - \beta' + \frac{\omega(D-d)\sin\theta}{2}\right). \tag{2.15}$$

This result depends on the assumption that the parallel component $(D - d)\cos(\theta)$ can be ignored. Since d has to vary over a large range for a population response curve, equation 2.15 is not as good an approximation as equation 2.14. For vertical RF orientation, θ equals 90 degrees, and equations 2.14 and 2.15 reduce to equations 2.6 and 2.7, respectively.

Similar to the discussion following equations 2.6 and 2.7, equations 2.14 and 2.15 also indicate that only the population response of the phase-shift mechanism to small stimulus disparity D (compared with the RF size) can be reliable and accurate. A new feature is that the $\sin\theta$ factor will make the tuning curves and population response curves broader as the RF orientation θ deviates further from vertical. In the extreme case of horizontal orientation ($\theta = 0$), the curves will be flat with infinite width. For phase-shift RFs with orientation θ, the preferred disparity expression should be generalized from equation 2.8 to:

$$D_{pref}^{pha} \approx \frac{\Delta\phi}{\omega\sin\theta} \equiv \frac{\Delta\phi}{\omega_x}, \tag{2.16}$$

and the detectable disparity range becomes $(-\pi/\omega\sin\theta, \pi/\omega\sin\theta)$ or $(-\pi/\omega_x, \pi/\omega_x)$ where $\omega_x = \omega\sin\theta$ (Qian, 1994; Mikaelian & Qian, 2000;

Matthews, Meng, Zu, & Qian, 2003). This range is smallest for the vertically oriented RFs and increases when the RF orientation θ deviates from vertical. Here ω is the preferred spatial frequency along the axis perpendicular to the RF orientation, and ω_x is the preferred spatial frequency along the horizontal axis; they equal each other when the RF is vertically oriented. For the position-shift RFs, equation 2.9 remains the same since d is assumed to be a horizontal shift regardless of the RF orientation. (If d is assumed to be the shift orthogonal to the RF orientation, then equation 2.9 will become $D_{pref}^{pos} \approx d/\sin\theta$.)

The simulated disparity population responses and the peak-location histograms of complex cells are shown in Figure 4 for a bar with a small horizontal disparity of 1 pixel (marked by the vertical line in each panel) and an orientation of 67.5 degrees. Eight RF orientations evenly distributed in the entire 180 degree range were considered. Since the complex cells with horizontal orientation are not sensitive to horizontal disparity and generate flat curves, we present only simulations from the seven nonhorizontal preferred orientations. The results for the phase- and position-shift RF models are shown in Figures 4A and 4B of the figure, respectively. For all seven preferred orientations, the population responses with the phase-shift mechanism are more reliable than those with the position-shift mechanism. For the phase-shift mechanism (see Figure 4A), the peak location of the population response depends on the difference between the RF and bar orientations. Only when the RF orientation matches the bar orientation (67.5 degrees), does the peak location agree with the actual stimulus disparity. Otherwise, the peak location underestimates the stimulus disparity magnitude. In particular, when the preferred orientation is 157.5 degrees, orthogonal to the bar orientation, the peak locates at zero disparity. As predicted by the analysis, the width of the curves in Figure 4 varies with the RF orientation.[4] The maximum response in each panel (max) is shown at the top of the panel. Not surprisingly, the largest response occurs when the RF orientation matches the bar orientation.

We have also done similar simulations with a random dot stereogram. The results (not shown) are similar except that since the stimulus is nonoriented, the maximum responses across all RF orientations do not differ much, and there is no orientation-dependent underestimation of the small disparity.

We now consider pooling across cells with different orientations. The above results suggest that one should first average together the population response curves from all orientations and then use the peak location of the

[4] However, for the position-shift RF model (see Figure 4B), the sharpest response curves occur when the cells' orientation matches the bar orientation (67.5 degrees) instead of when it is vertical (90 degrees). This is not predicted by the approximate result of equation 2.15.

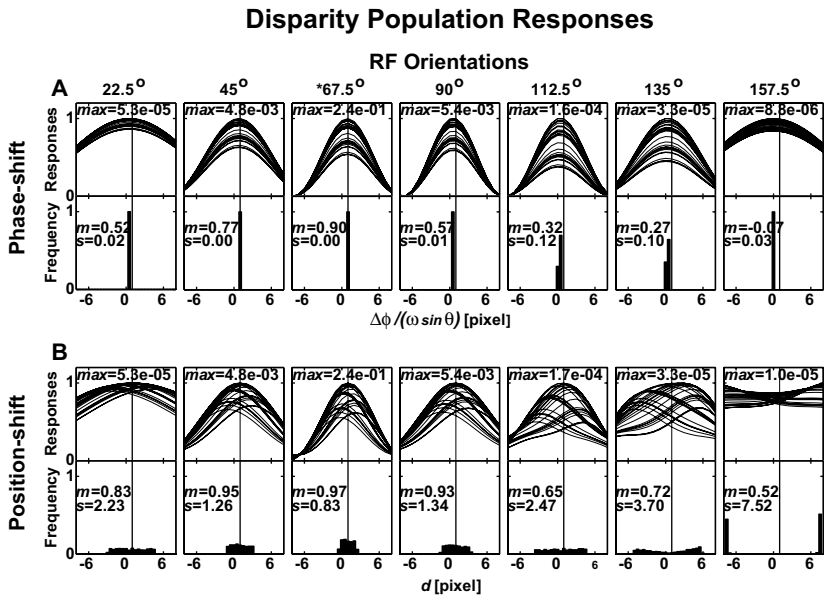


Figure 4: Population responses to bar stereograms from model complex cells with different preferred orientations and with (A) the phase-shift and (B) the position-shift RF mechanisms. The bar stereograms have a fixed disparity of 1 pixel and a fixed orientation of 67.5 degrees. The preferred orientations of the cells are indicated above the first row of the curves, and the case where the preferred orientation matches the bar orientation is marked by an asterisk. All simulation parameters (except orientation) and presentation format are the same as those for the population responses in Figures 1A and 1B. The number over the curves in each panel represents the maximum response to the 1000 stimuli with an arbitrary unit. Although the disparity range covered by a family of cells increases when the preferred orientation is closer to horizontal (see text), for the ease of presentation and comparison, we confined the disparity range of all cell families to that of the vertically oriented cells.

averaged curve to estimate disparity.[5] For oriented stimuli such as bars, the response from cells with the matched orientation is the most accurate, and it will dominate the average because it is also the strongest. For nonoriented stimuli such as random dots, cells with different orientations respond similarly, and the averaging helps reduce noise. An alternative method is to average the estimated disparities from all orientations weighted by the

[5] This procedure is analogous to what we did previously with spatial pooling (Zhu & Qian, 1996; Qian & Zhu, 1997).

