Reliable disparity estimation through selective integration

MICHAEL S. GRAY,[1,2] ALEXANDRE POUGET,[3] RICHARD S. ZEMEL,[4]
STEVEN J. NOWLAN,[5] AND TERRENCE J. SEJNOWSKI[2,6]

[1]Department of Cognitive Science, University of California, San Diego, La Jolla
[2]Howard Hughes Medical Institute, Computational Neurobiology Laboratory, The Salk Institute, San Diego
[3]Georgetown Institute for Cognitive and Computational Sciences, Georgetown University, Washington, DC
[4]Department of Psychology, University of Arizona, Tucson
[5]Lexicus, Inc., Palo Alto
[6]Department of Biology, University of California, San Diego, La Jolla

Abstract

A network model of disparity estimation was developed based on disparity-selective neurons, such as those found in the early stages of processing in the visual cortex. The model accurately estimated multiple disparities in regions, which may be caused by transparency or occlusion. The selective integration of reliable local estimates enabled the network to generate accurate disparity estimates on normal and transparent random-dot stereograms. The model was consistent with human psychophysical results on the effects of spatial-frequency filtering on disparity sensitivity. The responses of neurons in macaque area V2 to random-dot stereograms are consistent with the prediction of the model that a subset of neurons responsible for disparity selection should be sensitive to disparity gradients.

Keywords: Binocular disparity, Stereopsis, Disparity gradient, Selection, Macaque area V2

Introduction

A wide variety of computational models have been proposed to explain how binocular disparity is computed from left–right image pairs (Blake & Wilson, 1991; Weinshall & Malik, 1995). Disparity estimation is made difficult by the inconsistency of local disparity information, which may be sparse and noisy. Hence, there are two conflicting demands when estimating disparity in an image region: the need to spatially average to get an accurate estimate, and the problem of not averaging over discontinuities. Research in computer vision has often focused on a two-stage process to solving this problem: (1) find exact correspondence between matching points in the two images, and (2) compute the disparity based on the patterns of correspondences.

In standard approaches to the first-stage, finding correspondences (e.g. edge-based and correlation-based models), it is assumed that the goal is to provide an accurate disparity estimate for *every* region of the image. These approaches typically do not emphasize the *reliability* of the disparity estimate, which may be reduced in conditions where there is substantial occlusion, transparency, or noise. Also, many of these computational methods are iterative (e.g. Marr & Poggio, 1976, 1979; Yuille et al., 1991) which may be a drawback for a system attempting to respond dynamically to the world in real time.

This paper presents a model of disparity estimation based on a modular neural network architecture known as a *mixture-of-experts* architecture (Jacobs et al., 1991), in which a selection pathway gates the output of local disparity information from small adjacent regions of space. This model is motivated by the observation that local correlation measures are insufficient because they convey no information about the reliability of a particular disparity measurement. By contrast, our model uses a separate *selection* mechanism to determine which locations of the visual input have consistent disparity information. This is especially important in viewing situations in which disparity estimation is not straightforward, such as when multiple objects are present in the image, and particularly if one object occludes another. Reliability estimates may also be useful when dealing with stimuli with high background noise, poorly defined edges, or transparent surfaces. In this paper, we assess the importance of such an approach by generating several sets of stimuli in which one or several of these conditions occur. We compared the results for the mixture-of-experts model against two more standard approaches, and found that the mixture-of-experts model outperformed these alternatives on all the data sets tested.

The paper is organized as follows. First, the mixture-of-experts model is described in detail, along with the two more standard approaches. A procedure for obtaining disparity estimates from our model is outlined, and a detailed specification of the data sets used to optimize and test the model is also provided. Second, the results of the different models on a variety of data sets are reported. Third, a more detailed analysis of the response properties of the mixture-of-experts model is reported in Analysis of Model Phenomena.

Correspondence and reprint requests to: Michael S. Gray, Howard Hughes Medical Institute, Computational Neurobiology Laboratory, The Salk Institute, P.O. Box 85800, San Diego, CA 92186-5800, USA.

Finally, the results are discussed in the context of related work in stereopsis.

Model descriptions and methods

Mixture-of-experts model

The goal of the current model is to estimate the disparities present in a *small patch* of the image. It is assumed that processing related to *spatial localization* of objects occurs in other regions of the visual cortex. In other words, the model trades off spatial accuracy to obtain disparity accuracy.

The model of stereopsis used here is based on a filter model for motion detection in area MT (Nowlan & Sejnowski, 1994, 1995). The motion model was adapted to stereopsis by changing the time domain of the motion model to the left/right image domain for stereopsis. The stereo model consisted of several stages, and computed its output using only feedforward processing. The model has a mixture-of-experts architecture (Jacobs et al., 1991). This is a system of separate networks (expert nets) that specialize in different input patterns. A separate gating network learns which expert is best for the different kinds of input patterns. In the first stage of the model, the input was convolved with a set of disparity energy filters. The output of the filters then became the input to two different secondary pathways: (1) the local disparity (expert) networks, and (2) the selection (gating) networks. The output of the model was a disparity value that was the product of the outputs of the two secondary pathways. Because it was not known *a priori* what kinds of disparity signals would be valuable for both disparity estimation and segmentation, an optimization algorithm was used to find the best parameter values for the model. A schematic diagram of the model is shown in Fig. 1. The four important parts of the model will be described in detail: the retinal layer and disparity energy filters, the local disparity networks, the selection networks, and the output layer.

Retina and disparity energy filters

The retinal layer in the model consisted of two one-dimensional arrays 82 pixels in length for the right eye and left eye images. One-dimensional retinas were used for computational simplicity. The model would generate similar results for two-dimensional images because the spatial enhancement mechanism of the selection networks (described in Selection networks) generalizes directly to two-dimensional image representations (as shown in Nowlan & Sejnowski, 1994, 1995). The one-dimensional images were the inputs to disparity energy filters (Ohzawa et al., 1990, 1996, 1997), which generalized the motion energy filters first proposed by Adelson and Bergen (1985). Although phase-based disparity filters were chosen for preprocessing the input, similar output from the model would also be expected using filters based on shifted receptive fields (Fleet et al., 1996; Zhu & Qian, 1996).

At the energy filter layer, there were 51 receptive-field locations which received input from overlapping regions of the retina. At each of these receptive-field locations, there were 30 complex cells (three spatial frequencies × 10 phase differences), and each complex cell received input from four simple cells (two in phase, two quadrature) that were linearly combined. The two in-phase simple cells, as well as the two quadrature simple cells differed from each other by 180 deg. Each of these simple cells received input from a pair of subunits that can be described mathematically as Gabor functions (Gabor, 1946; Daugman, 1985) differing by a phase parameter ϕ:
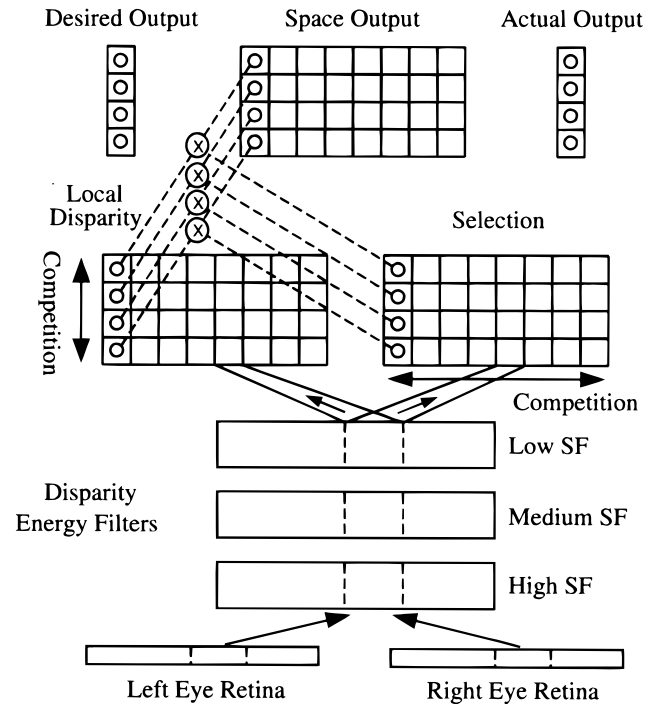

Fig. 1. Schematic diagram of the stereo model. Patterns of activation on the retina were combined in the disparity energy units at several different spatial frequencies and phases. The output of the disparity energy units were the input to both the local disparity pathway and the selection pathway. The outputs of the two pathways were combined multiplicatively to generate estimates of disparity across space (Space Output). By summing across space for each of the four disparity values in the output, the actual output of the model was obtained. For the local disparity, selection, and output parts of the model, the vertical axis represented disparity and the horizontal axis was space.

$$g(x,\phi) = \frac{1}{\sqrt{2}\pi^{(3/2)}\sigma_x} \exp\left(-\frac{x^2}{2\sigma_x^2}\right)\sin(2\pi\omega_x x + \phi) \quad (1)$$

where σ_x was the size of the Gaussian window, and ω_x was the filter center frequency. A simple cell pair (differing in phase by 90 deg) is shown in the upper panels of Fig. 2. The right panel of this figure shows schematically how simple cell output is combined at the complex cell. These disparity filters were implemented in the same manner as Ohzawa et al. (1990). The output of a simple cell (at a given phase and spatial frequency) was computed by convolving the right eye image with the right filter, the left eye image with the left filter, and then adding them. Simple cell output was then half-wave rectified (truncated and squared). Two of the four simple cell pairs were in phase, while the other two were in quadrature phase. The three spatial frequencies were each separated by an octave: 0.25, 0.125, and 0.0625 cycles/pixel (c/p). The 10 phase differences were equally spaced over a range between 0 and $\pi/2$. The output of the energy filters in response to a sample stimulus is shown in Fig. 3. The outputs of these complex cells (or disparity energy filters) should be considered the true input to the network.

The Gaussian windows of the filters in the model were inversely proportional to the spatial frequency. The variances of the windows were 1.0 for 0.25 c/p, 4.0 for 0.125 c/p, and 16.0 for
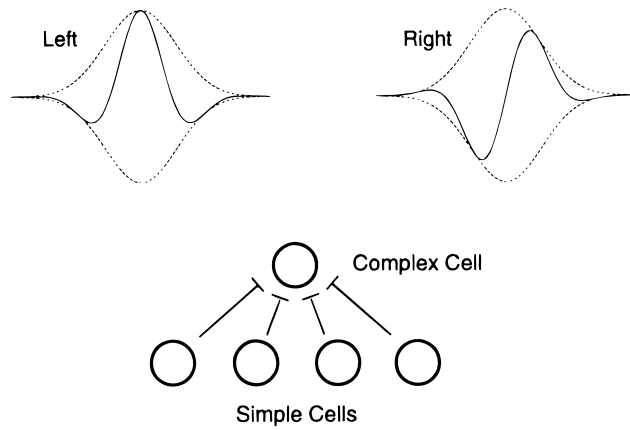
Fig. 2. The upper panels show the left and right eye filters for a binocular simple cell. These filters differed in phase by 90 deg. The lower panel shows how simple cell output was combined into a complex cell.

0.0625 c/p. Because there is evidence that cortical cells respond to relative contrast in a scene, rather than absolute contrast (Bonds, 1991), the outputs of the disparity energy units were normalized using a soft-max nonlinearity (Bridle, 1989):

$$\hat{E}_i(x) = \frac{\exp[E_i(x)]}{\sum_j \exp[E_j(x)]} \tag{2}$$

where j indexed the 10 complex cells with different phase shifts at a single spatial location within a single spatial-frequency band, $E_i(x)$ was the initial output of the complex cell, and $\hat{E}_i(x)$ was the normalized output. This normalization occurred at each spatial location within each frequency band. Activity in other parts of the model was also normalized with this soft-max computation.

