Biol. Cybern. 86, 117–136 (2002)
DOI 10.1007/s004220100280

Biological Cybernetics

A Bayesian model of stereopsis depth and motion direction discrimination

J. C. A. Read

University Laboratory of Physiology, Parks Road, Oxford OX1 3PT, UK

Received: 2 March 2001 / Accepted in revised form: 5 July 2001

Abstract. The extraction of stereoscopic depth from retinal disparity, and motion direction from two-frame kinematograms, requires the solution of a correspondence problem. In previous psychophysical work [Read and Eagle (2000) Vision Res 40: 3345–3358], we compared the performance of the human stereopsis and motion systems with correlated and anti-correlated stimuli. We found that, although the two systems performed similarly for narrow-band stimuli, broad-band anti-correlated kinematograms produced a strong perception of reversed motion, whereas the stereograms appeared merely rivalrous. I now model these psychophysical data with a computational model of the correspondence problem based on the known properties of visual cortical cells. Noisy retinal images are filtered through a set of Fourier channels tuned to different spatial frequencies and orientations. Within each channel, a Bayesian analysis incorporating a prior preference for small disparities is used to assess the probability of each possible match. Finally, information from the different channels is combined to arrive at a judgement of stimulus disparity. Each model system – stereopsis and motion – has two free parameters: the amount of noise they are subject to, and the strength of their preference for small disparities. By adjusting these parameters independently for each system, qualitative matches are produced to psychophysical data, for both correlated and anti-correlated stimuli, across a range of spatial frequency and orientation bandwidths. The motion model is found to require much higher noise levels and a weaker preference for small disparities. This makes the motion model more tolerant of poor-quality reverse-direction false matches encountered with anti-correlated stimuli, matching the strong perception of reversed motion that humans experience with these stimuli. In contrast, the lower noise level and tighter prior preference used with the stereopsis model means that it performs close to chance with anti-correlated stimuli, in accordance with human psychophysics. Thus, the key features of the experimental data can be reproduced assuming that the motion system experiences more effective noise than the stereoscopy system and imposes a less stringent preference for small disparities.

1 Introduction

To perceive depth in stereograms, or motion in two-frame kinematograms, the brain must solve a correspondence problem, deducing which point in the left retina or first frame corresponds to which in the right retina or second frame. Even for random-dot stimuli, in which each dot in one image has a multitude of possible matches in the other, our brains can often solve this problem instantaneously and effortlessly. However, the problem is far from trivial. It has, in general, no unique solution (Fig. 1). In order to select a possible solution, the brain must apply additional constraints, representing its inherent or acquired assumptions about the outside world. For instance, in the double-nail illusion of Fig. 1, the perception is of two spheres at the same distance from the observer – presumably because we more often see flat surfaces face on than edge on. A common constraint used by modellers is that of smoothness or continuity: solutions assuming a sudden jump in disparity should be avoided in favour of solutions in which disparity varies smoothly across the image, reflecting the fact that the real world tends to be composed of discrete objects with continuous surfaces. The psychophysical evidence suggesting a preference for matches with small disparity (McKee and Mitchison 1988) is also consonant with a smoothness constraint, since it implies an assumption that objects in the vicinity of the fixation point are all at a similar distance from the viewer.

A Bayesian model (Knill and Richards 1996) provides a convenient way of framing these constraints. In this approach, each possible solution of the correspondence problem is assigned a probability of being correct. This

Correspondence to: J. C. A. Read
(e-mail: jcr@lsr.nei.nih.gov)

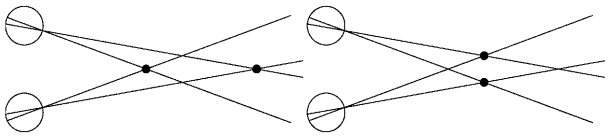

Fig. 1. A classic example of the fact that the correspondence problem has no unique solution. The *left* and *right* diagrams show different alignments of two spheres which both create identical situations on the retinae. The large *open circles* represent the eyeballs, seen from above. The *filled circles* represent spheres in front of the viewer. If the sizes of the spheres are chosen appropriately, both configurations create exactly the same stimulus at the retina. There is no way to distinguish between these situations from stereoscopic disparity alone

probability depends not only on the retinal images themselves, but also on the brain's assumptions about the world, encoded as the a priori probabilities accorded to each solution. For instance, a preference for small disparities can be achieved by according a lower probability to solutions involving large-disparity matches. In this paper, I develop one possible application of Bayesian theory to the correspondence problem.

I am interested in building a biologically realistic model. The brain performs extensive processing on the retinal images before tasks such as the correspondence problem are attempted. A multitude of psychophysical and physiological evidence suggests that the brain analyses images in a set of different spatial frequency and orientation channels, each optimally sensitive to a different Fourier component of the image (Campbell and Robson 1968; Blakemore and Campbell 1969; de Valois et al. 1982a,b; Mansfield and Parker 1993; Eagle 1997; Prince et al. 1998). In previous psychophysical work (Read and Eagle 2000), we investigated this aspect of the correspondence problem by using an illusion which can be easily explained in terms of Fourier channels, but which is hard to explain otherwise (Sato 1998). This is the "reverse phi" motion (Anstis 1970) observed with anti-correlated two-frame kinematograms. In these stimuli, the second frame is not only displaced from the first, it is also anti-correlated; that is, its polarity is inverted, so that previously black pixels become white, and vice versa. The reversed perception obtained with such stimuli is hard to understand in terms of feature-matching mechanisms, since if the mechanism is sensitive to polarity, it will fail, whereas if it is not, it will report veridical motion. However, a qualitative understanding is simple if we assume that the phenomenon involves perceptual channels which have a finite orientation and spatial frequency bandwidth, each sensitive to a different region of the Fourier spectrum of the stimulus.

We can qualitatively understand many aspects of our and others' results by considering the cross-correlation function (CCF) of images after filtering by these channels (Cleary and Braddick 1990; Prince and Eagle 2000b; Read and Eagle 2000). There is evidence that such a calculation is carried out in the brain: the firing rates of disparity- and motion-sensitive complex cells is described well by assuming that they carry out a local cross-correlation of filtered retinal images (Adelson and Bergen 1985; Ohzawa et al. 1990, 1997); the local cross-correlations obtained with filters in quadrature phase are then summed to obtain local contrast energy.

Figure 2 shows the CCF for correlated and anti-correlated one-dimensional (1-D) noise stimuli, filtered

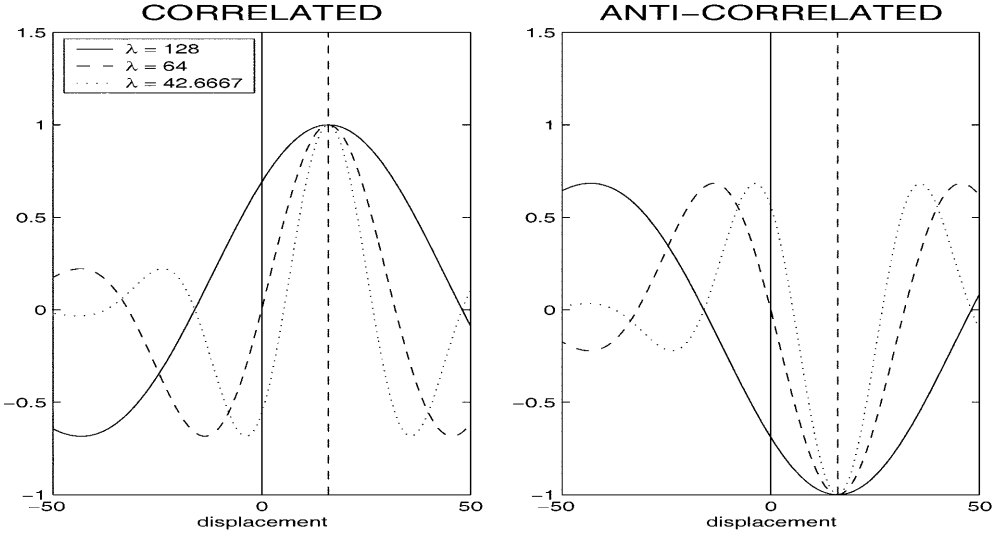

Fig. 2. Cross-correlation functions (CCFs) for a filtered disparate image pair, either correlated (left-hand plot) or anti-correlated (right-hand plot). The images are white noise with a disparity of 16 pixels, following filtering by model spatial frequency channels. Each channel is represented by a Gabor filter with bandwidth 1.5 octaves. The graphs show the mean CCFs (normalised to unit amplitude) which would be obtained by averaging over a large number of random images. In each case, the CCF of the anti-correlated image pair is the inverse of that for the correlated image-pair. When the images are correlated, the CCF has its central peak at the true displacement for all filters (16 pixels, indicated by the dashed vertical line). When the images are anti-correlated, the CCF has its central trough at 16 units, but its side-peaks occur at different positions ($d \pm \lambda/2$) for the different filters. The filters illustrated all have $\lambda > 2d$, so the side-peak closest to the origin is on the left of the origin. Thus, if these channels reported the position of the peak closest to the origin, they would agree on the sign of the displacement (which would be the reverse of the true value), but would differ on its magnitude

by 1-octave channels centred on three different spatial frequencies. The finite bandwidth of the channels removes the ambiguity found with sine-wave gratings. For correlated stimuli, the CCF always has its largest peak at the correct displacement d, irrespective of the preferred spatial period and orientation of the channel. If the channel reports the position of the largest peak, it always gives the correct displacement. However, there is evidence for perceptual reversal at large displacements of narrow-band stimuli (Cleary and Braddick 1990; Prince and Eagle 2000a; Read and Eagle 2000). This suggests that, for larger d, the preference for matches with small displacement wins out over the preference for larger peaks: the subsidiary peak which is closest to the origin – although on the wrong side – is chosen in preference to the largest, correct peak. The existence of an upper limit $D_{\max}$ on direction discrimination can be explained by arguing that, eventually, the subsidiary peaks close to the origin become small enough to be lost in noise, while the large veridical peak is considered too far from the origin to be accepted. One prediction of this qualitative model is that, for a single channel, $D_{\max}$ is proportional to the preferred period λ. This is consistent with experimental evidence across a range of spatial frequencies (Chang and Julesz 1983; de Bruyn and Orban 1989; Cleary 1990; Cleary and Braddick 1990; Smallman and McLeod 1994).

For anti-correlated stimuli, the CCF is inverted. It has two peaks corresponding to the troughs on either side of the central maximum for the correlated stimulus, at $d \pm \lambda/2$. If the peak nearest the origin is the most likely to be chosen, and $d < \lambda/2$, this is in the wrong direction. Quantitative models based on these ideas can explain a wide range of data (Prince and Eagle 2000b).

Fourier's theorem provides us not only with a formula for decomposing an image into its constituent components, but also with an inverse transform for recovering the image from a knowledge of its components. An analogous process must occur within the brain to provide us with our unified perception of the visual world. Thus, to explain the results obtained with broad-band stimuli, we need to understand not only how each individual channel arrives at a perception of displacement, but also how information from different channels is combined. We have previously investigated this in a series of psychophysical experiments using broad-band anti-correlated stereograms and kinematograms (Read and Eagle 2000). The value of anti-correlated stimuli is that they are expected to produce conflicting responses from channels tuned to a different spatial period λ and orientation θ, since peaks in the cross-correlation function occur at horizontal displacements of $d + (2m + 1)\lambda/(2\cos\theta)$, where m is an integer. This is in contrast to the situation for correlated stimuli. Here, the largest peak occurs at the correct position d for every channel, even though the position of the subsidiary peaks $d + m\lambda/\cos\theta$ differs between channels. Studying the responses to anti-correlated stimuli can yield insight into how conflicting responses are combined.

In our psychophysical work (Read and Eagle 2000), we studied conflict between different spatial frequency channels by using 1-D anti-correlated filtered noise with a bandwidth of 5 octaves. Since these images contain only vertical orientations, they are expected to activate just one orientation channel. However, their broad spatial frequency bandwidth should activate a wide range of spatial frequency channels. We found that such broad-band anti-correlated images produced only weak reversed perception. For motion as well as stereopsis, the performance was much closer to chance than for anti-correlated narrow-band images. This suggested that conflict between different spatial frequency channels was impairing, but not completely preventing, a coherent perception. Because the stereopsis and motion results were very similar for these 1-D stimuli, we suggested that the stereopsis and motion systems probably use a similar means of combining information from different spatial frequency channels.

We next presented our subjects with broad-band 2-D images, containing the full range of orientations. We expected these to produce still more cross-channel conflict, since now there is disagreement between channels tuned to different orientations as well as those tuned to different spatial frequencies. In accordance with this expectation, stereoscopic performance was slightly closer to chance than for the 1-D stimuli. However, to our surprise, the perception of reversed motion was greatly enhanced. Most subjects now reached performance close to 0%, indicating a reliable perception of reversed motion (Anstis 1970; Sato 1989). We suggested that this might indicate a difference in how the stereo and motion systems combine information across orientation channels. According to the model developed above, for small displacements ($d < \lambda/(2\cos\theta)$), different orientation channels agree on the *direction* of the displacement, but not on its magnitude. We argued that our results could be explained if, in motion, conflicting reports from different channels do not impede a clear perception of displacement provided the values reported from different channels agree on the direction of the displacement, whereas in stereopsis, a clear perception can be formed only if different channels agree both on the magnitude and sign of disparity.

