Vision Research 101 (2014) 11–31

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Neural population models for perception of motion in depth

Qiuyan Peng, Bertram E. Shi [*]

Department of Electronic and Computer Engineering and Division of Biomedical Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 21 August 2013
Received in revised form 28 April 2014
Available online 9 May 2014

Keywords:
Motion in depth
Changing disparity
Interocular velocity difference
Visual motion
Stereo disparity
Neural development

ABSTRACT

Changing disparity (CD) and interocular velocity difference (IOVD) are two possible mechanisms for stereomotion perception. We propose two neurally plausible models for the representation of motion-in-depth (MID) via the CD and IOVD mechanisms. These models create distributed representations of MID velocity as the responses from a population of neurons selective to different MID velocity. Estimates of perceived MID velocity can be computed from the population response. They can be applied directly to binocular image sequences commonly used to characterize MID perception in psychophysical experiments. Contrary to common assumptions, we find that the CD and IOVD mechanisms cannot be distinguished easily by random dot stereograms that disrupt correlations between the two eyes or through time. We also demonstrate that the assumed spatial connectivity between the units in these models can be learned through exposure to natural binocular stimuli. Our experiments with these developmental models of MID selectivity suggest that neurons selective to MID are more likely to develop via the CD mechanism than the IOVD mechanism.

1. Introduction

Motion in depth (MID) or stereomotion, refers to motion approaching or receding from the observer. Two mechanisms have been proposed to explain MID perception: changing disparity (CD) and interocular velocity difference (IOVD) (Rashbass & Westheimer, 1961a, 1961b). The CD mechanism is assumed to combine binocular information first to estimate static disparity and to estimate the change in disparity over time second. The IOVD mechanism is assumed to estimate monocular motion first and to estimate the difference between the motion in the left and right eyes second.

The extent to which these two mechanisms are employed by human observers is still an open question. Researchers typically attempt to address this question using modifications of the random dot stereogram (RDS). The standard RDS is created by generating a random dot pattern for the left eye, and presenting the same pattern to the right eye shifted by a given disparity. Over time, the random dots move coherently. This stimulus contains both CD and IOVD cues. To evaluate the potential contributions of the two mechanisms, researchers create stimuli that disrupt correlations between the two eyes or over time.

The dynamic RDS (DRDS) breaks the correlation over time. A new dot pattern is generated independently for every frame, but left and right eye patterns are identical except for a shift by the desired disparity. Since there is no coherent motion over time, the DRDS stimulus contains only the CD cue. It is commonly assumed that the ability to perceive MID from a DRDS stimulus is evidence for the use of a CD mechanism, and that degradation in perception of MID for DRDS stimuli in comparison with RDS stimuli is evidence for the use of an IOVD mechanism. However, this conclusion relies heavily upon the assumption that accurate estimates of disparity for each frame are available to the CD mechanism. Given the low temporal and spatial resolution of stereopsis (Norcia & Tyler, 1984; Regan & Beverley, 1973; Tyler, 1971), this may not be true of the neural representation of disparity.

The uncorrelated RDS (URDS), also called the time correlated RDS (TCRDS), breaks the correlation between the two eyes. The dot patterns presented to the left and right eyes are generated independently, but move coherently over time. The anticorrelated RDS (ARDS) stimulus is similar to the URDS except that instead of presenting independent dot patterns to the left and right eyes, anticorrelated (i.e. contrast reversed) dot patterns are presented. The URDS and ARDS stimuli contain spurious binocular correlations between the left and right images, which may evoke MID perception via a CD mechanism (Allison, Howard, & Howard, 1998). However, the CD cues in these stimuli are weaker than in the RDS. Thus, it is commonly assumed that the ability to perceive

[*] Corresponding author.
E-mail address: eebert@ust.hk (B.E. Shi).

http://dx.doi.org/10.1016/j.visres.2014.04.014
0042-6989/© 2014 Elsevier Ltd. All rights reserved.

MID from the URDS and ARDS stimuli is evidence for the use of an IOVD cue. Quantifying the degree to which these spurious correlations contribute to MID perception clearly depends upon the neural representation of disparity.

Fully addressing these assumptions in order to interpret the results of psychophysical experiments correctly will require the development of neurally plausible computational models of MID perception to proceed hand in hand with experimental work. To date, work in this area has been limited. We are aware of only two models of MID perception using cortically inspired operators: one using the IOVD mechanism and one using the CD mechanism. Sabatini et al. have demonstrated how a detector selective to MID can be constructed from motion/disparity energy units commonly used to model cortical neurons (Sabatini & Solari, 2004; Sabatini et al., 2001, 2003). This detector uses the IOVD mechanism, since it takes the difference between the two estimates of monocular velocity computed using the normalized opponent motion energy. Peng and Shi (2010) have proposed the CD energy model to construct units selective to MID using the CD mechanism. This model first constructs a distributed representation of binocular disparity using a population of binocular disparity energy units, in which peak response shifts as a function of disparity. The model then uses a modification of the motion energy model to construct units tuned to different speeds and directions of the shift in the peak response.

This paper presents a comprehensive comparison of the responses of populations of model units that are selective for MID using biologically plausible IOVD and CD mechanisms. The inputs to these models are the visual stimuli commonly used in psychophysical studies intended to tease out the relative contributions of these mechanisms to perception. We also present a developmental mechanism through which these units may emerge in response to visual stimuli. This work makes several contributions. First, to our knowledge, it is the first comparison of the responses of neurally plausible models of the IOVD and CD mechanism to visual stimuli. These models are comparable in that they consist of similar processing stages, only reversed in order. Surprisingly, we find that both CD and IOVD model predict the same performance degradation trend for various types of stimuli (RDS, URDS, DRDS, ARDS). This is inconsistent with common hypotheses that the CD and IOVD mechanisms can be distinguished by breaking the correlations in binocularity or time in stimuli (Brooks, 2001, 2002a; Brooks & Stone, 2004, 2006; Cumming & Parker, 1994; Fernandez & Farell, 2005; Harris, McKee, & Watamaniuk, 1998; Harris & Rushton, 2003; Regan & Beverley, 1973; Shioiri et al., 2009, 2008; Shioiri, Saisho, & Yaguchi, 2000). Second, it extends prior modeling work, which considered only the responses of single units or pairs of units tuned to MID, to the construction of entire populations of units tuned to different motions in depth. This enables us for the first time to derive estimates MID velocity from computational models of neuronal MID selectivity. Finally, it is the first developmental model of MID selectivity. Here we find evidence suggesting that units with the CD mechanism are easier to develop than units with the IOVD mechanism.

2. The CD and IOVD energy models

This section presents biologically plausible models for population responses of neurons selective to MID via the CD and IOVD mechanisms. Because these models are constructed using the same mathematical operations as used by the disparity energy (Ohzawa, DeAngelis, & Freeman, 1990) and motion energy models (Adelson & Bergen, 1985; Watson & Ahumada, 1985), which model the responses of disparity and motion selective complex cells in the primary visual cortex, we refer to these two models as the CD and IOVD energy models. The first subsection gives an overview of the CD and IOVD energy models. The second subsection defines the mathematical operations used in each step of the models. The third subsection derives a closed form expression for the outputs of the models. The fourth subsection describes how MID can be estimated from the outputs. The fifth subsection describes the parameter settings used in our numerical experiments. The final subsection describes the input stimuli applied to the models.

2.1. Overview

Fig. 1 shows the structure of the CD and IOVD energy models. Both models consist of two stages. The models can be considered dual in the sense that the two stages are similar, but their order is reversed. In the CD energy model, disparity is encoded first by combining inputs from the left and right eyes. Temporal changes in disparity are encoded second by combining inputs from current and delayed versions of the first stage disparity encodings. In the IOVD energy model, velocity in the left and right eyes is encoded first by monocularly combining current and delayed inputs. Velocity differences are encoded second by combining the left and right eye velocity encodings from the first stage.

At each image location, both models take as input binocular input sequences and produce as outputs the responses of populations of neurons tuned to different MID. Left and right eye inputs are passed through a spatial high pass filter (SHPF) with a center-surround kernel, which approximates the processing taking place in the retina and LGN.

As shown in Fig. 1(a), the CD model first constructs a distributed representation of the spatial disparity between the left and right eye images as the responses of a population of spatial disparity units shown by the blocks labeled SDU. This stage is biologically plausible, as the primary visual cortex of primates contains neurons tuned to different disparities (Barlow, Blakemore, & Pettigrew, 1967; Cumming & DeAngelis, 2001; Pettigrew, Nikara, & Bishop, 1968) and it seems likely that the neural system encodes stereo disparity as the distributed activity across a population of such units (Qian, 1994). The response of each SDU combines information from the left and right eye images according to the disparity energy model, which has been used to model the responses of disparity tuned complex cells (Ohzawa, DeAngelis, & Freeman, 1990). The population is constructed by varying the phase parameter Ψ, which determines the disparity tuning of the unit, from $-\pi$ to π. The original disparity energy model considered only the instantaneous binocular input, and did not include the temporal low pass filter (TLPF). Since we consider time varying input, we include the temporal filter to account for the finite temporal kernels of neurons in V1 (DeAngelis et al., 1999; DeAngelis, Ohzawa, & Freeman, 1993; Ohzawa, DeAngelis, & Freeman, 1996). The normalization stage makes the responses invariant to image contrast.

Our analysis below shows that at each spatial location and time instant, the population response of the first stage varies smoothly over phase Ψ, and has a single peak, whose location changes with the disparity between the left and right eyes. Thus, the shift in the peak location over time is an indicator of MID via a CD mechanism. The second stage of the CD energy model encodes this shift as the distributed activity across a population of phase disparity units (PDUs). Each PDU is tuned to a different phase shift between the current and delayed versions of the normalized SDU population response by a phase parameter Θ, which varies from $-\pi$ to π. The current and delayed versions of the normalized SDU population response are obtained by two temporal filters (TCPF/TSPF) that have similar kernels but differ in phase (CP = cosine phase, SP = sine phase). The computations performed by the PDU are similar to those of the SDU, except that the PDU combines information from units that vary by phase, rather than spatial location. The

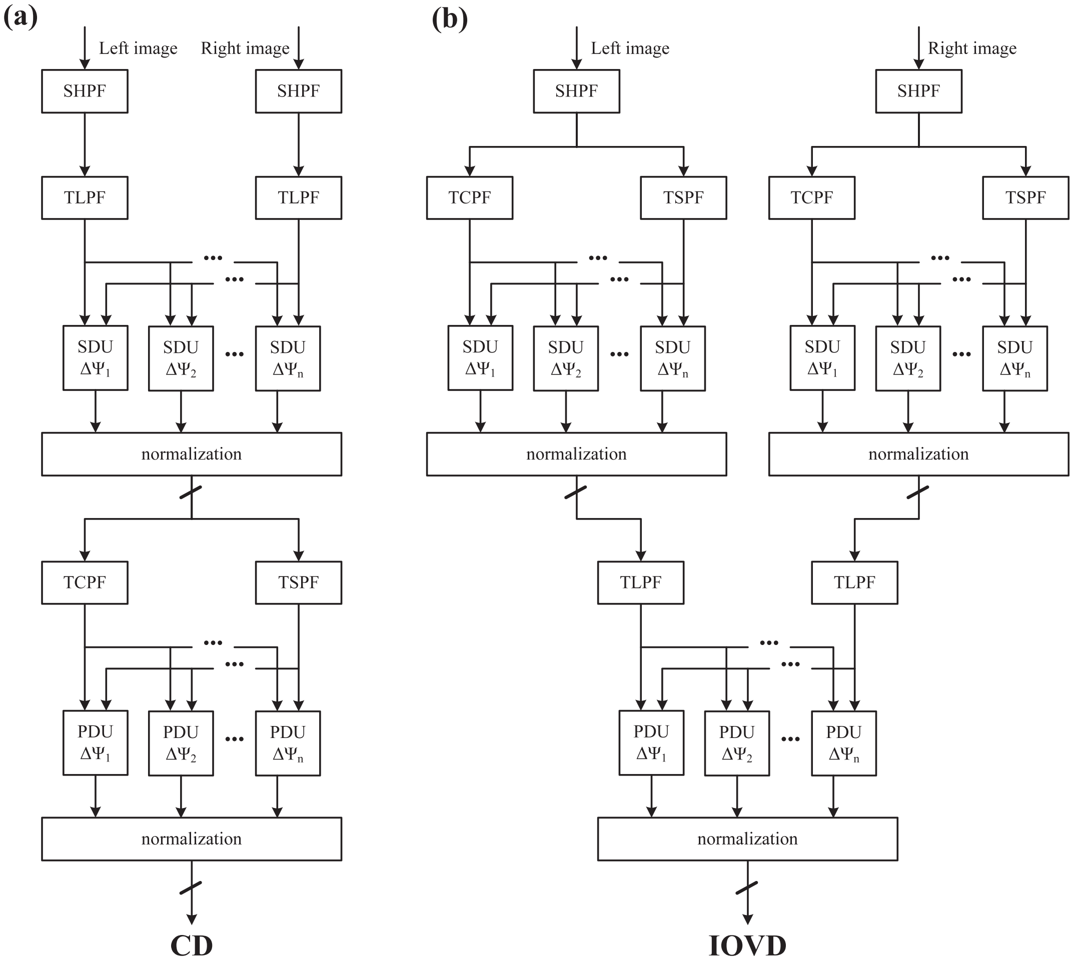

Fig. 1. The structure of the CD and IOVD energy models. (a) The CD energy model. The first stage constructs a distributed representation of binocular disparity by combining left and right eye images in a population of spatial disparity units (SDUs). The second stage encodes temporal changes in disparity by combining current and delayed versions disparity encoding from the first stage. Current/delayed versions are obtained by temporal filtering (TCPF/TSPF). (b) The IOVD energy model. The first stage constructs distributed representation of monocular velocity by SDUs selective to spatial shifts (disparities) between current and delayed monocular inputs. The second stage encodes interocular velocity differences by combining the first stage velocity encodings for the left and right eyes. Abbreviations: SHPF = Spatial High Pass Filter. TLPF = Temporal Low Pass Filter. TCPF = Temporal Cosine Phase Filter. TSPF = Temporal Sine Phase Filter. SDU = Spatial Disparity Unit. PDU = Phase Disparity Unit.

magnitude and direction of MID can be estimated from the peak location in the PDU population.