Local disparity networks

In the local disparity pathway, there were eight receptive-field locations, and each received a weighted input from the 30 complex cells at each of nine disparity energy locations in the preceding layer (270 inputs total). Input to each of these eight receptive-field locations overlapped by three locations in the disparity energy layer. Weights were shared across all receptive-field locations for each disparity. Each receptive-field location at the local disparity layer contained a pool of four disparity-tuned units, and functioned as a single network. These four disparity-tuned units each received 270 inputs from the disparity energy layer. Each receptive-field location was intended to provide strong support for only one disparity. This constraint was enforced as the result of competition
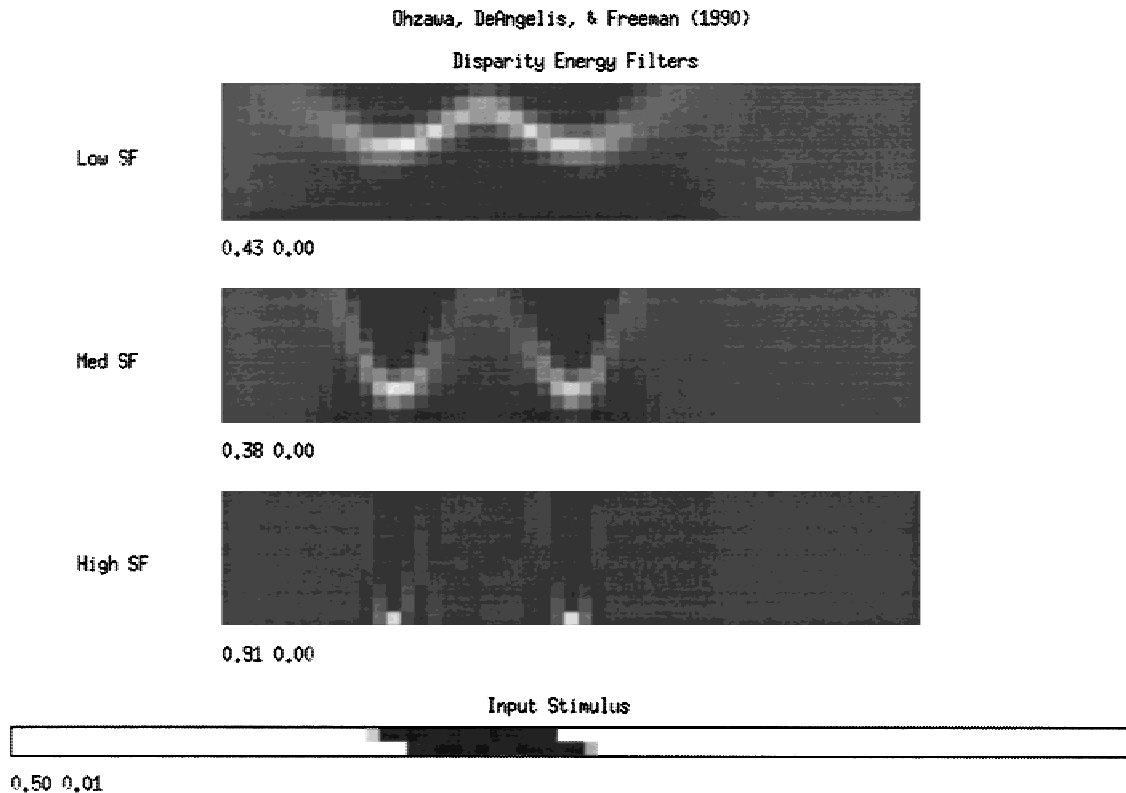
Fig. 3. Disparity energy filter output in response to a single object at a disparity of approximately 2 pixels. At the bottom of the figure is the input stimulus with the right eye image in the top row and the left eye image in the bottom row. The three regions above the stimulus represented the output of the disparity energy filters at three different spatial frequencies. Within each spatial-frequency band, the horizontal axis represented space, while the vertical axis represented 10 phase differences, equally spaced between 0 and $\pi/2$ radians. The top row of each band had a 0-radian phase difference, while the bottom row had a phase difference of $\pi/2$. The maximum disparity to which the low SF(spatial frequency) band was responsive was 4 pixels (at the $\pi/2$ radians phase difference). The medium SF band responded up to 2 pixels, and the high SF band responded to disparities of 1 pixel or less. In this example, both edges of the object gave strong signals in each of the SF bands. The pair of numbers below each frequency band were the maximum and minimum values within that band. White indicated the highest value, black the smallest.

among the four disparity-tuned units at each receptive-field location using the soft-max nonlinearity [Eq. (2)]. This competition induced by the soft-max can be considered analogous to the uniqueness constraint of Marr and Poggio (1976).

In summary, each local disparity network corresponded to a specific spatial location in the input (Fig. 1). The soft-max competition occurred *within* each local disparity network, and insured that only one disparity was strongly activated.

Selection networks

As in the local disparity networks, the selection networks were organized into a grid of eight receptive-field locations with a pool of four disparity-tuned units at each location, and weights were shared across all receptive-field locations for each disparity. The four units at each location in the selection layer represented the local support for each of the different disparity hypotheses.

Since the goal of this pathway was to select the spatial locations with the most reliable evidence for each disparity, it is useful to think of the selection networks as four separate layers that responded to a specific disparity across all regions of the image. As in the disparity energy and local disparity pathways, the outputs of the selection networks were normalized with the soft-max operation. This competition, however, occurred separately for each of the four disparities in a global fashion *across space*—that is, across all spatial locations (Fig. 1). In summary, each selection network corresponded to a *specific disparity*, and its goal was to find the spatial location with the best support for that disparity. In comparison, the local disparity networks normalized responses locally (across disparities), as described above.

Output layer

The output of the model (as shown in Fig. 1, and in subsequent figures) had a spatial map generated by pairwise multiplying the activity of units in the local disparity and selection pathways. The global output was generated from the space output by summing these products across all spatial locations for each of the four disparities:

$$O_k = \sum_x L_k(x) S_k(x) \qquad (3)$$

where O_k was the global (space-independent) evidence for disparity k, $L_k(x)$ was the local disparity output for disparity k at location x, and $S_k(x)$ was the selection output for disparity k at location x. This resulted in a distributed representation for disparity that was independent of the spatial location of the disparity in the image.

Training

The weights from the retina to the disparity energy layer were fixed. The weights in the local disparity and selection pathways, however, were initialized with small random values, and then optimized using the mixture-of-experts learning algorithm (Jacobs et al., 1991). Training was stopped when the performance of the model stopped improving. The difference between the activities of the local disparity units and the known disparities in the image provided a measure of performance of the local disparity pathway. Those local disparity units which had activity levels close to the known disparities in the real scene adjusted their weights to improve their prediction even more. The selection units, on the other hand, were trained to predict what kinds of features in the image were likely to lead to good disparity estimates. This functional division of labor was reflected in the common objective function

for the network. Specifically, the weights to the local disparity and selection units were adjusted according to the following error function (Jacobs et al., 1991):

$$E_k = -\log \sum_x S_k(x) \exp\left[-\frac{1}{2} \|D_k(x) - L_k(x)\|^2 \right] \qquad (4)$$

where E_k represented the error on a single case for the output unit tuned to disparity k, $S_k(x)$ was the output of the selection network for spatial location x and disparity k, $L_k(x)$ was the output of the local disparity network for location x and disparity k, and $D_k(x)$ was the target output. The learning rule for each pathway was formed by taking the derivative of this error function with respect to the activities in the local disparity and selection pathways [see Nowlan & Sejnowski (1994) for further details].

Comparison models

For comparison with the mixture-of-experts model, a single-pathway model trained with backpropagation (Rumelhart et al., 1986) and a cross-correlation model were developed. The single-pathway model had 32 hidden units between the same input filters as in the mixture-of-experts and the same output layer. Units in the hidden layer had localized receptive fields identical to those in the mixture-of-experts model; weights were shared across the different receptive-field locations. In addition, each output unit received connections only from those hidden units that became tuned to the same disparity. All units in the hidden and output layers had logistic activation functions:

$$a_i = \frac{1}{1 + \exp(-net_i)} \qquad (5)$$

The cross-correlation (Stevenson et al., 1991; Cormack et al., 1991; Gonzalez & Woods, 1992) of a particular stimulus was defined for the four integral pixel disparity values ($d = 0, 1, 2,$ and 3 pixels):
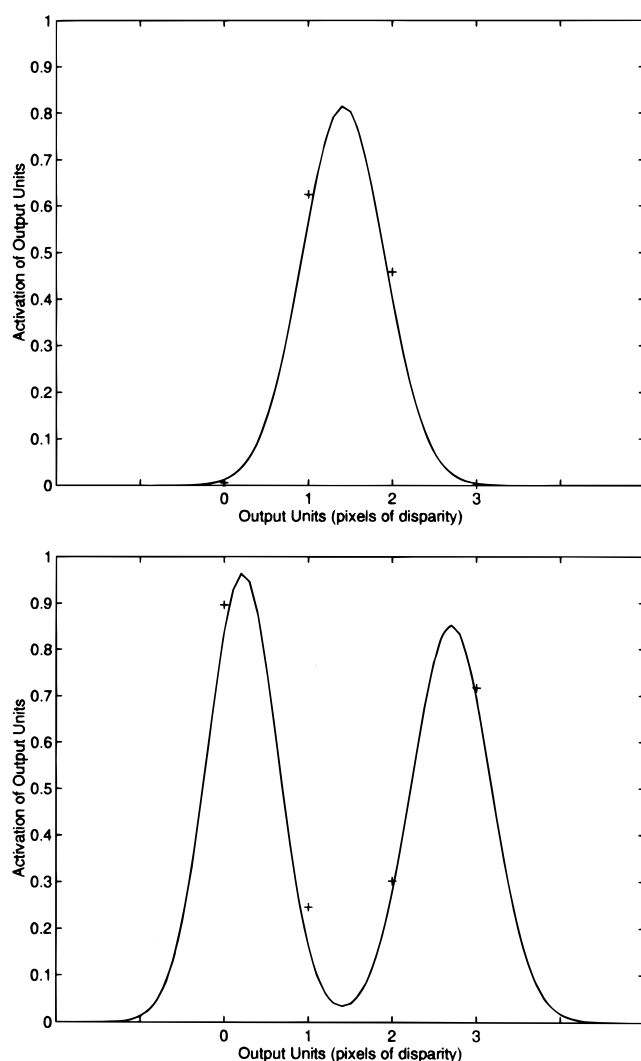
$$C_{L,R}(d) = \frac{\displaystyle\sum_{x=1}^{M-d} I_L(x+d) I_R(x)}{M-d} \qquad (6)$$

where d was the disparity (in pixels) between the left image I_L and the right image I_R, x indexed the spatial locations in the image, and M was the length of the image in pixels. This correlation was *unbiased*—it was normalized by the number of terms that contributed to the sum. This normalization was important for obtaining accurate disparity estimates. The four resulting cross-correlation values, $C_{L,R}(d)$ for $d = 0,1,2,3$, were then linearly normalized to sum to 1.0. These four values typically did not differ greatly in magnitude because most of the image (82 pixels in length) was constant luminance background, which contributed a constant value to the sum in the numerator of eq. (6). To make this disparity signal more salient, a soft-max operation [eq. (2)] was performed on these four output values. Disparity estimates were obtained by fitting these outputs with a Gaussian using the procedure described in Disparity estimation.

Disparity estimation

To compare the model's outputs with the known disparities of the objects present in the input, the model's estimate of disparity needs

to be computed from the four output activation values (labeled "Actual Output" in figures illustrating the model). For stimuli containing a single object, a Gaussian was fit to these outputs using a least-squared-error minimization procedure in MATLAB. A Gaussian was chosen because the disparity tuning curves of the outputs were Gaussian. The mean and variance of the Gaussian were iteratively adjusted to find the least-squares fit with the four data points in the output. The initial value of the mean was 1.0, and the initial value of the variance was 1.0. An example of this fitting procedure is illustrated in the upper panel of Fig. 4. For this particular stimulus, the input disparity was 1.45 pixels. The model's estimate of disparity, as indicated by the mean of the Gaussian, was 1.42 pixels. Through the remainder of this paper, references to the model's disparity estimate are based on this Gaussian-fitting procedure. When two objects were present in the input, the four

activation values of the output were fit with the sum of two Gaussians, as shown in the lower panel of Fig. 4. In this case, the means were initialized to 1.0 and 2.0 pixels, and the variances were initialized to 1.0. In this example, the input stimulus disparities of the two objects were 0.24- and 2.69-pixel disparity. The model estimates of disparity for this stimulus, based on the sum of Gaussians fit of the output activations, were 0.21- and 2.70-pixel disparity. There is no mechanism in the current model to indicate whether one object or two objects are present in the input. This Gaussian-fitting procedure is, of course, subject to local minima. However, given the accuracy of disparity estimation exhibited by the model (as described below), local minima were not believed to be a problem, and no precautions were taken to avoid them.