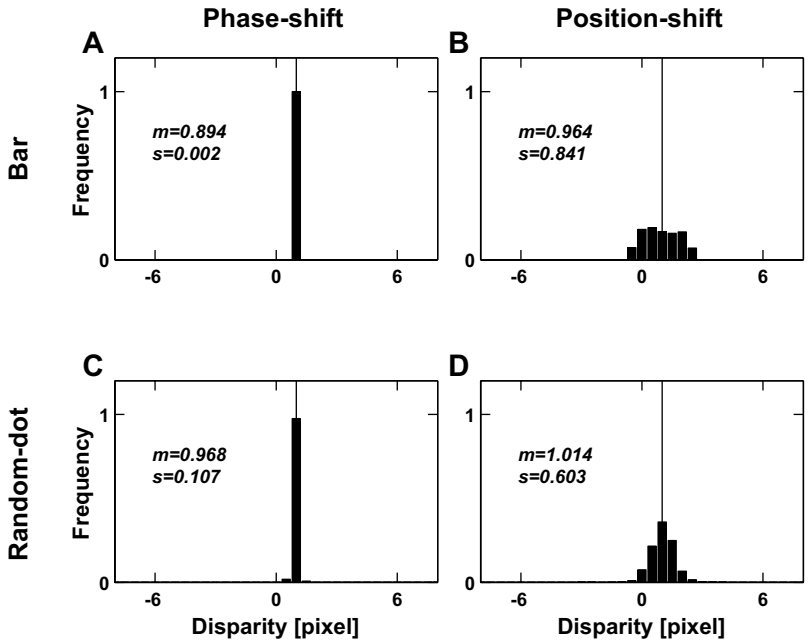


Figure 5: Distribution histograms of the estimated disparity after orientation pooling. Both phase- and position-shift mechanisms were applied to bar and random dot stimuli. The bar stimuli are same as in Figure 4, and the random dot stimuli are identical to those for panel Ad or Bd of Figure 2. The seven preferred orientations in Figure 4 were used in the pooling.

corresponding peak responses. We found that this method is usually not as good and will report simulation results only with the first method.

Figure 5 shows the distribution histograms of the estimated disparity with orientation pooling applied to bar and random-dot stimuli. The bar stimuli are same as in Figure 4. As expected, the distribution for the phase-shift mechanism in Figure 5A is as good as the case where the RF orientation matches the bar orientation (67.5 degrees) in Figure 4A. The random dot stimuli are identical to those for Figure 2Ad, and the distribution in Figure 5C is much more reliable due the pooling. In contrast, the orientation pooling is much less effective for the position-shift RF model (Figures 5B and 5D). Note that we used a relatively small stimulus disparity D; if D is increased to about half of the maximum value allowed by the phase-shift model, the difference between the two RF models will disappear (results not shown).

2.3.2 Scale Pooling and a Coarse-to-Fine Algorithm. In addition to orientation, disparity-selective cells are also tuned to spatial frequency. Cells in each frequency band (or scale) can be used to compute disparity (Marr & Poggio, 1979; Sanger, 1988; Qian, 1994), and an important question is how to combine information from different scales. Obviously, the scale pooling can be done according to the same methods for orientation pooling. The first method is to average all the population response curves from different scales and then estimate the disparity.[6] Alternatively, one could estimate one disparity from each scale and then average the estimates with proper weighting factors (Sanger, 1988; Qian & Zhu, 1997; Mikaelian & Qian, 2000). However, although these approaches work reasonably well for orientation pooling, they may not be adequate for scale pooling because the large and small scales have different problems that are unlikely to cancel each other through averaging. Cells at large scales have large RFs, and they tend to mix together different disparities in the image area covered by the RFs. This will make transitions at disparity boundaries less sharp than our perception (Qian & Zhu, 1997) and render disparities in small image regions difficult to detect. At small scales, the detectable disparity range by the cells is correspondingly smaller (Sanger, 1988; Qian, 1994), and large disparities in a stereogram may lead to completely wrong estimations. If one simply averages across the scales, the resulting disparity map will not be accurate unless the majority of the scales included perform reasonably well. With too many large scales included, the computed disparity map will lose sharp details, and with too many small scales included, disparity estimations at some locations may be totally wrong. If one knew the true stimulus disparities beforehand, one could pick a few appropriate scales and average across them only. However, the purpose of a stereo algorithm is to compute disparity maps from arbitrary stereograms with unknown disparities.

A method known to alleviate these problems is coarse-to-fine tracking across scales. This method has been applied to stereovision previously (Marr & Poggio, 1979), but to our knowledge, its role in the disparity energy model with the phase- and position-shift RF mechanisms has not been explored. It is most natural to introduce the coarse-to-fine technique into our iterative algorithm presented in section 2.2. The only modification is to start at a large scale and then reduce the scale of the RFs through the iterations. By starting at a large scale, the algorithm can cover a large range of disparities. With each iteration, the disparity will be reduced by a constant position shift, and the residual disparity can thus be estimated by the phase mechanism at a smaller scale that sharpens the disparity boundaries. This procedure can be continued until a fine disparity map is obtained at the smallest scale.

[6] This approach has been applied to disparity tuning curves (Fleet et al., 1996) and can be extended to population responses for disparity computation.

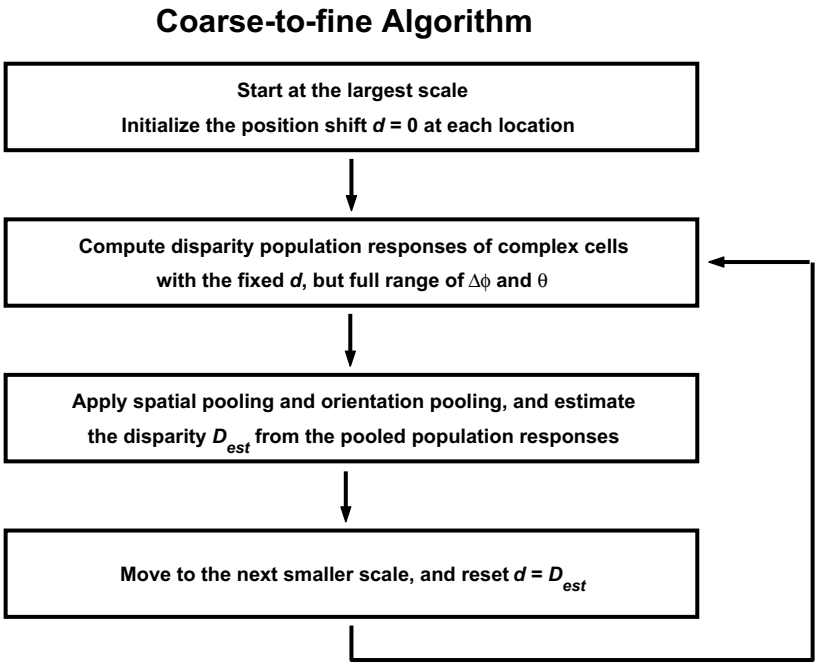


Figure 6: The coarse-to-fine algorithm with both phase- and position-shift RFs.