The IOVD energy model, shown in Fig. 1(b), uses the same building blocks, but in the opposite order. Its first stage constructs distributed representations of the monocular motion in the left and right eye images. Motion in each eye is represented by a population of SDUs, which are selective to different spatial disparities between current and delayed versions of the input images. As in the second stage of the CD energy model, these current and delayed versions are obtained by passing the images through two temporal filters (TCPF/TSPF). As in the first stage of the CD energy model, the population is constructed by varying a phase parameter Ψ from $-\pi$ to π.

The two population responses in the first stage each have a single peak, whose location changes with the left and right eye image velocities. Thus, an IOVD appears as a shift in the peak location between the left and right eye SDU populations. As in the CD energy model, the second stage of the IOVD energy model encodes this shift using a population of PDUs. Unlike the CD energy model, where the PDUs combine current and delayed versions of the SDU population responses, in the IOVD energy model, the PDUs combine SDU populations responses from the left and right eyes.

2.2. Mathematical definitions

This section details the mathematical operations performed by each of the blocks shown in Fig. 1.

The spatial high pass filters (SHPF) are implemented by taking the original image and subtracting a spatially low pass filtered version of the image. The kernel of the spatial low pass filter is a circular Gaussian with variance of σ_x^2 in both directions.

The temporal low pass filter (TLPF) has kernel

$$\mathcal{G}(t|\alpha,\tau) = \frac{1}{\Gamma(\alpha)\tau^{\alpha}} t^{\alpha-1} e^{-\frac{t}{\tau}} u(t) \qquad (1)$$

where $\Gamma(\alpha)$ is the standard Gamma function and $u(t)$ is the unit step function. The parameters α and τ are known as the shape parameter and time constant. In our experiments, the TLPFs in the first stage of the CD energy model and the second stage of the IOVD model share the same shape parameter $\alpha = 1$, but may differ in their time constants, which are denoted by τ_{CD1} and τ_{IOVD2}.

The kernels of the temporal cosine phase filter (TCPF) and temporal sine phase filter (TSPF) are the real and imaginary parts of a complex valued kernel $\mathcal{G}(t|\alpha,\tau)e^{j\Omega_t t}$. Intuitively, the TCPF output depends largely on the current input, since if $\alpha = 1$, then $\mathcal{G}(t|\alpha,\tau) = \tau^{-1}e^{-t/\tau}u(t)$, and the kernel of the TCPF,

$\mathcal{G}(t|\alpha,\tau)\cos(\Omega_t t)$, has maximum value at $t = 0$. The TSPF output depends largely on the past input, since its kernel, $\mathcal{G}(t|\alpha,\tau)$ sin $(\Omega_t t)$, has maximum value at a time $t > 0$, which depends upon the values of τ and Ω_t and approaches $t = \pi/(2\Omega_t)$ for large τ.

The spatial disparity unit is selective to the spatial disparity between its two inputs. It combines two image inputs (left eye/right eye or cosine phase/sine phase) using the disparity energy model (Ohzawa, DeAngelis, & Freeman, 1990). The output of each spatial disparity unit is called the disparity energy, and depends upon space, time and a phase parameter Ψ, which varies from $-\pi$ to π in the population. Denote the two inputs to this unit by $W_1(x, y, t)$ and $W_2(x, y, t)$, The SDU first combines the two inputs linearly over space according to

$$\begin{bmatrix} Z_1(x,y,t,\Psi) \\ Z_2(x,y,t,\Psi) \end{bmatrix} = \iint \begin{bmatrix} b_{1,1}(x-\xi,y-\eta) & b_{1,2}(x-\xi,y-\eta,\Psi) \\ b_{2,1}(x-\xi,y-\eta) & b_{2,2}(x-\xi,y-\eta,\Psi) \end{bmatrix} \times \begin{bmatrix} W_1(\xi,\eta,t) \\ W_2(\xi,\eta,t) \end{bmatrix} d\xi d\eta \tag{2}$$

The kernels of the spatial filters are Gabor functions with a phase shift of Ψ between the kernels applied to the first and second inputs. The spatial filters used for computing Z_1 and Z_2 are identical, but in quadrature phase.

$$\begin{bmatrix} b_{1,1}(x,y) & b_{1,2}(x,y,\Psi) \\ b_{2,1}(x,y) & b_{2,2}(x,y,\Psi) \end{bmatrix} = \mathcal{N}(x,y|0,\mathbf{C}) \begin{bmatrix} \cos(\Omega_x x) & \cos(\Omega_x x+\Psi) \\ \sin(\Omega_x x) & \sin(\Omega_x x+\Psi) \end{bmatrix} \tag{3}$$

The spatial envelope $\mathcal{N}(x,y|0,\mathbf{C})$ is the 2D Gaussian kernel with zero mean and covariance matrix **C**. We assume **C** to be a diagonal matrix with elements σ_x^2 and $\sigma_y^2 = (2\sigma_x)^2$ so that the envelope is longer in the vertical than in the horizontal direction. The parameter σ_x^2 determines the spatial frequency bandwidth. The parameter Ω_x determines the center spatial frequency. The disparity energy is given by

$$E(x,y,t,\Psi) = (Z_1(x,y,t,\Psi))^2 + (Z_2(x,y,t,\Psi))^2 \tag{4}$$

This operation makes the disparity energy less sensitive to small shifts in the position of the stimulus.

The phase disparity units are similar to the spatial disparity units, except that they integrate information over the phase parameter Ψ in the population, rather than over space. By analogy with the above, we refer to their outputs as phase energy. If we denote the two inputs to the unit by $W_1(x, y, t, \Psi)$ and $W_2(x, y, t, \Psi)$, the phase energy is given by

$$E(x,y,t,\Theta) = (Z_1(x,y,t,\Theta))^2 + (Z_2(x,y,t,\Theta))^2 \tag{5}$$

where Z_1 and Z_2 sum the outputs of phase filters applied the two inputs, i.e.,

$$\begin{bmatrix} Z_1(x,y,t,\Theta) \\ Z_2(x,y,t,\Theta) \end{bmatrix} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \begin{bmatrix} b_{1,1}(\Psi) & b_{1,2}(\Psi,\Theta) \\ b_{2,1}(\Psi) & b_{2,2}(\Psi,\Theta) \end{bmatrix} \begin{bmatrix} W_1(x,y,t,\Psi) \\ W_2(x,y,t,\Psi) \end{bmatrix} d\Psi \tag{6}$$

and

$$\begin{bmatrix} b_{1,1}(\Psi) & b_{1,2}(\Psi,\Theta) \\ b_{2,1}(\Psi) & b_{2,2}(\Psi,\Theta) \end{bmatrix} = \begin{bmatrix} \cos(\Psi) & \cos(\Psi+\Theta) \\ \sin(\Psi) & \sin(\Psi+\Theta) \end{bmatrix} \tag{7}$$

Eq. (7) is simpler than Eq. (3) for the SDU. There is no need for the spatial frequency parameter, which converts from spatial coordinates to phase. There is also no need for the Gaussian envelope, since phase is restricted to lie between $-\pi$ and π.

The two normalization stages remove the effect of image contrast. These normalize the energy at each unit by the average energy across the population and the local average over space and time. If the inputs to the normalization units are $E(x, y, t, \Psi)$, the outputs are $\widetilde{E}(x,y,t,\Psi) = E(x,y,t,\Psi)/\overline{S}(x,y,t)$ where the overbar indicates pooling over a spatial and temporal neighborhood.

$$\overline{S}(x,y,t) = \frac{1}{2\pi}\int\iiint \mathcal{N}(\xi,\eta|0,\sigma_n^2\mathbf{I})\cdot\mathcal{G}(\gamma|\alpha,\tau_n) \cdot E(x-\xi,y-\eta,t-\gamma,\Psi)d\xi d\eta d\gamma d\Psi \tag{8}$$

The spatial neighborhood is modeled by a circularly symmetric Gaussian with variance σ_n^2. The temporal neighborhood is modeled by a Gamma distribution envelope with parameters α and τ_n.

For both CD and IOVD models, the final output energy at stage two is pooled over a spatial neighborhood modeled by a circularly symmetric Gaussian with variance σ_p^2 to account for the larger receptive fields of complex cells and to improve estimation performance (Fleet, Wagner, & Heeger, 1996; Heeger, 1987; Zhu & Qian, 1996).

2.3. Analysis of the population responses

This section gives mathematical expressions for the population responses of the two stages of the models, and describes how these population responses vary with the input disparity, motion or motion in depth.

At each location (x,y) and time t, the SDU population response as a function of phase has a stereotypical form, with a single peak whose location varies with the input disparity. Defining $Z = Z_1 + jZ_2$ and applying Eqs. (2) and (3),

$$Z(x,y,t,\Psi) = \iint \mathcal{N}(x-\xi,y-\eta|0,\mathbf{C})e^{j\Omega_x(x-\xi)}\begin{bmatrix} 1 & e^{j\Psi} \end{bmatrix} \times \begin{bmatrix} W_1(\xi,\eta,t) \\ W_2(\xi,\eta,t) \end{bmatrix} d\xi d\eta = V_1(x,y,t) + e^{j\Psi}V_2(x,y,t) \tag{9}$$

where

$$V_h(x,y,t) = \iint \mathcal{N}(x-\xi,y-\eta|0,\mathbf{C})e^{j\Omega_x(x-\xi)}W_h(\xi,\eta,t)d\xi d\eta \tag{10}$$

for $h \in \{1, 2\}$. Applying these equations to Eq. (4), we obtain

$$E(x,y,t,\Psi) = \|Z(x,y,t,\Psi)\|^2 = S(x,y,t) + P(x,y,t)\cos(\Phi(x,y,t)-\Psi) \tag{11}$$

where (omitting the dependency on (x,y,t) to avoid clutter)

$$\begin{aligned} S &= |V_1|^2 + |V_2|^2 \\ P &= 2|V_1 \cdot V_2^*| \\ \Phi &= \arg(V_1 \cdot V_2^*) \end{aligned} \tag{12}$$

Thus, the population response is a cosine in phase Ψ with magnitude P, offset S and achieves its peak value at Φ.

For the first stage of the CD energy model, the two inputs to the SDU population, W_1 and W_2, come from the left and right eyes. Chen and Qian (2004) show that the peak location Φ gives a reliable estimate for the horizontal spatial disparity between the left and right eyes according to

$$d_{\text{est}} = \frac{\Phi}{\Omega_x} \tag{13}$$

The two inputs to the SDUs in the first stage of the IOVD model are like current and delayed versions of the left or right images. The outputs are similar to the motion energy described by Adelson and Bergen (1985). In motion energy model, the spatio-temporal RF profiles were spatio-temporal Gabor functions. In the model presented here, the spatio-temporal RF profiles are temporally causal variants of spatio-temporal Gabor functions when $\Psi = \frac{\pi}{2}$ or $\Psi = -\frac{\pi}{2}$. The corresponding motion energy units respond maximally to sine wave gratings with spatial frequency Ω_x moving with speed Ω_t/Ω_x. The direction of motion resulting in the maximal response depends on the sign of the phase shift. Changing the magnitude of the phase shift does not change the velocity tuning, but rather the direction selectivity of the units. If $\Delta\Psi = 0$, the corresponding motion energy unit is still tuned to respond maximally to sine wave gratings with spatial frequency Ω_x moving with speed Ω_t/Ω_x, but responds equally well to gratings moving to the left and to the right.

The peak location Φ in the population shifts with the image velocity, and can be used to estimate the stimulus velocity reliably (Meng & Shi, 2009). The velocity estimate changes monotonically, although nonlinearly, with Φ. The nonlinearity arises because the phase difference between the frequency response of the TCPF and TSPF varies nonlinearity with frequency. If we replace the cosine phase and sine phase paths by the current and previous frame, then the estimate of the input stimulus velocity would change linearly with Φ. However, an ideal delay would not be consistent with the temporal responses of cortical cells.

For the PDUs in the second stages of the models, an analysis similar to that above shows that the population response is also given by Eqs. (11) and (12) with Ψ replaced by Θ and a slight change in the definitions of V_1 and V_2,

$$V_h(x,y,t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{j\Psi} W_h(x,y,t,\Psi)\,d\Psi \tag{14}$$

The peak location in the population is a reliable estimate for the phase disparity between the two inputs, W_1 and W_2.

2.4. Estimation of MID velocity

The analysis above suggests that the population responses of the CD and IOVD energy models are a distributed representation of the MID velocity. If populations of such units were used by the brain to encode MID velocity, the accuracy of MID perception would depend upon two factors: (1) the accuracy of the MID information contained in the population and (2) the mechanism used to extract this information. Since this study is primarily concerned with the first, we use the concept of ideal observer analysis (Geisler, 2003) to estimate the limit of perception given this representation. In particular, we estimate the MID v_d from the peak location Θ^p in the population responses of the PDUs in the second stages of the CD and IOVD models by maximum-a posteriori (MAP) estimation:

$$v_{d,\text{est}} = \arg\max_{v_d} p(v_d|\Theta^p) \tag{15}$$

The MAP estimate is the optimal estimate assuming a 0–1 loss function penalizing incorrect estimates. Thus, the quality of this estimate of MID velocity is a measure of the amount of information about the MID contained within the population.