In fact, although a quantitative model based on these ideas is capable of providing a good match to experimental results, it is not *necessary* to postulate that the two systems differ in how they combine information from different orientations in order to explain the experimental results. In the present paper, I show that the difference between the stereopsis and motion results can, in fact, be explained economically by using stereopsis and motion models which are essentially identical structurally. The different pattern of results is accounted for by the different values of two free parameters: the amount of noise and the strength of the preference for small disparities. By choosing these parameters appropriately for each model, we can reproduce the experimental finding that the two systems perform similarly for 1-D stimuli, both correlated and anti-correlated, while differing strongly in their response to 2-D anti-correlated stimuli. Thus, our original suggestion that the two systems adopted a different algorithm for combining

information from different orientation channels, while plausible, is not the simplest way of explaining the data.

The model proposed here is closely based on the known physiological properties of cells in striate cortex. The initial filtering of the retinal image is carried out by simple cells with receptive fields of different spatial frequencies and orientations. These feed into complex cells according to the energy model (van Santen and Sperling 1984; Adelson and Bergen 1985; Ohzawa et al. 1990, 1997). The extraction of displacement is based on the output of these complex cells. In this, the model is close to previous approaches (Sanger 1988; Qian 1994; Fleet et al. 1996; Zhu and Qian 1996; Qian and Zhu 1997; Prince and Eagle 2000b). However, it differs in the use of Bayesian analysis to calculate the probability of each possible match between the two images. The Bayesian approach provides a natural way to incorporate the preference for smaller displacements, and to combine information across different channels.

For each system (stereopsis/motion), the model is tested with twelve different sets of data: correlated and anti-correlated stimuli with six different spectral profiles. With two free parameters, representing the level of retinal noise and the strength of the prior preference for small displacements, all twelve sets of experimental results can be matched reasonably well. In particular, the model reproduces the observation that, with anti-correlated stereograms containing power at all orientations and spatial frequencies, subjects perform close to chance, whereas with the same stimuli presented as kinematograms, they consistently report the wrong answer.

2 Methods

2.1 Images

The images used are the same as in our psychophysical work (Read and Eagle 2000). At the viewing distance used, the 128×128 pixels of the image took up $1.7^\circ \times 1.7^\circ$ on the retina. Our images were constructed to maximise cross-channel conflict by supplying the same power to each spatial frequency and orientation channel. Spatial frequency channels have been estimated to have roughly equal octave bandwidths across a range of frequencies (de Valois and de Valois 1988). We therefore constructed our images to have equal spectral power in equal octaves, a property approximately exhibited by some natural images. For 1-D patterns, this meant that the Fourier power had to scale inversely with frequency; for 2-D patterns, as the inverse-square of frequency. We used a variety of spatial frequency and orientation bandwidths. In each case the central frequency was 3.2 cycles per degree. The contrast of images with different bandwidths was set so as to preserve the property of equal power in equal octaves, so that the 5-octave images had $\sqrt{5}$ times the contrast of the 1-octave images. Similarly, we wished the channels tuned to vertical orientations to be activated equally by our 1-D and 2-D stimuli. Since there is both physiological and psychophysical evidence that the bandwidth of orientation channels is approximately 30° (Campbell and Kulikowski 1966; de Valois et al. 1982b; Mansfield and Parker 1993), we gave each 30° orientation band of our 2-D images the same power as a 1-D image of the same octave bandwidth.

At every trial, the image was "shuffled" in the same way as for human observers (Read and Eagle 2000): a random number of columns of pixels were removed from the right edge of the stored image and added back onto the left edge; then, a random number of rows were removed from the top of the image and added onto the bottom. In the psychophysical experiments, this shuffling was to avoid providing a monocular cue to disparity by having particular features recurring in the same place. In the computer model, shuffling ensured that the position of particular features in the stored images was randomised with respect to the retinal array of receptive field centres (shown in Fig. 5).

In both psychophysical experiments and computer simulations, the shuffled image formed the stimulus to the left eye/first frame. This was then displaced by wrapping columns of pixels around to form the stimulus for the right eye/second frame. This preserved the Fourier spectrum of the band-pass stimuli without introducing luminance discontinuities (Read and Eagle 2000).

2.2 Experimental procedure

In both motion and stereopsis psychophysical experiments, we used a two-interval forced-choice protocol. Observers pressed a keypad button to indicate which interval contained rightward motion/crossed disparity. Subjects performed 80 trials at each of 7 displacements. The motivation for using a two-interval protocol was to circumvent any observer bias, for example a tendency to perceive crossed disparity. We assumed that, in the event of perceptual ambiguity, subjects would report which interval seemed to contain the *most* crossed disparity/rightward motion. The model is constructed accordingly. In each interval, the model is presented with two images, corresponding to left/right retinae for the stereo model, and first/second frames for the motion model. It compares these two retinae/frames and arrives at an estimate for the global displacement in the horizontal direction in that interval. It then reports which interval contained the most crossed disparity/fastest rightward motion. If the same displacement was estimated for both intervals, the model reports either interval at random. The method of obtaining the estimated global horizontal displacement is described in the following sections.

2.3 Initial processing

The Bayesian probability analysis is based on the output of displacement-tuned complex cells described by the energy model put forward by van Santen and Sperling (1984) and Adelson and Bergan (1985) in the context of motion, and adapted for stereopsis by Ohzawa et al.

(1990, 1997) and Fleet et al. (1996). The stereopsis and motion versions of the energy model are essentially the same. In each case, the complex cell receives input from a set of four simple cells, which in turn receive input from a matched pair of receptive fields (RFs). In the stereopsis version of the model, the simple cells are binocular: they have RFs in each retina. In the original motion model, the two RFs represent different temporal phases of the response. One set of RFs responds to the first frame of the kinematogram, and the other to the second frame. Physiologically, this could be achieved by a dynamic RF function, which varies with time as well as space. I do not model the temporal profile in detail, but simplify it into two snapshots, representing the average response to the first and second frames. In what follows I describe the stereopsis version of the energy model, indicating occasionally how the interpretation would be modified for motion-sensitive cells.

There is evidence that simple cells exhibit nearly linear spatial summation (Hubel and Wiesel 1962; Movshon et al. 1978; Anzai et al. 1999a), although including an output non-linearity. The model assumes that simple cells respond to the difference $I(x,y)$ between the luminance at a given point (x,y), and the mean luminance across the whole retina. They calculate the convolution of this difference with their RF function $\rho(x,y)$, which gives the cell's response to a spot of light at position (x,y) in the retina. Negative values of $\rho(x,y)$ represent an OFF region, in which bright stimuli tend to suppress firing. The spatial frequency and orientation tuning depends on the RF profile. In the model, simple cell RFs are represented by Gabor functions (Appendix A), since most real simple cell RFs can be fitted well using a Gabor function with appropriate parameters.

Because simple cells have a low baseline firing rate, they cannot signal negative values, and so output a half-wave rectified version of this convolution. The output of a simple cell with RF centred on (x_0,y_0) is therefore

$$S(x_0,y_0) = \mathrm{Pos}\left(\int\int \mathrm{d}x\,\mathrm{d}y\; I(x,y)\rho(x-x_0,y-y_0)\right) \equiv \mathrm{Pos}(v(x_0,y_0)) \quad , \tag{1}$$

where Pos represents half-wave rectification and we define v to represent the convolution of an image with an RF. However, the energy model side-steps this half-wave rectification by assuming that simple cells occur in pairs whose RF functions are inverses of each other, so that every ON region in one cell is matched by an OFF region in the other. The effect of this is to remove the half-wave rectification (because $\mathrm{Pos}(x) - \mathrm{Pos}(-x) \equiv x$). Thus, the value of the convolution v is encoded by the difference in the firing rates of two simple cells, only one of which fires in response to any given image.

The simple cells which feed into each complex cell are divided into two types, with different receptive field symmetry. One set has a RF with even symmetry – for example, a central ON region flanked by two weaker OFF regions. The other set has odd-symmetric RFs – for example, an ON region next to an OFF region of the same strength (Fig. 3). I shall refer to these as odd and even simple cells. The choice of purely even and odd cells is convenient but arbitrary (Ohzawa et al. 1990, 1997). The only requirement of the energy model is that the simple cells feeding into each complex cell should be divided into two groups differing in phase by $\pi/2$.

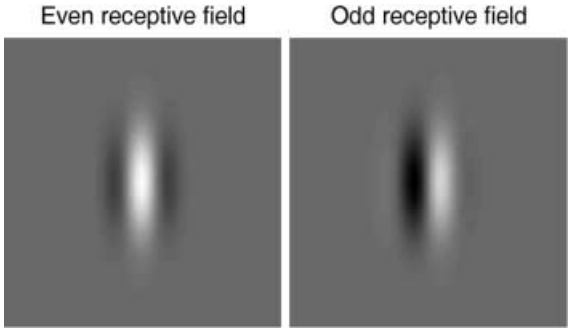

Fig. 3. Even and odd receptive fields (RFs). Both RFs are Gabor functions with a spatial frequency bandwidth of 1.5 octaves and orientation bandwidth of 30° (see Appendix A). The model retina shown is 128 × 128 pixels, and the preferred spatial period of both RFs is 32 pixels. The same greyscale is used in plotting both RFs. Bright regions of the plot represent ON regions of the RF; dark regions represent OFF regions

A binocular simple cell is achieved by combining four such monocular simple cells. I assume that all four have same-symmetry RFs (all even or all odd), with the same orientation and spatial-frequency tuning. Two of the monocular cells excite the binocular cell, and two inhibit it. The firing rate of the binocular cell is the half-wave rectified sum of its inputs:

$$\begin{aligned} B &= \mathrm{Pos}\{\mathrm{Pos}[v_\mathrm{L}(x_{0\mathrm{L}},y_{0\mathrm{L}})] - \mathrm{Pos}[-v_\mathrm{L}(x_{0\mathrm{L}},y_{0\mathrm{L}})] \\ &\quad + \mathrm{Pos}[v_\mathrm{R}(x_{0\mathrm{R}},y_{0\mathrm{R}})] - \mathrm{Pos}[-v_\mathrm{R}(x_{0\mathrm{R}},y_{0\mathrm{R}})]\} \\ &= \mathrm{Pos}\{v_\mathrm{L}(x_{0\mathrm{L}},y_{0\mathrm{L}}) + v_\mathrm{R}(x_{0\mathrm{R}},y_{0\mathrm{R}})\} \quad . \end{aligned} \tag{2}$$

Four such binocular simple cells feed into each complex cell. The output of the complex cell is assumed to reflect the sum of the square of its inputs. This squaring non-linearity is the reason for the name "energy" model. Again, I assume that binocular simple cells are combined in pairs whose receptive fields are inverses of one another. Since $[\mathrm{Pos}(x)]^2 + [\mathrm{Pos}(-x)]^2 \equiv x^2$, the effect of this is to remove the half-wave rectification in (2). Finally, then, the output of the complex cell can be written

$$\begin{aligned} C(x_{0\mathrm{L}},y_{0\mathrm{L}},x_{0\mathrm{R}},y_{0\mathrm{R}}) &= \left[v_\mathrm{L}^\mathrm{even} + v_\mathrm{R}^\mathrm{even}\right]^2 + \left[v_\mathrm{L}^\mathrm{odd} + v_\mathrm{R}^\mathrm{odd}\right]^2, \\ &= [v_\mathrm{L}^\mathrm{even}]^2 + [v_\mathrm{R}^\mathrm{even}]^2 + [v_\mathrm{L}^\mathrm{odd}]^2 + [v_\mathrm{R}^\mathrm{odd}]^2 \\ &\quad + 2\left[v_\mathrm{L}^\mathrm{even} v_\mathrm{R}^\mathrm{even} + v_\mathrm{L}^\mathrm{odd} v_\mathrm{R}^\mathrm{odd}\right] \quad . \end{aligned} \tag{3}$$

Equation (3) represents the response of a complex cell tuned to horizontal disparity $(x_{0\mathrm{L}} - x_{0\mathrm{R}})$ and vertical disparity $(y_{0\mathrm{L}} - y_{0\mathrm{R}})$. The motion energy model involves entirely analogous expressions. Instead of the subscripts L and R (for left and right retinae), we have 1 and 2 (for first and second frames). The cross-terms represent a local cross-correlation between filtered versions of the first and second frames. The difference between the

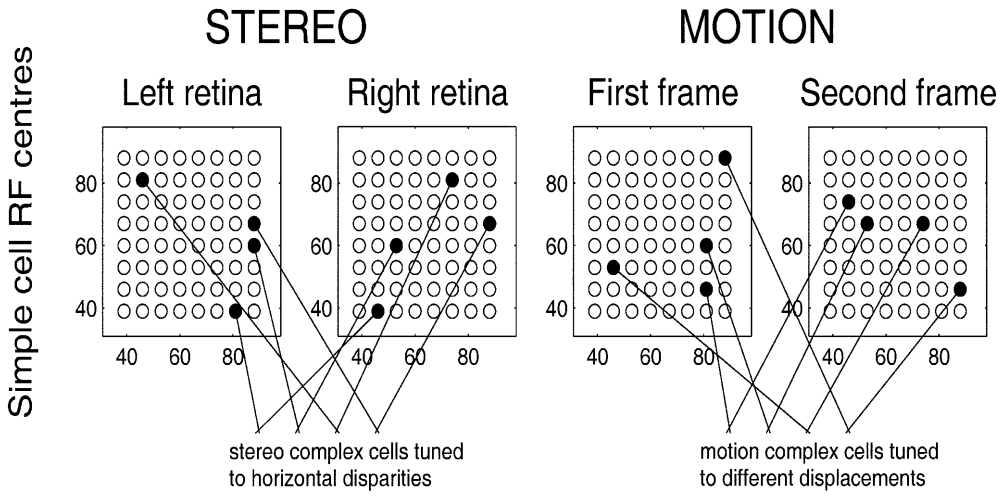

Fig. 4. Schematic wiring diagram of disparity-sensitive complex cells (*left*) and motion-sensitive complex cells (*right*). Complex cells receive input from simple cells with RFs at different positions in each image. The array of circles represent the possible locations of simple cell RFs on the retina. The circles represent the centre of each receptive field (there is one such array for each spatial period λ and orientation θ). In the stereopsis model, complex cells receive input from simple cells at the same vertical position on each retina. The results shown in this paper used an 9×9 array of retinal RF positions. Thus, within each spatial frequency/orientation channel, there are $9^3 = 729$ complex cells tuned to zero vertical disparity, of which four are shown schematically here. In the motion model, there are $9^4 = 6561$ complex cells, tuned to displacements in all directions. Four are shown schematically, receiving input from simple cell RFs at (in general) different horizontal and vertical positions in each frame

positions of the RFs in the two images determines the displacement/disparity tuning of the complex cell. That is, my models are based on position rather than phase disparity, and incorporate only tuned-excitatory cells (Poggio and Fischer 1977).