The full algorithm, with the spatial pooling and orientation pooling procedures incorporated, is shown schematically in Figure 6. At each iteration, the spatial pooling step combines quadrature pair responses in a local area to improve the reliability of the population responses (Zhu & Qian, 1996; Qian & Zhu, 1997), and the orientation pooling further improves the quality of the estimated disparity as described in section 2.3.1.

Based on the above considerations, the largest scale should be chosen according to the largest disparity the system should be able to extract, and the smallest scale should correspond to the finest details of disparity the system should be able to recover. For the simulations reported below, we employed a set of scales whose σ's follow a geometric series with a ratio of $\sqrt{2}$. To keep the cells' frequency bandwidth constant at different scales, we fixed the product $\omega\sigma_\perp = \pi$ for all scales, where $\sigma_\perp$ is again the gaussian width in the direction orthogonal to the RF orientation. At each scale, there are several sets of cell populations, each with a constant position shift d and the full range of $\Delta\phi$. Only the population whose d is closest to the estimated disparity in the previous scale will actually be used. We simply let the range of d's be the same across all scales and equal to the disparity range of the phase-shift mechanism at the largest scale. For spatial pooling,

we used a 2D gaussian function to combine the responses of quadrature pairs around each location (Zhu & Qian, 1996; Qian & Zhu, 1997). Finally, to reduce computational time, we used five RF orientations (30, 60, 90, 120, and 150 degrees from horizontal) to perform orientation pooling instead of the seven orientations in Figure 5.

Figure 7C is an example of applying our coarse-to-fine algorithm to a random dot stereogram. To test the model's performance for both large and small stimulus disparities, we picked a far disparity of 5 pixels for the central region and a near disparity of -1 pixel for the surround. The panels in Figure 7C show the estimated disparity maps at each of the five iterations or scales (with one iteration per scale). At the largest scale, the transition at the disparity boundaries is poor, and the disparity magnitude of both the center and the surround are underestimated due to the zero disparity bias of the phase mechanism (Qian & Zhu, 1997).[7] However, since this disparity map is generally in the right direction, the subsequent smaller scales were able to refine it. At the smallest scale, the map has sharp transition boundaries and accurate disparity values for both center and surround regions. The final map is much better than those computed from any individual scale with either phase- or position-shift mechanism.

We also compared the coarse-to-fine algorithm with the simple method of averaging population responses across scales mentioned above. Since at small scales, the phase-shift RF model can cover only very small ranges of stimulus disparity, we used the position-shift model with the scale averaging method. The results of applying the method to the same random dot stereogram are shown in Figure 7D. The spatial pooling and orientation pooling were applied as in Figure 7C. In Figure 7D, the panels from left to right show the computed disparity maps by gradually including more scales in the averaging process, with the left-most panel showing the result from the largest scale alone and the right-most panel the result of averaging all five scales. As expected, in the left two panels where the scales are large, the disparity map is fuzzy at the transition boundaries. With more smaller scales included, the estimated disparity at some locations is totally wrong (the black and white spots in the right panels of Figure 7D).

In the simulations of different scales, we fixed the stimuli and generated the RFs of different sizes. Alternatively, we could fix the RFs and scale the stimuli appropriately. With this approach, both the coarse-to-fine and the scale-averaging methods generate results similar to that in Figure 7C. This suggests that the scale-averaging method is more sensitive to implementation details. Also note that for the simulations in Figure 7, at each scale, only eight complex cells were used for the coarse-to-fine simulation but 33 cells

[7] As we noted earlier, without scale pooling, the single-scale disparity maps computed with 2D filters in this article are much worse than those computed with 1D filters reported previously (Qian, 1994; Qian & Zhu, 1997) because 2D filters tend to mix up disparities in a larger region.