Using the mean of the Gaussian obtained from this least-squared error-fitting procedure, the model's ability to discriminate between different disparities can be determined. The discrimination threshold, which can be measured both psychophysically in primates and computationally in a model, is defined as the disparity difference at which one can correctly detect a difference in depth 75% of the time. The disparity-discrimination threshold for the model was determined using signal-detection theory (Green & Swets, 1966). An alternative, more physiologically plausible way to estimate the disparity from a noisy population code is to use a network with lateral interactions and a separate inhibitory population (Pouget & Zhang, 1997).
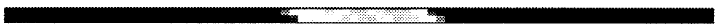
Data sets

Five data sets were used to train and test the stereo models described in this paper. In this section, each of these data sets is described in detail. In each data set, a single training pattern consisted of a right eye image, a left eye image, and desired output values. Sample images from each of the different data sets are shown in Table 1. There are training and test stimuli for each data set. The number of training stimuli is indicated in the description below. All test stimuli data sets contained 100 exemplars, unless otherwise indicated.

1. *Single Object*. This data set contained 100 stimuli generated in the following manner. The images all started with a zero-disparity background (of constant luminance 0.5). Then, a randomly chosen object between 10 and 25 pixels in size was included in the image at a real-valued location with a real-valued (nonintegral) disparity between 0.0 and 3.0 pixels. At the edge of the object, that is, between integral pixel locations, luminance values were linearly interpolated. The luminance values in this single object were randomly chosen from a uniform distribution, and were either in the range [0.0, 0.1] or [0.9, 1.0]. This object always appeared in the fronto-parallel plane; that is, the disparity did not change at any point on the object.

2. *Multiple Objects*. This data set contained 250 stimuli. Half of the patterns contained a single object in the image, while the other half contained two objects. The single object stimuli were generated as described above in the Single Object data set. The training patterns with two objects also had a constant luminance (0.5) zero-disparity background and the objects had a size that was randomly chosen between 10 and 25 pixels. One object had luminance values in the range [0.0, 0.1], while the other had luminance values from [0.9, 1.0]. The differences in luminance were for visualization

Fig. 4. Upper panel: The Gaussian fit for a stimulus containing a single object. The plus symbols ('+') indicate the activation values of the four output units tuned to 0-, 1-, 2-, and 3-pixel disparity. For this particular test stimulus, the input disparity was 1.45 pixels; the model's estimate (the mean of the Gaussian) was 1.42 pixels. Lower panel: The sum of Gaussians fit for a stimulus containing two objects. The test stimulus disparities were 0.24 and 2.69 pixels. The model estimates of disparity were 0.21 and 2.70 pixels.

Table 1. *Sample stimuli for each of the data sets*

Stimulus Type	Sample Stimulus
Single	
Double	
Noise	
Random Dot	
Transparent	
Real	

purposes—model performance did not depend on this. The locations of the two objects in the image were real-valued and were randomly chosen subject to the constraint that they differed by at least 6 pixels. Because the locations were chosen independently, often one of the two objects occluded the other one. The disparities of the two objects were also randomly and independently chosen, but were included in the data set only if the disparities of the two objects differed by at least 1.5 pixels. The psychophysical finding of disparity averaging (Parker & Yang, 1989) was the primary motivating factor for requiring this disparity difference between the two objects.

3. *Noise*. This data set contained 250 stimuli. First, a uniform random background (in the range [0.0, 1.0]) was written on both the left and right eye images. This background was uncorrelated between the left and right eyes. Then, an object (10–25 pixels in length) was generated from the same uniform random distribution (in the range [0.0, 1.0]). This object was written in the image at a randomly chosen real-valued location with a real-valued disparity in the range [0.0, 3.0]. This data set differed from the Single Object data set in that it had an uncorrelated noise background, and the luminance edge between the object and the background was difficult to detect. Extracting disparity information from stimuli in this data set was an especially difficult task.

4. *Binary*. This data set had binary random-dot stereograms (40% dot density) that included transparent surfaces. Luminance values were limited to the integral values of 0 and 1. Transparent random-dot stimuli were particularly challenging for models of stereopsis so this data set contained 1400 stimuli. 40% of the stimuli contained a single integral disparity shift of 0, 1, 2, or 3 pixels. The remaining stimuli represented transparent stimuli and consisted of dots at two different disparities. Specifically, 50% of the dots were shifted by an integral disparity value (0, 1, 2, or 3 pixels) while the remaining dots were shifted by a different disparity. The disparity difference between the two surfaces was always at least 2 pixels. Thus, the following three combinations of disparity (for the two transparent surfaces) were present in the data set: (1) 0- and 2-pixel disparity, (2) 0- and 3-pixel disparity, and (3) 1- and 3-pixel disparity. Dots at different disparities were not separated spatially; they were interleaved across space. Psychophysically, these stimuli corresponded to transparent fronto-parallel surfaces at different depths.

5. *Real*. This data set contained 16 stimuli from real-world stereoscopic images. These were obtained from the Cali-

brated Imaging Laboratory (www.cs.cmu.edu/afs/cs.cmu.edu/project/cil/ftp/html/cil-ster.html) at Carnegie Mellon University, and contain ground truth information for the pixel location of specific features in the images. The data set used was the "Planar texture" files (data set CIL-0002).

The desired values for all training patterns (with real-valued disparities) were generated under the assumption that each of four output units had a fixed-variance Gaussian disparity tuning curve, centered at a disparity of 0.0, 1.0, 2.0, and 3.0 pixels, respectively. This resulted in a distributed representation for each output. It should be noted that this output representation was space independent. In other words, during training the error signal indicated only what disparities were present in the image. The network received no information as to where those particular disparities were located in the image. It had to extract this spatial information through learning.

Results

Qualitative overview of model performance

In Fig. 5, the activation of the model is shown in response to a stimulus with two objects in the image. The object on the left in the image (the lighter one) was at a disparity of approximately zero pixels, while the darker object on the right was at a disparity of around 2 pixels. The darker object was at a closer crossed disparity than the lighter object, and occluded it. In the disparity energy layer of the model, the two outside edges of each object produced a visible localized signal of moderate amplitude (as indicated by the grayscale values). The strongest signal in the disparity energy layer, however, came at the position where the dark object occluded the lighter object. This edge (at the left side of the dark object) produced essentially the same pattern of activation in the disparity energy layer as the right edge of the dark object. The difference was that the left edge of the dark object produced a stronger amplitude signal because the response of the binocular complex cells was a function of contrast in addition to disparity. At the left edge of the dark object, the contrast was greater than at the right edge.

The pattern of activation in the local disparity networks in response to this two object stimulus appears complicated, but can be examined systematically. In the leftmost local disparity network, activation was strongest for a disparity of zero pixels. Based on the topographic mapping of the energy filters onto the local disparity networks (and the selection networks), it is apparent that this activation in the leftmost local disparity network was due to
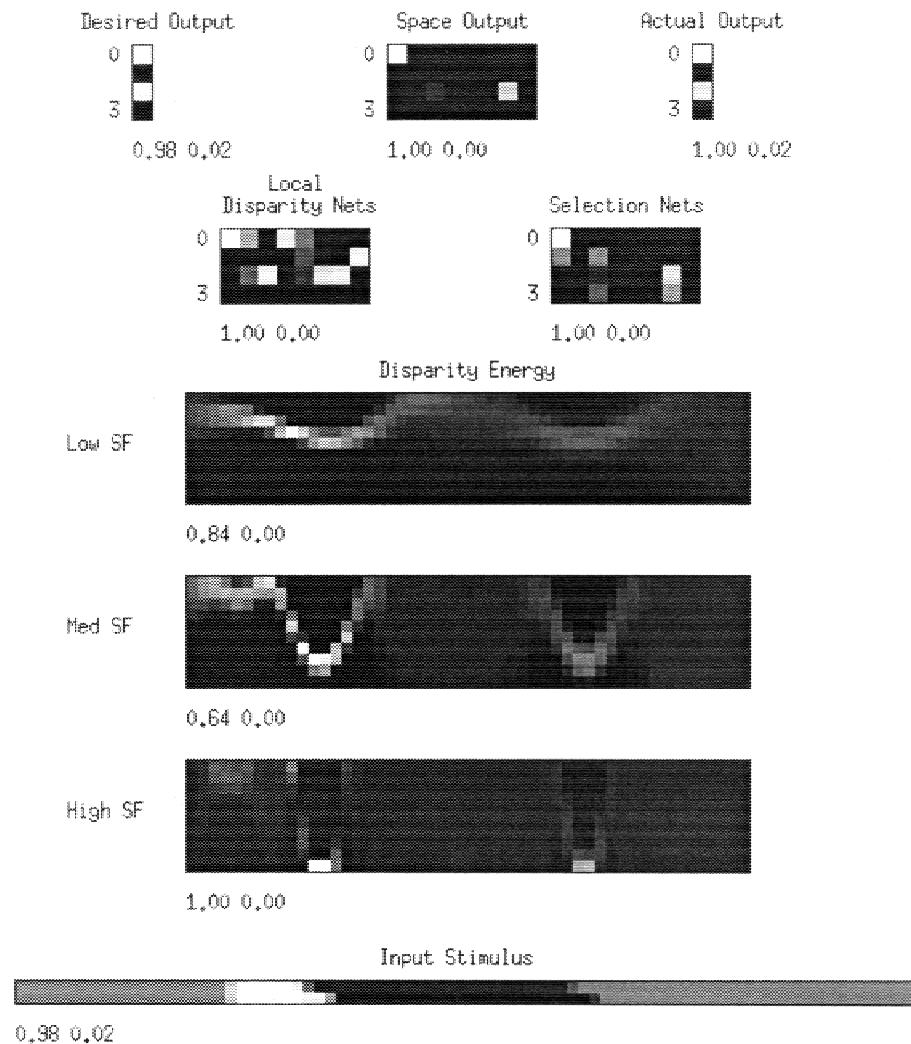
Fig. 5. The activity in the mixture-of-experts model in response to an input stimulus containing two objects—the lighter one on the left at a disparity of approximately 0 pixels and the darker one on the right at a disparity of approximately 2 pixels. This figure is identical in layout to the previous figure. In this example, the dark object on the right was at a crossed disparity (closer than the fixation plane) and occluded the lighter object on the left. The object on the left was approximately at the plane of fixation—that is, at zero-pixel disparity.

the left edge of the lighter object. The next local disparity network to the right was more confused—activation was shared between the unit for 0-pixel disparity and the unit for 2-pixel disparity. In the next local disparity network (third from the left), activation was strongest for a 2-pixel disparity. This signal was due to the left edge of the darker object. On the right side of the local disparity pathway, the networks that were second and third from the right edge were most active at a disparity of 2 pixels, corresponding to the right edge of the dark object.

The activity in the selection networks reveals much about how the network solved the disparity-estimation task. In the top row (tuned to 0-pixel disparity), activation was concentrated in the leftmost unit, corresponding to the left edge of the lighter object. The next row down (tuned to find evidence for a 1-pixel disparity) showed a more diffuse pattern of activation. The 2-pixel disparity network (second row from the bottom) showed strong activation at the position that is close to the right edge. One interesting aspect about the activation in this row is that looking in the disparity

energy layer, the left edge of this darker object had a much stronger amplitude signal. However, in the selection network, this pattern of activation from the left edge of the object was apparently not as reliable as the information from the right edge of the object. The proximity between the left edge of the lighter object and the left edge of the darker object may account for the lack of strong activation in the left part of the 2-pixel disparity selection network. That is, the two left edges have contaminated each other in the disparity energy output, and thus did not provide reliable information for disparity estimation. The bottom row (for disparity of 3 pixels), like the row for 1-pixel disparity, also showed a more diffuse activation pattern.