The monocular terms represent the local intensity of the filtered images. For instance, $v_{\mathrm{L}}^{\mathrm{even}}(x_{0\mathrm{L}}, y_{0\mathrm{L}})$ represents the intensity, at the retinal position $(x_{0\mathrm{L}}, y_{0\mathrm{L}})$, of the left image after filtering by an even Gabor function. The cross-terms represent a local cross-correlation between the filtered images in each eye. Thus, the energy complex cell computes the contrast energy of the two images, after filtering them through a band-pass filter tuned to a particular spatial frequency and orientation.

The receptive fields of the component simple cells endow the complex cell with its spatial frequency and orientation tuning. The spatial period and orientation of the carrier cosine grating correspond to the preferred spatial period and orientation of the simple cell. The spatial extent of the envelope determines its spatial frequency and orientation bandwidths. Evidently, simple cells with wider envelopes relative to the carrier cosine have narrower bandwidths (Appendix A).

The energy model captures several features of real complex cells (Ohzawa et al. 1990, 1997; Anzai et al. 1999b). The response is unchanged by inverting the contrast in both images (e.g. the cell responds equally to two bright bars as to two dark bars). It responds best to features with a particular displacement, irrespective of precisely where they occur within its receptive field. The energy model does not capture the reduced response of complex cells to anti-correlated stimuli (Ohzawa et al. 1990; Cumming and Parker 1997; Read et al. 2000). However, if we assume that all complex cells have their responses reduced by the same factor for anti-correlated stimuli, this complication should make little difference to the results of the modelling.

In the model, a particular spatial frequency/orientation "channel" corresponds to a population of complex cells tuned to spatial period λ and orientation θ. Psychophysical and physiological evidence suggests that the orientation bandwidth of both stereopsis and motion channels is around 30° (Campbell and Kulikowski 1966; Mansfield and Parker 1993). I therefore set the orientation bandwidth to 30° for all channels, although in reality this may be an over-simplification – for instance, channels tuned to vertical orientations may have smaller bandwidth than those tuned to off-orientations. Thus, six orientation channels, tuned to 0°, 30°, 60°, 90°, 120° and 150°, cover the full range. Psychophysical and physiological evidence suggests that, across a wide range of spatial frequencies, different channels have roughly the same octave bandwidth, around 1–2 octaves (de Valois and de Valois 1988). I therefore set the octave bandwidth to the same value – 1.5 octaves – for all channels, and spaced the spatial frequency channels 1 octave apart. Again, this is an over-simplification, because in fact there is evidence that higher frequency channels do have smaller octave bandwidths. The range of spatial frequencies easily visible to humans spans around 5 octaves (de Valois and de Valois 1988), so most relevant frequencies could be covered with five channels, tuned to 0.6, 1.2, 2.4, 4.8 and 9.6 cycles per degree (1, 2, 4, 8 and 16 cycles per image).

In each retina (for stereopsis) and frame (for motion), we have an array of $n \times n$ monocular RFs. The model thus has the potential for complex cells tuned to n^4 different displacements (Fig. 4). I now assume that the stereopsis system uses only horizontal disparity to reconstruct depth [vertical disparities (Mayhew and

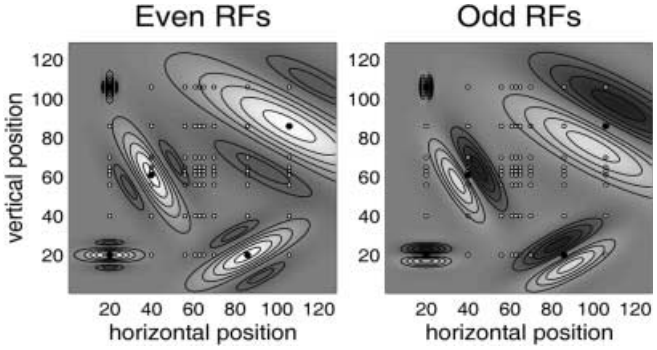

Fig. 5. This figure shows some of the receptive fields used in the model. The model retina is 128×128 pixels. The array of RF centres is indicated by the *small dots*. In each plot, five example RFs are shown. The *left plot* shows even-symmetric RFs and the right plot odd-symmetric ones. The RFs have spatial period 64, 32, 16 and 8 pixels. (No example is shown of an RF with the largest spatial period, 128 pixels, because it would cover almost the whole retina and thus obscure the other RFs.) *Dark shading* represents the OFF regions of the RF, and *pale shading* the ON regions. *Contour lines* are also superimposed. For clarity, the RF functions shown here all have the same amplitude. In fact, in the model, the higher frequency RF functions are given larger amplitude, so that despite the smaller extent of these RFs, all simple cells respond with the same firing rate to their own optimal stimulus. The model includes 4860 such RF functions: odd/even RFs at each of the 9×9 possible positions on the retina, with six different orientations and five different spatial frequencies

Longuet-Higgins 1982) do not in any case contribute to the front/back discrimination task modelled here]. Thus the stereo model includes only the n^3 complex cells tuned to horizontal disparities. Complex cells have RFs centred at the same vertical position y_0 in each retina, although the horizontal positions $x_{0\mathrm{L}}$ and $x_{0\mathrm{R}}$ are different in the left and right retina. In contrast, the motion system must be sensitive to movement in any direction. The motion model thus includes the full n^4 complex cells. In general, complex cell have RFs centred at different positions in the first frame (x_{01}, y_{01}) and the second frame (x_{02}, y_{02}).

Within each channel, complex cells have the same spatial frequency and orientation tuning, but differ in the location of their retinal receptive fields. With the assumption of constant octave bandwidth, the spatial extent of each receptive field scales linearly with the preferred spatial period. Thus, in order to ensure equivalent coverage in each spatial frequency channel, it would be logical to scale the spacing of the RFs according to their spatial frequency tuning. In practice, this is difficult to implement. The spacing of the simple-cell RFs determines the possible disparities to which a complex cell may be tuned. If the spacing is different for different channels it is hard to compare different channels. Interpolating between the different disparities is unsatisfactory, because for some stimuli random noise causes different disparities to be assigned very different probabilities. I therefore set the array spacing the same for all channels, irrespective of RF extent. Effectively, then, the model retina is covered more densely by RFs tuned to lower spatial frequencies (Fig. 5).

The computer time and memory required increases very rapidly with the number of RFs (Sect. 2.10) – a 9×9 grid was the largest that was practicable to run. Yet, if the maximum separation between RF centres is too small, large displacements will not be correctly perceived. Conversely, if the minimum separation between RFs is too large, the model will fail for small displacements. I needed to build an RF array suitable for the data I wished to model, which includes displacements ranging from 2 pixels (1.6 arcmin) to 45 pixels (36.4 arcmin). In order to cover this range of displacements with the smallest possible number of RFs, I used the irregular array shown in Fig. 5. The nine RF centres are located at 20, 40, 56, 61, 63, 65, 70, 86 and 106 pixels across the 128-pixel retinal image. The differences between these horizontal positions define the disparities to which the model is sensitive. With this choice of RF centres, the model is sensitive to 41 different disparities: $0, \pm 2, \pm 4, \pm 5, \pm 7, \pm 9, \pm 14, \ldots, \pm 41, \pm 43, \pm 45, \pm 46, \pm 50, \pm 66$ and ± 86 (representing all possible differences between pairs of RF centres, whose positions are given above). This range (2–86 pixels = 1.4–69 arcmin) is suitable for the data I wish to model. Other data might require finer spacing between RF centres, with the associated computational overhead.

I did not include any RFs centred on positions outside the 128×128 pixels of the stimulus; i.e. I did not explicitly model simple cells whose receptive fields fell predominantly on the grey background region of the screen. In the psychophysical experiments, attention was directed to the target region. Furthermore, simple cells outside the target region have low firing rates and thus are not expected to contribute to the task, since the model automatically assigns greater significance to cells with higher firing rates (Sect. 2.5).

2.4 Noise

The sources of biological noise are currently uncertain. Opinions vary as to the reliability of cortical neurons (Stevens 1994; Rieke et al. 1997). I started with the assumption that all the noise in the visual system arises at the retina (for instance, due to "shot" noise from the finite number of photons arriving at each receptor), with cortical calculations introducing no additional noise. Thus, if $\tilde{I}(x, y)$ is the "perfect" image as presented to the retina and $I(x, y)$ is the noisy image available to the brain, then

$$I(x, y) = \tilde{I}(x, y) + \zeta \epsilon(x, y) \ , \tag{4}$$

where ζ is the standard deviation of retinal noise and $\epsilon(x, y)$ is a random variable drawn from a standard normal distribution (mean 0, variance 1).

Since simple cells compute the convolution of the image with their receptive field, the effect of the retinal noise is to add normally distributed random noise to the output of the simple cells. The standard deviation of the noise affecting each simple cell reflects the integral-squared of its receptive field function. For the Gabor receptive fields used here (Appendix A), this turns out to be

$$\xi_{\text{even/odd}} = \zeta \sqrt{\frac{1}{8\pi\sigma_x\sigma_y}\left[1 \pm \exp\left(-\frac{4\pi^2\sigma_x^2}{\lambda^2}\right)\right]} \ , \qquad (5)$$

where the plus sign holds for even receptive fields and the minus for odd. The model channels have constant octave bandwidth, so their spatial extent σ_x, σ_y scales with their period λ (Eqs. A6 and A8). Thus from (5), the amplitude of the noise on the convolution scales as $1/\lambda$.

When implementing the model computationally, noise is in fact added directly to the convolutions calculated by the simple cells, rather than to the retinal images themselves. This means that our model has to include only the 128×128 pixels of the experimental stimulus itself, and not a surrounding region representing the grey background, even though the receptive fields of some simple cells extend into this background. The grey region contributes nothing to the mean firing rate of such cells, and the additional noise it contributes is taken into account in the square of the integral.

2.5 Bayesian analysis

Each spatial frequency/orientation channel consists of a set of complex cells, tuned to different horizontal and (in the case of motion) vertical displacements. Consider a motion-tuned complex cell that receives input from simple cells with RFs centred on (x_{01}, y_{01}) in the first frame for motion, and (x_{02}, y_{02}) in the second frame. The complex cell is thus tuned to a horizontal displacement of $(x_{01} - x_{02})$ and a vertical displacement of $(y_{01} - y_{02})$. Effectively, it is considering the possibility that the region around (x_{01}, y_{01}) in the first frame corresponds to the region around (x_{02}, y_{02}) in the second. Each complex cell in a given channel is considering a different possible match. Evidently, a stimulus of a particular disparity will preferentially activate complex cells tuned to that disparity. Thus, one obvious way to judge the disparity of the stimulus from the activity of the complex cells is to see which complex cells are responding most vigorously (Qian 1994; Fleet et al. 1996; Zhu and Qian 1996; Qian and Zhu 1997; Prince and Eagle 2000b). However, I found that a modification of this approach, incorporating a Bayesian probability analysis, was more successful at capturing the observed psychophysical response to broad-band and anti-correlated stimuli. In what follows, I discuss the Bayesian analysis in terms of the motion model, since that is more general. The discussion can easily be adapted to the stereopsis model (''first/second frame'' becomes ''left/right retinal image'' etc.), with the additional restriction to zero vertical disparity.

The first step in the Bayesian analysis is to calculate the relative probability of the possible matches considered by the whole population of complex cells in each channel. First of all, I invoke the smoothness or continuity constraint. I assume that the disparity varies only slowly across the image, so that across each receptive field, the local disparity is approximately constant. Since the octave bandwidth is approximately constant across different channels, the size of the receptive field scales linearly with its preferred spatial period. This means that different channels ''smooth'' out disparity variations over a scale depending on their spatial period. Some form of smoothness or continuity constraint like this is necessary in order to arrive at a unique solution of the fundamentally ill-posed correspondence problem (Marr 1982). We can then argue that, if the feature at (x_{01}, y_{01}) in the first frame really does correspond to the feature at (x_{02}, y_{02}) in the second, the first and second frames are locally related by

$$I_1(x, y) = I_2(x - x_{01} + x_{02}, y - y_{01} + y_{02}) \ . \qquad (6)$$

It is then easy to prove that, in the absence of retinal noise, the convolution of the first image over each simple cell's first RF would be the same as the convolution of the second image over the corresponding second RF:

$$\begin{aligned} v_1^{\text{even}} &= \int\int \mathrm{d}x\,\mathrm{d}y \rho^{\text{even}}(x - x_{01}, y - y_{01}) I_1(x, y) \\ &= \int\int \mathrm{d}x\,\mathrm{d}y \rho^{\text{even}}(x - x_{02}, y - y_{02}) I_2(x, y) \\ &= v_2^{\text{even}} \ . \end{aligned} \qquad (7)$$

The importance of this observation is that it enables the response of the motion-sensitive complex cell to be predicted from a knowledge only of the convolutions of the *first* image. The difference between this prediction and the actual firing rate of the model complex cell enables us to assess the validity of the assumption that the match considered by this complex cell is correct. Some difference is expected, due to the noise. However, the greater the difference, the less likely it is that this match is correct. We can quantify this notion mathematically.