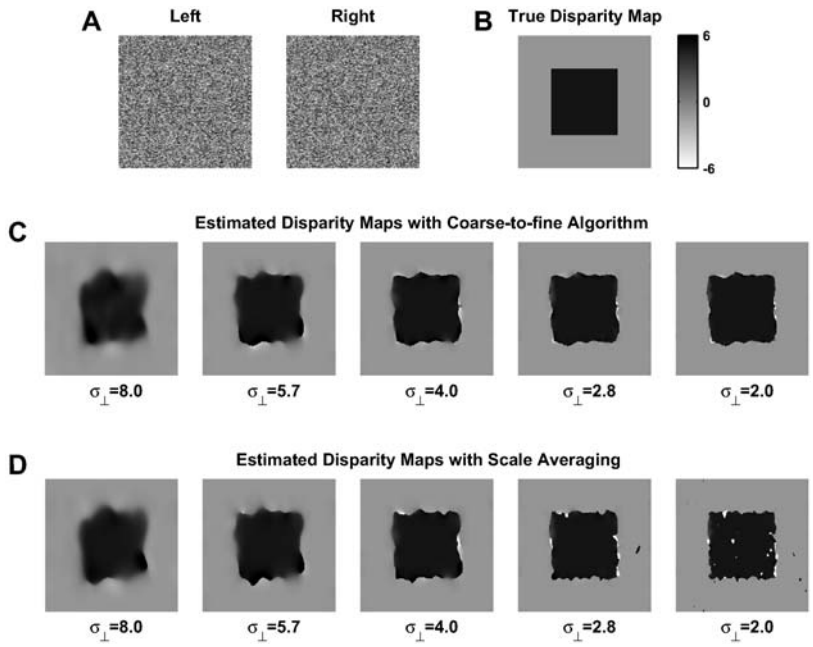


Figure 7: The coarse-to-fine and scale-averaging algorithms applied to a random dot stereogram. (A) A 200×200 random dot stereogram with a dot density of 50% and dot size of 1 pixel. The central 100×100 area has a disparity of 5 pixels, while the surround has a disparity of -1 pixel. (B) True disparity map. The white and black colors represent near and far disparities, respectively. (C) Estimated disparity maps at five scales and iterations obtained with the coarse-to-fine algorithm. The spatial pooling function was a 2D gaussian with both standard deviations equal to $\sigma_\perp$ in each scale. Orientation pooling covered five orientations from 30 to 150 degrees in steps of 30 degrees. The other parameters for the RFs in the left-most panel were the same as those in Figure 3. The scales of other panels were successively reduced from left to right by a factor of $\sqrt{2}$, as indicated by the $\sigma_\perp$ value under each panel. Note that for all five scales, $\omega\sigma_\perp = \pi$; thus, the frequency bandwidths of all scales were fixed at 1.14 octaves. For each scale, the position shift d always varied from -8 pixels to 8 pixels in a step of 0.5 pixel, while the phase shift $\Delta\phi$ covered a period from $-\pi$ to π with a step of $\pi/4$. (D) Estimated disparity maps with the scale-averaging procedure and the position-shift mechanism. From left to right, a smaller scale with the indicated $\sigma_\perp$ was added to the average at each panel. Thus, the left-most panel shows the result from the largest scale alone, while the right-most panel is the average result of all five scales. The spatial pooling and orientation pooling were also applied as in *C*. The RF parameters were same as those in *C*, except that $\Delta\phi$ was kept at 0.

were used for the scale-averaging simulation. The coarse-to-fine method requires fewer cells because at each finer scale, the cells used are more focused around the stimulus disparity. No such adjustment is present in the scale-averaging method, and if Figure 7D used only eight cells, the results (not shown) would be much worse. The scale-averaging method is also more sensitive to the frequency contents of a stereogram; it performs worse for narrowband stimuli (results not shown). In summary, the scale-averaging method is not as robust as the coarse-to-fine method.

2.4 Application of the Coarse-to-Fine Algorithm to More Complex Stereograms. We also applied the coarse-to-fine algorithm to more complex synthetic stereograms and to real-world stereograms. Figure 8 shows the results for a disparity ramp and a Gabor disparity profile. The stereograms were created by starting with a reference random dot image and then shifting each dot horizontally in opposite directions for the left and right images by half of the disparity value prescribed to the dot. Gray-level interpolation was used to represent subpixel disparities. Figures 8C and 8F demonstrate that the coarse-to-fine algorithm works well on these stereograms. Except for the slightly blurred disparity boundaries, the estimated disparity maps closely match the true disparity maps for both stereograms, with most errors within ± 0.25 pixel (true for 89% and 93% of the total pixels for the ramp and Gabor stereograms, respectively). For comparison, we also show in Figure 8 the results from the scale-averaging method used for Figure 7D; again, the estimated disparities at some locations are completely wrong.

Figure 9 shows three real-world stereograms and the estimated disparity maps with our coarse-to-fine algorithm and the scale-averaging algorithm. Since the true disparity maps are unknown, we can access the performance only qualitatively. The results computed with our coarse-to-fine algorithm seem to be quite reasonable for the Pentagon and Tree stereograms, but less accurate for the Shrub stereogram. In general, the method works well on image areas with relatively high-contrast textures (e.g., the grass ground of the Tree stereogram), but fails at low-contrast regions (the foreground pavement of the Shrub stereogram). This problem is not surprising as the low-contrast areas generate only weak complex-cell responses that are more prone to noise. Another problem is exemplified by the small black spot in the Pentagon disparity map: if a large scale reports a very wrong disparity, the smaller scales usually cannot correct it. Solving these problems may require more global but selective interactions of disparity information at different locations and a bidirectional information flow among the scales (see section 3). With the scale-averaging method, the problem of spots with wrong disparities is more pronounced, similar to Figures 7 and 8. Moreover, the estimated disparity maps appear more blurred than those obtained with the coarse-to-fine algorithm (e.g., the signpost in Shrub).

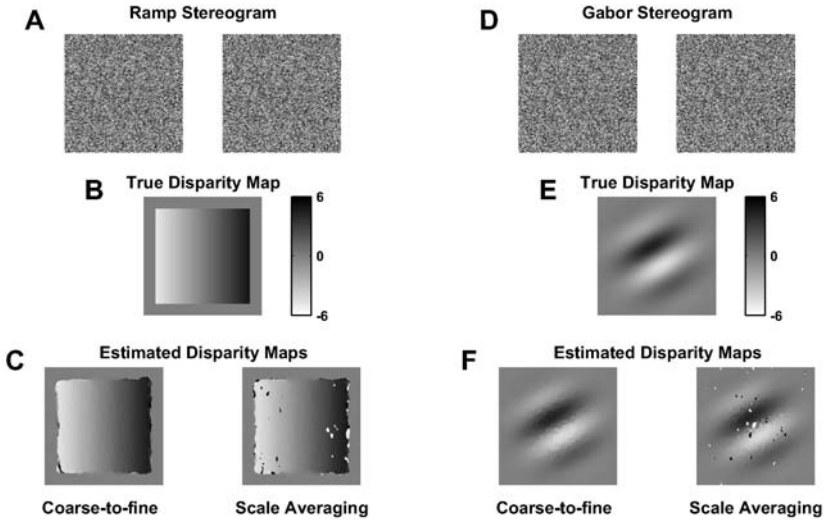


Figure 8: More complex synthetic stereograms. (A) A ramp stereogram with a size of 200×200 pixels. In the central 160×160 area, the disparity varies linearly from -5 pixels to 5 pixels, while the surround has a zero disparity. The gray level of a pixel is randomly chosen between 0 and 1. (B) True disparity map for the ramp stereogram. The white and black colors represent near and far disparities, respectively. (C) Estimated disparity map with the coarse-to-fine algorithm (left panel) and the scale-averaging algorithm (right panel). (D) Gabor stereogram with the same size as the ramp stereogram. The disparity map is created according to a Gabor function: $D(x, y) = D_{\max} \exp(-\frac{x^2+y^2}{2\sigma_D^2}) \cos(\omega_D \sin(\theta_D)x + \omega_D \cos(\theta_D)y + \phi_D)$. The parameters of the Gabor function are: $D_{\max} = 5$ pixels, $\omega_D/2\pi = 1/80$ cyc/pixel, $\sigma_D = 40$ pixels, $\phi_D = 1.39$, and disparity orientation $\theta_D = 30^\circ$. (E) True disparity maps for the Gabor stereogram. (F) Estimated disparity map with the coarse-to-fine algorithm (left panel) and the scale-averaging algorithm (right panel). All the RF parameters applied to both ramp and Gabor stereograms are same as those in Figure 7.