At the space output level, there was strong activation at only the two locations where the selection network indicated reliable evidence for a disparity, and the local disparity net confirmed that that disparity was present in its receptive field. At the leftmost position, the unit responsive for a disparity of 0 pixels was highly active, corresponding to the lighter object on the left part of the retina. At

the second location from the right, there was high activation for a disparity of 2 pixels. This corresponded to the right edge of the dark object. Comparing the global output to the desired output, there was a close match. For each output unit, the actual value was within 0.05 of its desired value. A quantitative comparison of the performance of the network to two objects across a range of disparities is given in Fig. 6 (lower left panel).

Stereo hyperacuity and performance on stereograms

Stereo hyperacuity

Humans can discriminate differences in depth stereoscopically in the range of a few arc seconds. This discrimination threshold is much smaller than both the width of a photoreceptor and the width of the narrowest disparity tuning curves, but it can be accounted for in a distributed population of statistically independent neurons (Lehky & Sejnowski, 1990).

After training the model on the Multiple Objects data set, it was tested to see if it also demonstrated stereo hyperacuity. Using the disparity-estimation procedure (described in Disparity estimation), the disparity threshold was 0.23 pixels on the Single Object data set (Fig. 6, upper left panel). This value is substantially less than the input resolution of the model (1 pixel) and is thus indicative of stereo hyperacuity. In addition, the model had a fairly low bias in its estimates of -0.05 pixels, determined as the y-intercept of the best-fitting line through the model disparity estimates. Because the Gaussian-fitting procedure can result in disparity estimates that are outside the input range of 0.0 to 3.0 pixels, estimates were clipped
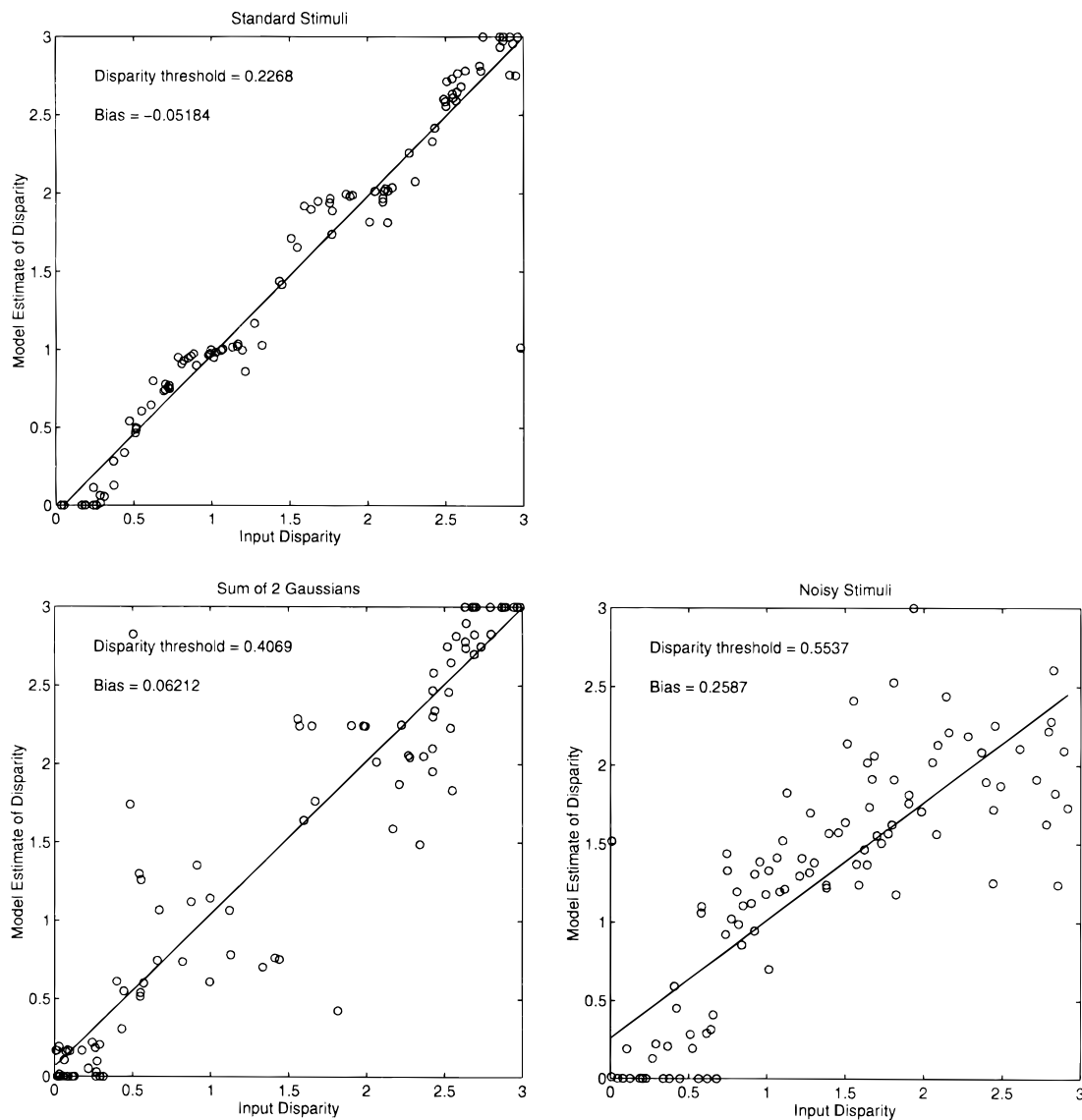
Fig. 6. Upper left: Mixture-of-experts model performance for the Single Object data set. The model's estimates are plotted as a function of the input disparity. Using signal detection theory, the disparity threshold was 0.23 pixels. The bias (the y-intercept of the best-fitting line) was -0.05. Lower left: Model performance in estimating disparity of double-object stimuli. The test set contained 50 stimuli, each with two objects. This figure shows disparity estimates for all 100 individual objects in this set of 50 double-object stimuli. Lower right: Mixture-of-experts model performance on 100 novel Noise stereograms.

at 0.0 and 3.0 pixels when this occurred. The data appear to approximate a step function (with steps at 0-, 1-, 2-, and 3-pixel disparity) because the tuning curves for the output units are centered at these values. This caused the model to estimate, for example, a disparity of 1.0 pixels when the input was in the range of 0.8–1.2 pixels. It is also apparent from this figure that the model overestimated at high disparities and underestimated at low disparities. This is likely due to the fact that the Gaussian-fitting problem is severely underconstrained when there are only four output units (preferred disparity values). With a more dense map of disparity (10–20 output units), this problem should be substantially reduced.

The model was also tested on a novel set of 50 Multiple Object stimuli (Fig. 6, lower left panel). The sum of two Gaussians was fit to the output activation values, as described in Disparity estimation. The disparity threshold calculated using the model's disparity estimate for each of the 100 objects (50 stimuli × 2 object/stimuli) was 0.41 pixels, and relatively unbiased. The paucity of objects with input disparities roughly between 1.0 and 2.0 pixels (see Fig. 6, lower left panel) is due to the constraint, in generating the stimuli, that the disparities must differ by 1.5 pixels. This tended to push the input disparities out toward the limiting values of 0- and 3-pixel disparity.

Stereogram performance

Noise stereograms. In the experiments described thus far, all stimuli contained objects, defined as a contiguous array of similar luminance values that were significantly different from the background luminance. The model was also trained and tested on two kinds of stereograms (Julesz, 1971): the Noise data set and the Binary data set. In the Noise data set, as noted in the description in Data sets, the strong luminance edge between the object and the background (found in the Multiple Objects data set) was no longer present. The disparity signal of the object remained, but it was much more difficult to detect because of the variation in luminance values and the uncorrelated random background (see Fig. 7).

The model's response to a novel set of 100 noisy random-dot stereograms is shown in Fig. 6 (lower right panel). The model's disparity threshold was 0.55 pixels, and still demonstrated stereo hyperacuity. The increased threshold may be attributed to the distracting uncorrelated noise background, as well as to the loss of the strong luminance edge at the border of the object and the background.

Binary random-dot stereograms. After training the model on a data set of 1400 binary random-dot stereograms (40% dot density) that included transparent surfaces, the model was tested on novel
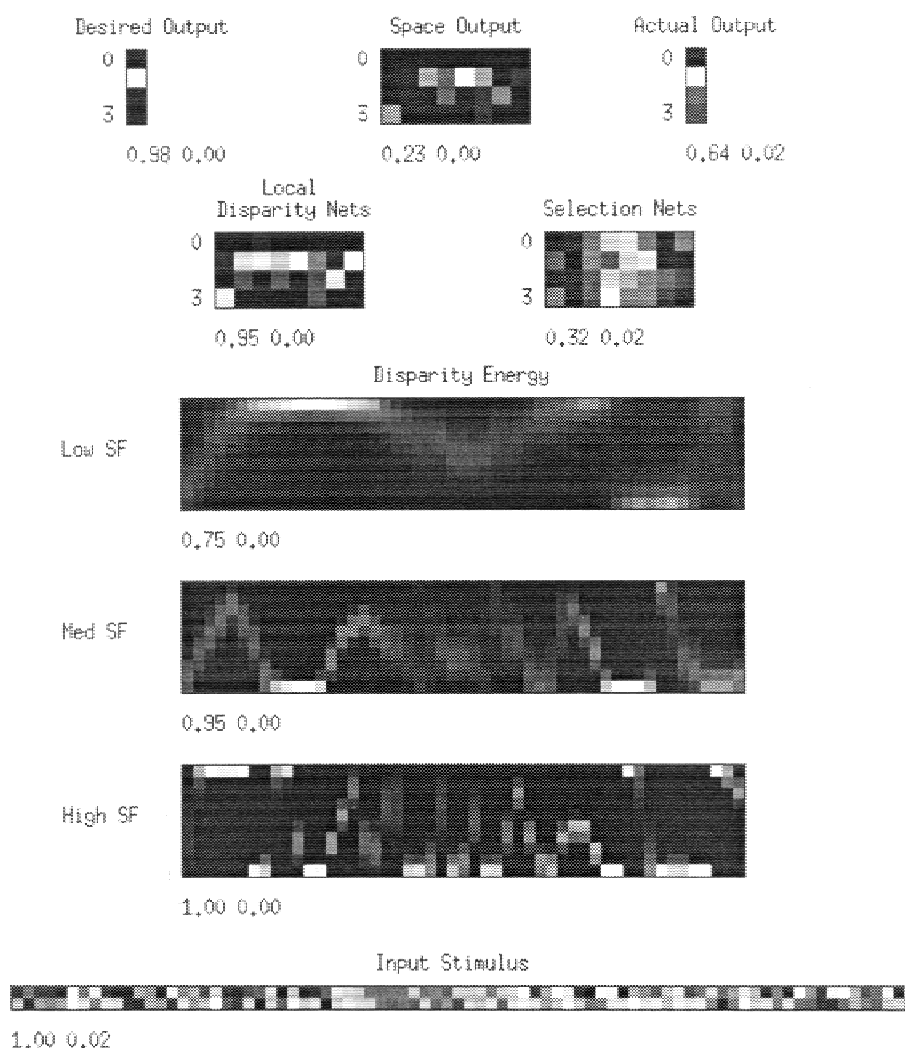

Fig. 7. The response of a model trained on noisy stereograms to a novel noisy stereogram.

stimuli—100 standard single-disparity random-dot stereograms and 50 transparent stimuli that defined two surfaces at different disparities. The model performed well on the standard random-dot stereograms, with a disparity threshold of 0.36 pixels. For the stimuli containing two transparent surfaces, the threshold rose substantially to 0.83 pixels. An example of the model's response to a test stimulus after training is shown in Fig. 8.

Real-world images

The results described thus far for the stereo model are based on a variety of synthetic stimuli. Although these stimuli differ in a number of ways (e.g. luminance profile, spatial-frequency content, and contrast), they may still contain certain statistical regularities that are not representative of natural visual stimuli found in the world.