According to Bayes rule,

$$p(v_d|\Theta^p) = \frac{p(\Theta^p|v_d)p(v_d)}{\sum_i p(\Theta^p|v_{di})p(v_{di})} \tag{16}$$

We approximate the conditional density $p(\Theta^p|v_d)$ with a von Mises distribution

$$p(\Theta^p|v_d) = \frac{1}{2\pi I_0(\sigma(v_d)^{-2})}\exp\left(\frac{\cos(\Theta^p - \mu(v_d))}{\sigma(v_d)^2}\right) \tag{17}$$

where $I_0(\cdot)$ is a modified Bessel function of order 0. The functions $\mu(v_d)$ and $\sigma(v_d)$ are found by first estimating their values at a discrete set of v_d from data, and then interpolating and extrapolating over a wider range of v_d by a least squares fit of the estimates to functions

$$\mu(v_d) = k_1 \arctan(k_2 v_d) \tag{18}$$

and

$$\sigma(v_d) = k_3 + k_4 \arctan|k_5 v_d + k_6| \tag{19}$$

We assume that the prior $p(v_d)$ is uniformly distributed between ±16 deg/s. Given these parameter fits and the assumption about the prior, we can map each peak location Θ^p to an estimate of the MID velocity. See the experimental results section for more details.

2.5. Parameter settings

Our simulations discretize space and time into pixels and frames. One arc minute of visual angle corresponds to one pixel. One second corresponds to 120 frames.

The spatial filters used in the SDUs of the CD and IOVD models (equation (3)) are identical. We use vertically oriented Gabor functions with center spatial frequency tuning of 3.75 cycles per degree and choose the spatial bandwidth to be 1.95 octaves with aspect ratio of two. In discrete space, this corresponds to $\Omega_x = 2\pi/16$ radians per pixel, $\sigma_x = 5$ pixels and $\sigma_y = 10$ pixels. The other spatial parameters are based on the choice of σ_x. The spatial standard deviation used to compute the normalization factor $\overline{S}(x,y,t)$ is $\sigma_n = 3\sigma_x$. The Gaussian used to spatially pool the second stage outputs has $\sigma_p = 2\sigma_x$.

Temporal filter parameters were chosen by setting the center temporal frequencies ($\Omega_{t,\text{IOVD1}}$ and $\Omega_{t,\text{CD2}}$) and the time constants (τ_{IOVD1} and τ_{CD2}) of the TCPF/TSPF block in the first stage of the IOVD model and the second stage of the CD model as free parameters. The other temporal filter parameters were scaled based on these. The time constants of the TLPF are $\tau_{\text{CD1}} = 0.6\tau_{\text{CD2}}$ and $\tau_{\text{IOVD2}} = 0.6\tau_{\text{IOVD1}}$. For the normalization factor $\overline{S}(x,y,t)$, the temporal parameters are $\alpha = 1$and $\tau_n = 1.6\tau_{\text{IOVD1}}$ or $1.6\tau_{\text{CD2}}$.

The free temporal filter parameters were chosen to minimize the root mean squared error (RMSE) in the estimates of MID for RDS stimuli with direct trajectories, while also ensuring that the performance of two models were similar for these stimuli. In this study, we are primarily interested in comparing the degradation in the accuracy of the information about MID contained in the two representations when temporal or inter-occular correlations in the input are removed, rather than comparing the absolute accuracy of the estimates from the CD or IOVD models. Equalizing the performance of the two models for RDS stimuli ensures that the two models start from similar baseline levels of performance. We made this assumption in an attempt to make the comparison of the two mechanisms unbiased, rather than based on any assumption about the relative accuracy of these two mechanisms in human observers.

Further details about how the parameters were chosen are given in Appendix A. Here, we summarize the final parameter choices. In order to ensure that the range of MID selectivity is the same in both models, we chose $\Omega_{t,\text{CD2}} = 2\Omega_{t,\text{IOVD1}}$. The factor of two appears because the first stage of the IOVD model must be selective to monocular velocities, while the second stage of the CD model must be selective to their difference, which can be twice as large for direct trajectories. Specifically, the temporal frequencies used are $\Omega_{t,\text{IOVD1}} = 5$ Hz and $\Omega_{t,\text{CD2}} = 10$ Hz. The time constants were chosen so that the relative time constant was $\Omega_t\tau = 0.2$ cycles ($\tau_{\text{IOVD1}} = 40$ ms and $\tau_{\text{CD2}} = 20$ ms). In units of discrete time, $\Omega_{t,\text{IOVD1}} = 2\pi/24$ radians per frame, $\tau_{\text{IOVD1}} = 4.8$ frames, $\Omega_{t,\text{CD2}} = 2\pi/12$ radians per frame and $\tau_{\text{CD2}} = 2.4$ frames.

The parameters chosen are consistent with data published with cortical cells in macaque (Dayan & Abbott, 2001; De Valois, Albrecht, & Thorell, 1982; Foster et al., 1985). They are also consistent with those used in our previous work on CD energy model (Peng & Shi, 2010).

2.6. Stimuli

Input stimuli used to characterize the model were generated from random dot patterns covering 2.1 * 2.1 degrees of visual space (128 * 128 pixels) with 50% dot density, where 50% indicates that the stimulus contains 50% white and 50% black elements. Dot size is 3 arcmin in diameter (3 * 3 pixels). Stimuli last for 1 s. We add Gaussian white noise to all stimuli. The noise is independent across space, time and eye with standard deviation equal to 2% percent of the maximum value.

For the random dot stereogram (RDS) stimuli, left and right eye image velocities, v_L and v_R, ranged between -2 and 2 deg/s in steps of 0.1 deg/s, depending upon the desired MID trajectory. Negative velocities correspond to leftward motion. The MID velocity is the difference $v_d = v_L - v_R$. The lateral (frontoparallel) velocity is determined by the average. For direct trajectories, where the surface approaches or recedes from the observer directly, the left and right eye image velocities have the same magnitude but opposite sign, $v_L = -v_R$ and there is no lateral component to the velocity. For oblique trajectories, the lateral component is nonzero. We characterize the obliqueness by the ratio between the monocular image velocities $\gamma = v_R/v_L$. For direct trajectories, $\gamma = -1$. For stimuli moving directly toward the right eye, $\gamma = 0$. Trajectories with $\gamma \leqslant 0$ ($\gamma > 0$) are called Hit (Miss) trajectories (Brooks & Stone, 2004; Cynader & Regan, 1978; Regan, 1993).

The initial disparity between the left and right stereograms is set so that the disparity 67 ms before the end of the stimulus is equal to the desired disparity pedestal. As the disparity pedestal increases beyond the size of the spatial RF of the first stage, the left and right eye inputs become more and more spatially uncorrelated and the input approaches the uncorrelated random dot stereogram (URDS). We generate URDS stimuli by generating the left and right eye inputs independently and translating them according to the desired left and right eye image velocities.

We generate dynamic random dot stereograms (DRDS) by generating a new random dot pattern independently at each frame and setting the horizontal shift between the left and right images according to the expected disparity. Since there is no coherent monocular motion for these stimuli, they cannot have oblique trajectories. However, they do have a well defined MID velocity, which is determined by the rate of change of the disparity.

We generate stimuli with varying degrees of monocular motion coherence by superimposing RDS and DRDS with different dot densities. For example, a stimulus with 50% density and 80% coherence is the combination of an RDS with 40% density and a DRDS with 10% density.

3. Results of parametric CD and IOVD energy models

3.1. Single unit tuning characteristics and population responses

The PDUs in the second stages of both the CD and IOVD energy models are selective to MID. Fig. 2(a) and (b) show the tuning characteristics of the PDU with $\Theta = \pi/2$ in the CD and IOVD models in response to RDS with different combinations of left and right image velocities (v_L and v_R). Both units exhibit selectivity to MID, since the response is maximal along the line $v_L - v_R = 2.2$ deg/s for the CD model and $v_L - v_R = 1.6$ deg/s for the IOVD model, and decreases with distance from the line. For pure MID selectivity, the tuning characteristics responses of the units should be constant along diagonal lines slanting from upper left to lower right, which correspond to points with the same MID velocity but different lateral velocities. Comparing the two heat maps, we observe that the CD unit exhibits better invariance to oblique trajectories. However, for large enough lateral components, the response of the unit does decay due to the temporal low pass filtering applied to the monocular inputs (Fig. 2 in Peng and Shi (2010)).

On the other hand, the responses of the IOVD unit decay more quickly as the lateral component of the 3D velocity increases. The faster decay with the lateral velocity is due to the nonlinear variation of the peak locations in the SDU populations of the first stage with monocular input velocity, which is shown in Fig. 3. The peak location saturates at phases of $\pm\pi$ for large velocities. Consider two trajectories (v_L, v_R) and (v'_L, v'_R), with the same MID velocity, $v_d = v_L - v_R = v'_L - v'_R$, but different lateral velocities: $(v_L + v_R)/2 = 0$ (direct) and $(v'_L + v'_R)/2 > 0$ (oblique). Fig. 3 shows that the difference in the peak locations for the oblique trajectory is smaller than the difference for the direct trajectory. Since the PDU in the second stage is tuned to a particular difference in peak location, the response decays as the difference changes.

The peak locations in the PDU populations vary smoothly with changes in the monocular image velocities in the left and right eyes. Fig. 2(c) and (d) show the variation of average peak location Θ^p in the CD and IOVD population with input image velocity. The peak locations increase smoothly and monotonically with MID along lines with constant lateral motion, suggesting that Θ^p may provide a good estimate of input MID velocity. The peak location in the CD energy model population exhibits better invariance to lateral translation. The poorer invariance exhibited by the IOVD model is primarily due to the nonlinear variation of the peak locations in the SDU populations of the first stage with monocular input velocity as discussed earlier.

3.2. MID estimation

As discussed in Section 2.4, the accuracy of MID estimates obtained from the location of the peak response is a measure of the information about MID velocity contained in each representation. We use MAP estimates derived from a probabilistic model of the variation in the peak response with MID velocity to obtain an optimal estimate of the MID velocity from the population response. Fig. 4(a) and (b) show that the empirical distribution of the peak location for direct trajectories can be well fit by the parameterized density $p(\Theta^p|v_d)$ in (17) for $v_d = 4$ deg/s for both the CD and IOVD energy models. Fits with similar quality were observed at other values of v_d. Fig. 4(c)–(f) compare empirically estimated values of $\mu(v_d)$ and $\sigma(v_d)$ with fits to Eqs. (18) and (19). Fig. 4(g) and (h) show the relationships between the peak location Θ^p and the MID velocity $v_{d,\text{est}}$ estimated according to (15). The estimated MID velocity saturates at 16 deg/s, the limit of the uniform prior over v_d.

Fig. 2(e) and (f) show the mean estimated MID velocity from the CD and IOVD models for RDS inputs. Since the estimates are obtained from the peak location, the heat maps are qualitatively similar to those in Fig. 2(c) and (d) except for a change in the color maps. The two models give similar estimates for direct trajectories, which fall along the diagonal line from the lower left corner to the upper right corner. Ideally, the heat maps should be constant along lines of constant MID (diagonal lines slanting from upper left to lower right). This is true for the CD model. However, the IOVD model systematically underestimates the MID velocity as the lateral velocity increases. This underestimation is due to the nonlinear variation of the peak location in the SDU populations of the first stage with monocular input velocity as discussed earlier.

Fig. 5 shows the mean and standard deviations of the MID velocities estimates from the two models for RDS, DRDS, URDS and ARDS stimuli with different MID velocities. Trajectories of the RDS, URDS and ARDS stimuli were direct. Since DRDS stimuli have no coherent monocular motion, they cannot have oblique