We wish to assess the probability that (A) the retinal position (x_{01}, y_{01}) in the first image really does correspond to position (x_{01}, y_{01}) in the second image, given (B) the firing rate of the complex cell, $C(x_{01}, y_{01}, x_{02}, y_{02})$, together with the convolutions of the first image over even and odd receptive fields, $v_1^{\text{even}}(x_{01}, y_{01})$ and $v_1^{\text{odd}}(x_{01}, y_{01})$. We use Bayes' theorem, which states that $\mathscr{P}\{A|B\} = \mathscr{P}\{B|A\}\,\mathscr{P}\{A\}/\mathscr{P}\{B\}$, where $\mathscr{P}\{A\}$ represents the prior probability of event A, and $\mathscr{P}\{A|B\}$ represents the probability of event A, given that event B has occurred.

One possibility would be to consider the complex-cell firing rate C directly, and compare this to the value expected in the absence of noise and assuming that the match is correct. From (3), this expected value is $4([v_1^{\text{even}}]^2 + [v_1^{\text{odd}}]^2)$. It turns out to be more convenient to normalise the complex cell firing rate by dividing by the expected value. I refer to this normalised value as K_1, where the subscript indicates that the expected value was estimated using the values of convolutions in the first frame:

$$K_1(x_{01}, y_{01}, x_{02}, y_{02}) = \frac{C(x_{01}, y_{01}, x_{02}, y_{02})}{[v_1^{\text{even}}(x_{01}, y_{01})]^2 + [v_1^{\text{odd}}(x_{01}, y_{01})]^2} \ . \qquad (8)$$

If the complex cell really is tuned to the correct match, then we would expect this normalised firing rate K_1 to be 1. The further it departs from unity, the less likely it is that the cell is, in fact, tuned to the correct match.

Mathematically, we are calculating the Bayesian posterior:

$$\begin{aligned}&\mathscr{P}\{(x_{01},y_{01})\Leftrightarrow(x_{02},y_{02})\big|K_1(x_{01},y_{01},x_{02},y_{02}),\\&\quad v_1^{\text{even}}(x_{01},y_{01}),v_1^{\text{odd}}(x_{01},y_{01})\}\\&\quad=\mathscr{P}\{K_1(x_{01},y_{01},x_{02},y_{02})\big|(x_{01},y_{01})\Leftrightarrow(x_{02},y_{02}),\\&\qquad v_1^{\text{even}}(x_{01},y_{01}),v_1^{\text{odd}}(x_{01},y_{01})\}\\&\qquad\times\frac{\mathscr{P}\{(x_{01},y_{01})\Leftrightarrow(x_{02},y_{02})\big|v_1^{\text{even}}(x_{01},y_{01}),v_1^{\text{odd}}(x_{01},y_{01})\}}{\mathscr{P}\{K_1(x_{01},y_{01},x_{02},y_{02})\big|v_1^{\text{even}}(x_{01},y_{01}),v_1^{\text{odd}}(x_{01},y_{01})\}}\ ,\end{aligned}\tag{9}$$

where the symbol $\Leftrightarrow$ means "corresponds to" or "is the correct match for". In the following sections, each term on the right-hand side of this equation is now discussed in detail. By making appropriate assumptions regarding the brain's experience of the visual world, each term on the right can be evaluated. Hence, we can derive the probability that the match under consideration is correct. For brevity, from now on this probability will be written simply as

$$\begin{aligned}&\mathscr{P}_{\theta\lambda}\{(x_{01},y_{01})\Leftrightarrow(x_{02},y_{02})\}\\&\quad\equiv\mathscr{P}\{(x_{01},y_{01})\Leftrightarrow(x_{02},y_{02})\big|K_1(x_{01},y_{01},x_{02},y_{02}),\\&\qquad v_1^{\text{even}}(x_{01},y_{01}),v_1^{\text{odd}}(x_{01},y_{01})\}\ ,\end{aligned}\tag{10}$$

where I have dropped the explicit reminders that this probability depends on the normalised complex-cell firing rate and monocular convolutions, in favour of a more compact subscript $\theta\lambda$. This reminds us that this is the *local single-channel probability*, derived from the output of a single complex cell, looking at a particular region of the retina, and tuned to a particular spatial period λ and orientation θ. In Sect. 2.9, I discuss how to combine information from different channels.

2.6 The prior

The numerator in (9), $\mathscr{P}\{(x_{01},y_{01})\Leftrightarrow(x_{02},y_{02})\big|v_1^{\text{even}}(x_{01},y_{01}),v_1^{\text{odd}}(x_{01},y_{01})\}$, is the Bayesian prior. Mathematically, it describes the a priori probability that retinal position (x_{01},y_{01}) in the first frame corresponds to retinal position (x_{02},y_{02}) in the second frame, given the monocular convolutions $v_1^{\text{even}}(x_{01},y_{01})$ and $v_1^{\text{odd}}(x_{01},y_{01})$. *The single-frame convolutions can contain no information about the correspondence, so the values* v_1^{even} *and* v_1^{odd} *are irrelevant*. However, not all possible matches are considered equally probable. This is where we build in the brain's preference for small displacements. The prior probability of a match is assumed to depend on the magnitude of the displacement (not its direction):

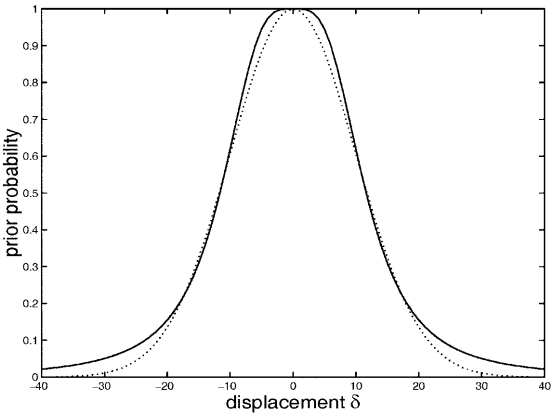

Fig. 6. The Bayesian prior, incorporating a preference for small disparities. The *solid line* shows the priors employed in the models. These are Helmholtz-coils functions (13) with $D = 10$ pixels, 7.5 arcmin. For comparison, the *dotted lines* show a Gaussian with standard deviation D. The functions are very similar, but the Gaussian is more sharply peaked at the origin

$$\delta=\sqrt{(x_{01}-x_{02})^2+(y_{01}-y_{02})^2}\ .\tag{11}$$

Thus

$$\begin{aligned}&\mathscr{P}\{(x_{01},y_{01})\Leftrightarrow\ (x_{02},y_{02})\big|v_1^{\text{even}}(x_{01},y_{01})v_1^{\text{odd}}(x_{01},y_{01})\}\\&\quad=\mathscr{P}\{\delta\}\ ,\end{aligned}\tag{12}$$

where $\mathscr{P}\{\delta\}$ is a prior probability function describing the brain's assessment of how intrinsically likely it is to encounter a match with displacement δ. This, presumably, depends on the brain's experience of the visual world built up throughout life. As discussed, it appears that the brain considers small displacements intrinsically more likely. In the model, I used the prior probability function shown in Fig. 6:

$$\begin{aligned}\mathscr{P}\{\delta\}=&\left[D^2+(\delta-D/2)^2\right]^{-3/2}\\&+\left[D^2+(\delta+D/2)^2\right]^{-3/2}\ ,\end{aligned}\tag{13}$$

where D is a scale length, defining which disparities count as "small" (the prior probability falls to roughly half its maximum when $\delta = D$). I chose the form given in (13) because it is very flat at the origin; the first four derivatives are zero. (The inspiration came from classical electromagnetism: the function in (13) in fact describes the magnetic field in between two Helmholtz coils, an arrangement designed to give a region of extremely uniform field.) This property was chosen so that the model brain would prefer small disparities to larger ones, but would not show such a strong preference for zero disparity as, for example, a Gaussian. In addition, the Helmholtz function, having broader shoulders, is more tolerant of large disparities than a Gaussian. However, the precise form of the function is not important. I have experimented with a Gaussian ($\exp(-\delta^2)$) and obtained similar results; Prince and Eagle (2000b) used an

exponential ($\exp(-|\delta|)$) for a similar purpose (although not actually a Bayesian prior). The key property is that the function declines monotonically with the magnitude of the displacement.

2.7 The likelihood

The first probability on the right-hand side of (9),

$$\mathscr{P}\{K_1(x_{01},y_{01},x_{02},y_{02})\big|(x_{01},y_{01})\Leftrightarrow(x_{02},y_{02}),\; v_1^{\text{even}}(x_{01},y_{01}),v_1^{\text{odd}}(x_{01},y_{01})\}\ ,$$

is the Bayesian likelihood. It is the probability of obtaining a particular normalised complex-cell firing rate, given the convolutions of the first image over odd and even cells, and the assumption that the postulated match is correct. This probability distribution can be derived analytically using the smoothness constraint discussed in Sect. 2.5.

We have seen (7) that in the absence of noise, given that the match is correct and disparity is locally constant, the convolution of the first and second images across corresponding receptive fields are equal. That is, $\tilde{v}_2^{\text{even}}=\tilde{v}_1^{\text{even}}$ and $\tilde{v}_2^{\text{odd}}=\tilde{v}_1^{\text{odd}}$ (7), where we use a tilde to denote the "ideal" convolutions which would be obtained if the retinal image were free of noise. In fact, the brain has access only to the noisy values ($v_1^{\text{even}}=\tilde{v}_1^{\text{even}}+\xi^{\text{even}}\epsilon$ etc.), where ξ is the amplitude of noise on the convolution (5), and we are using ϵ to represent a random variable drawn from a standard normal distribution. We assume that the noise is uncorrelated between frames and in different retinae. The noisy convolutions are therefore related by

$$\begin{aligned} v_2^{\text{even}}-v_1^{\text{even}}&=\epsilon\xi^{\text{even}}\sqrt{2},\\ v_2^{\text{odd}}-v_1^{\text{odd}}&=\epsilon\xi^{\text{odd}}\sqrt{2}\ . \end{aligned} \tag{14}$$

Note that the *average* difference between the noisy convolutions is still zero, as it would be in the absence of noise.

The firing rate of the complex cell is (3)

$$C(x_{01},y_{01},x_{02},y_{02})=\left[v_1^{\text{even}}(x_{01},y_{01})+v_2^{\text{even}}(x_{02},y_{02})\right]^2+\left[v_1^{\text{odd}}(x_{01},y_{01})+v_2^{\text{odd}}(x_{02},y_{02})\right]^2\ . \tag{15}$$

If the postulated match is correct, we can substitute for v_2^{even} and v_2^{odd} in (15) from (14):

$$C(x_{01},y_{01},x_{02},y_{02})=\left[2v_1^{\text{even}}(x_{01},y_{01})+\epsilon\xi^{\text{even}}\sqrt{2}\right]^2+\left[2v_1^{\text{odd}}(x_{01},y_{01})+\epsilon\xi^{\text{odd}}\sqrt{2}\right]^2\ . \tag{16}$$

In the Bayesian analysis, we normalise the complex cell firing rate by dividing by the value which would have been obtained in the absence of noise:

$$\tilde{C}(x_{01},y_{01})=\left[2v_1^{\text{even}}(x_{01},y_{01})\right]^2+\left[2v_1^{\text{odd}}(x_{01},y_{01})\right]^2\ . \tag{17}$$

Thus we define $K_1\equiv C/\tilde{C}$ (8). We can then use (16) to derive an analytical expression for the probability density function (PDF) of K_1, describing what noise-related fluctuations in K_1 can be expected under the null hypothesis that the complex cell is tuned to the correct match. This turns out to be

$$\begin{aligned} f(K_1,v_1^{\text{even}},v_1^{\text{odd}})=&\frac{\tilde{C}}{2\pi\xi^{\text{even}}\xi^{\text{odd}}}\\ &\times\int_{-\pi/2}^{\pi/2}\mathrm{d}w\left\{\exp\left[-\frac{1}{[\xi^{\text{even}}]^2}\left(v_1^{\text{even}}-\sqrt{\frac{\tilde{C}K_1}{2}(1+\sin w)}\right)^2\right]\right.\\ &\left.+\exp\left[-\frac{1}{[\xi^{\text{even}}]^2}\left(v_1^{\text{even}}+\sqrt{\frac{\tilde{C}K_1}{2}(1+\sin w)}\right)^2\right]\right\}\\ &\times\left\{\exp\left[-\frac{1}{[\xi^{\text{odd}}]^2}\left(v_1^{\text{odd}}-\sqrt{\frac{\tilde{C}K_1}{2}(1-\sin w)}\right)^2\right]\right.\\ &\left.+\exp\left[-\frac{1}{[\xi^{\text{odd}}]^2}\left(v_1^{\text{odd}}+\sqrt{\frac{\tilde{C}K_1}{2}(1-\sin w)}\right)^2\right]\right\}, \end{aligned} \tag{18}$$

with $\tilde{c}$ given in (17). This must be evaluated numerically for each complex-cell response (since the PDF depends on three variables, it was not practical to store it as a look-up table). I used the trapezoidal rule with 200 steps. If neither even nor odd simple cells were firing, so that $\tilde{c}=0$, the normalised complex cell response is undefined. If this occurred, the PDF was set equal to zero. However, this was unlikely to occur, since the retinal noise means that simple cells maintain a small firing rate.