Using the parameters obtained after optimizing the model on the Multiple Objects data set, the model was tested on real-world stereoscopic image pairs (Real data set described in Data sets). The disparity threshold for these real-world stimuli was 0.30 pixels, with a bias of -0.02 pixels (Fig. 9, upper left panel). Thus, the model generalized well from the statistics of the synthetic images to the real-world images.

Ensemble performance on a varied set of stimuli

The simulations that we have described thus far in the paper are the result of training and testing on a single kind of stimulus (with the exception of Real-world images). But it is not clear how well the model would generalize when trained on several different kinds of stimuli. In this section, we describe how the mixture-of-experts model performs when trained on a set of stimuli containing: (1) 150 single objects, (2) 150 double objects, (3) 150 Noise stereograms (from the Noise data set), and (4) 150 stereograms from the Binary data set. These stimuli were generated in the same way as described in Data sets, and illustrated in Table 1.

After training on this ensemble of stimuli, we computed disparity thresholds for four testing sets—one for each of the four kinds of stimuli. For single-object stimuli, the model performed extremely well, with a disparity threshold of 0.16 pixels. This performance is even better than when the model is trained on *only* single and double objects. It may be that the added diversity of this
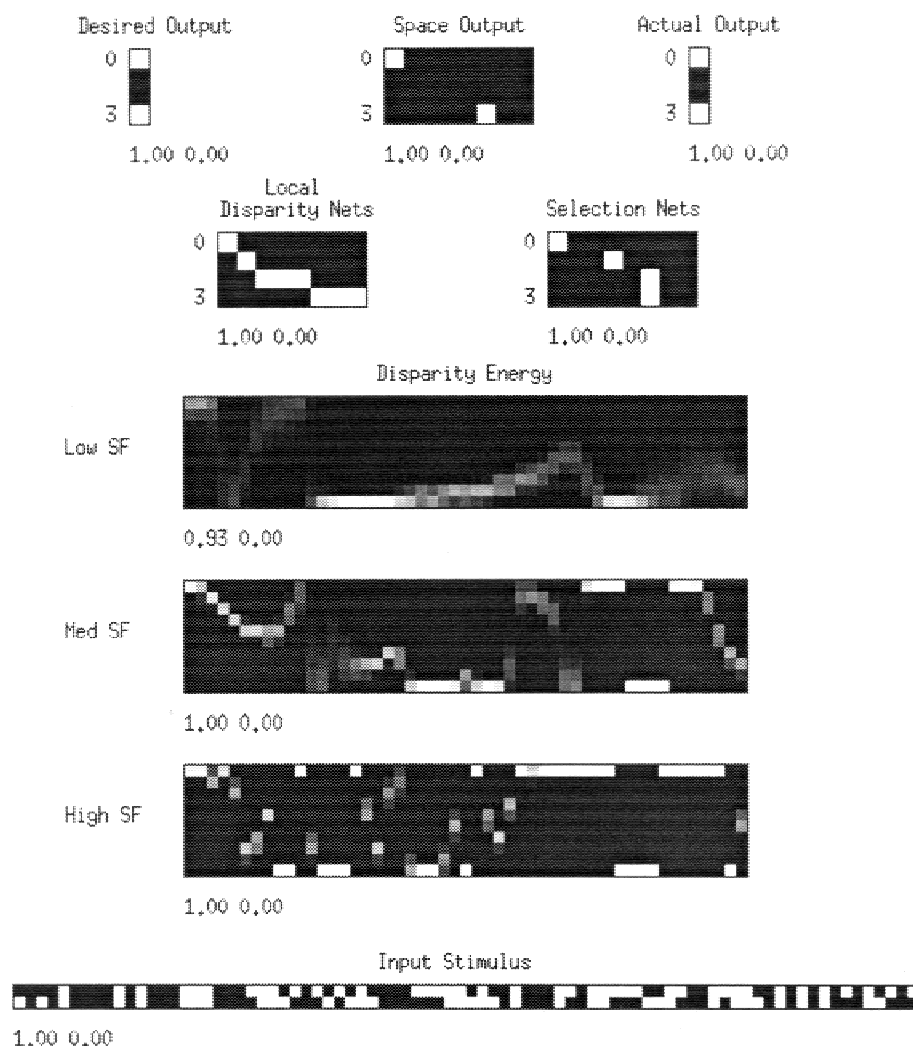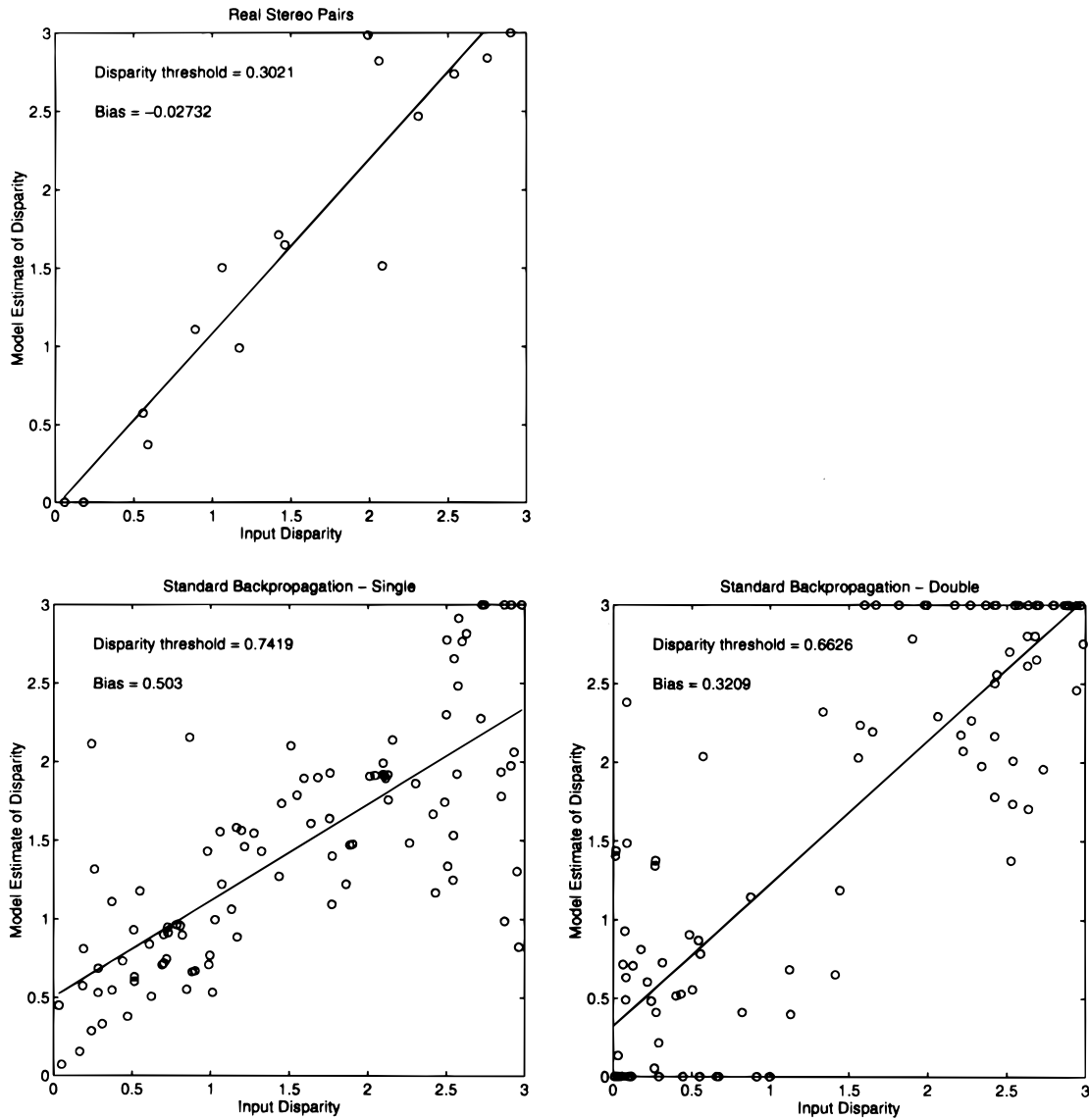

Fig. 8. Model response to a novel binary stereogram with two transparent surfaces—at 0-pixel and 3-pixel disparity.

Fig. 9. Upper left: Disparity estimates for the Real data set using the model trained on the Multiple Objects data set. Lower left: Performance of the single-pathway model for the Single Object data set after training on the Multiple Objects data set. Lower right: Performance of the single-pathway model for 50 novel stimuli each containing two objects, after training on the Multiple Objects data set.

training set (with many different kinds of objects) accounts for the model's outstanding generalization ability. This disparity threshold rose to 0.45-pixel disparity for stimuli containing two objects. For the Noise stereograms, the threshold increased substantially to 1.09-pixel disparity with a bias of 0.55 pixels. This is nearly double the threshold of the model when trained on Noise stimuli alone (see Stereogram performance). The poor performance on these stimuli is likely due to the fact that they are, statistically, a difficult disparity signal to estimate, and that they only account for 25% of the stimuli in the training set. In a similar manner, the model did not generalize well to novel stimuli containing two transparent surfaces. The threshold for these stimuli was 0.84 pixels, with a bias of 0.41 pixels. Discovering the kinds of energy filter responses that are indicative of two transparent surfaces is difficult, especially when these kind of stimuli also comprise only 25% of the stimuli on which the model is trained.

Comparison to the single-pathway model

A single-pathway model (trained with backpropagation) was used for comparison with the two pathway (local disparity and selection) mixture-of-experts model. A difference in performance between these two models provides an estimate of the contribution made by the selection pathway (Table 2). The single-pathway model (described in Comparison models) was trained using several different data sets. The first training set was the Multiple Objects data set. Disparity estimates were computed for the Single Object data set using the Gaussian-fitting technique described in Disparity estimation. This test set is the same as the one used for the mixture-of-experts disparity threshold results shown in Fig. 6 (upper left panel). The disparity threshold was 0.74-pixel disparity in this single-pathway model (Fig. 9, lower left panel), compared to 0.23 for the mixture-of-experts model. When tested on a set of 50

Table 2. *Disparity thresholds (in pixels) for each data set for the mixture-of-experts model (ME), the single-pathway model (SP), and the cross-correlation model (CC)[a]*

| Stimulus type	ME	SP	CC
Single	0.23	0.74	0.46
Double	0.41	0.66	NT
Noise	0.55	NT	1.28
Random dot	0.36	NT	NT
Transparent	0.83	NT	NT
Real	0.30	NT	0.28

[a]NT means the model was not tested on that data set. For all data sets, the model was tested and trained on the same kind of stimuli, with one exception: *Double*-object stimuli were tested after training on the Multiple Objects data set, which included Single Object and Double Object stimuli. The cross-correlation model was not trained.

Multiple Object stimuli (100 objects altogether), the threshold of the model was 0.66 pixels (Fig. 9, lower right panel).

The single-pathway model was also trained on the Single Object data set, and tested on a different set of 100 Single Object stimuli. The computed disparity threshold of the single-pathway network for these novel stimuli was 0.33 pixels (see Fig. 10, upper left), substantially lower than when multiple objects were present in the training set.