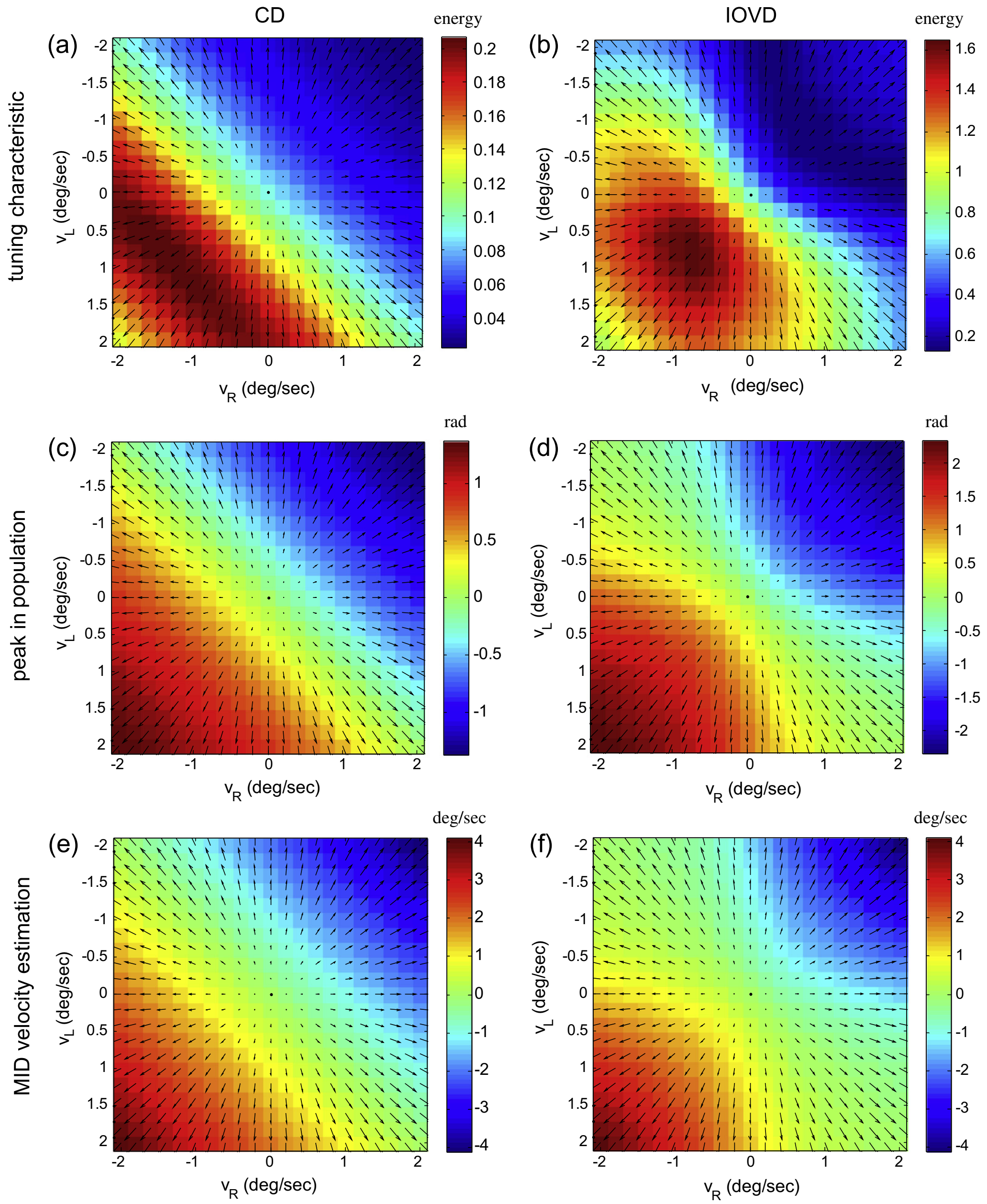

Fig. 2. (a and b) Heat maps of the average responses of the PDU with $\Theta = \frac{\pi}{2}$ for the CD (a) and IOVD (b) energy models for RDS inputs with varying left (v_L) and right (v_R) image velocities. (c and d) Heat maps of the average peak location in the CD (c) and IOVD (d) PDU populations. (e and f) Heat maps of the MID estimates obtained from the peak location in the CD (e) and IOVD (f) PDU populations. Arrows indicate interocular velocities, where negative velocities correspond to leftward motion. In each subplot, points in the lower left correspond to motion toward the observer. Points on the upper right correspond to motion away from the observer. Points corresponding to direct MID trajectories ($v_L = -v_R$) lie on the diagonal line connecting the lower left corner to the upper right corner. Other points correspond to oblique trajectories. Lines slanting diagonally from upper left to lower right correspond to points with the same MID but different lateral velocities.

trajectories. Statistics were calculated using data from a 17 $*$ 17 rectangular array of 289 neurons whose RF centers are spaced by 5 arcmin for 67 ms (8 frames) before and after the binocular images coincide at fixation or the desired disparity pedestal and from 10 trials (49,130 data points). As expected, the RDS stimuli result in the best estimates. Surprisingly, neither model exhibits much degradation for both URDS and ARDS inputs. For the DRDS input, the performance of both models degrade, but in different ways. The estimates from the CD model exhibit a smaller bias (difference between the mean estimate and the true value) but larger variability (as measured by the standard deviation). The estimates from the IOVD model have a larger bias, but smaller variability.

As a single quantitative measure of the performance of the estimators, we compute the root mean squared error (RMSE) between the estimated and true values of the direct component of the MID velocity The RMSE combines the bias and the standard deviation as the square root of the sum of their squares. Since we consider only estimates of the direct component of MID velocity, the RMSE is well defined for both RDS and URDS stimuli (which contain both lateral and direct components) and for DRDS stimuli (which contain only the direct component).

Fig. 6(a) and (b) show the RMSE of the CD and IOVD models for RDS inputs with direct trajectories, zero disparity pedestal and MID velocities varying from −4 to +4 deg/s. Different curves show the results for RDS inputs with different monocular motion coherence levels ranging from 100% (pure RDS) to 0% (pure DRDS). The RMSE curves as a function MID velocity generally have a “V” shape, and all intersect at a MID velocity of zero. When the MID velocity is

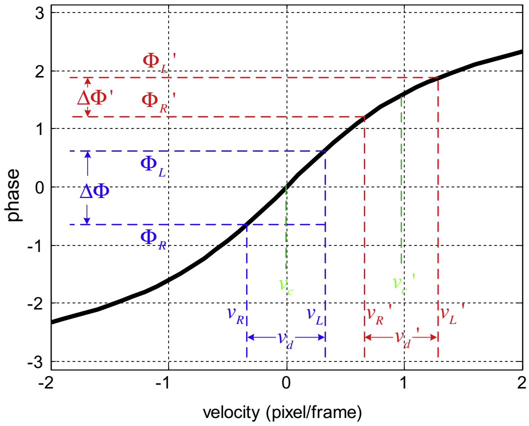

Fig. 3. The relationship between the peak location in the SDU population of the first stage of the IOVD energy model and the monocular image velocity.

zero, the stimulus is stationary in three dimensions. Thus the two inputs to the left and the right eyes are identical, except for the small amount of additive noise in the inputs. For the CD model, the population response of the SDU stage will have a peak that remains near zero for all time, except for fluctuations introduced by the additive noise. For the IOVD model, the population responses of the two SDU stages encoding monocular image velocity will be nearly identical. For pure RDS inputs, the monocular motion is zero, so the peak locations in both populations will remain near zero. For pure DRDS inputs, since there is no coherent monocular motion, the peak locations will vary randomly over time, but will be nearly identical at all times in the left and right eye populations. In both cases, the phase difference seen by the second stage will be close to zero, and thus the estimates of MID velocity will be close to zero. Thus, degradations in performance with monocular motion coherence will be evident only at nonzero MID velocities.

Estimation error increases as the coherence decreases for both models. The increase in RMSE is larger for the IOVD model than that for the CD model. At an MID input velocity of 1 degree per second, the RMSE of the IOVD model for the DRDS is 7.6 times larger than the RMSE for the RDS. For the CD model, the RMSE increase is 4.9 times. This larger increase is expected, since the DRDS stimuli disrupt the monocular motion cues which the IOVD model depends upon.

It might be surprising that the estimates from the CD model degrade at all, since each frame of the DRDS has a well defined binocular disparity. This degradation is due in part to the temporal filtering in the first stage of the CD model, which we included to model the finite temporal kernels of neurons in V1. Fig. 6(c) shows the RMSE of the CD energy model with the temporal dynamics in the first stage of the CD model removed by setting $\tau_{\mathrm{CD1}} = \tau_{n,\mathrm{CD1}} = 0$. The RMSE decreases in comparison with Fig. 6(a). On the other hand, removing the temporal dynamics in the binocular combination stage of the IOVD energy model by setting $\tau_{\mathrm{IOVD2}} = \tau_{n,\mathrm{IOVD2}} = 0$ leads to an increase in the RMSE (Fig. 6(d)).

Fig. 7(a) and (b) show the RMSE of the MID velocity estimates from the CD and IOVD models for RDS stimuli all with direct trajectories but at different disparity pedestals. As disparity pedestal increases, the RDS stimuli approach the URDS stimulus. Thus, we also include the result for the URDS stimulus. The RMSE for both models changes little with changes in the disparity pedestal. For small MID, the RMSE for the CD model with URDS stimuli is only 10% larger than the RMSE with RDS stimuli at fixation. The RMSE is essentially unchanged for the IOVD model.

Fig. 7(c) and (d) show the RMSE of the MID velocity estimates for the two models for RDS inputs with various oblique trajectories. As expected from the results shown in Fig. 2, the estimates from the IOVD model degrade faster than those from the CD model as the trajectory becomes more oblique, due to a stronger bias toward underestimating MID.

3.3. Decision making task

We also compare predictions of the CD and IOVD energy models with the results from psychophysical experiments reported in (Brooks & Stone, 2004). In these experiments, observers were presented with a pair of standard and test stimuli and asked to determine which stimulus was moving faster in MID. Brooks & Stone fitted the psychometric curve with the Gaussian cumulative distribution function, and computed the point of subjective equality (PSE), just-noticeable difference (JND) and Weber fraction from the fits. They found the Weber fraction for DRDS stimuli to be higher than the Weber fraction for RDS stimuli, which they suggested as evidence for IOVD processing.

We replicate this experiment with our models by adding a decision making stage to the outputs. To compute one point on the psychometric curve, we choose a standard stimulus MID velocity and a test stimulus MID velocity. We present the models with RDS stimuli with the standard and test MID velocities, and compare the MID velocities estimated by the models to determine which stimulus appears to be faster in each pair. We then calculate the percentage of time the test stimuli is perceived as faster than the standard stimuli. We repeat this process for test stimuli with varying MID velocity to generate the psychometric curve shown by the blue circles in Fig. 8. We then compute the Weber fraction using the same procedure as (Brooks & Stone, 2004). We fit the blue points with a Gaussian cumulative distribution function (shown as the red curve in Fig. 8), and calculated the PSE, JND and Weber fraction. Since our experimental and model settings are consistent with those used in (Brooks & Stone, 2004), our results should be comparable.

In order to match the targeted psychophysical experiments, we choose the standard stimulus to be 0.6 deg/s and swept the MID velocity of the test stimulus from −1 to 1 deg/s. Fig. 9 plots the speed discrimination thresholds from the two models, expressed as Weber fraction, for RDS and DRDS stimuli at different disparity pedestals. For RDS and DRDS with zero disparity pedestal, the Weber fractions produced by both the CD and IOVD model are quite similar to the experimental results presented in Fig. 3 of (Brooks & Stone, 2004), i.e. around 0.2 for RDS and higher (0.5–0.8) for DRDS. Brooks and Stone reported that "there is a clear effect of stimulus type (RDS vs. DRDS), but no obvious general effect of relative disparity pedestal." The results from both models for the RDS stimuli are consistent with this finding, as the Weber fractions are quite constant across all disparity pedestals. However, neither model replicates this finding for the DRDS stimuli. Instead, the Weber fractions increase with disparity threshold, indicating poorer speed discrimination for these DRDS stimuli as disparity pedestals increase.

4. Developmental models of MID selective neurons

In the models described above, the strengths b of the receptive field structures in the SDUs and PDUs were defined explicitly in Eqs. (3) and (7). These receptive field structures may be implemented by plastic synaptic connections between neurons. We propose here a developmental model for these connections. This developmental model enables us to address the question of whether the patterns of connectivity required to implement MID selectivity might be learned by a network in response to binocular motion-in-depth stimuli.

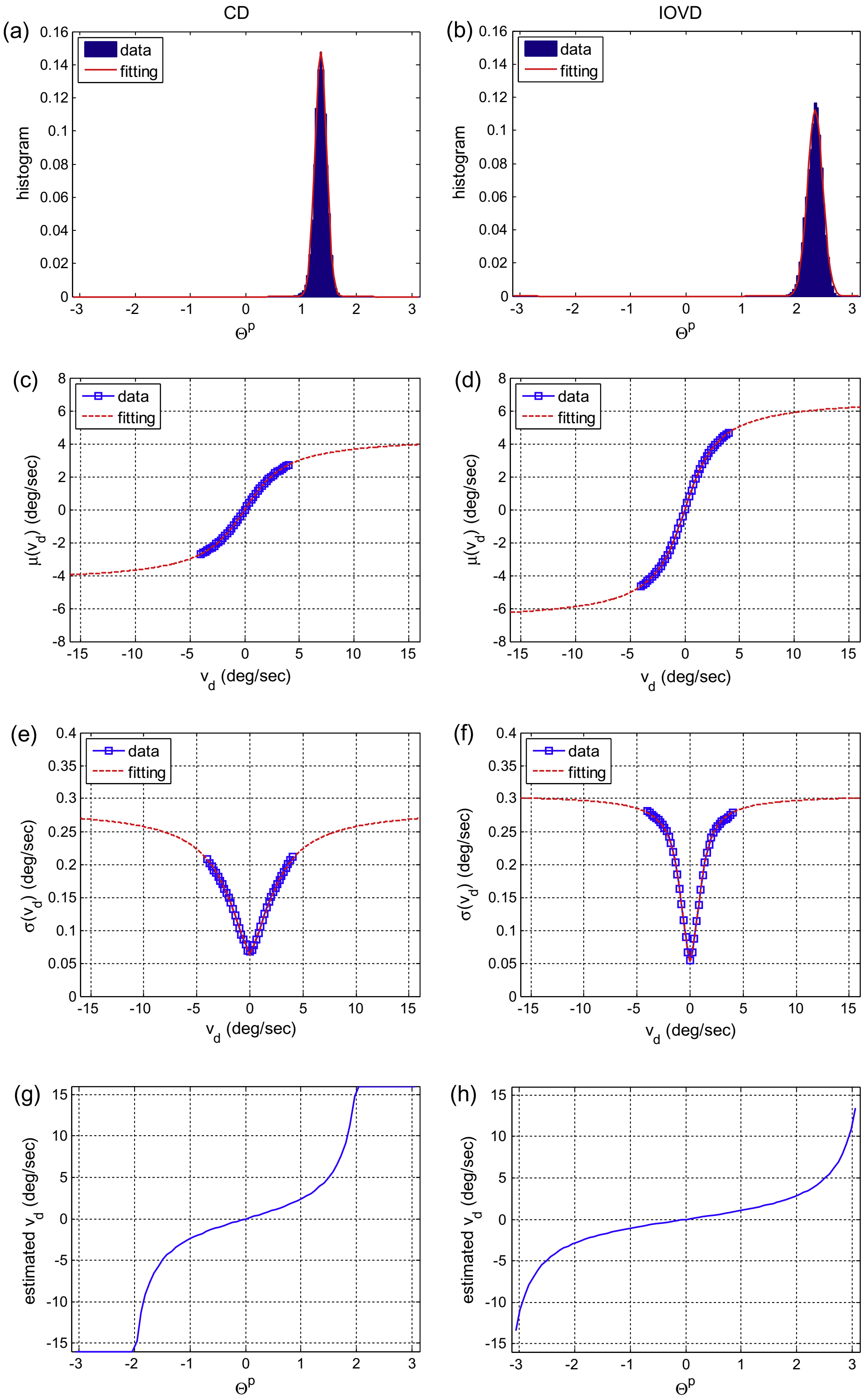

Fig. 4. Illustration of MID velocity estimation from the peak location in the phase-tuned population in response to RDS inputs. First row: two examples of distribution fitting. The blue bars show the distribution of peak phases given an input MID velocity equal to 4 deg/s. The red curves are fits to the conditional probability density function $p(\Phi^p|v_d)$ given in Eq. (17). In these two examples, $\mu(v_d = 4) = 1.3546$ and $\sigma(v_d = 4) = 0.1059$ for CD model, and $\mu(v_d = 4) = 2.3334$ and $\sigma(v_d = 4) = 0.1391$ for IOVD model. Second row: curve fitting of $\mu(v_d)$ with Eq. (18), where $k_1 = 1.4189$ and $k_2 = 0.7015$ for CD model, and $k_1 = 2.1725$ and $k_2 = 0.9154$ for IOVD model. Third row: curve fitting of $\sigma(v_d)$ with Eq. (19), where$k_3 = 0.0547$, $k_4 = 0.0571$, $k_5 = 0.7940$ and $k_6 = -0.4349$ for CD model, and where $k_3 = 0.0662$, $k_4 = 0.0557$, $k_5 = -2.3338$ and $k_6 = -0.8818$ for IOVD model. Fourth row: transfer function from the simulated Θ^p to estimated v_d. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