The likelihood itself is then $f(K_1,v_1^{\text{even}},v_1^{\text{odd}})\mathrm{d}K_1$. However, since the normalised units are the same for all channels and all contrasts, the $\mathrm{d}K_1$ term simply introduces an overall scaling factor, and can therefore be neglected (this, in fact, was the motivation for using the normalised response).

2.8 Complex cells

The last term on the right-hand side of (9) is the denominator $\mathscr{P}\{K_1(x_{01},y_{01},x_{02},y_{02})\big|v_1^{\text{even}}(x_{01},y_{01}),v_1^{\text{odd}}(x_{01},y_{01})\}$. In principle, this can be deduced from (9) by summing both sides over all possible situations and requiring the result to equal unity. In the summation, we must take into account not only all possible matches (x_{02},y_{02}) in the second frame for the point (x_{01},y_{01}) under consideration in the first frame, but also the possibility that (x_{01},y_{01}) actually has no such match. In the stereopsis case, there are many everyday stimulus configurations in which certain features are visible to

only one eye; in motion, one object may pass behind another and hence disappear. We therefore require, for each spatial frequency and orientation channel,

$$\mathscr{P}\{(x_{01},y_{01})\Leftrightarrow\emptyset\,\big|\,K_1(x_{01},y_{01},x_{02},y_{02}),v_1^{\text{even}}(x_{01},y_{01}),\ v_1^{\text{odd}}(x_{01},y_{01})\}+\sum_{(x_{02},y_{02})}\mathscr{P}\{(x_{01},y_{01})\Leftrightarrow(x_{02},y_{02})\ \big|K_1(x_{01},y_{01},x_{02},y_{02}),.\ v_1^{\text{even}}(x_{01},y_{01}),v_1^{\text{odd}}(x_{01},y_{01})\}=1\ , \tag{19}$$

where the notation "$(x_{01},y_{01})\Leftrightarrow\emptyset$" means "the point (x_{01},y_{01}) has no matching point in the other frame". Using this constraint in (9) and rearranging, we obtain

$$\begin{aligned}\mathscr{P}\{K_1(x_{01},y_{01},x_{02},y_{02})\,\big|\,v_1^{\text{even}}(x_{01},y_{01}),v_1^{\text{odd}}(x_{01},y_{01})\}&=\mathscr{P}\{K_1(x_{01},y_{01},x_{02},y_{02})\,\big|\,(x_{01},y_{01})\\&\Leftrightarrow\emptyset,v_1^{\text{even}}(x_{01},y_{01}),v_1^{\text{odd}}(x_{01},y_{01})\}\mathscr{P}\{(x_{01},y_{01})\Leftrightarrow\emptyset\}\\&+\sum_{(x_{02},y_{02})}\mathscr{P}\{K_1(x_{01},y_{01},x_{02},y_{02})\,\big|\,(x_{01},y_{01})\\&\Leftrightarrow(x_{02},y_{02}),v_1^{\text{even}}(x_{01},y_{01}),v_1^{\text{odd}}(x_{01},y_{01})\}\mathscr{P}\{(x_{01},y_{01})\\&\Leftrightarrow(x_{02},y_{02})\}\ .\end{aligned} \tag{20}$$

To calculate this, we would have to make further postulates. We need the brain's a priori assumptions about the firing rates of simple and complex cells (in order to assign probabilities to the possible values of K_1 when the complex cell is not tuned to the correct match), and about the probability of occlusion (that is, $\mathscr{P}\{(x_{01},y_{01})\Leftrightarrow\emptyset\}$). Rather than introducing these complications, I simply *assume* that the function $\mathscr{P}\{K_1(x_{01},y_{01},x_{02},y_{02})\,\big|\,v_1^{\text{even}}(x_{01},y_{01}),v_1^{\text{odd}}(x_{01},y_{01})\}$ is the same for each spatial frequency/orientation channel. It then becomes an overall constant, whose value is irrelevant. If this assumption is wrong, it will have the effect of differentially weighting the various spatial frequency/orientation channels, giving anomalously high weight to those channels which should have a low value of $\mathscr{P}\{K_1(x_{01},y_{01},x_{02},y_{02})\,\big|\,v_1^{\text{even}}(x_{01},y_{01}),\ v_1^{\text{odd}}(x_{01},y_{01})\}$. However, the fact that the complex-cell firing rates have been normalised – so that, for noise-free complex cells tuned to the stimulus disparity, $K_1=1$ irrespective of the channel orientation or frequency – at least makes the assumption plausible.

Let us review what this rather involved calculation is telling us. We have started with a particular point in the first frame, (x_{01},y_{01}). We know the local convolution of the noisy, filtered retinal image in the vicinity of this point. We then look at all the potential matches in the second frame. Assuming that each match in turn is correct, we are able to predict the normalised output of the complex cell tuned to that match. We can thus assign a probability to the normalised output actually observed from that complex cell. It is of course equally possible to pursue this calculation from the other direction: that is, to start with a particular point in the second frame, (x_{02},y_{02}), and assign a probability to the output of each complex cell using K_2 instead of K_1. In the computer program, I took the average of both these probabilities. This was particularly motivated by the stereopsis model, where it seemed desirable to treat both retinae equally.

2.9 Global judgement of displacement direction

Putting together these expressions for each term in the right-hand side of (9), we arrive at the probability that the match under consideration is correct, given the firing rate of the complex cell and the monocular convolutions. So, finally, we arrive at the local single-channel match probability

$$\mathscr{P}_{\theta\lambda}\{(x_{01},y_{01})\Leftrightarrow(x_{02},y_{02})\}=\mathscr{P}\{\delta\}\big[f(K_1,v_1^{\text{even}},v_1^{\text{odd}})+f(K_2,v_2^{\text{even}},v_2^{\text{odd}})\big]\ , \tag{21}$$

with f given in (18). Recall that this is the probability reported by a complex cell tuned to a single spatial period λ and orientation θ. The model brain then averages the different local single-channel probabilities so as to arrive at a global judgement of the overall direction of displacement in each interval of a two-interval forced-choice experiment. It should be stressed that this averaging process is purely heuristic; it is not based on statistical theory. Thus, its justification depends on the extent to which it can reproduce experimental data. A full mathematical treatment, remaining within the Bayesian framework, would require us to consider the responses of all complex cells together. Rather than the PDF of the complex cell response, (18), describing the probability of obtaining a particular firing rate from one complex cell if the stimulus locally has the disparity to which the complex cell is tuned, we would require the *joint* PDF describing the probability of getting a particular *set* of firing rates from the entire population of complex cells, given that the stimulus has the global disparity δ_x. Since the responses of different complex cells are evidently not independent, this joint PDF could not be resolved into a product of single PDFs like (18). As well as being highly intractable, this analysis might also be less satisfactory as a model, since it would give individual channels the power of veto. Logically, if the value from any complex-cell is inconsistent with a global stimulus disparity (in the sense that the probability of obtaining the observed complex-cell firing rate, given the outputs of monocular simple cells in one eye, is zero), then that global stimulus disparity cannot be correct and must be assigned probability zero. Thus, a model of this form would not be misled by false matches unless they were considered plausible by *all* channels. Such a model might well be an ideal disparity detector. However, it is hard to see how such a model could yield systematically wrong answers for anti-correlated kinematograms, as observed from human subjects. The present model was designed to be misled by false matches in the same way as humans apparently are.

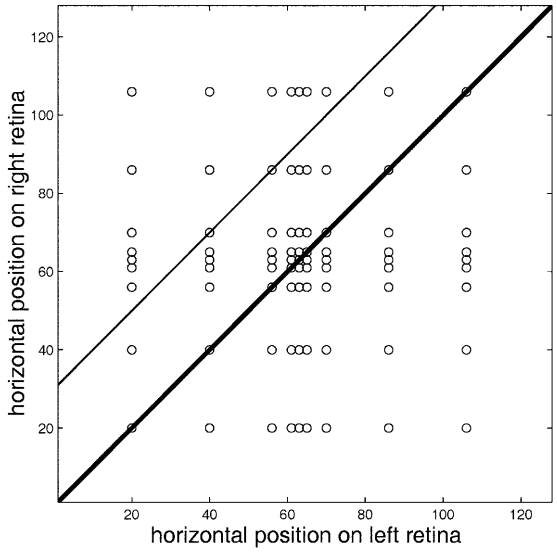

Fig. 7. The square model retina results in more complex cells tuned to zero disparity. Each *dot* represents a different complex cell; its position on the grid shows the horizontal position on left and right retinae of the RF centres of the component monocular simple cells (in pixels). The *thick line* marks complex cells tuned to zero disparity, where the simple cell RFs are at corresponding positions in left and right retinae. The *thin line* marks complex cells with a disparity of 30 pixels. Evidently there are more of the former than the latter. Specifically, $M(\delta_x) = 9$ for $\delta_x = 0$; $M(\delta_x) = 2$ for $\delta_x = 30$ pixels (22)

Thus, the probability of a particular horizontal disparity δ_x is estimated by averaging the single-channel match probability $\mathscr{P}_{\lambda\theta}\{(x_{01}, y_{01}) \Leftrightarrow (x_{02}, y_{02})\}$ over all spatial frequency and orientation channels, and over all vertical positions in the receptive field. Finally, we average over all horizontal positions x_{01}, x_{02} with the specified disparity: $x_{01} - x_{02} = \delta_x$. Here, we have to account for the fact that there are more complex cells tuned to zero disparity than to larger disparities, as illustrated in Fig. 7. We thus incorporate a final normalisation factor $M(\delta_x)$, describing how many complex cells are tuned to the horizontal disparity δ_x under consideration. That is,

$$\mathscr{P}\{\delta_x\} = \frac{1}{M(\delta_x)} \sum_{y_{01}} \sum_{y_{02}} \sum_{\substack{x_{01}, x_{02}: \\ x_{02} - x_{01} = \delta_x}} \times \sum_{\theta} \sum_{\lambda} \mathscr{P}_{\theta\lambda}\{(x_{01}, y_{01}) \Leftrightarrow (x_{02}, y_{02})\} \quad . \tag{22}$$

For the stereo model, the averaging occurs analogously (with the labels 1, 2 replaced with L, R). The only difference is that the stereo model does not include complex cells tuned to non-horizontal disparities, so the double sum over y_{01}, y_{02} in (22) is replaced by a single sum over vertical position y_0.

Having obtained the global match probability $\mathscr{P}\{\delta_x\}$, the global horizontal displacement Δ_x is taken to be the value of δ_x for which $\mathscr{P}\{\delta_x\}$ is a maximum: $\Delta_x = \text{argmax}(\mathscr{P}\{\delta_x\})$. If this maximum is not unique because several values of δ_x have equally large estimated probabilities $\mathscr{P}\{\delta_x\}$, the smallest disparity is chosen. This approach is analogous to that adopted by Fleet et al. (1996), who also took the average across all spatial scales and orientations, although they were summing neural firing rates rather than probability.

2.10 Implementation

The model is computationally very intense. The results shown in this paper use an 9×9 array of simple cell RFs on a retina of 128×128 pixels, with six orientation and five spatial frequency channels. These RFs occupy around 500 Mb of computer memory. Since there are both odd and even simple cells, whose response must be evaluated four times for each trial (two intervals each containing two images), at each trial $9^2 \times 6 \times 5 \times 2 \times 4 = 19\,440$ convolutions must be evaluated. The stereo model includes a total of $9^3 \times 6 \times 5 = 21\,870$ complex cells, and the motion model $9^4 \times 6 \times 5 = 196\,830$. The Bayesian likelihood of each complex cell response must be calculated, which means evaluating an integral (18) 196 830 times for a single trial with the motion model. We ran 80 trials at each of seven displacements for six different stimulus sets, in both correlated and anti-correlated conditions. Initial versions of the model, in only one dimension, were developed on a Macintosh, using MATLAB (Mathworks, Natick, Mass.) to allow easy visualisation of model behaviour. After the basic Bayesian approach had been formulated, the model was rewritten in C for running on Oxford University's supercomputer OSCAR. It was generalised to two dimensions, allowing the incorporation of different orientation channels, and parallelised, with each processor handling a different set of stimuli. The model was then run on six processors in parallel, all sharing a model with the same choice of retinal noise and Bayesian prior, and each running a different simulated experiment. Under these circumstances, the stereopsis model took around 3 days to run a complete set of simulations (Figs. 8, 9), and the motion model track over a week.

For each model (stereopsis/motion), the two free parameters – retinal noise ζ and bias towards small disparities D – were adjusted so as to achieve a good fit to the experimental data. Because the models took so long to run, an automated fitting procedure was impractical. Instead, trial and error was used to find a reasonable fit, using only a small number of trials under each condition. Then, for the best parameters found, the model was run again with the full 80 trials at each disparity to obtain the results shown in Figs. 8 and 9. With a more thorough exploration of the parameter space, it is possible that slightly better fits could have been achieved.

3 Results

Figures 8 and 9 compare the results obtained from the best models with those obtained from a human subject.