Comparison to the cross-correlation model

The disparity estimates for the Single Object data set using the cross-correlation model are shown in Fig. 10 (upper right panel). The disparity threshold was 0.46 pixels. When tested on 100 stimuli from the Noise data set, the threshold rose to 1.28 pixels of disparity, with a large bias (Fig. 10, lower left panel). The disparity signal in these Noise stimuli was much more difficult to detect
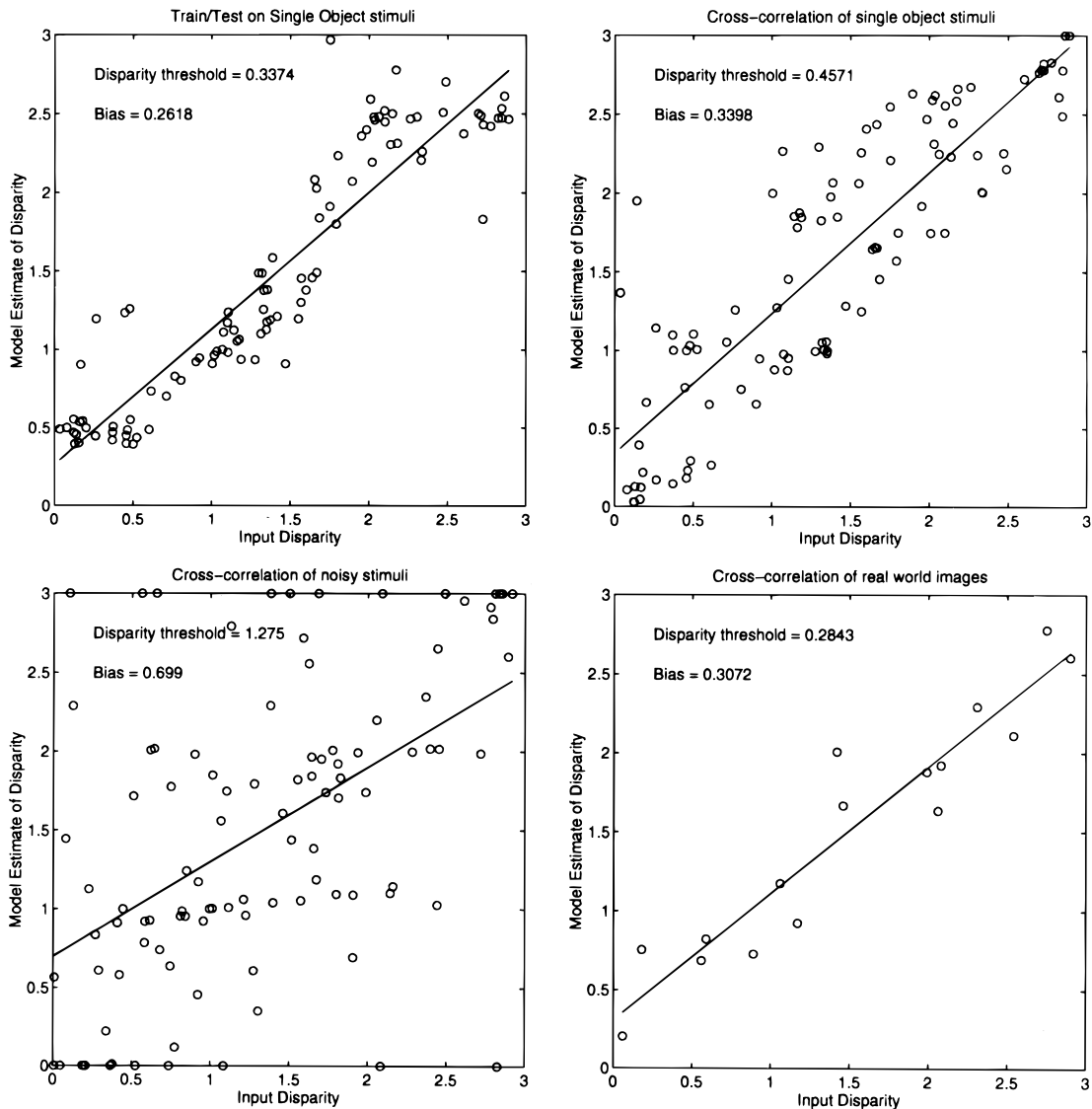
Fig. 10. Upper left: Single-pathway model performance for 100 different single-object test stimuli after training on the Single Object data set. Upper right: Cross-correlation model performance on the Single Object data set. Lower left: Cross-correlation model performance on 100 stimuli from the Noise data set. Lower right: Cross-correlation model performance on the data set of real images.

with the cross-correlation approach. Because the background luminance was no longer constant (as it was for the Single Object data set), the background did not contribute a constant value to the sum in eqn. (6). Instead, the background contributed a different amount for each disparity value *d*. So, although the object in the stimulus contained a well-defined disparity signal, this information was swamped by the noisy disparity cues in the background. The cross-correlation model was also tested on the Real data set, a set of real-world images with known (ground truth) disparities. The disparity threshold was 0.28 pixels (Fig. 10, lower right panel). For performance comparisons with the mixture-of-experts and single-pathway models, see Table 2.

Analysis of model phenomena

Receptive-field properties

The receptive-field properties of units in the local disparity and selection pathways were examined to gain insight into the function of the mature network. As noted in Qualitative overview of model performance, the model became sensitive to the disparity edges of the objects as a result of the optimization of model parameters. To further explore how the model responded to edges, the receptive field of a unit in each pathway was mapped systematically.

The receptive-field response of a unit depended on the architecture, receptive-field layout, and weight-sharing properties of the model. The receptive field of a unit in either the local disparity or the selection pathway covered nine spatial locations (horizontally in the figures) in the disparity energy filters. At each of these nine spatial locations there were 30 disparity energy outputs (vertically—10 at each of three spatial frequencies). Weights in each pathway were shared *across space*. For example, the weights to each of the eight local disparity units tuned to 3-pixel disparity (the bottom row of the local disparity pathway) were the same. The same was true for other rows (disparities) of the local disparity pathway, and for the selection pathway. Thus, all units tuned to the same disparity in the same pathway computed the same function of the input.

The responses of a single unit were mapped at each of four disparities in each pathway as a function of space and disparity. A single high-contrast edge was moved systematically across the receptive field of the unit in increments of approximately 0.5 pixels. At each of these spatial locations, the disparity of the edge was varied between 0.0 and 3.0 pixels in increments of approximately 0.15 pixels. In this way, a dense response map was generated as a function of space (within the receptive field) and disparity for each of the four disparities in both the local disparity and selection pathways.

Local disparity units

The response of a unit in the local disparity pathway tuned to 0-pixel disparity is shown in the upper left part of Fig. 11. The unit responded strongly when the edge was at a 0-pixel disparity in the middle of its receptive field. The response decreased when moved away from 0-pixel disparity (vertically, in the figure), and when the edge was moved toward the side of the receptive field. Similar responses were found for units tuned to 1-, 2-, and 3-pixel disparity in the local disparity pathway (Fig. 11). Within each spatial location in the local disparity pathway, there was a soft-max competition across the four disparities. The responses in Fig. 11 reflect this competition.

Selection units

The responses of units in the selection pathway as a function of space and disparity, shown in Fig. 12, were not as stereotyped as those of units in the local disparity pathway. The selection unit tuned to 0-pixel disparity (upper left part of Fig. 12) responded strongly to a change in disparity (from approximately 0- to 1-pixel disparity) moving from left to right across the receptive field. The selection unit tuned to 1-pixel disparity (upper right) showed a more complex response pattern. Note, however, that it shared (with selection unit 0) the property that it responded to changes in disparity across space. This pattern also held for the selection unit tuned to a pixel disparity of 2 (lower left). Unlike the units in the local disparity pathway, the selection pathway units were *not* trained to find the best disparity estimate within their local receptive field. Instead, units in the selection pathway were trained to find reliable patterns in the disparity energy layer that were indicative of a given disparity, regardless of spatial location. This selection was enhanced by competition across space in the selection pathway.

The spatial receptive-field structure of these selection units suggests that they may be sensitive to *disparity contrast*. In other words, the selection units have learned that a reliable indicator for a given disparity was a change in disparity across space. These units responded only at the edge of an object (not in the middle), even when there was a disparity signal present in the middle of the object. This selection-unit activity can be interpreted as indicating where the *continuity constraint* has been violated (Marr & Poggio, 1976). The *continuity constraint* suggests that surfaces generally change smoothly in depth except at object boundaries. The operation of the selection units occurred not only across space, but in depth as well. These selection units could thus provide valuable information in the construction of a three-dimensional (3-D) model of the world.

Recent neurophysiological data from von der Heydt et al. (1995) is consistent with selection-unit activity. They found that neurons of awake, behaving monkeys in area V2 responded to edges of random-dot stereograms. Because random-dot stereograms have no monocular form cues, these neurons must be responding to edges in depth. This behavior is analogous to that observed in the selection pathway of the model. The units were responsive to changes in disparity across space—in other words, they were sensitive to edges in a depth map.

Psychophysics

The model was tested on a psychophysical task for which human experimental data were available for comparison. Disparity sensitivity is affected by the spatial-frequency content of an image. Westheimer and McKee (1980) found, in human psychophysical experiments, that disparity thresholds increased for any kind of spatial-frequency filtering of line targets, although disparity sensitivity was more adversely affected by high-pass filtering than by low-pass filtering. The disparity-estimation performance of the model was assayed after manipulations of the spatial-frequency content of the input. The effects of this spatial-frequency filtering on model performance depended on the frequency responses of the disparity energy filters that pre-process the input to the model. These frequency responses (of simple cells in the energy filters) are shown in Fig. 13a. The specific amplitude values of the frequency response are not important because these simple cell responses were normalized and combined to generate the complex cell (disparity energy) response. In addition, the soft-max operation across disparities within each spatial location of each filter
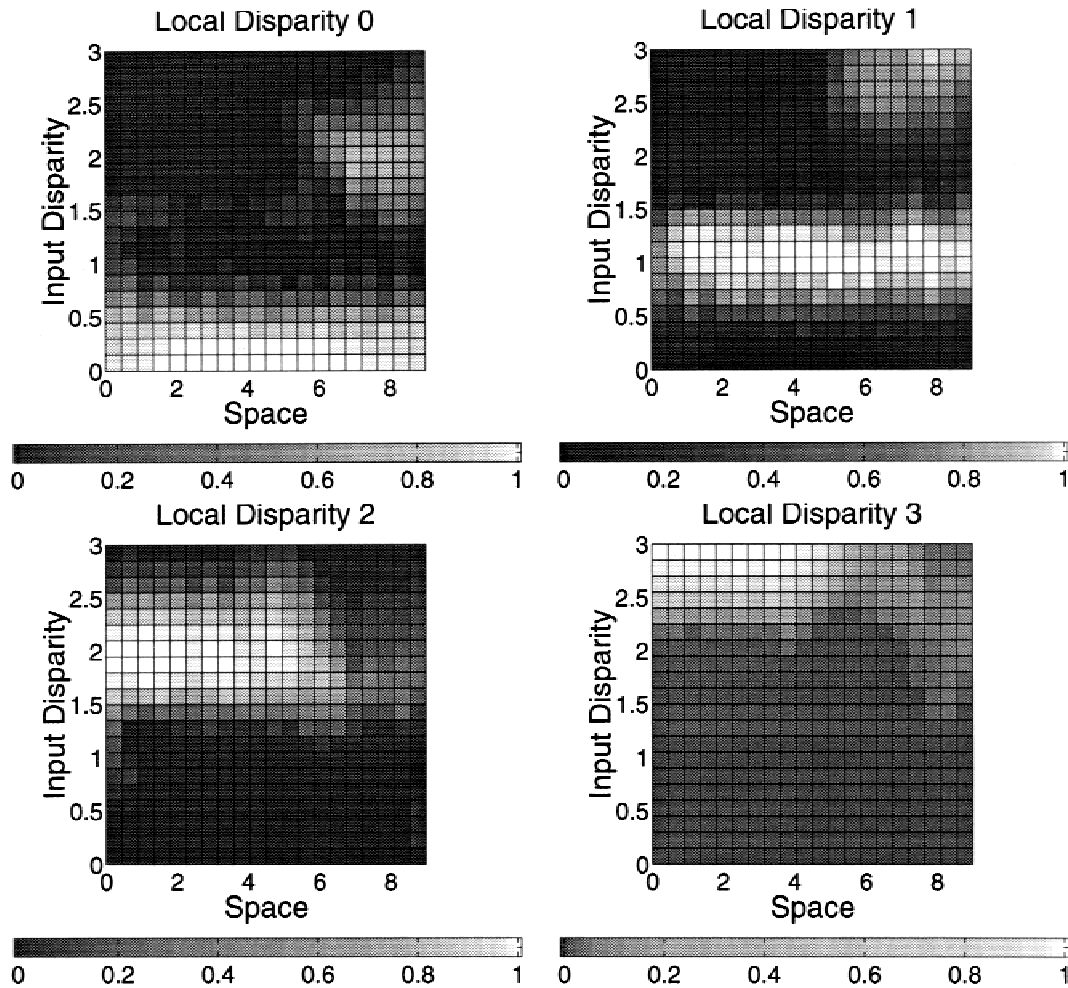
Fig. 11. The activation of units in the local disparity pathway tuned to 0-, 1-, 2-, and 3-pixel disparity (starting in upper left) in response to a single edge, as a function of space and disparity.

bank normalized responses by boosting the highest signal. (See Retina and disparity energy filters and Ohzawa et al. (1990) for more details).