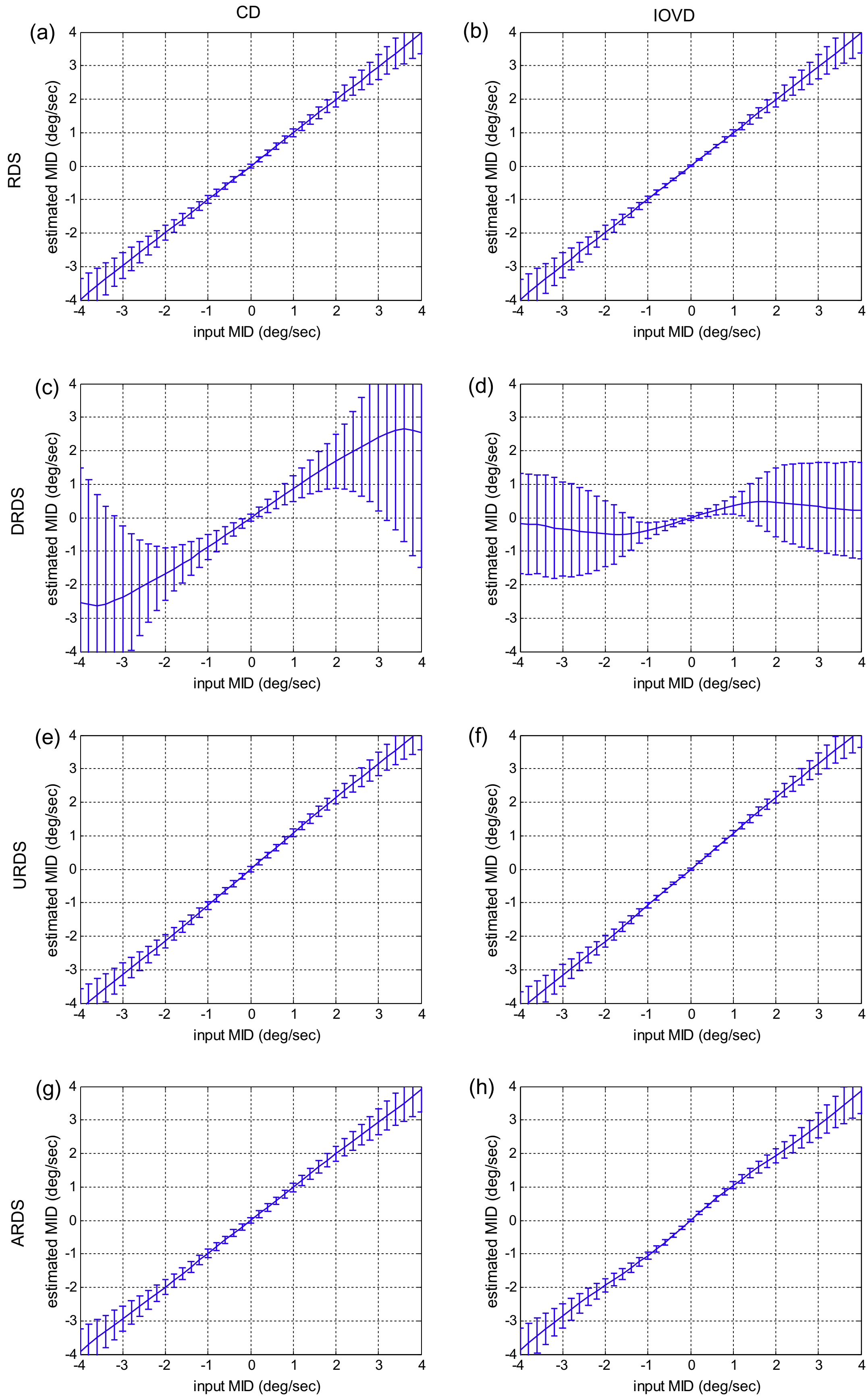

Fig. 5. Plots of the average MID velocity estimates by the CD and IOVD energy models for four different stimuli (RDS, DRDS, URDS, ARDS) as a function of the input MID velocity. Error bars show one standard deviation in the MID velocity estimates. Ideal estimates correspond to a diagonal line with unit slope. Trajectories for RDS, URDS and ARDS stimuli were direct.

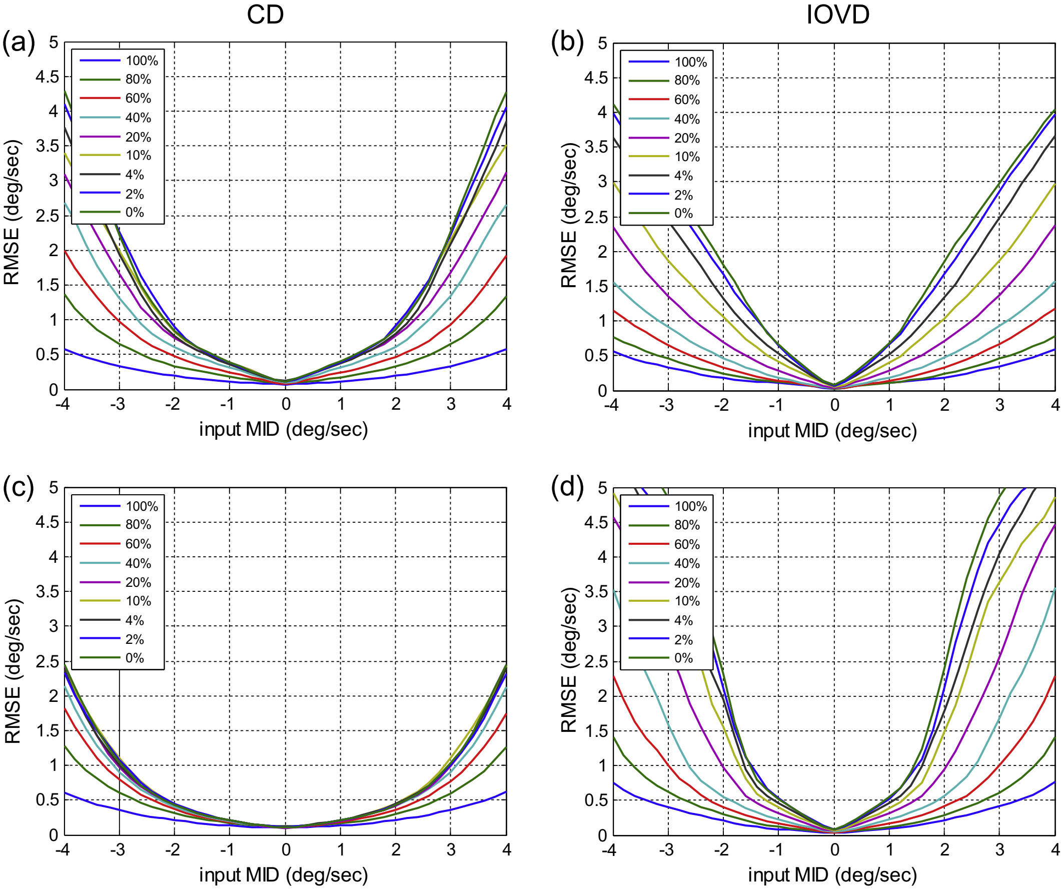

Fig. 6. (a and b) RMSE of the CD (a) and IOVD (b) models for RDS stimuli with different values of motion coherence. Stimuli with 0% motion coherence are DRDS. (c and d) Similar to (a and b) except that the TLPF and temporal filter for the normalization factor in the stages integrating information from the two eyes has been removed.

4.1. Model description

Our developmental model replaces the SDU/PDU populations with models with Adaptive Subspace Self-Organizing Maps (ASSOM) (Kohonen, Kaski, & Lappalainen, 1997). The ASSOM seeks to find a set of low dimensional subspaces that reflect collectively the statistics of a set of N dimensional input vectors. Let M be the number of subspaces. Each subspace, $\mathcal{L}^{(i)}$ where $i \in \{1, 2, \ldots, M\}$, has dimension H and is defined by H orthonormal basis vectors $\mathbf{b}_h^{(i)} \in \mathcal{R}^N$ where $h = 1, 2, \ldots, H$. Each subspace exhibits invariance along a stimulus dimension, e.g. translation, in the sense that it models input vectors that are similar except for a variation along the invariant stimulus dimension. Thus, each subspace is analogous to a complex cell in the visual cortex, which exhibits selectivity along some stimulus dimensions, e.g. orientation, but invariance along others, e.g. position. The collection of subspaces is organized topographically. Neighboring subspaces are constrained to be close to each other by using similar techniques as the traditional self organizing map (Kohonen, 1990).

Each SDU or PDU corresponds to one subspace, where we assume $H = 2$. The projected length of an input pattern at time t, $\mathbf{w}(t) \in R^N$, onto basis $h \in \{1, 2\}$ of subspace $L^{(i)}$ is

$$z_h^{(i)}(t) = \mathbf{b}_h^{(i)T}\mathbf{w}(t), \tag{20}$$

Eq. (20) corresponds to Eqs. (2) and (6). For the first stage, the input vectors $\mathbf{w}(t)$ contain the values of $W_1(x, y, t)$ and $W_2(x, y, t)$ in a small spatial patch. The basis vectors $\mathbf{b}_h^{(i)}$ contain the coefficients $b_{h,1}(x, y)$ and $b_{h,1}(x, y, \Psi)$ weighting information in those patches (Eq. (3)) over space. For the second stage, the input vectors $\mathbf{w}(t)$ contain the values of $W_1(x, y, t, \Psi)$ and $W_2(x, y, t, \Psi)$ at varying phases Ψ. The basis vectors $\mathbf{b}_h^{(i)}$ contain the coefficients $b_{h,1}(\Psi)$ and $b_{h,2}(\Psi, \Theta)$ weighting information over phase (Eq. (7)).

The projection of the input vector onto the subspace is

$$\hat{\mathbf{w}}^{(i)}(t_p) = \sum_h z_h^{(i)}(t_p)\mathbf{b}_h^{(i)}. \tag{21}$$

Since the basis vectors are orthonormal, the squared length of the projection is given by

$$E^{(i)}(t) = (z_1^{(i)}(t))^2 + (z_2^{(i)}(t))^2, \tag{22}$$

and corresponds to the energy in Eq. (4), (5). Input from the first stage to the second stage normalized through a soft max function,

$$\widetilde{E}^{(i)}(t) = \frac{\exp(E^{(i)}(t)/K)}{\sum_j \exp(E^{(j)}(t)/K)}, \tag{23}$$

where $K > 0$ is a "temperature" parameter controlling the entropy the soft-max function's output. For large K, all outputs are nearly equal (large entropy). As K approaches 0, the output corresponding to the largest input tends to one, and the remaining outputs to zero (small entropy).

The ASSOM is an attractive developmental model here because it organizes the units topologically so that adjacent neurons have similar tunings. This is consistent with experimental measurements of the organization of neurons tuned to orientation (Hubel & Wiesel, 1963; Maldonado et al., 1997; Ohki et al., 2006), disparity and ocular dominance (Chen, Lu, & Roe, 2008; DeAngelis & Newsome, 1999; Kara & Boyd, 2009) in primates and cats.