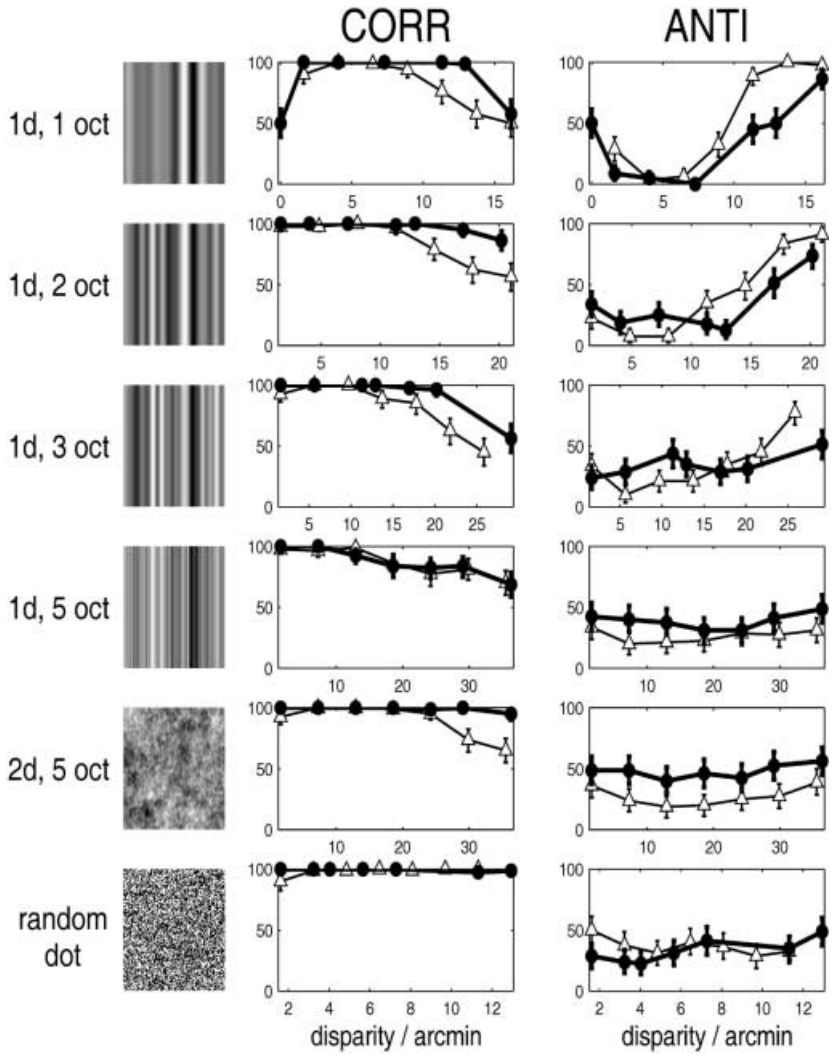

Fig. 8. Performance of the stereo model compared with a human observer (R.A.E.). Psychophysical functions are presented for stereograms with six different spectral profiles, as illustrated on the *left*. In the *left-hand set of results*, the stimuli were correlated; in the *right*, anti-correlated. *Solid lines* (*circles*) show results obtained with the stereo model for six different stimuli (illustrated on the left). The *thin lines* (*triangles*) show the results for observer R.A.E. from Read and Eagle (2000). For both model and human, data points show the percentage p of correct responses, out of 80 trials at each disparity. The error bars show the 95% confidence intervals, assuming a binomial distribution with 80 trials and success probability p. The model parameters are as described in Sect. 2, with the spread of the prior function $D = 2.4$ arcmin (3 pixels). The contrast of the retinal noise is 1% of that of the 1 octave stimulus

The stereo data were best fitted with a very narrow prior (the spread D being just 3 pixels, or 2.4 arcmin) and very low noise levels: the contrast of the simulated retinal noise was only 1% of that of the lowest-contrast image (the 1-D image with a bandwidth of 1 octave). For the motion model, a good match to experimental data required both much higher noise levels (20% of the lowest-contrast image) and a wider prior ($D = 7$ pixels, or 5.6 arcmin). The figures show the psychophysical functions (percentage of correct responses) as a function of disparity for the model (thick lines; circles) and one of our human observers (R.A.E., Read and Eagle 2000: thin lines; triangles). Note that the models could not be tested at precisely the same disparities as the humans. Because of computational limitations, the models sample the retinal image rather sparsely, and are thus sensitive to a limited number of disparities (see Sect. 2). For example, the model presented here contains no complex cells tuned to a disparity of 12 pixels, and is thus likely to perform poorly when tested at this disparity, even though it can "perceive" 9 and 14 pixels perfectly well. In order to make a fair assessment of the model, it was tested only at disparities to which it was sensitive. Note that this restriction is imposed by computational limitations, and is not a fundamental feature of the model. Ideally, the model would incorporate a very large number of simple cells, positioned at random across the retina, preferably in accordance with an experimentally determined distribution.

3.1 Stereopsis model

For the stereopsis model, the best results were obtained with noise equal to 1% of the contrast of the lowest-power image (1-D, 1 octave; see Sect. 2), with a bias towards disparities lower than $D = 2.4$ arcmin (13). Figure 8 shows the results obtained with these parameters. The model reproduces the key features of the experimental data reasonably well. For correlated stimuli, the model performs close to 100%, as required, falling to chance at large disparities, where the prior bias towards small disparities effectively blinds it to the correct match. The major discrepancy is that the fall towards chance occurs at larger disparities ("$D_{\max}$") for the model than for the human observer. This is especially

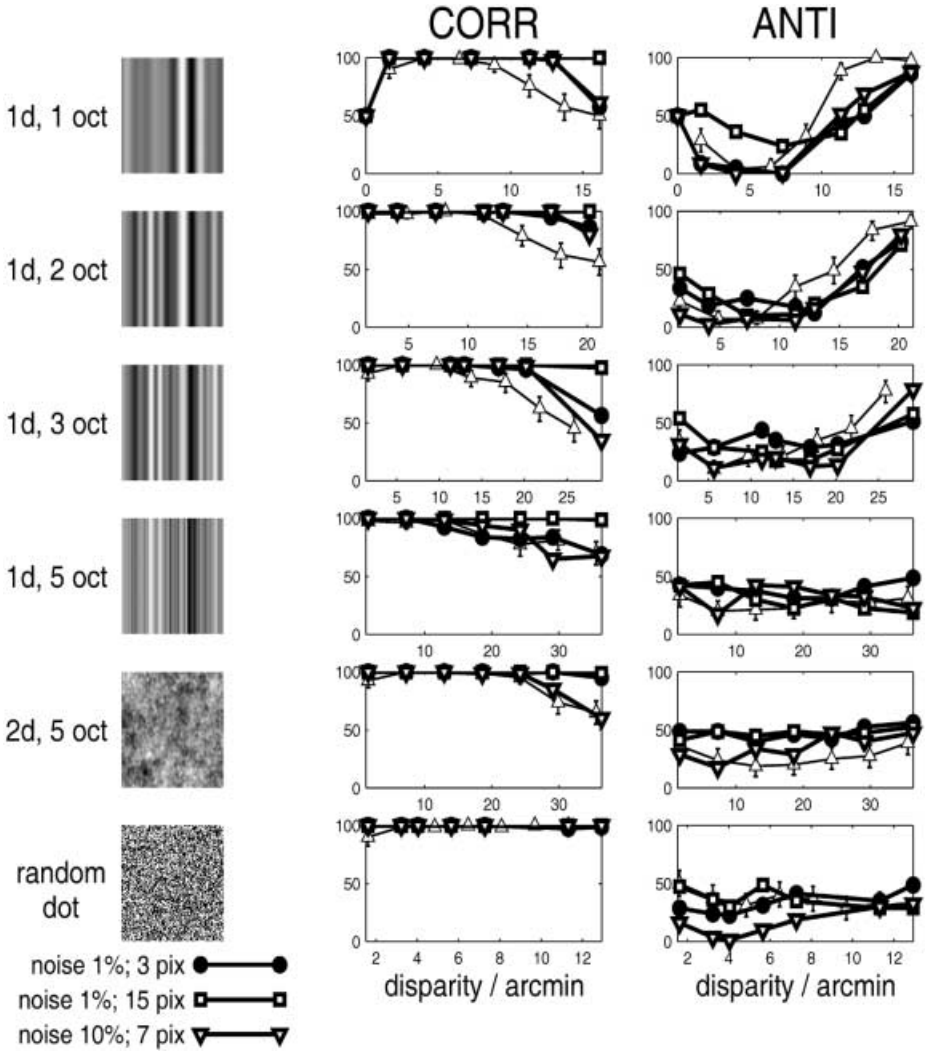

Fig. 9. Performance of the stereo model with three different sets of parameters. As in Fig. 8, the *thin lines* (*triangles*) show the results for observer R.A.E. and the *solid lines* show results obtained with the model. The *filled circles* show model results with $D = 3$ pixels, noise = 1% of the contrast of the 1-octave stimulus, as in Fig. 8. *Unfilled squares* show the results with a wider prior, $D = 15$ pixels, while *unfilled downward-pointing triangles* show the results with higher noise (noise = 10% and $D = 7$ pixels)

marked in the response to 2-D, 5-octave stereograms, where the model is still performing close to 100% at large disparities where the human performance is well down towards 50%. Human performance with 1-D band-pass stereograms tends to scale with the longest spatial period present in the image (Chang and Julesz 1983; de Bruyn and Orban 1989; Cleary 1990; Cleary and Braddick 1990; Read and Eagle 2000). The model is constructed to have a similar tendency, because for displacements much larger than the longest spatial period, there will be alternative, false matches at smaller absolute disparities, which are considered a priori more probable.

2-D (isotropic) stereograms actually contain power at arbitrarily long spatial periods, contributed by oblique orientations. We then expect performance to scale with the longest spatial period to which the detector is sensitive. In the present model, this was 1 cycle per degree. The 5-octave 1-D stereograms also contain power at 1 cycle per degree, so the longest available period was in fact the same for the 1-D and 2-D 5-octave stereograms. However, the 2-D 5-octave stereograms contain the same power in each 30° slice as the 1-D 5-octave stereograms do in total. The model's better performance for 2-D stereograms probably reflects this increased power. Non-vertically oriented model complex cells are more strongly stimulated by the 2-D images, thus reducing the effective level of noise and improving accuracy. A similar effect occurs in our human observers (Read and Eagle 2000); their performance begins to decline from 100% at smaller disparities for 1-D 5-octave stimuli than for 2-D 5-octave stimuli. However, this effect is much less marked in humans than in the model, which may indicate that the model is giving undue weight to the longer spatial periods. A plausible reason for this is that the model retina is covered more densely (in units of spatial period λ) with RFs tuned to large λ (Sect. 2; Fig. 5).

For anti-correlated stimuli, the model again describes human responses reasonably well. For small displacements of narrow-band stimuli, both model and human systematically report the wrong answer, resulting in scores close to 0%. For larger displacements, performance moves towards 100%. As the bandwidth is increased, the reversed depth becomes less pronounced, with the minimum score closer to 50%. The model performs completely at chance level for both the 2-D 5-octave and the random-dot stereograms (except for a very weak reversed depth effect with random-dot patterns at the smallest disparities). This is very much in line

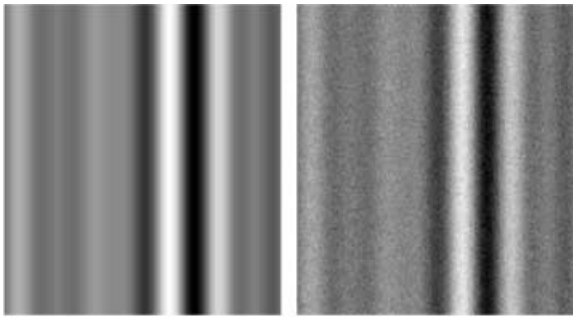

Fig. 10. The effect of retinal noise. The figure shows an example 1-D band-pass noise stimulus, bandwidth 1 octave. The image on the *right* has been degraded with "retinal" noise at the amplitude used in obtaining the results in Fig. 9. The model retinal noise is Gaussian, with contrast 20% of this 1-octave stimulus. Since the experimental stimuli have equal power in equal octaves, this 1-octave stimulus is the most affected by a given level of retinal noise. The same level of retinal noise represents only 1.5% ($20/\sqrt{5}/6$) of the contrast of the 2-D 5-octave stimulus

with human data (Cogan et al. 1993; Cumming et al. 1997; Read and Eagle 2000); the observer R.A.E. is unusual amongst our subjects in showing even weak reversed depth for the 2-D 5-octave stereogram.

The effect of the two parameters is explored in Fig. 10. Here, the unfilled squares show results obtained with a wider prior, and the downward-pointing triangles show results with a higher noise level. First consider the response to correlated stimuli. With a low noise level (1%), a very tight prior (3 pixels) is necessary in order to reproduce the observed fall-off in performance as disparity increases. With a wider prior (15 pixels, squares in Fig. 10), the model continues to perform at 100% out to large disparities beyond the maximum disparity reliably perceived by human observers. With more noise, a looser prior still produces a good match to human observers (downward triangles in Fig. 10). These results are readily understandable. Where there is very little noise, the Bayesian likelihood is sharply peaked at the correct match. Thus, to prevent the model from selecting the correct match even at large disparities, a tight prior must be applied, so that correct large-disparity matches are punished severely enough so as not to be selected despite their high likelihood. If the noise level is higher, the Bayesian likelihood is much more tolerant to false matches. The correct match is assigned a lower probability, since it is considered just one of several possibilities. Thus, less "punishment" from the prior is necessary to deflect the model from the veridical match.

The same effect of increased noise is visible in the results for anti-correlated stimuli. Again, increasing the noise level makes the model more tolerant to the false matches presented by anti-correlated stimuli. If the noise level is low, the model can tell that these matches cannot actually be correct; no match is consistently accorded high probability, and the model performs close to chance level. If the noise level is high, false matches in the wrong direction are routinely accepted, leading to performance close to 0%.