Westheimer and McKee (1980) filtered their line targets with low-pass and band-pass filters. When their stimuli were low-pass filtered, the disparity thresholds increased by a factor of 1.1–1.4, relative to the unfiltered stimuli. After band-pass filtering, disparity thresholds had increased by a factor of 1.7–3.5. The low-pass filtering of Westheimer and McKee was matched using filter A, as shown in Fig. 13b, and filter B for the band-pass filter. Filter A retained most of the information in the low and medium SF filters, while losing some of the signal in the high SF filters. Relative to filter A, filter B lost much of the low SF information.

With the weights obtained during training on the Multiple Objects data set, the model's estimates of disparity for the Single Object data set filtered using filters A (low-pass) and b (band-pass) were determined. Consistent with Westheimer and McKee (1980), the disparity threshold of low-pass filtered images had risen to 0.33-pixel disparity (a factor of 1.46) relative to unfiltered test stimuli (with a threshold of 0.23 pixels). The disparity threshold of band-pass filtered images (filter B) rose to 0.60-pixel disparity, a factor of 2.64 relative to unfiltered stimuli. This result is also consistent with Westheimer and McKee (1980).

Discussion

The difficulty in disparity estimation under less than ideal circumstances (e.g. transparency, occlusion, noise) is knowing which disparity cues are reliable and should be integrated. The approach to this problem taken here was to model the disparity estimation process that would be needed to account for a *small patch* of the image. A multiscale neurophysiologically realistic implementation of binocular cells was used for the input, combined with a neural network model to learn reliable cues for disparity estimation. This integration of neurally realistic modeling and a sophisticated statistical learning model represents a powerful framework for approaching other open problems in computational neuroscience. A number of related computational models have also emphasized the importance of data reliability for estimation tasks (Derin & Elliot, 1987; Földiák & Young, 1995; Grzywacz et al., 1995). The performance of the model, its relationship to neurophysiology and behavior, and the limitations of the model are discussed below.

Performance

The mixture-of-experts model performed well overall on a variety of disparity-estimation tasks, compared to the single-pathway and
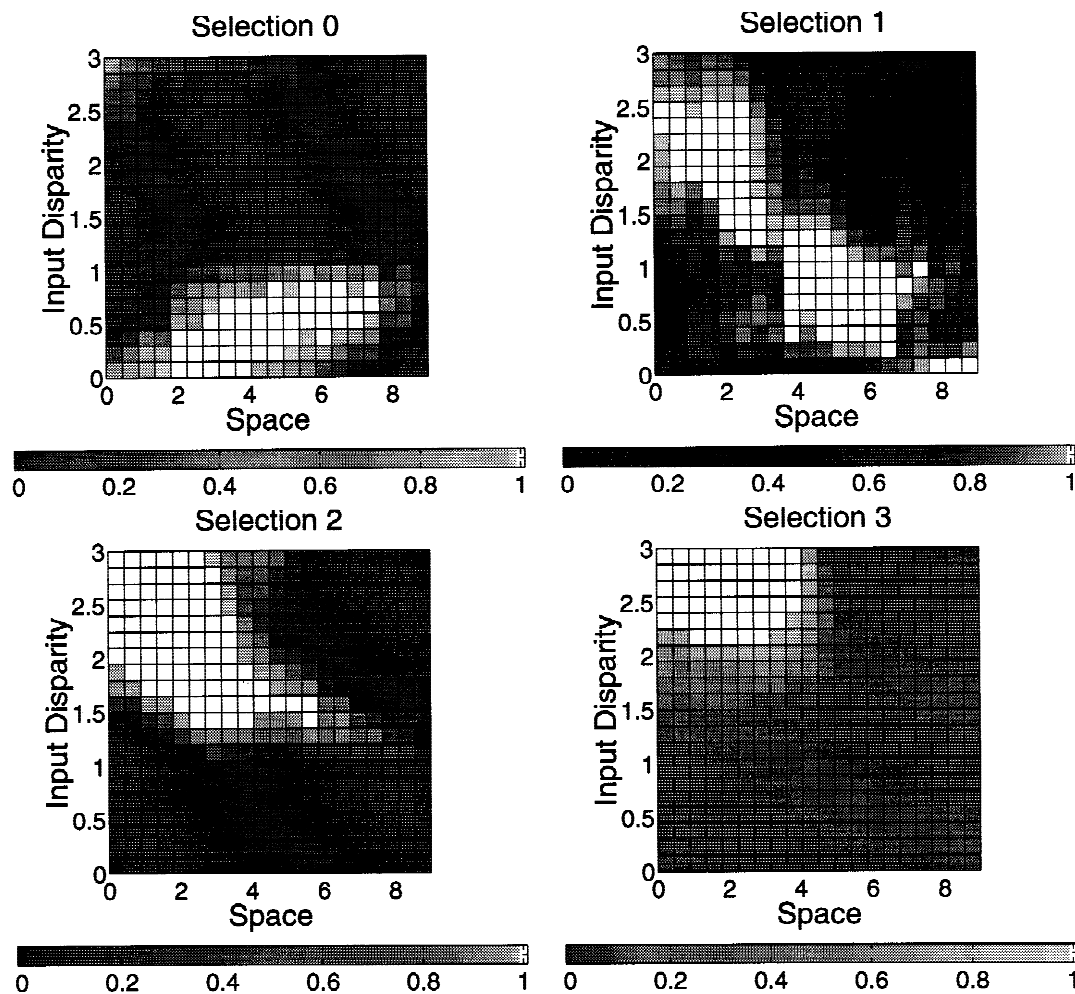
Fig. 12. The activation of units in the selection pathway tuned to 0-, 1-, 2-, and 3-pixel disparity (starting in upper left) in response to a single edge, as a function of space and disparity.

cross-correlation models. The performance of these models for a variety of data sets is summarized in Table 2. One explanation for the poorer performance of the mixture-of-experts model on the Double Object stimuli (relative to the Single Object stimuli) may be the difficulty in fitting the sum of two Gaussians using only four data points. More data points (i.e. a larger number of output units) would improve this fitting process substantially. Disparity thresholds for the mixture-of-experts model rose as the data sets became more complicated. Especially challenging was the Transparent data set, where the threshold rose to 0.83 pixels. For all data sets, however, the mixture-of-experts model achieved stereo hyperacuity performance.

In addition to the Multiple Objects data set, the single-pathway model was also trained and tested on Single Object stimuli (not shown in Table 2). In this case, the disparity threshold improved to 0.33 pixels. This low threshold for single-object stimuli suggests that a selection mechanism was not necessary when only one disparity was present in the image, and when there were high-contrast disparity edges at the object boundaries. The lack of a selection mechanism in the single-pathway model, however, limited performance when more than object was present (Table 2). The selection pathway of the mixture-of-experts model, on the other hand, per-

formed the crucial task of selecting reliable energy filter output when multiple objects were present, and avoided averaging over depth discontinuities. This comparison with a single-pathway network illustrated the advantages of an architecture that separated the disparity-estimation and selection (integration) tasks into two separate pathways.

The cross-correlation model had no mechanism for estimating the disparity of more than one object, and so could not be tested on several of the data sets. Without a means to *spatially select* reliable disparity information (like the single-pathway model), the cross-correlation approach fared poorly, relative to the mixture-of-experts model, on both the Single Object stimuli and the Noise stimuli. It estimated accurately, however, the disparities present in the Real data set.

Another model using the multiscale approach was developed by Marr and Poggio (1979). The strategy in this model is often described as "coarse-to-fine." Spatial filters with low spatial frequencies were presumed to trigger vergence eye movements. When the eyes verged to within a quarter-cycle disparity of the smaller (high-frequency) receptive fields, disparity estimates would be computed by summing disparities from all spatial scales. There are two important differences between the model developed by Marr

a)

b)
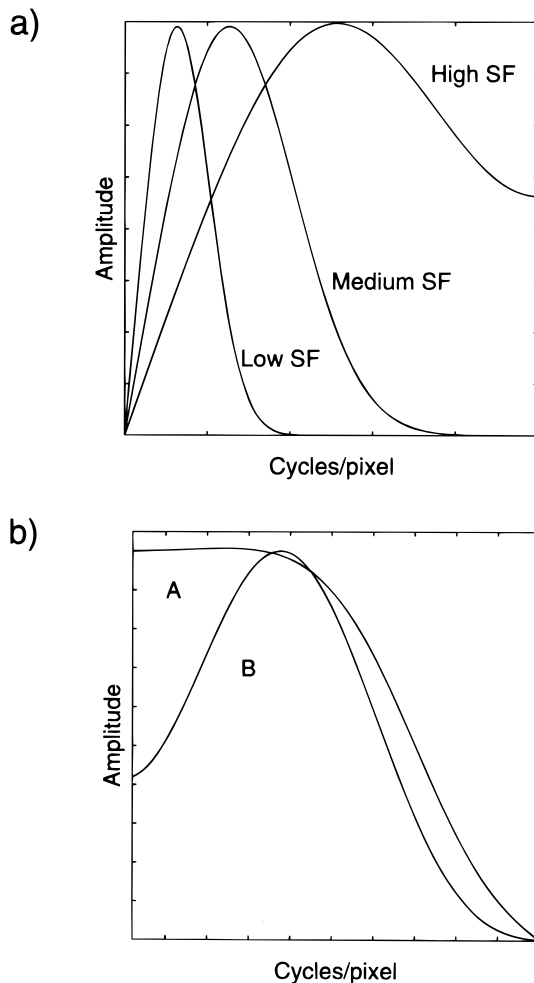
Fig. 13. Frequency response of simple cells and spatial filters. (a) Frequency response of simple cells (in each frequency band) that were the basis of the disparity energy filters. (b) Spatial-frequency filters analogous to those used by Westheimer and McKee (1980). 'A' is the low-pass filter, 'B' is the band-pass filter. The units of frequency were cycles/pixel.

and Poggio (1979) and the current model. First, Marr and Poggio (1979) matched disparities at each scale sequentially (in time). The current model had only a single time step. Information from all scales (high, medium, and low spatial frequencies) was integrated simultaneously. Since these computations could be performed in parallel, the current model might be expected to compute more quickly than Marr and Poggio (1979). In addition, the selection pathway allowed the disparity to be robustly estimated. Psychophysical evidence against the coarse-to-fine strategy (Smallman, 1995; Mallot et al., 1996; Mowforth et al., 1981) would also favor a parallel architecture like the present one.

The disparity energy filtering of the input image was a fixed computation in the model—no adaptation or learning was involved. In subsequent stages (the local disparity and selection pathways), however, parameters were adjusted according to an error signal. The mixture-of-experts model was optimized, *by example*, to determine what features in the disparity energy layer were important. The advantage of this particular network architecture was that it dissociated the disparity-estimation process (local disparity pathway) from the task of determining which cues in the

disparity energy output are likely to be reliable for estimating disparity (selection pathway). The selection pathway determined what patterns of activity in the complex cell output were consistent indicators of disparity and should be selected. The response properties of units in the selection pathway qualitatively matched responses of binocular cells in monkey visual cortex area V2 (von der Heydt et al., 1995), which also were selectively activated by disparity gradients. Although it was not clear *a priori* what response properties these selection units should have, after learning it was found that units that responded to disparity contrast were particularly valuable. The hard-coding of these response characteristics is a clear next step for future models.