3 Discussion

We have demonstrated through analyses and simulations that in the framework of the disparity energy model, the phase-shift RF mechanism is better suited for disparity computation than the position-shift mechanism. Although the two RF mechanisms generate similar tuning curve properties, the phase-shift mechanism provides more reliable population response curves than the position-shift mechanism does. The phase-shift model, however, has its own limitations, such as the restricted range of detectable disparity (Qian, 1994) and a zero-disparity bias (Qian & Zhu, 1997). To overcome

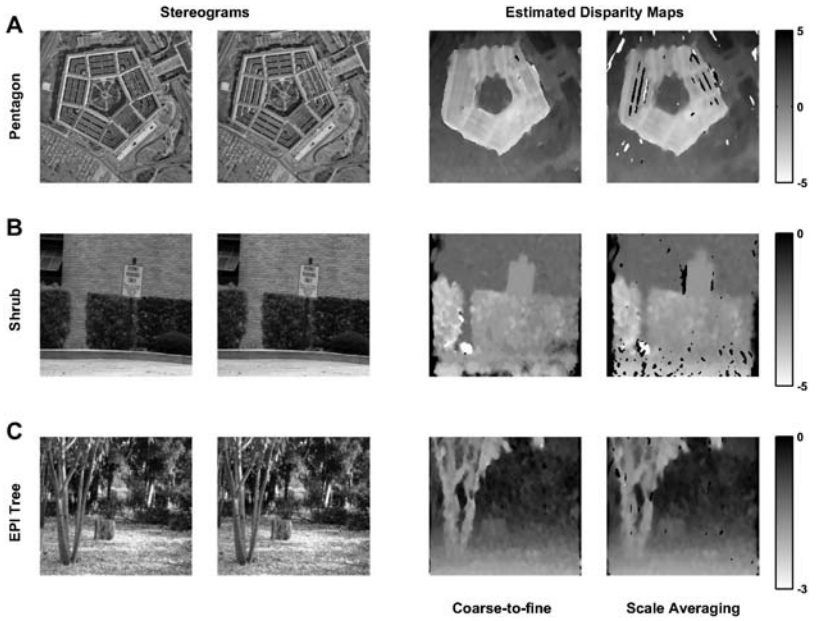


Figure 9: Real-world stereograms and the estimated disparity maps with the coarse-to-fine and the scale-averaging algorithms. (A) Pentagon stereogram. The size of the stereogram is 256 × 256 pixels. (B) Shrub stereogram. The size is 256×240 pixels. (C) Tree stereogram. The size is 256×233 pixels. The stereograms have all been scaled to the same size in the figure for ease of presentation. All the RF parameters are the same as those in Figure 8. For attenuating the interference of the DC component in the stimuli, we deducted the mean luminance from each stereogram before applying the filters. All three stereograms are obtained from the Carnegie Mellon University Image Database.

these problems, we have proposed an iterative procedure in which disparity is always estimated with the phase-shift mechanism but the magnitude of disparity is gradually reduced by introducing a constant position-shift parameter to all cells at each iteration. We then considered integrating information across different RF orientations and spatial scales and found that although a simple averaging procedure (used previously for spatial pooling; Zhu & Qian, 1996; Qian & Zhu, 1997) seems sensible for orientation pooling, it may not be a good approach for scale pooling. Instead, it is better to treat the scale pooling as a coarse-to-fine process. Such a coarse-to-fine process can be easily combined with the iterative procedure with a smaller scale used at each new iteration. The final algorithm is an iterative coarse-to-fine procedure with spatial and orientation pooling incorporated and with both position-shift and phase-shift RF components.

We have applied the algorithm to a variety of stereograms. Our simulations demonstrate that the algorithm can recover both sharp disparity boundaries and gradual disparity changes in the stimuli. However, a couple of problems of the algorithm are also revealed by real-world stereograms. Unlike synthetic patterns, real-world images tend to have regions of very low contrast. Model cells whose RFs cover only a low-contrast area will not respond well, and the results will tend to be very unreliable. The solution to the problem might involve introducing long-range spatial interactions to allow disparity information to propagate from high-contrast regions to low-contrast regions. The challenge is that the interactions should also be selective in order not to smear out real disparity boundaries. It is not clear how to introduce such interactions in a physiologically plausible manner without resorting to a list of ad hoc rules. Another problem with the current algorithm is that if a large scale gives a completely wrong estimation of stimulus disparity, it is impossible for the subsequent smaller scales to correct it. This does not happen often due to the inclusion of spatial and orientation pooling, which greatly improves the reliability of estimation at any scale, but when it does happen, the estimated disparity can be far from the true value. This problem may be solved by a bidirectional information flow across the scales instead of the current unidirectional flow from large to small scales. Indeed, there is psychophysical evidence suggesting that scale interaction is bidirectional (Wilson, Blake, & Halpern, 1991; Smallman & MacLeod, 1994). The scale-averaging approach may be viewed as bidirectional. However, it does not appear to be adequate because the wrong-disparity problem becomes worse with such a method. In general, we found that the coarse-to-fine algorithm is superior to the scale-averaging method, particularly with complex or natural stereograms. But the scale averaging is much easier to implement and may be useful when the precision of the disparity map is not critical.

It is interesting to compare the effects of varying RF orientation and varying RF scale to disparity computation. When the RF orientation is farther away from the vertical, the cells' detectable disparity range increases. This is similar to an increase of the RF scale. However, as the orientation gets closer to horizontal, there will be fewer cycles of RF modulation along the horizontal dimension, tuning curves and population response curves will become broader, and the disparity computation will become less meaningful. This problem does not occur when one uses vertically oriented cells at larger scales to cover a wider disparity range, as long as the spatial-frequency bandwidth (and thus the number of RF modulation cycles) is kept constant across scales. At a fixed scale, cells with different orientations cover the same stimulus area, whereas cells with different scales cover very different stimulus areas. These differences suggest that the orientation pooling and scale pooling should be treated differently, as we did in this article. In particular, since cells with different orientations at a given scale all cover the same image patch, their responses can be averaged together to represent the total

disparity energy of that patch. In contrast, cells with different scales cover different image sizes, and it appears more sensible to use a coarse-to-fine algorithm that can gradually recover disparity details. Also note that the averaging procedure of orientation pooling appears to be more effective for the phase-shift model than for the position-shift model, at least for small disparities (see Figure 5).