3.2 Motion model

The structure of the motion model is slightly different from that of the stereopsis model, in that it was constructed to be sensitive to displacements in all directions, not just horizontal. This was motivated by the desire for biological plausibility, since there is no reason why the brain's motion system should emphasise horizontal displacements. However, this structural difference has little effect on the results presented here, since the models were asked to report only the horizontal direction of motion of horizontally moving stimuli. Thus, for a particular choice of noise and prior, the stereopsis and motion models give similar results. The results of Sect. 3.1 (Fig. 10) therefore already indicate that a higher retinal noise level will be needed to reproduce the reversed motion reported by human observers. Figure 9 shows results from the motion model with parameters which were found to give reasonable results: a retinal noise level equal to 20% of the contrast of the lowest-contrast image, and a relatively wide prior ($D = 5.6$ arcmin). Once again, the model agrees reasonably well with experimental data. In particular, it reproduces the results found with anti-correlated stimuli: for 1-D, broadband anti-correlated stimuli, performance is close to chance, whereas widening the orientation bandwidth to 2-D elicits a "reverse phi" phenomenon. The major discrepancy is the larger $D_{\max}$ displayed by the model, especially for the narrow-band, 1-D stimuli.

4 Discussion

Our previous psychophysical work (Read and Eagle 2000) revealed an intriguing difference between the responses to stereograms and to kinematograms. For both stereopsis and motion, small displacements of anti-correlated 1-D stimuli with narrow spatial frequency bandwidth produce reversed perceptions. For 2-D anti-correlated stimuli, however, small displacements produce a strong perception of reversed motion and little or no perception of reversed depth.

In this paper, I have presented a model that can reproduce these results. It is based on the known physiology of primary visual cortex, albeit simplified. The image is initially processed within channels tuned to a particular spatial frequency and orientation. Each channel represents a population of complex cells in cortical area V1. Similar approaches have been used before (Sanger 1988; Qian 1994; Fleet et al. 1996; Zhu and Qian 1996; Qian and Zhu 1997; Prince and Eagle 2000b) in discussing the correspondence problem. Most previous workers have deduced the stimulus disparity from the peak firing rate of a population of complex cells. While there is an extensive literature on decoding channel-coded systems, this paper appears to be the first to combine a Bayesian approach with an analysis into different spatial frequency/orientation channels.

4.1 Advantages of a Bayesian approach

The Bayesian approach has several advantages. First, it provides a natural way to build in the constraints which are necessary to arrive at a solution to the ill-posed correspondence problem. Here, I have used the prior to encode a preference for small disparities. The calculation of the Bayesian likelihood in terms of simple and complex cells, rather than directly from the retinal images, also naturally leads the model to include a smoothness constraint, over a scale appropriate for each channel (Sect. 2.5). The other key advantage of a Bayesian approach is that, by converting firing rates from different channels into the common language of probability, it simplifies the combination of information from different channels. If firing rates are compared directly, we need to postulate some scaling rule to convert between channels – for example, by requiring that simple cells in different channels respond with the same firing rate to their optimal sine-wave grating. This is avoided in the present approach by using a form of contrast normalisation in the Bayesian analysis, in which the firing rate of each complex cell is divided by the firing rate of some of the same simple cells that feed into it. Perhaps paradoxically, this contrast normalisation actually has the sensible effect of encouraging the model to pay greater attention to regions of the image where the contrast is high, rather than to dim grey regions which match equally well at any disparity. The value of the normalised response is that, in the absence of noise, it would be unity for any complex cell tuned to the correct disparity, irrespective of the cell's spatial frequency or orientation tuning, and irrespective of the contrast and spectral content of the stimulus. However, random retinal noise causes small departures from unity. These fluctuations are smaller for larger values of the stimulus contrast within the cell's receptive field. Thus, the probability of getting unity, assuming that the cell actually is tuned to the correct match, increases with stimulus contrast. This means that – other things being equal – greater significance is attached to a report of a correct match coming from a high-contrast region of the stimulus, since this is less likely to be a spurious match occurring at random in the retinal noise. Divisive contrast normalisation has been employed to good effect in several models (Heeger 1993; Thomas and Olzak 1997; Tolhurst and Heeger 1997), although there the divisor represents the pooled activity of a large number of cells, from different channels.

4.2 Effect of noise

The Bayesian analysis means that the effect of noise can be slightly counter-intuitive. It is natural to envisage noise as always limiting performance. For instance, in Prince and Eagle (2000b), increasing noise always pushes the model's performance towards chance level. Within the present Bayesian framework, noise has a more complicated effect. The brain's estimate of the noise levels on its data affects its calculation of the probability that a particular solution of the correspondence problem is correct. When implementing the present model, I assumed that this assessment was accurate, reflecting a lifetime of visual experience. That is, the noise used in the model's calculation of the match probability is equal to the noise actually affecting each simple cell (ξ^{even} and ξ^{odd} in Eq. 5). One consequence of this is that changing noise levels can have rather an unpredictable effect on performance. For instance, increasing retinal noise does not necessarily degrade the performance of the model. This is because the model is programmed to seek the correct match on the assumption that, apart from noise, the left retinal image is exactly a horizontally shifted version of the right, whereas in our experiments, to avoid providing monocular cues, the images were displaced within a fixed window on the screen. This means that even those complex cells tuned to the notionally correct disparity do not see a perfect match. If retinal noise is very low, the model will expect very precise matching, and so may reject the notionally "correct" match. Increasing retinal noise makes the model more tolerant of inexact matches, and so may actually improve performance by enabling the model to accept the best match on offer. This effect could be avoided by making the model allow for higher noise levels than are actually present. This might be valuable in real life if it enabled a more robust response to interocular discrepancies caused, for instance, by occlusion. This variant of the model is not explored here.

The decision to add noise at the retinal level, rather than centrally, means that the amplitude of the noise added to simple cells is different for each spatial frequency channel, and for each RF type (odd/even) within this channel, reflecting differences in the integral-squared of the receptive field. If the noise arose centrally, it would be natural to add noise of the same amplitude to every simple cell. With retinal noise, it turns out (5) that the high-frequency channels experience more noise, because they have smaller receptive fields and so average over a smaller number of photoreceptors. Note that because account is taken of the higher noise in the Bayesian probability calculation, this does *not* mean that the higher frequency channels contribute less to detecting disparity. On the contrary, they may contribute more because they are more tolerant of poor matches.

4.3 Differences between the stereopsis and motion models

I have developed two versions of the model, one adapted for the stereo problem, the other for motion. For stereopsis, I assume that horizontal disparities are both generally larger than vertical disparities, and more significant in judging the sign of depth. Thus, the stereopsis model uses complex cells tuned to horizontal disparities only. In contrast, for the motion model I assume that stimuli are equally likely to be moving in any direction. Thus, the motion model includes complex cells tuned to displacements in all directions. In fact, as might be expected, these extra complex cells make little difference to the model's performance when tested with

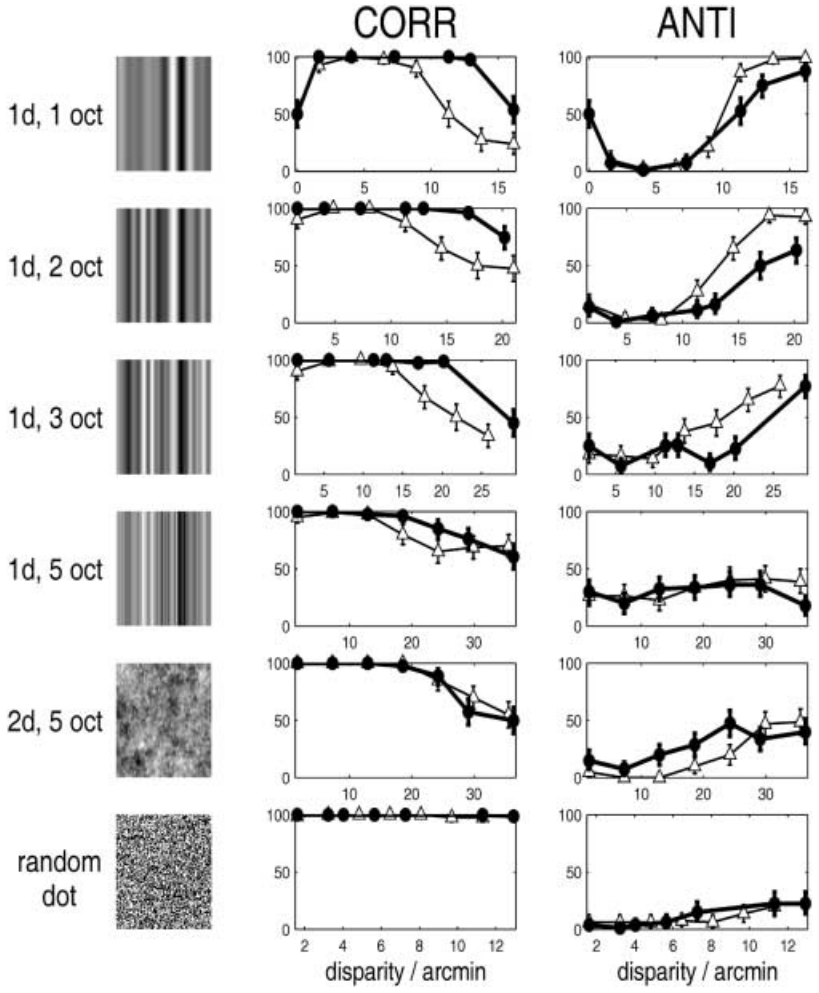

Fig. 11. Performance of the motion model compared with a human observer, (R.A.E.). All details are as in Fig. 8, except that the stimuli were presented to the human observer as two-frame kinematograms rather than stereograms, and the computer simulation used the motion model described in Sect. 2, with the spread of the prior function $D = 5.6$ arcmin (7 pixels) and the contrast of the retinal noise set to 20% of the 1-octave stimulus

horizontal displacements. Both models use position rather than phase disparity, and both use complex cells tuned to all orientations, not just vertical. It has been suggested (Anzai et al. 1999) that neurons can encode disparities only orthogonal to their RF orientation, but other workers (Ohzawa and Freeman 1986; Prince et al. 2001) find cells tuned to all orientations which are highly sensitive to horizontal disparity. Certainly, from a theoretical point of view, even cells that are tuned to horizontal orientations can encode horizontal disparities which are larger than their receptive field size.

4.4 Model parameters

Several parameters were required to specify the models. Where possible, these were determined from experimental data (e.g. the channels' spatial frequencies and orientation bandwidths). Sometimes, they were set by computational limitations; for instance, simple cell RFs are placed on a 9×9 grid (Fig. 5), whereas ideally the model would include RFs at many more different positions on the retina, scattered at random according to a distribution observed experimentally. Only two parameters for each model were "free" in the sense of being systematically altered so as to obtain the best fit (although computational limitations precluded systematic optimisation). These free parameters were the retinal noise ζ and the extent of the bias towards small disparities D for each model system. The different values adopted for these parameters are the only significant difference between the stereopsis and motion models. With suitable values of these two parameters, each model produces a reasonable match to twelve data sets each containing seven data points. This success may indicate that the models capture some of the essential properties of the human visual system.

Although both of the free parameters have a simple interpretation, the model cannot be taken as predicting real biological values. For instance, noise was modelled as occurring purely in the retina. The main motivation for this was parsimony: adding noise separately to the retina, simple cells and complex cells would have resulted in more free parameters. Thus, the noise level incorporated in the model, although nominally retinal, has to represent all possible biological noise sources. Furthermore, the appropriate level of model noise may depend on other aspects of the model, for instance, how closely retinal fields were spaced across the retina (Fig. 5).

In addition, different levels of retinal noise were used in the two forms of the model. The retinal noise applied to the motion model was an order of magnitude larger than

that used in the stereopsis model. In the motion model, the high level of noise was important not because it degraded the image (as Fig. 11 shows, it had relatively little effect on even the lowest-contrast stimuli), but because it made the motion model more tolerant to poor matches. Thus, for anti-correlated stimuli, it accepted the relatively poor reversed-motion matches, whereas the stereopsis system could not find any acceptable match and so performed at chance level. One interpretation of these results is that the retinal noise level is really the same in both stereopsis and motion, but that the motion system is more tolerant to poor matches. This could be expressed in the motion model by allowing for a larger noise level in calculating the match probability (18) than is actually present in the retina. Alternatively, the effective retinal noise may actually be greater for the motion system. One factor which has been neglected in this study, but which might contribute to higher effective noise levels for motion, is the temporal properties of the input neurons [18]. The simplistic modelling of the motion-sensitive complex cell as responding only to two frames does not address this complexity.

Finally, it is striking just how low the noise is. Figure 11 shows the effect of the noise in the motion model on the lowest-contrast image. For the higher-contrast stimuli, or the lower noise levels incorporated in the stereopsis model, the degradation is even smaller. This agrees with the view that the brain adds remarkably little noise to its inputs (Rieke et al. 1997).

4.5 Future work

Because the model was developed to explain the results of a two-interval forced-choice psychophysical experiment, it was constructed to output a single estimate of global stimulus disparity, by pooling information from across the image. The model could instead be used to derive an estimate of the disparity at each point in the image, to see whether it could construct an accurate disparity map of the stimulus. This has already been carried out in a related (non-Bayesian) model (Qian 1994; Zhu and Qian 1996; Qian and Zhu 1997).

Secondly, it would be interesting to exploit the probabilistic nature of this model further. Because it seeks initially to assign probability to each potential match, rather than homing in on a single correct answer, it might be well suited to deal with problems such as ambiguity (Fig. 1), transparency or occlusion. For instance, Fig. 12 shows Panum's limiting case, where one object in the right retina matches two in the left retina. Such a situation would cause problems for an algorithm which seeks a unique partner for each retinal position. However, it might be handled well by a probabilistic model, which could assign equal probability to the case that position A matches position C and position A matches position B.

4.6 Summary

I present a physiologically realistic pair of models which together can reproduce the response of human observers to correlated and anti-correlated stereo- and kinematograms with a range of spatial frequency and orientation bandwidths. The model predicts that the motion system experiences higher effective noise levels than the stereopsis system, making the motion system more tolerant of poor-quality correspondences. The model is constructed in Bayesian probabilistic terms. This provides a natural way to incorporate the visual system's prior knowledge and assumptions, and has the potential to be exploited further to address complex problems such as ambiguity.