Comparison to psychophysical data

The performance of the model matched psychophysical data of Westheimer and McKee (1980) on the effects of spatial-frequency filtering on disparity thresholds in humans. These results can be understood in terms of the frequency response of the model. Consider the dynamic range of the disparity energy filters for the different spatial frequencies. As described in the caption of Fig. 3, the phase differences of the low spatial-frequency filters covered a range of 0- to 4-pixel disparity, while the medium and high spatial-frequency filters had a much narrower response range. Because the test stimuli included disparities between 0 and 3 pixels, it is understandable that the network performed better with low-frequency information present. It should be noted, however, that although the low-frequency information did lead to reasonable stereoacuity, performance of the network was much improved if information from all spatial-frequency bands was present.

The sensitivity to depth contrast that was found in the selection units in the model may be a general mechanism for segmenting objects in depth. This response profile is consistent with a Laplacian-like center-surround operating on a depth map (Mitchison, 1993), and may explain psychophysical results reported by Westheimer (1986). Westheimer studied stereoscopic depth contrast, in which disparity estimates of vertical line targets presented psychophysically are modulated by the disparity of neighboring features. This modulation depended on the spatial separation between targets. When targets were closely spaced, they were perceived at a similar depth—this is often referred to as disparity pooling. At larger spatial separations, however, the line targets were perceived as separated in depth, called a *repulsion* effect. These phenomena could be explained by neurons with short-range excitatory connections and long-range inhibitory ones, as suggested by Lehky and Sejnowski (1990).

Limitations

The model was designed to selectively integrate reliable patterns of activation in the local disparity pathway which are combined to produce a space-independent measure of the disparities present in a small local patch of an image. One limitation of the current model is this focus on only a small region of space. The integration of the local disparity estimates (produced at the output level of our model) across a whole image scene is beyond the scope of the current model. In practice, the visual field would be tiled with networks performing local operations similar to those performed by our model, in a columnar organization as found in visual cortex.

A related limitation of the model is that there is no representation of *surfaces* present in a visual scene, as may be found in the brain (see Nakayama & Shimojo, 1992). This surface representa-
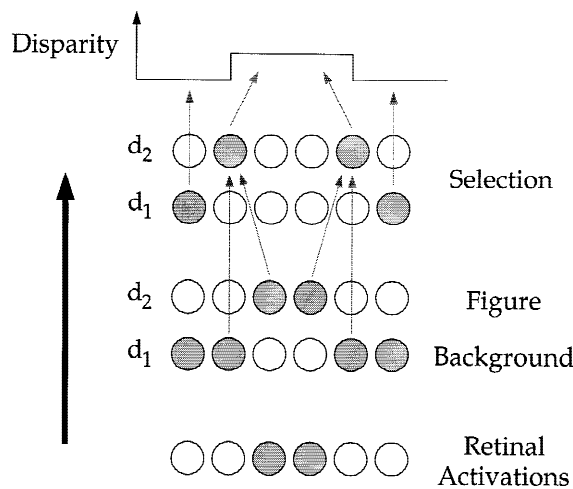
Fig. 14. Schematic summary of the current stereo model. Gray circles indicate active units, and white circles are inactive units. The large arrow on the left indicates the direction of processing. Retinal activations (bottom) represent a stimulus causing two units to be active. At the next level of processing, units sensitive to local disparity (disparities d_1 and d_2) indicate activations for figure and ground in the image. Selection units indicate where reliable disparity information is present. The top of the figure shows a map of disparity across the image.

tion can be interpreted as a form of segmented depth map across space. Such a representation did not explicitly exist in the model. However, the disparity estimates at the output level of the model could serve as a good initial estimate for an algorithm that smoothed and segmented surfaces in depth. The Bayesian model developed by Yuille et al. (1991) would work well in this regard.

The model can only accurately represent the disparity of fronto-parallel surfaces. The representation of slanting surfaces (tilted in depth) would require an interpolation mechanism (Mitchison & McKee, 1985) such as that suggested above to account for smooth changes in depth across space.

Finally, the model does not attempt to include temporal dynamics of stereopsis. Thus, it cannot account for the results of Norcia and Tyler (1984), who found temporal processing limits for the binocular percepts of apparent depth motion and depth pulsation. Related work by McKee and colleagues (1997) has shown that binocular disparity information is most valuable for static images, but the slow response of the system makes it less useful for detecting moving targets. Nonetheless, the model has provided evidence that assessing the reliability of local disparity estimates can improve the performance of feedforward networks in a way that is compatible with the physiological properties of neurons in the visual cortex and psychophysical results from humans. A selection mechanism along the lines studied here could be included in more sophisticated models that incorporated spatial integration and temporal dynamics.

Conclusions

The problem of estimating disparity under less than ideal circumstances, such as those that occur under conditions of transparency, occlusion, and noise, is knowing which disparity cues are reliable and should be integrated, and which are unreliable and should be ignored. This problem was approached here with a feedforward

network model using a multiscale neurophysiologically realistic implementation of binocular cells for the input, combined with a method for discovering the properties of reliable cues for disparity estimation. The gradient of the disparity, which is computed in the visual cortex, emerged as an important variable in an intermediate stage of the model that was responsible for selecting the most reliable disparity estimates. This is illustrated in Fig. 14 which provides a schematic summary of the current model. The model achieved excellent performance on a range of different input stimuli. This same approach could be used to integrate motion with disparity information, and could be combined with further stages of processing that dynamically represent surfaces and boundaries of objects.

References

Adelson, E.H. & Bergen, J.R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A* **2**, 284–299.

Blake, R. & Wilson, H.R. (1991). Neural models of stereoscopic vision. *Trends in Neurosciences* **14**, 445–452.

Bonds, A.B. (1991). Temporal dynamics of contrast gain in single cells of the cat striate cortex. *Visual Neuroscience* **6**, 239–255.

Bridle, J.S. (1989). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neuro-Computing: Algorithms, Architectures, and Applications*, ed. Fougelman-Soulie, F. & Herrault, J., pp. 227–236. London: Springer-Verlag.

Cormack, L.K., Stevenson, S.B. & Schor, C.M. (1991). Interocular correlation, luminance contrast and cyclopean processing. *Vision Research* **31**, 2195–2207.

Daugman, J.G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A* **2**, 1160–1169.

Derin, H. & Elliot, H. (1987). Modeling and segmentation of noisy and textured images using gibbs random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **9**, 39–55.

Fleet, D.J., Wagner, H. & Heeger, D.J. (1996). Neural encoding of binocular disparity: Energy models, position shifts and phase shifts. *Vision Research* **36**, 1839–1857.

Földiák, P. & Young, M.P. (1995). Sparse coding in the primate cortex. In *The Handbook of Brain Theory and Neural Networks*, ed. Arbib, M.A., pp. 895–898. Cambridge, Massachusetts: MIT Press.

Gabor, D. (1946). Theory of communication. *Journal of the Institute of Electrical Engineers* **93**, 429–549.

Gonzalez, R.C. & Woods, R.E. (1992). *Digital Image Processing*. New York: Addison-Wesley.

Green, D.M. & Swets, J.A. (1966). *Signal Detection Theory and Psychophysics*. New York: John Wiley and Sons.

Grzywacz, N.M., Watamaniuk, S.N.J. & McKee, S.P. (1995). Temporal coherence theory for the detection and measurement of visual motion. *Vision Research* **35**, 3183–3203.

Jacobs, R.A., Jordan, M.I., Nowlan, S.J. & Hinton, G.E. (1991). Adaptive mixtures of local experts. *Neural Computation* **3**, 79–87.

Julesz, B. (1971). *Foundations of Cyclopean Perception*. Chicago, Illinois: University of Chicago Press.

Lehky, S.R. & Sejnowski, T.J. (1990). Neural model of stereoacuity and depth interpolation based on a distributed representation of stereo disparity. *Journal of Neuroscience* **10**, 2281–2299.

Mallot, H.A., Gillner, S. & Arndt, P.A. (1996). Is correspondence search in human stereo vision a coarse-to-fine process? *Biological Cybernetics* **74**, 95–106.

MARR, D. & POGGIO, T. (1976). Co-operative computation of stereo disparity. *Science* **194**, 283–287.

MARR, D. & POGGIO, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society B* (London) **204**, 301–328.

MCKEE, S.P., WATAMANIUK, S.N.J., HARRIS, J.M., SMALLMAN, H.S. & TAYLOR, D.G. (1997). Is stereopsis effective in breaking camouflage for moving targets? *Vision Research* **37**, 2047–2055.

MITCHISON, G.J. (1993). The neural representation of stereoscopic depth contrast. *Perception* **22**, 1415–1426.

MITCHISON, G.J. & MCKEE, S.P. (1985). Interpolation in stereoscopic matching. *Nature* **315**, 402–404.

MOWFORTH, P., MAYHEW, J.E. & FRISBY, J.P. (1981). Vergence eye movements made in response to spatial-frequency-filtered random-dot stereograms. *Perception* **10**, 299–304.

NAKAYAMA, K. & SHIMOJO, S. (1992). Experiencing and perceiving visual surfaces. *Science* **257**, 1357–1363.

NORCIA, A.M. & TYLER, C.W. (1984). Temporal frequency limits for stereoscopic apparent motion processes. *Vision Research* **24**, 395–401.

NOWLAN, S.J. & SEJNOWSKI, T.J. (1994). Filter selection model for motion segmentation and velocity integration. *Journal of the Optical Society of America A* **11**, 3177–3200.

NOWLAN, S.J. & SEJNOWSKI, T.J. (1995). A selection model for motion processing in area MT of primates. *Journal of Neuroscience* **15**, 1195–1214.

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. *Journal of Neurophysiology* **75**, 1779–1805.

OHZAWA, I., DEANGELIS, G.C. & FREEMAN, R.D. (1997). The neural coding of stereoscopic depth. *Neuroreport* **8**, iii–xii.

PARKER, A.J. & YANG, Y. (1989). Spatial properties of disparity pooling in human stereo vision. *Vision Research* **29**, 1525–1538.

POUGET, A. & ZHANG, K. (1997). Statistically efficient estimations using cortical lateral connections. In *Advances in Neural Information Processing Systems, Vol. 9*, ed. MOZER, M.C., JORDAN, M.I. & PETSCHE, T., pp. 97–103. Cambridge, Massachusetts: MIT Press.

RUMELHART, D.E., HINTON, G.E. & WILLIAMS, R.J. (1986). Learning internal representations by error propagation. In *Parallel Distributed Processing, Vol. 1, Chap. 8*, ed. RUMELHART, D.E., MCCLELLAND, J.L. & the PDP Research Group, pp. 318–362. Cambridge, Massachusetts: MIT Press.

SMALLMAN, H.S. (1995). Fine-to-coarse scale disambiguation in stereopsis. *Visual Neuroscience* **35**, 1047–1060.

STEVENSON, S.B., CORMACK, L.K. & SCHOR, C.M. (1991). Depth attraction and repulsion in random-dot stereograms. *Vision Research* **31**, 805–813.

VON DER HEYDT, R., ZHOU, H., FRIEDMAN, H. & POGGIO, G.F. (1995). Neurons of area V2 of visual cortex detect edges in random-dot stereograms. *Society for Neuroscience Abstracts* **21**, 18.

WEINSHALL, D. & MALIK, J. (1995). Review of computational models of stereopsis. In *Early Vision and Beyond*, ed. PAPATHOMAS, T.V., CHUBB, C., GOREA, A. & KOWLER, E., pp. 33–41. Cambridge, Massachusetts: MIT Press.

WESTHEIMER, G. (1986). Spatial interaction in the domain of disparity signals in human stereoscopic vision. *Journal of Physiology* **370**, 619–629.

WESTHEIMER, G. & MCKEE, S.P. (1980). Stereoscopic acuity with defocus and spatially filtered retinal images. *Journal of the Optical Society of America* **70**, 772–777.

YUILLE, A.L., GEIGER, D. & BÜLTHOFF, H.H. (1991). Stereo integration, mean field theory, and psychophysics. *Network* **2**, 423–442.

ZHU, Y. & QIAN, N. (1996). Binocular receptive field models, disparity tuning, and characteristic disparity. *Neural Computation* **8**, 1611–1641.