We finally discuss the physiological relevance of our model. Our algorithm is based on the disparity energy model, which has been found to provide a good approximation of real complex-cell responses although some discrepancies have also been noted (Freeman & Ohzawa, 1990; Ohzawa et al., 1990, 1997; DeAngelis et al., 1991; Anzai, Ohzawa, & Freeman, 1999b, 1999c; Chen et al., 2001; Livingstone & Tsao, 1999; Cumming, 2002). In addition, experimental evidence indicates that both the phase- and position-shift RF mechanisms are employed by the binocular cells for disparity representation (Hubel & Wiesel, 1962; Bishop & Pettigrew, 1986; Poggio, Motter, Squatrito, & Trotter, 1985; Ohzawa et al., 1990, 1997; DeAngelis et al., 1991; Anzai et al., 1999a; Prince et al., 2000). We (Zhu & Qian, 1996; Qian & Zhu, 1997) have pointed out previously that spatial pooling of quadrature pair responses to construct complex-cell responses is consistent with the fact that the RFs of complex cells are, on average, larger than those of simple cells (Hubel & Wiesel, 1962; Schiller, Finlay, & Volman, 1976). It is also reasonable to incorporate orientation pooling since physiological studies have demonstrated that obliquely oriented cells are tuned to disparity (Poggio & Fischer, 1977; Ohzawa et al., 1996, 1997) and thus must contribute to disparity estimation. This is further supported by the psychophysical finding that binocular correspondence appears to be solved by oriented filters along the contours in a stimulus (Farell, 1998) and does not follow the epipolar constraint (Stevenson & Schor, 1997). On the other hand, there is no evidence for (or against) our specific proposal of using the phase- and position-shift mechanisms in an iterative, coarse-to-fine process. In particular, we do not know if real binocular cells rely on the phase-shift mechanism to estimate disparity and use the position-shift mechanism to reduce the disparity magnitude that the phase mechanism has to process.

Our coarse-to-fine procedure is similar to that proposed by Marr and Poggio (1979), but with one important difference. Marr and Poggio assumed that the large disparity is reduced by vergence eye movement before being processed by smaller scales. Since vergence eye movement shifts stimulus disparity globally, a particular vergence state will not be able to reduce stimulus disparities at all spatial locations, and different vergence state will have to be assumed for each image location. Our model requires only a single vergence state that brings stimulus disparity within a reasonable range; further disparity reduction during the estimation process is carried out by the position-shift mechanism locally at each point. This is consistent with the psychophysical observation that the scale interaction does not depend

on eye movement (Rohaly & Wilson, 1993).[8] In addition, the model is consistent with the finding that with vergence minimized, the fusional range decreases with spatial frequency (Schor, Wood, & Ogawa, 1984) because higher-frequency stimuli benefit less from the guidance of the larger scales. A recent report indicates that individual V1 cells may undergo a coarse-to-fine process over time independent of eye movement (Menz & Freeman, 2003). Therefore, it may not be necessary to implement our coarse-to-fine algorithm with several different populations of cells as we did in our simulations; instead, a single cell population progressively reducing their scale over time might be sufficient.

Appendix: Derivation of Equations 2.6 and 2.7

Based on the disparity energy model, the response of simple cell to a stereo image pair $I(x, y)$ and $I(x + D, y)$ can be written as (Ohzawa et al., 1990; DeAngelis, Ohzawa, & Freeman, 1993; Qian, 1994; Anzai et al., 1999b; Chen et al., 2001):

$$\begin{aligned} r_s &= \left[\iint_{-\infty}^{\infty} \{g_l(x, y)I(x, y) + g_r(x, y)I(x + D, y)\}dxdy \right]^2 \\ &= \left[\iint_{-\infty}^{\infty} \{g_l(x, y)I(x, y) + g_r(x - D, y)I(x, y)\}dxdy \right]^2, \end{aligned} \tag{A.1}$$

where D is the stimulus disparity, and $g_l(x, y)$ and $g_r(x, y)$ are the left and right RFs of simple cell. Note that the full squaring used here is equivalent to a push-pull pair of half-squaring simple cells.

We define the linear filtering of the left and right images as

$$r_{sl} = \iint_{-\infty}^{\infty} g_l(x, y)I(x, y)dxdy \tag{A.2}$$

$$r_{sr} = \iint_{-\infty}^{\infty} g_r(x - D, y)I(x, y)dxdy, \tag{A.3}$$

[8] The model may also be consistent with the finding that the stereo threshold elevation with base disparity is not eliminated by the addition of a low-frequency component (Rohaly & Wilson, 1993) because the low-frequency component itself has an elevated threshold with base disparity and thus becomes unreliable at large base disparity.

so that $r_s = (r_{sl} + r_{sr})^2$. The left RF of a simple cell has the same form for the phase- and position-shift models, and we rewrite equations 2.2 and 2.4 as

$$g_l(x, y) = \cos(\phi) g_{\cos}(x, y) - \sin(\phi) g_{\sin}(x, y), \tag{A.4}$$

where

$$g_{\cos}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \cos(\omega x)$$
$$g_{\sin}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \sin(\omega x).$$

Then, r_{sl} for both phase- and position-shift RF models becomes

$$\begin{aligned} r_{sl} &= \cos(\phi) \iint_{-\infty}^{\infty} g_{\cos}(x, y) I(x, y) dx dy - \sin(\phi) \iint_{-\infty}^{\infty} g_{\sin}(x, y) I(x, y) dx dy \\ &= A \cos(\alpha + \phi), \end{aligned} \tag{A.5}$$

where

$$A = \sqrt{A_1^2 + A_2^2}, \quad \alpha = \arctan(A_2/A_1) \tag{A.6}$$
$$A_1 = \iint_{-\infty}^{\infty} g_{\cos}(x, y) I(x, y) dx dy$$
$$A_2 = \iint_{-\infty}^{\infty} g_{\sin}(x, y) I(x, y) dx dy.$$