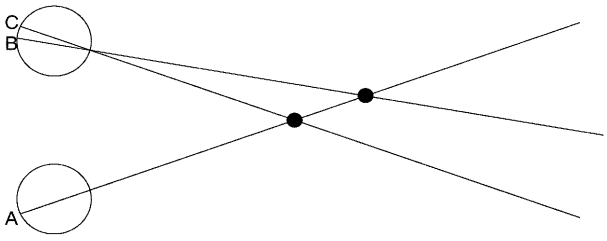

Fig. 12. As in Fig. 1, the *large open circles* represent the eyeballs, seen from above. The *filled circles* represent spheres in front of the viewer. The further sphere is hidden from the right eye by the nearer sphere. However, the left eye can see both spheres. To deduce the distance of both spheres, the brain would have to match position *A* in the right retina with position *B and* position *C* in the left retina

Acknowledgements. This work was begun in collaboration with the late Richard Eagle, and greatly benefited from his insight. I thank Bruce Cumming, Simon Prince and Andrew Glennerster for helpful comments on an earlier version of the paper. The simulations were run with support from the Oxford Supercomputing Centre. I am funded by a Training Fellowship in Mathematical Biology from the Wellcome Trust.

Appendix A: Properties of the Gabor receptive field

The model simple cell RF has the general form (Fig. 3)

$$\rho(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{[x\cos\theta + y\sin\theta]^2}{2\sigma_x^2}\right) \times \exp\left(-\frac{[y\cos\theta - x\sin\theta]^2}{2\sigma_y^2}\right) \times \cos\left(\frac{2\pi}{\lambda}[x\cos\theta + y\sin\theta] - \phi\right) , \quad \text{(A1)}$$

where λ and θ are the spatial period and orientation on the retina ($\theta = 0$ is vertical). The exponential terms describe a 2-D Gaussian envelope, limiting the cell's receptive field to a finite region of the retina. For the RF in (A1), this region is centred on the origin; to obtain an RF centred on (x_0, y_0), we write $\rho(x - x_0, y - y_0)$. σ_x controls the spatial extent of the RF perpendicular to its preferred orientation, and σ_y that parallel to it. Together with the spatial period, these set the spatial frequency and orientation bandwidth of the RF, as explained in the Section A.1. ϕ describes the phase of the carrier cosine relative to the Gaussian envelope: $\phi = 0$ for even RFs and $\pi/2$ for odd.

A.1 Spatial frequency and orientation bandwidth of a Gabor filter

The Fourier transform of the Gabor receptive field in (A1), $\tilde{\rho}(\tilde{\lambda},\tilde{\theta})$, represents the response of the simple cell to a sine-wave grating with spatial period $\tilde{\lambda}$ oriented at angle $\tilde{\theta}$ to the vertical. Mathematically, it is given by

$$\tilde{\rho}(\tilde{\lambda},\tilde{\theta}) = \frac{1}{2}\exp\left\{-2\pi^2\sigma_y^2\frac{\sin^2(\tilde{\theta}-\theta)}{\tilde{\lambda}^2}\right\} \times \left[e^{i\phi}\exp\left\{-2\pi^2\sigma_x^2\left[\frac{\cos(\tilde{\theta}-\theta)}{\tilde{\lambda}}+\frac{1}{\lambda}\right]^2\right\} + e^{-i\phi}\exp\left\{-2\pi^2\sigma_x^2\left[\frac{\cos(\tilde{\theta}-\theta)}{\tilde{\lambda}}-\frac{1}{\lambda}\right]^2\right\}\right] . \quad (A2)$$

To obtain the spatial frequency bandwidth, consider how the Fourier transform varies as a function of the stimulating spatial period $\tilde{\lambda}$, when the orientation of the stimulus is matched to the RF ($\tilde{\theta}=\theta$). The octave spatial bandwidth β is defined in terms of the lowest and highest spatial periods passed by the filter:

$$\beta = \frac{\ln(\lambda_{\rm lo}/\lambda_{\rm hi})}{\ln 2} . \quad (A3)$$

Of course, a Gabor filter has no sharp cut-off. I define the limits of the filter, $\lambda_{\rm lo}$ and $\lambda_{\rm hi}$, to be where the power of the Fourier spectrum drops to half its maximum value. To an excellent approximation, these are given by the solutions of

$$\exp\left\{-2\pi^2\sigma_x^2\left[\frac{1}{\tilde{\lambda}}-\frac{1}{\lambda}\right]^2\right\} = \frac{1}{\sqrt{2}} , \quad (A4)$$

so that

$$\frac{1}{\lambda_{\rm hi/lo}} = \frac{1}{\lambda} \pm \frac{1}{2\pi\sigma_x}\sqrt{\ln 2} . \quad (A5)$$

This yields the following relationship between bandwidth as defined in (A3) and the spatial extent of the RF:

$$2^{\beta} = \frac{2\pi\sigma_x + \lambda\sqrt{\ln 2}}{2\pi\sigma_x - \lambda\sqrt{\ln 2}}, \quad \sigma_x = \sqrt{\ln 2}\,\frac{\lambda}{2\pi}\frac{2^{\beta}+1}{2^{\beta}-1} . \quad (A6)$$

Due to the asymmetry of the Gabor filter in octave space, the preferred spatial period λ is a smaller number of octaves from $\lambda_{\rm lo}$ than $\lambda_{\rm hi}$.

In the model, I assume that all channels have the same octave bandwidth, so that the spatial extent of the receptive fields scales with the preferred spatial period. Together with the overall division by σ_x and σ_y in (A1), this also means that all simple cells respond with the same firing rate to a sine-wave grating at their preferred spatial frequency, orientation and phase (as we see by using Eq. 24 to calculate the absolute value of the Fourier transform with $\tilde{\lambda}=\lambda$ and $\tilde{\theta}=\theta$).

Similarly, I define the orientation bandwidth α as the range of angles over which the filter passes components at greater than half-maximal power. The limiting angles are given by the solutions of

$$\exp\left\{-\frac{2\pi^2\sigma_y^2}{\lambda^2}\sin^2(\tilde{\theta}-\theta)\right\} \times \exp\left\{-\frac{2\pi^2\sigma_x^2}{\lambda^2}\left[\cos(\tilde{\theta}-\theta)-1\right]^2\right\} = 2^{-1/2} . \quad (A7)$$

This comes from (A2), for a component at the preferred spatial period. $\tilde{\theta}$ here represents the angle at which the modulus-squared of the Fourier transform falls to half of the maximum attained for $\tilde{\theta}=\theta$. Thus $\alpha = 2|\tilde{\theta}-\theta|$. If we assume that the orientation bandwidth is small, we can use the small-angle approximations for sine and cosine. Retaining powers of α up to second order, we obtain the following relationship between the orientation bandwidth in radians α and the spatial extent of the receptive field along its preferred orientation, σ_y:

$$\alpha = \frac{\lambda}{\pi\sigma_y}\sqrt{\ln 2}, \quad \sigma_y = \frac{\lambda}{\pi\alpha}\sqrt{\ln 2} . \quad (A8)$$

References

Adelson EH, Bergen JR (1985) Spatiotemporal energy models for the perception of motion. J Opt Soc Am A 2: 284–299

Anstis SM (1970) Phi movement as a subtraction process. Vision Res 10: 1411–1430

Anzai A, Ohzawa I, Freeman RD (1999a) Neural mechanisms for procesing binocular information. I. Simple cells. J Neurophys 82: 891–908

Anzai A, Ohzawa I, Freeman RD (1999b) Neural mechanisms for procesing binocular information. II. Complex cells. J Neurophys 82: 909–924

Blakemore C, Campbell FW (1969) On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. J Physiol 203: 237–260

Bruyn B de, Orban GA (1989) Discrimination of opposite directions measured with stroboscopically illuminated random-dot patterns. J Opt Soc Am A 6: 323–328

Campbell FW, Kulikowski JJ (1966) Orientational selectivity of the human visual system. J Physiol 187: 436–445

Campbell FW, Robson JG (1968) Application of Fourier analysis to the visibility of gratings. J Physiol 197: 551–566

Chang JJ, Julesz B (1983) Displacement limits for spatial frequency filtered random-dot cinematograms in apparent motion. Vision Res 23: 1379–1385

Cleary R (1990) Contrast dependence of short-range apparent motion. Vision Res 30: 463–478

Cleary R, Braddick OJ (1990) Direction discrimination for band-pass filtered random dot kinematograms. Vision Res 30: 303–316

Cogan AI, Lomakin AJ, Rossi AF (1993) Depth in anti-correlated stereograms: effects of spatial density and interocular delay. Vision Res 33: 1959–1975

Cumming BG, Parker AJ (1997) Responses of primary visual cortical neurons to binocular disparity without depth perception. Nature 389: 280–283

Cumming BG, Shapiro SE, Parker AJ (1998) Disparity detection in anti-correlated stereograms. Perception 27: 1367–1377

DeAngelis GC, Ohzawa I, Freeman R (1995) Receptive-field dynamics in the central visual pathways. Trends Neurosci 18: 451–458

Eagle R (1997) Independent processing across spatial frequency in moving broadband patterns. Perception 26: 961–976
Fleet DH, Wagner H, Heeger DJ (1996) Neural encoding of binocular disparity: energy models, position shifts and phase shifts. Vision Res 36: 1839–1857
Heeger DJ (1993) Modeling simple-cell direction selectivity with normalized, half-squared, linear operators. J Neurophys 70: 1885–1898
Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160: 106
Knill D, Richards W (1996) Perception as Bayesian inference. Cambridge University Press, Cambridge
Mansfield JS, Parker A (1993) An orientation-tuned component in the contrast masking of stereopsis. Vision Res 33: 1535–1544
Marr D (1982) Vision: a computational investigation into the human representation processing of visual information. Freeman, San Francisco
Mayhew JE, Longuet-Higgins HC (1982) A computational model of binocular depth perception. Nature 297: 376–378
McKee SP, Mitchison GJ (1988) The role of retinal correspondence in stereoscopic matching. Vision Res 28: 1001–1012
Movshon J, Thompson I, Tolhurst DJ (1978) Spatial summation in the receptive fields of simple cells in the cat's striate cortex. J Physiol 283: 53–77
Ohzawa I, Freeman RD (1986) The binocular organization of simple cells in the cat's visual cortex. J Neurophysiol 56: 221–242
Ohzawa I, DeAngelis G, Freeman R (1990) Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. Science 249: 1037–1041
Ohzawa I, DeAngelis G, Freeman R (1997) Encoding of binocular disparity by complex cells in the cat's visual cortex. J Neurophysiol 77: 2879–2909
Poggio GF, Fischer B (1977) Binocular interaction and depth sensitivity of striate and prestriate cortex of behaving rhesus monkey. J Neurophysiol 40: 1392–1405
Prince SJD, Eagle RA, Rogers BJ (1998) Contrast masking reveals spatial-frequency channels in stereopsis. Perception 27: 1345–1355
Prince SJD, Eagle RA (2000a) Stereo correspondence in one-dimensional Gabor stimuli. Vision Res 40: 913–924
Prince SJD, Eagle RA (2000b) Weighted directional energy model of human correspondence. Vision Res 40: 1143–1155
Prince SJD, Pointon AD, Cumming BG, Parker AJ (2001) Quantitative analysis of the Responses of V1 neurons to horizontal disparity in dynamic random dot stereograms. (In press, J Neurophysiol)
Qian N (1994) Computing stereo disparity and motion with known binocular cell properties. Neural Comput 6: 390–404
Qian N, Zhu Y (1997) Physiological computation of binocular disparity. Vision Res 37: 1811–1827
Read JCA, Eagle RA (2000) Reversed stereo depth and motion direction with anti-correlated stimuli. Vision Res 40: 3345–3358
Read JCA, Cumming BG, Parker AJ (2000) Local models can account for the reduced response of disparity-tuned V1 neurons to anti-correlated images. Soc Neurosci Abstr Vol 26 p.1845
Rieke F, Warland D, de Rugter van Steveninck RR, Bialek W (1997) Spikes: exploring the neural code. MIT Press, Cambridge, Mass
Sanger TD (1988) Stereo disparity computation using Gabor filters. Biol Cybern 59: 405–418
Santen JP van, Sperling G (1984) Temporal covariance model of human motion perception. J Opt Soc Am A 1: 451–473
Sato T (1989) Reversed apparent motion with random dot patterns. Vision Res 29: 1749–1758
Sato T (1998) Dmax: relations to low- and high-level motion processes. In: Watanabe T (ed) High-level motion processing. MIT Press, Boston
Smallman HS, McLeod DI (1994) Size-disparity correlation in stereopsis at contrast threshold. J Opt Soc Am A 11: 2169–2183
Stevens C (1994) Cooperativity of unreliable neurons. Curr Biol 4: 268
Thomas JP, Olzak LA (1997) Contrast gain control and fine spatial discriminations. J Opt Soc Am A 14: 2392–2405
Tolhurst DJ, Heeger DJ (1997) Comparison of contrast-normalization and threshold models of the responses of simple cells in cat striate cortex. Vis Neurosci 14: 293–309
Valois RL de, Valois KK de (1988) Spatial vision. Oxford University Press, Oxford
Valois RL de, Albrecht DG, Thorell LG (1982a) Spatial frequency selectivity of cells in macaque visual cortex. Vision Res 22: 545–559
Valois RL de, Yund EW, Hepler N (1982b) The orientation and direction selectivity of cells in macaque visual cortex. Vision Res 22: 531–544
Zhu YD, Qian N (1996) Binocular receptive field models, disparity tuning and characteristic disparity. Neural Comput 8: 1611–1641