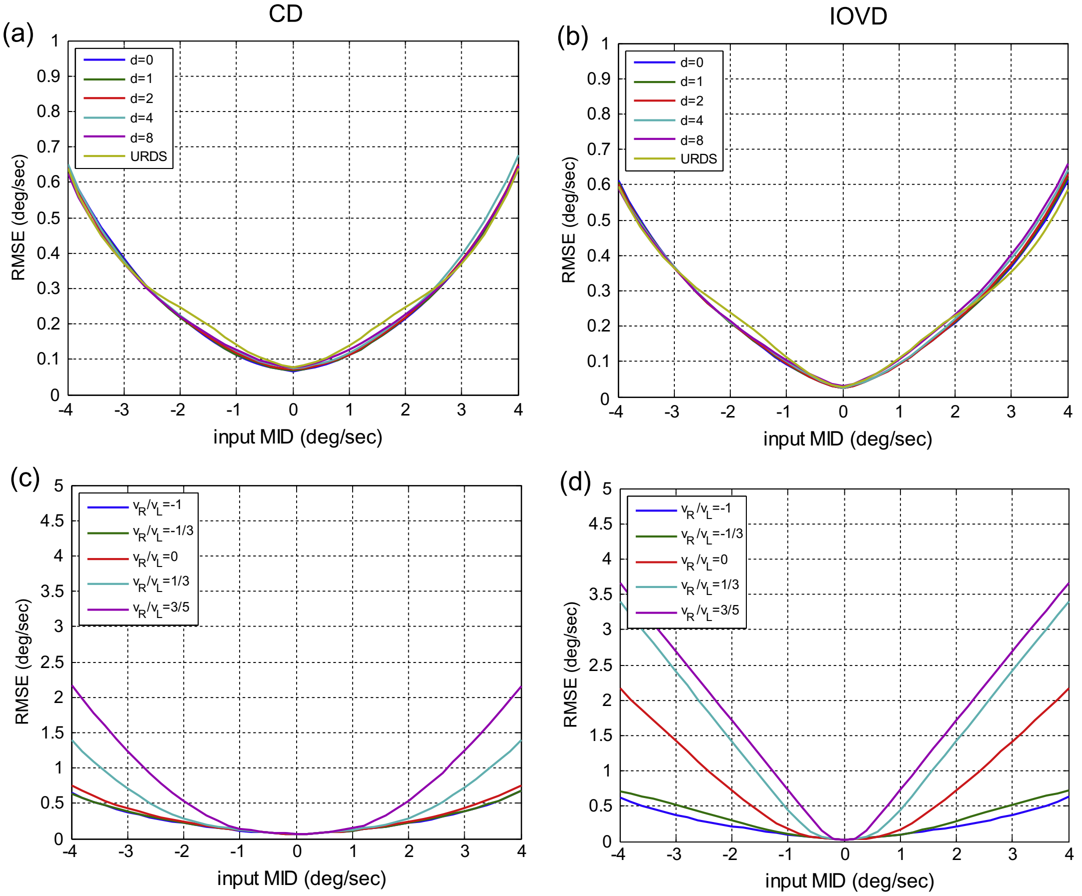

Fig. 7. (a and b) Effect of disparity pedestals. The plots show the RMSE of MID velocity estimates from the CD (a) and IOVD (b) models as a function of MID velocity for RDS inputs with different disparity pedestals ranging from 0 to 8 arcmin, and for URDS inputs. (c and d) The effect of oblique trajectories. The plots show the RMSE of MID velocity estimates from the CD (c) and IOVD (d) models as a function of MID velocity for RDS inputs with oblique trajectories covering direct ($v_R/v_L = -1$), hit ($v_R/v_L = -1/3$ or 0) and miss (v_R/v_L = 1/3 or 3/5) conditions.

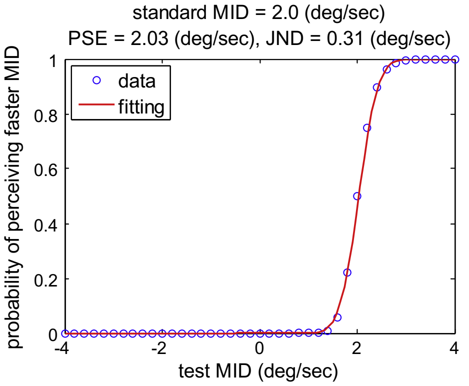

Fig. 8. Example of psychometric curve obtained from the decision making task performed by CD model. The standard stimulus has MID velocity of 2.0 deg/s. Blue circles indicate data generated by the decisions from the model. The red curve shows the Gaussian cumulative distribution function that is fit to the data.

In our experiments, the ASSOM units in the first stage are trained before those in the second stage. During training, the ASSOM subspaces are iteratively updated to maximize the lengths of the projections of input vectors onto the subspaces while also keeping neighboring subspaces similar. Input vectors are grouped into learning episodes, $S = \{t_p\}_{p=1}^{P}$, consisting of vectors that are similar but vary along the dimension (e.g. translation) of the desired invariance. For each episode, the algorithm proceeds in three steps.

Step 1: Find the winning subspace, which minimizes the total residual between the input vectors in the episode and their projections onto that subspace,

$$c = \arg\min_i \left\{ \sum_{t_p \in S} \|\tilde{\mathbf{w}}^{(i)}(t_p)\|^2 \right\}. \tag{24}$$

where the residual is given by

$$\tilde{\mathbf{w}}^{(i)}(t_p) = \mathbf{w}(t_p) - \hat{\mathbf{w}}^{(i)}(t_p). \tag{25}$$

Step 2: Update each basis vector $\mathbf{b}_h^{(i)}$ in the direction of the residuals $\tilde{\mathbf{w}}(t_p)$ according to

$$\Delta\mathbf{b}_h^{(i)} = \lambda h(i,c) \sum_{t_p \in S} \frac{\mathbf{b}_h^{(i)}(t)^T \mathbf{w}(t_p)}{\|\hat{\mathbf{w}}(t_p)\| \|\mathbf{w}(t_p)\|} \tilde{\mathbf{w}}^{(i)}(t_p). \tag{26}$$

where $\lambda > 0$ is a learning rate that decays during training (see Section 4.2). The size of the update decreases with the distance between the subspace and the winning subspace according to a Gaussian neighborhood function

$$h(i,c) = \exp\left(\frac{-\|\mathbf{r}_i - \mathbf{r}_c\|^2}{2\sigma^2} \right), \tag{27}$$

where $\mathbf{r}_i$ is the location of subspace i in the latent space. The width of the Gaussian, σ shrinks during training (see Section 4.2). The weighting factor inside the summation ensures that input patterns

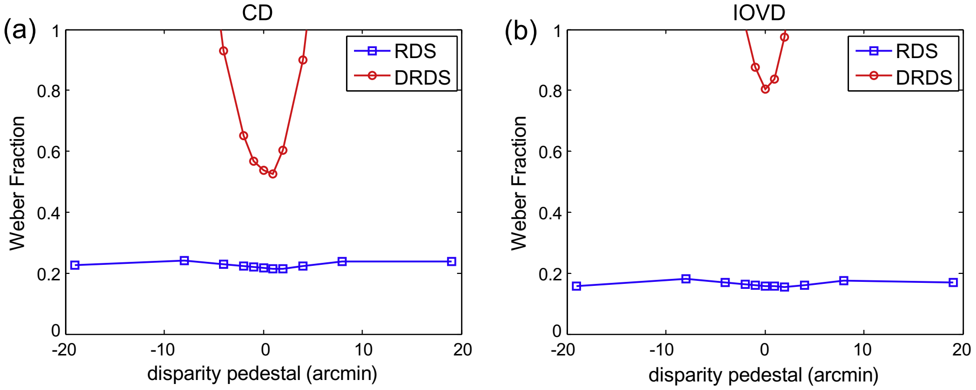

Fig. 9. Stereomotion speed discrimination thresholds, expressed as Weber fractions, for RDS and DRDS inputs with direct trajectories plotted as a function of the relative disparity pedestal.

closer to the basis vector contribute more to the update. This update minimizes a cost function given by

$$E = \sum_i h(i,c) \sum_{t_p \in S} \|\tilde{\mathbf{w}}^{(i)}(t_p)\|^2. \tag{28}$$

Step 3: Orthonormalize the basis vectors in each subspace.

The update rule we use here is slightly different from the original ASSOM algorithm in three ways. First, the original algorithm updates the basis vector in the direction of the data $\mathbf{w}(t_p)$, while we update it in the direction of the residual $\tilde{\mathbf{w}}^{(i)}(t_p)$. Since the basis vectors are always orthonormalized, both methods give similar numerical results. However, the direction of the residual is more consistent with the gradient of the cost function in Eq. (28)with respect to $\mathbf{b}_h^{(i)}$. Second, we do not use dissipation for the components of the basis vector. Dissipation was introduced in the original algorithm to improve the stability of learning, but we found it unnecessary in our application. Third, because we did not use dissipation, we can update the basis functions once per episode, rather than once per training pattern. These three modifications are similar to those used by the principal component analysis self-organizing map (PCASOM) algorithm (López-Rubio, Munoz-Pérez, & Gómez-Ruiz, 2004), another modification of the ASSOM algorithm.

4.2. Parameter settings

For both stages of the model, training lasts for 2×10^4 episodes. Each episode consists of $P = 9$ nine samples. The learning rate decreases according to

$$\lambda(m) = \frac{A_\alpha}{B_\alpha + m} \tag{29}$$

where m indexes the training episode, and $A_\alpha > 0$ and $B_\alpha > 0$. The width of the Gaussian neighborhood, σ, shrinks linearly from 5 to 0.5 during training.

For the first stages of the models, each ASSOM contains $M = 1024$ subspaces organized in a 32×32 grid. The parameters in (29) are $A_\alpha = 100$ and $B_\alpha = 500$. The temperature parameter in (23) is $K = 0.05$. The temporal frequency for the temporal filter is $\Omega_t = 7.5$ Hz ($2\pi/16$ radian/frame), and the time constant is $\tau = 42$ ms (5 frames). Inputs are 8x8 pixel image patches extracted from pre-whitened natural images provided at http://redwood.berkeley.edu/bruno/sparsenet/ (Olshausen & Field, 1997). For the CD model, the input disparity is uniformly distributed over ±8 pixels. We chose this disparity range to match the sizes of the image patches (8×8 pixels), which determine the spatial RF sizes. For disparities outside this range, the left and right inputs are uncorrelated. For the IOVD model, the input velocity is sampled from a uniform distribution between ±4 deg/s (±2 pixel/frame).

For the second stage, each ASSOM contains $M = 256$ subspaces organized in a 16×16 grid. The parameters in (29) are $A_\alpha = 200$ and $B_\alpha = 1000$. The temperature parameter in (23) is $K = 0.01$. The temporal filter parameters used in the second stage of the CD model are the same as in the first stage of the IOVD model. The input vectors are the concatenation of the two 1024 dimensional outputs of the first stage ($N = 2048$). For both networks, the monocular input velocities are independently chosen from uniform distributions between ±4 deg/s (±2 pixel/frame) in the horizontal direction. This assumption ensures that the model is equally exposed to all velocity pairs within the square regions in left–right image velocity space we used to characterize the tuning characteristics (e.g. Fig. 2(a and b) and Fig. 13), but may not be reflective of the actual statistics of image velocities encountered by the agent during development. These statistics depend upon both velocities in objects in the environment and the agent's behavior, i.e., eye movements, which co-develop with visual perception. Given the complexity of and lack of a clear consensus on modeling this problem, we have chosen the mathematically simple assumption of independence. It would certainly be worthwhile to re-examine this assumption as better characterizations of the statistics of visual input during development become available.

4.3. First stage of developmental models

Fig. 10 shows one of the two basis vectors that develop in the ASSOM subspaces in the first stages of the CD and IOVD developmental models. The other basis vectors (not shown) are similar, but are approximately in phase quadrature, as expected for translation invariance and orthogonality. Since we only consider the horizontal disparity and velocities, the two segments of the basis vectors corresponding to the left/right (TCPF/TSPF) inputs have similar shape and orientation but differ by a horizontal shift, which determines the disparity (motion) tuning preferences of the cell. The ASSOM algorithm arranges the basis vectors topographically, so basis vectors of neighboring subspaces are similar.

Fig. 11 shows the disparity tuning curves for subspaces in the first stage of the CD developmental model, which are computed by averaging responses given by (23) to 1000 RDS inputs for each disparity. A similar topographical organization is evident, where neighboring subspaces have similar tuning curves and preferred disparities. The motion tuning curves from the first stage of the IOVD developmental model are similar, and show a similar topographical organization (data not shown).

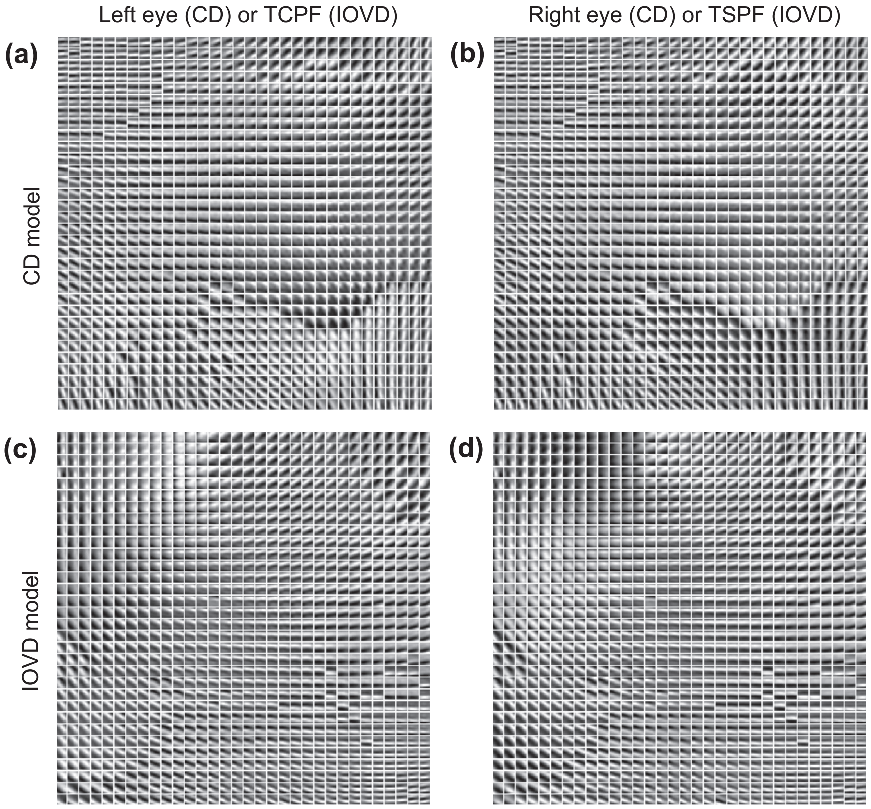

Fig. 10. The first basis vectors for the subspaces that develop in the first stage of the CD and IOVD developmental models. (a and b) Basis vectors for the first stage of the CD network shown as two separate images: one for the left eye (a) and one for the right eye (b). (c and d) Basis vectors for the first stage of the IOVD network shown as two separate images: one for the TCPF (c) and the other for the TSPF (d). Each image has 8 × 8 pixels, where the intensity is normalized to use the full range from black to white such that gray corresponds to zero.