Since the right RF of a simple cell has different forms for the two RF models, we consider them separately. In equation A.3, $g_r(x - D, y)$ can be written as

$$\begin{aligned} g_r^{pha}(x - D, y) &= \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{(x - D)^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \\ &\quad \times \cos(\omega(x - D) + \phi + \Delta\phi) \end{aligned} \tag{A.7}$$
$$\begin{aligned} g_r^{pos}(x - D, y) &= \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{(x - (D - d))^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \\ &\quad \times \cos(\omega(x - (D - d)) + \phi) \end{aligned} \tag{A.8}$$

for the phase- and position-shift mechanisms, respectively. In the previous modeling analyses (Qian, 1994; Zhu & Qian, 1996; Fleet et al., 1996; Qian

& Zhu, 1997; Chen et al., 2001), a simplifying assumption is that the horizontal envelope shifts (D in the gaussian term of equation A.7 and $D - d$ in the gaussian term of equation A.8) are small enough to be ignored. For the phase-shift mechanism, the assumption can be satisfied as long as the stimulus disparity D is small enough. However, for the position-shift mechanism, the above assumption requires the stimulus disparity D close to the position shift d. This is obviously a more stringent requirement. To compare the two mechanisms, the envelope shifts should be considered. Here, we take the first-order Taylor expansion in D/σ_x and $(D - d)/\sigma_x$ for the two mechanisms, respectively:

$$\exp\left(-\frac{(x-D)^2}{2\sigma_x^2}\right) \approx \exp\left(-\frac{x^2}{2\sigma_x^2}\right) + \frac{xD}{\sigma_x^2}\exp\left(-\frac{x^2}{2\sigma_x^2}\right) \tag{A.9}$$

$$\exp\left(-\frac{(x-(D-d))^2}{2\sigma_x^2}\right) \approx \exp\left(-\frac{x^2}{2\sigma_x^2}\right) + \frac{x(D-d)}{\sigma_x^2} \times \exp\left(-\frac{x^2}{2\sigma_x^2}\right). \tag{A.10}$$

Under these approximations, the right RFs in equiations A.7 and A.8 can be written as

$$g_r^{pha}(x-D,y) \approx \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{(x-D)^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \times \cos(\omega(x-D)+\phi+\Delta\phi)\left(1+\frac{xD}{\sigma_x^2}\right) \tag{A.11}$$

$$g_r^{pos}(x-D,y) \approx \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{(x-(D-d))^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \times \cos(\omega(x-(D-d))+\phi)\left(1+\frac{x(D-d)}{\sigma_x^2}\right). \tag{A.12}$$

Then, similar to the derivation for equation A.5, we have

$$r_{sr}^{pha} \approx A\cos(\alpha+\phi+\Delta\phi-\omega D) + \frac{D}{\sigma_x}B\cos(\beta+\phi+\Delta\phi-\omega D) \tag{A.13}$$

$$r_{sr}^{pos} \approx A\cos(\alpha+\phi-\omega(D-d)) + \frac{D-d}{\sigma_x}B\cos(\beta+\phi-\omega(D-d)) \tag{A.14}$$

for the phase- and position-shift models, respectively, with

$$B = \sqrt{B_1^2+B_2^2}, \quad \beta = \arctan(B_2/B_1) \tag{A.15}$$

$$B_1 = \iint_{-\infty}^{\infty} \frac{x}{\sigma_x} g_{\cos}(x, y) I(x, y) dx dy$$

$$B_2 = \iint_{-\infty}^{\infty} \frac{x}{\sigma_x} g_{\sin}(x, y) I(x, y) dx dy.$$

The final expressions for simple cell response are:

$$r_s^{pha} \approx \left[2A \cos\left(\alpha + \phi + \frac{\Delta\phi - \omega D}{2}\right) \cos\left(\frac{\Delta\phi - \omega D}{2}\right) + \frac{D}{\sigma_x} B \cos(\beta + \phi + \Delta\phi - \omega D) \right]^2 \tag{A.16}$$

$$r_s^{pos} \approx \left[2A \cos\left(\alpha + \phi - \frac{\omega(D-d)}{2}\right) \cos\left(\frac{\omega(D-d)}{2}\right) + \frac{D-d}{\sigma_x} B \cos(\beta + \phi - \omega(D-d)) \right]^2. \tag{A.17}$$

Based on the well-known quadrature pair method for the energy models (Adelson & Bergen, 1985; Watson & Ahumada, 1985; Pollen, 1981; Ohzawa et al., 1990; Emerson, Bergen, & Adelson, 1992; Qian, 1994), the complex cell receives the inputs from two simple cells, both with identical $\Delta\phi$, but their ϕ differing by $\pi/2$. The resulting complex-cell responses to the first-order approximation are:

$$r_q^{pha} \approx 4A^2 \cos^2\left(\frac{\omega D - \Delta\phi}{2}\right) + \frac{D}{\sigma_x} 4AB \cos\left(\frac{\omega D - \Delta\phi}{2}\right) \times \cos\left(\alpha - \beta + \frac{\omega D - \Delta\phi}{2}\right) \tag{A.18}$$

$$r_q^{pos} \approx 4A^2 \cos^2\left(\frac{\omega(D-d)}{2}\right) + \frac{D-d}{\sigma_x} 4AB \cos\left(\frac{\omega(D-d)}{2}\right) \times \cos\left(\alpha - \beta + \frac{\omega(D-d)}{2}\right) \tag{A.19}$$

These are equations 2.6 and 2.7.

For a relatively thin bar at x_o, $I(x, y) \approx \delta(x - x_o, y)$, and equations A.6 and A.15 show $\alpha \approx \beta$.

Acknowledgments

This work was supported by NIH grant MH54125. We thank the two anonymous reviewers for their helpful comments.

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, *2*, 284–299.

Anzai, A., Ohzawa, I., & Freeman, R. D. (1997). Neural mechanisms underlying binocular fusion and stereopsis: Position vs. phase. *Proc. Nat. Acad. Sci. USA*, *94*, 5438–5443.

Anzai, A., Ohzawa, I., & Freeman, R. D. (1999a). Neural mechanisms for encoding binocular disparity: receptive field position vs. phase. *J. Neurophysiol.*, *82*, 874–890.

Anzai, A., Ohzawa, I., & Freeman, R. D. (1999b). Neural mechanisms for processing binocular information: I. Simple cells. *J. Neurophysiol.*, *82*, 891–908.

Anzai, A., Ohzawa, I., & Freeman, R. D. (1999c). Neural mechanisms for processing binocular information: II. Complex cells. *J. Neurophysiol.*, *82*, 909–924.

Bishop, P. O., & Pettigrew, J. D. (1986). Neural mechanisms of binocular vision. *Vision Res.*, *26*, 1587–1600.