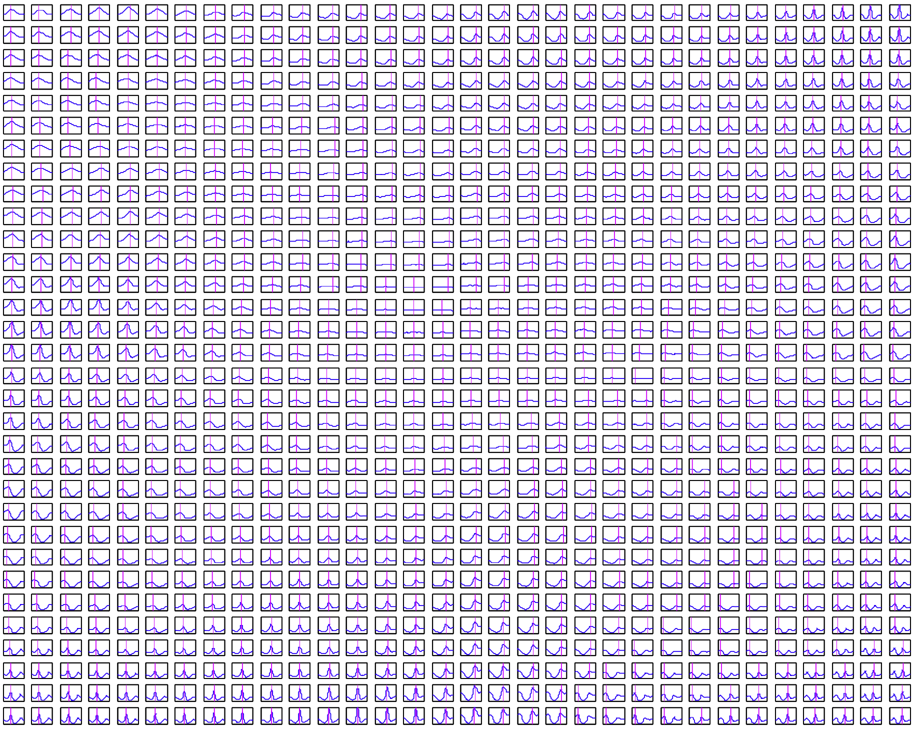

Fig. 11. Disparity tuning curves for subspaces in the first stage of the CD network are shown in blue. The magenta line indicates the preferred disparity. The horizontal axis corresponds to disparity values ranging from ±32 arcmin (pixels) in steps of 0.1 arcmin (pixel).

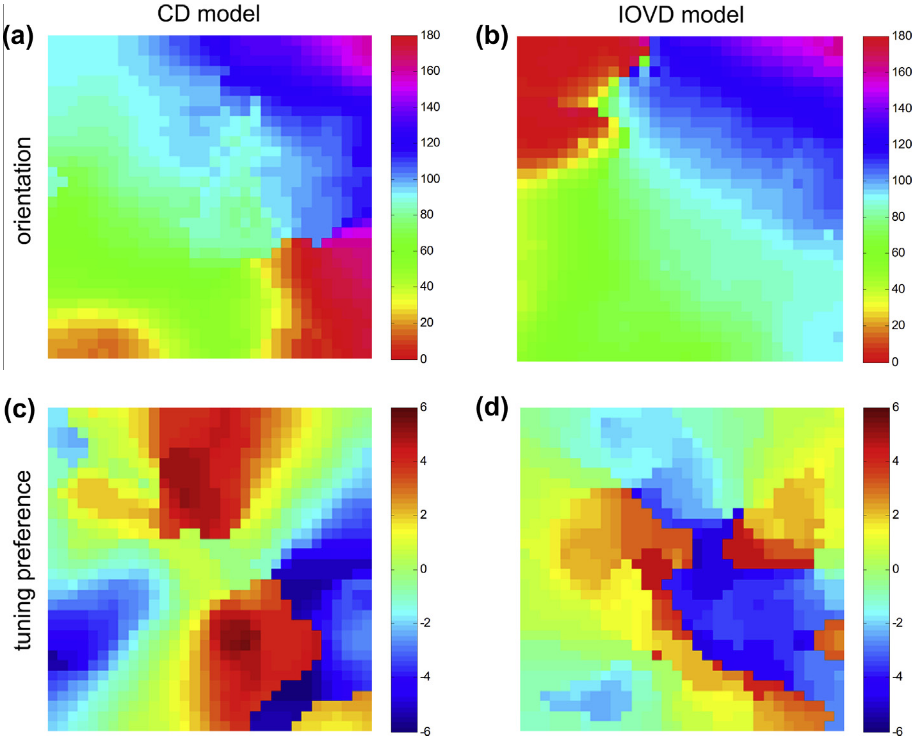

Fig. 12. (a and b) Heat maps of orientation selectivity of the subspaces learned by the first stages of the CD (a) and IOVD (b) developmental models. The color map legend shows the preferred orientation of the subspace in degrees. Orientations of 0° or 180° are vertical. Orientations of 90° are horizontal. (c) Heat map of preferred disparity for subspaces in the first stage of the CD developmental model. The color map legend gives the disparity preference of the subspace in arcmin (pixels). (d) Spatial map of preferred horizontal velocity for subspaces in the first stage of the IOVD developmental model. The color map legend gives preferred horizontal velocity of the subspace in pixels per frame. Multiply by two for degrees per second. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 12 shows that the orientation, disparity and horizontal motion selectivity of the subspaces vary smoothly in the latent space. Preferred orientations were estimated from the orientation of the best fitting Gabor to each basis function in Fig. 10. Preferred disparity and horizontal motion selectivity were estimated from the tuning curves as the input with the largest response.

4.4. Second stage of the developmental models

Fig. 13 shows the tuning characteristics of the ASSOM subspaces in the second stages of the CD and IOVD developmental models for RDS inputs with different pairs of monocular image velocities. Measured tuning characteristics are estimated by averaging responses to RDS stimuli at binocular fixation over 1000 trials. Many of the subspaces in the CD model are tuned to MID. Their tuning characteristics are very similar to those of the CD energy model in Fig. 2(a). Note that the *y*-axes in the heat maps of Fig. 13 are flipped in comparison with the heat map in Fig. 2(a). Thus, lines of constant MID run from lower left to upper right. Their offset from the origin determines the preferred MID velocity. The broad tuning along these diagonal line indicates the subspaces in the CD model exhibit invariance to oblique trajectories. On the other hand, the tuning characteristics for the IOVD network in Fig. 13(c) are not consistent with MID velocity selectivity. The maximum responses are either highly localized in a small regions, indicating selectivity to a particular combination of left and right eye image velocities, or lie along horizontal or vertical lines, indicating selectivity to image velocity in one eye, but invariance to the image velocity in the other. While the subspaces with localized tunings to a particular pair of unequal left and right eye velocities do respond preferentially to inputs with a non-zero MID, they do not exhibit the broader invariance to oblique trajectories observed in the tuning curve of the IOVD energy model (Fig. 2(b)).

In order to characterize the tuning characteristics across the population of ASSOM subspaces, we fit the measured tuning characteristics with 2D Gaussian functions with a bias:

$$\tilde{r} = A\exp\left(-\frac{x_1^2}{2\sigma_1^2} - -\frac{x_2^2}{2\sigma_2^2}\right) + B \tag{30}$$

where $\sigma_1 > \sigma_2$, and

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} v_L - \bar{v}_L \\ v_R - \bar{v}_R \end{bmatrix} \tag{31}$$

These 2D Gaussians are centered at image velocities $(\bar{v}_L, \bar{v}_R)$, and oriented along an axis with angle θ with respect to the horizontal. The fits are shown in Fig. 13(b) and (d).

Fig. 14 shows the distribution of the fitting parameters. Fig. 14(a) and (c) show the peak location $(\bar{v}_L, \bar{v}_R)$ obtained from the Gaussian fitting for CD and IOVD networks respectively. The joint distribution looks quite similar for both networks, with what appears to be a slight bias toward equal preferred left/right eye velocities for subspaces in the IOVD model. For ideal MID selectivity, the population should exhibit tuning to a variety of MID velocities $\bar{v}_d = \bar{v}_L - \bar{v}_R$ with $\theta = 45°$. The joint distribution of $(\bar{v}_d, \theta)$ for the CD developmental model shown in Fig. 14(b) exhibits this desired distribution. The orientation is concentrated around 45 degrees, and the MID selectivity across the population is broadly spread. On the other hand, the joint distribution of $(\bar{v}_d, \theta)$ for the IOVD developmental model shown in Fig. 14(d) does not exhibit this trend. The preferred MID velocities are concentrated around zero. The orientations exhibit some clustering around 45°, but the majority of subspaces exhibit other orientations, with a

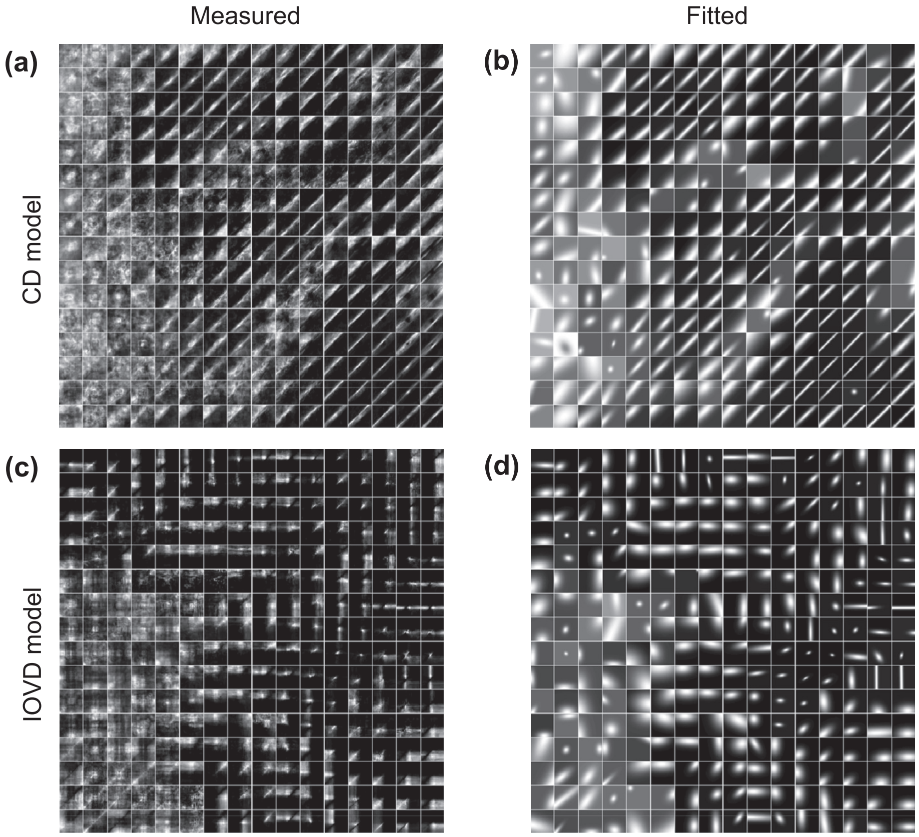

Fig. 13. Measured (a and c) and best fit (b and d) tuning characteristics of the CD (a and b) and IOVD (c and d) developmental models in response to RDS stimuli with different monocular image velocities. The tuning characteristics of each subspace are presented as 2D heat maps arranged according to their positions in the latent space. Horizontal and vertical axes indicate left and right eye image velocities ranging between ±4 deg/s (±2 pixel/frame) in steps 0.2 deg/s (±0.1 pixel/frame). The *y* axis of each heat map is flipped in comparison with those of Fig. 2. Black indicates zero response. White indicates the maximum response over the range tested.