Chen, Y., Wang, Y., & Qian, N. (2001). Modeling V1 disparity tuning to time-varying stimuli. *J. Neurophysiol.*, *86*(1), 143–155.

Cumming, B. G. (2002). An unexpected specialization for horizontal disparity in primate primary visual cortex. *Nature*, *418*, 633–636.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am. A*, *2*, 1160–1169.

DeAngelis, G. C., Ohzawa, I., & Freeman, R. D. (1991). Depth is encoded in the visual cortex by a specialized receptive field structure. *Nature*, *352*, 156–159.

DeAngelis, G. C., Ohzawa, I., & Freeman, R. D. (1993). Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II. Linearity of temporal and spatial summation. *J. Neurophysiol.*, *69*, 1118–1135.

Emerson, R. C., Bergen, J. R., & Adelson, E. H. (1992). Directionally selective complex cells and the computation of motion energy in cat visual cortex. *Vision Res.*, *32*, 203–218.

Farell, B. (1998). Two-dimensional matches from one-dimensional stimulus components in human stereopsis. *Nature*, *395*, 689–692.

Fleet, D. J., Wagner, H., & Heeger, D. J. (1996). Encoding of binocular disparity: Energy models, position shifts and phase shifts. *Vision Res.*, *36*, 1839–1858.

Freeman, R. D., & Ohzawa, I. (1990). On the neurophysiological organization of binocular vision. *Vision Res.*, *30*, 1661–1676.

Hubel, D. H., & Wiesel, T. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *J. Physiol.*, *160*, 106–154.

Jones, J. P., & Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in the cat striate cortex. *J. Neurophysiol.*, *58*, 1187–1211.

Livingstone, M. S., & Tsao, D. T. (1999). Receptive fields of disparity-selective neurons in macaque striate cortex. *Nature Neurosci.*, *2*(9), 825–832.

Marr, D., & Poggio, T. (1979). A computational theory of human stereo vision. *Proc. R. Soc. Lond. B*, *204*, 301–328.

Matthews, N., Meng, X., Xu, P., & Qian, N. (2003). A physiological theory of depth perception from vertical disparity. *Vision Res.*, *43*, 85–99.

Menz, M. D., & Freeman, R. D. (2003). Stereoscopic depth processing in the visual cortex: A coarse-to-fine mechanism. *Nature Neurosci.*, *6*(1), 59–65.

Mikaelian, S., & Qian, N. (1997). Disparity attraction and repulsion in a two-dimensional stereo model. *Soc. Neurosci. Abs.*, *23*, 569.

Mikaelian, S., & Qian, N. (2000). A physiologically-based explanation of disparity attraction and repulsion. *Vision Res.*, *40*, 2999–3016.

Ohzawa, I., DeAngelis, G. C., & Freeman, R. D. (1990). Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science*, *249*, 1037–1041.

Ohzawa, I., DeAngelis, G. C., & Freeman, R. D. (1996). Encoding of binocular disparity by simple cells in the cat's visual cortex. *J. Neurophysiol.*, *75*, 1779–1805.

Ohzawa, I., DeAngelis, G. C., & Freeman, R. D. (1997). Encoding of binocular disparity by complex cells in the cat's visual cortex. *J. Neurophysiol.*, *77*, 2879–2909.

Poggio, G. F., & Fischer, B. (1977). Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey. *J. Neurophysiol.*, *40*, 1392–1405.

Poggio, G. F., Gonzalez, F., & Krause, F. (1988). Stereoscopic mechanisms in monkey visual cortex: Binocular correlation and disparity selectivity. *J. Neurosci.*, *8*, 4531–4550.

Poggio, G. F., Motter, B. C., Squatrito, S., & Trotter, Y. (1985). Responses of neurons in visual cortex (V1 and V2) of the alert macaque to dynamic random-dot stereograms. *Vision Res.*, *25*, 397–406.

Pollen, D. A. (1981). Phase relationship between adjacent simple cells in the visual cortex. *Nature*, *212*, 1409–1411.

Prince, S. J. D., Cumming, B. G., & Parker, A. J. (2000). Range and mechanism of encoding of horizontal disparity in macaque V1. *J. Neurophysiol.*, *87*, 209–221.

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

Qian, N. (1997). Binocular disparity and the perception of depth. *Neuron*, *18*, 359–368.

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

Rohaly, A. M., & Wilson, H. R. (1993). Nature of coarse-to-fine constraints on binocular fusion. *J. Opt. Soc. Am. A*, *10*, 2433–2441.

Sanger, T. D. (1988). Stereo disparity computation using Gabor filters. *Biol. Cybern.*, *59*, 405–418.

Schiller, P. H., Finlay, B. L., & Volman, S. F. (1976). Quantitative studies of single-cell properties in monkey striate cortex: I. Spatiotemporal organization of receptive fields. *J. Neurophysiol.*, *39*, 1288–1319.

Schor, C. M., & Wood, I. (1983). Disparity range for local stereopsis as a function of luminance spatial frequency. *Vision Res.*, *23*, 1649–1654.

Schor, C. M., Wood, I., & Ogawa, J. (1984). Binocular sensory fusion is limited by spatial resolution. *Vision Res.*, *24*(7), 661–665.

Smallman, H. S., & MacLeod, D. I. (1994). Size-disparity correlation in stereopsis at contrast threshold. *J. Opt. Soc. Am. A*, *11*, 2169–2183.

Stevenson, S. B., & Schor, C. M. (1997). Human stereo matching is not restricted to epipolar lines. *Vision Res.*, *37*, 2717–2773.
Teich, A. F., & Qian, N. (2003). Learning and adaptation in a recurrent model of V1 orientation selectivity. *J. Neurophysiol.*, *89*, 2086–2100.
Tsai, J. J., & Victor, J. D. (2003). Reading a population code: A multi-scale neural model for representing binocular disparity. *Vision Res.*, *43*, 445–466.
Watson, A. B., & Ahumada, A. J. (1985). Model of human visual-motion sensing. *J. Opt. Soc. Am. A*, *2*, 322–342.
Wilson, H. R., Blake, R., & Halpern, D. L. (1991). Coarse spatial scales constrain the range of binocular fusion on fine scales. *J. Opt. Soc. Am. A*, *8*, 229–236.
Zhu, Y., & Qian, N. (1996). Binocular receptive fields, disparity tuning, and characteristic disparity. *Neural Comput.*, *8*, 1611–1641.

Received July 30, 2003; accepted January 30, 2004.