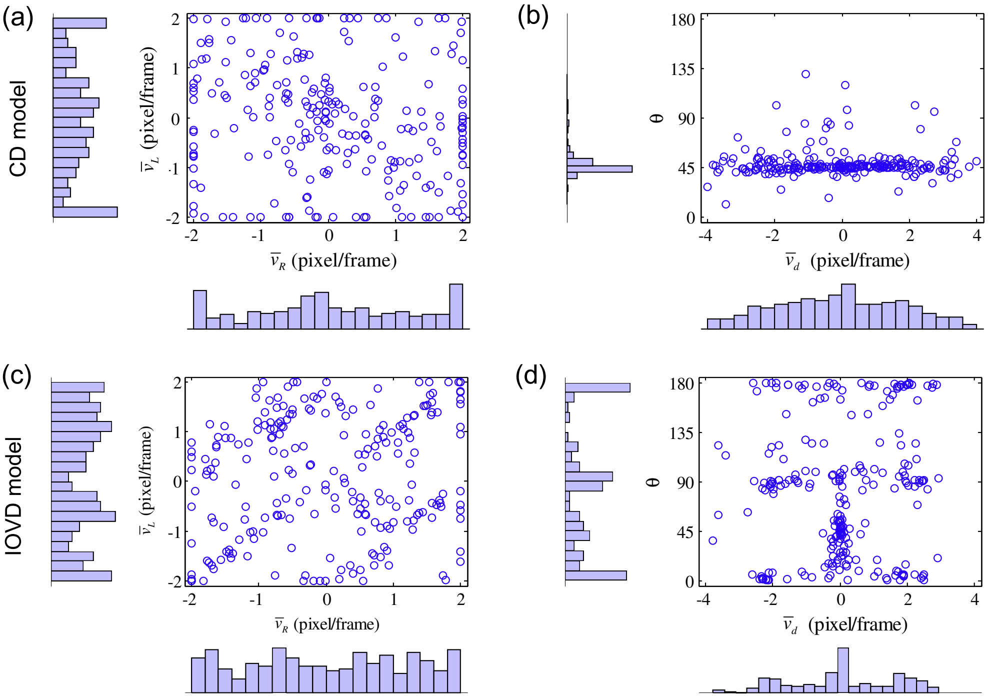

Fig. 14. Joint distributions of parameters used to fit tuning characteristics in the second stages of the CD and IOVD developmental models. (a) Joint distribution of the parameters $(\bar{v}_L, \bar{v}_R)$ determining the center of the Gaussian fit for the subspaces in the CD model. (b) Joint distribution of the preferred MID velocity $\bar{v}_d = \bar{v}_L - \bar{v}_R$ and the orientation of the Gaussian (θ). (c and d) Similar to (a and b) for the IOVD model. In all plots, the data from the ten percent of subspaces with the largest fitting error are excluded. Histograms along the horizontal and vertical axes indicate the marginal distributions.

preponderance of horizontal and vertical orientations, indicating the subspaces are tuned for velocity in one eye only.

Fig. 15 plots the heat map of the preferred MID velocity for the ASSOM units in the second stage of the CD model. The preferred MID velocity varies smoothly in the latent space.

5. Discussion

We have described two models for MID perception via CD and IOVD mechanisms, which are built using neurally plausible mechanisms acting directly upon binocular image sequences. The models studied here extend past work in several ways. First, the two models are comparable in that they are constructed from similar building blocks, only reversed in order. Second, the models construct distributed representations of MID velocity from which estimates of MID velocity can be extracted. Past work modeling joint selectivity to motion and disparity either did not model neurons tuned to MID or considered only one mechanism (either CD or IOVD). Past work has also considered models of single neurons or pairs of neurons. Third, we have shown that the required connectivity can be learned from exposure to binocular image sequences constructed from natural images. Past work has specified connectivity explicitly, with no consideration of how such connectivity might have developed.

Qian's stereomotion joint encoding model (Chen, Wang, & Qian, 2001; Qian & Andersen, 1997) is one of the earliest computational energy models to integrate stereopsis and motion. This model is a single stage model that replaces the monocular spatial RFs in the disparity energy model with monocular spatio-temporal RFs that are tuned to orientations in space–time (i.e. motion). These RFs are constructed by combining the spatial Gabor RF of the SDU in (3) with the temporal RF of the TCPF/TSPF. If the motion tuning in the two eyes is the same, units are tuned to stimuli at a particular disparity moving in a fronto-parallel direction. If two eyes are tuned to opposite or unequal motions, the units would be tuned to a specific trajectory that does contain a motion in depth component, but would not exhibit the invariance to oblique trajectories observed in the two stage IOVD and CD units proposed here.

Sabatini et al. have proposed a hierarchical model for MID velocity selectivity based on linearly combining the outputs of binocular units similar to those used by Qian with mismatches in ocular dominance (Sabatini & Solari, 2004; Sabatini et al., 2001, 2003). The units exhibit MID velocity tuning if the ocular dominance in the binocular units of the first stage (Layer 0) is not balanced, i.e. the ocular dominance index $\alpha \neq 0.5$. MID tuning is maximized if $\alpha = 0$ or 1. In this case, the first stage units reduce to monocular motion energy units, and the model output is the difference between the left and right opponent motion energies computed separately in the two eyes. Thus, this model is a neurally plausible implementation of the IOVD mechanism.

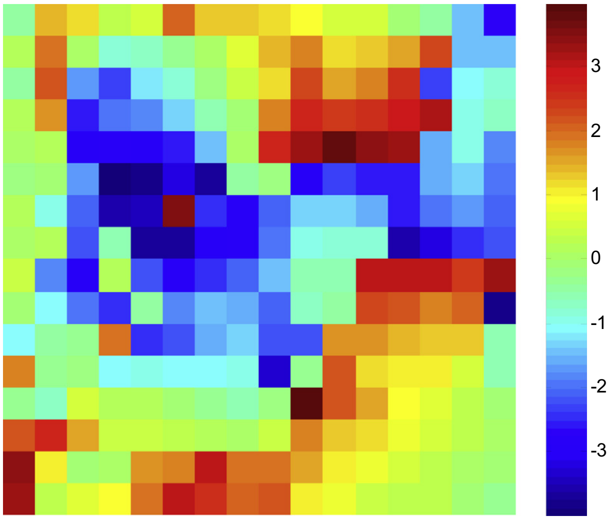

Fig. 15. Heat map of the preferred MID velocity of the ASSOM subspaces in the second stage of the CD model. The color map legend indicates the preferred MID velocity in units of pixels/frame.

Sabatini et al.'s model shares a similar hierarchical structure to the IOVD energy model described here, where motion information from each eye is first represented separately in the first stage before being combined binocularly in a later stage. However, it constructs only a single unit whose output varies with MID velocity, whereas the IOVD energy model constructs a distributed representation of MID velocity in the responses of a population of units tuned to different MID. The distributed encodings used by the IOVD energy model are more consistent with the population encodings of stimulus quantities found in the brain. Population-based representations provide more robust estimates than representations based on the responses of single neurons for input stimulus variables (Chen & Qian, 2004). In particular, interpreting the output of a single neuron is difficult, because its responses are confounded by selectivity along other stimulus dimensions, such as contrast and orientation. In addition, the opponent energy representation used by Sabatini et al. uses only two units to encode monocular motion, whereas the IOVD energy model encodes monocular image velocity as the distributed activity across a much larger population of units.

The hierarchical structure of the CD and IOVD energy models is consistent with what little is known about the possible neural substrate for MID perception. Visual processing in the brain is thought to be organized hierarchically (Felleman & Van Essen, 1991). The primary visual cortex contains populations neurons tuned to particular binocular disparities and spatio-temporal frequencies (Hubel & Wiesel, 1962, 1968). The responses of these neurons can be modeled using the mechanisms employed in the first stages of the CD and IOVD energy models (Adelson & Bergen, 1985; Ohzawa, DeAngelis, & Freeman, 1990; Watson & Ahumada, 1985). Neurons in higher stages have more complex selectivity, which is often modeled assuming input from V1 neurons, e.g. motion selective neurons in the middle temporal (MT) area (Simoncelli & Heeger, 1998), or neurons for object recognition in the inferotemporal (IT) cortex (Riesenhuber & Poggio, 1999). Recent evidence suggests that the computation of MID occurs in human occipito-temporal regions or hMT+ based on information extracted in earlier visual areas (Lamberty et al., 2008; Likova & Tyler, 2007; Rokers, Cormack, & Huk, 2009).

The IOVD energy model assumes that the IOVD computations are based directly on the outputs of motion selective neurons in V1. Although motion processing starts in V1, higher cortical areas further elaborate this processing, e.g. by introducing selectivity to pattern motion, as well as integrating information over larger spatial scales. Recent work suggests that the IOVD calculation is based upon computations performed at these later stages and larger scales (Rokers et al., 2011). An interesting avenue for future development of these models would be to extend them to include models of neuronal processing at these stages. In particular, integrating information over multiple and larger spatial scales may help to resolve the mismatch between the results predicted by the model on the decision task with DRDS stimuli at disparity pedestals and the experimental results in (Brooks & Stone, 2004). Both IOVD and CD energy models predict a rapid increase in the speed discrimination threshold for DRDS stimuli with increasing disparity pedestals.

Because these models assume binocular image sequences, they can be applied directly to the input stimuli used in psychophysical experiments designed to determine relative contributions of the CD or IOVD mechanisms. The predictions of our proposed CD and IOVD models are both consistent with many findings of these psychophysical experiments.

Several authors have reported degradations in MID perception for DRDS in comparison with RDS, with detection thresholds for DRDS being larger than for RDS (Brooks & Stone, 2004; Harris, McKee, & Watamaniuk, 1998). Fig. 5(c) and (d) show degraded MID perception for DRDS stimuli by both the CD and IOVD energy models. Fig. 6(a) and (b) show the RMSE of MID estimates increases with reduced motion coherence, where 0% motion coherence corresponds to a DRDS stimulus. Brooks and Stone (2004) observed no clear effect of a disparity pedestal in their experiments for both RDS and DRDS stimuli. This is consistent with results from both models for RDS stimuli. Fig. 7(a) and (b) show no change in the RMSE for input at different disparity pedestals. Fig. 9 also shows no changes in the speed detection thresholds. However, predictions of the models are not consistent with this finding for DRDS stimuli. Fig. 9 shows rapid increases in threshold with disparity pedestal for both models.

Several studies have indicated that URDS stimuli can lead to the perception of MID (Allison, Howard, & Howard, 1998; Shioiri, Saisho, & Yaguchi, 2000). This is consistent with both the CD and IOVD energy models. Fig. 5(e) and (f) show that the velocity estimates from both models for URDS stimuli are comparable to estimates from RDS stimuli. Fig. 7(a) and (b) show that RMSE of MID velocity estimates for URDS and RDS stimuli are almost identical in both models.

Several studies have found that although binocular anti-correlation impairs depth perception, it does not affect MID estimation (Czuba et al., 2011, 2010; Rokers, Cormack, & Huk, 2008). This is consistent with our simulation results of both the CD and the IOVD energy models with ARDS input, shown in Fig. 5(g) and (f). In addition, we find the RMSE of MID velocity estimates for anticorrelated RDS to be identical to the RMSE for standard RDS. This data is not shown here, as the RMSE curves essentially overlap. We discuss the reasons that there is no degradation in RMSE for ARDS input below.

Researchers have found that subjects over-estimate trajectory angle for oblique stimuli (Brooks & Stone, 2006; Harris & Dean, 2003). This overestimation is consistent with an underestimation of the MID component of an oblique trajectory compared with the lateral motion component. The MID velocity estimates of the IOVD energy model systematically underestimate the true velocity for oblique trajectories (Fig. 2(f)). Consistent with Fig. 2 in Harris and Dean (2003), the IOVD energy model predicts that the amount by which the trajectory angle is overestimated will increase as the trajectories become more and more oblique, assuming veridical estimation of the lateral motion. The CD energy model also leads to a slight underestimation of MID velocity for oblique trajectories, but the effect is not as strong as for the IOVD energy model.

Table 1 summarizes by comparing the predictions of common hypothesis about the changes in MID perception governed by the CD or IOVD mechanisms for different stimuli types and the predictions of our models. Common hypotheses predict MID perception by the two mechanisms will exhibit different changes for different stimuli. Thus, our finding that both the CD and IOVD models both exhibit the same qualitative responses to these stimuli may be a bit surprising. In the following, we explain the reasons for these results.

DRDS stimuli are binocularly correlated, but temporally uncorrelated. Each frame carries a well defined disparity signal, but they contain no monocular motion cues. Thus, observed degradations in MID perception for DRDS stimuli have been taken to imply that human perception relies partially upon an IOVD mechanism. Consistent with this intuition, our simulations with the IOVD energy model do indicate degraded MID perception. However, our simulations with the CD energy model also indicate degraded perception. This is unexpected if we assume a CD mechanism that can estimate *instantaneous* disparity reliably. However, the temporal and spatial resolutions of stereopsis are relatively poor (Norcia & Tyler, 1984; Regan & Beverley, 1973; Tyler, 1971). We incorporate this effect by modeling the temporal responses of disparity selective neurons in V1 by adding a temporal RF to the spatial RF in the first stage of the CD energy model (Chen, Wang, & Qian, 2001). This finite temporal response is in part responsible for the degradation, as removing it significantly reduces the RMSE of the CD energy model (compare Fig. 5(c) and (a)). These results highlight the importance of taking into account the temporal dynamics of neural processing when attempting to tease out the relative contributions of IOVD and CD mechanisms for MID perception.

Table 1
Comparison between the predictions about changes in perception of MID velocity for DRDS and URDS stimuli in comparison to RDS stimuli by common hypotheses in the literature and by the models presented. ∼ indicates little or no change, ↓ indicates significantly degraded perception. Numbers for the model predictions indicate the RMSE of the MID velocity estimates for inputs with MID velocity 1 deg/s normalized by the RMSE of the estimates for RDS inputs with the same input MID velocity.

Stimuli	Model			
	Common hypotheses		Model predictions	
	CD	IOVD	CD	IOVD
DRDS	∼	↓	↓4.9	↓7.6
URDS	↓	∼	∼1.1	∼1.0
ARDS	↓	∼	∼1.0	∼1.0

URDS stimuli are binocularly uncorrelated, but temporally correlated. These stimuli have well defined monocular motions, but no consistent disparity between the left and right images. Spurious correlations between the two eyes do lead to some CD cues in these stimuli, but these cues are weaker than in the standard RDS. Thus, the perception of MID in these signals has been taken for evidence for the use of an IOVD mechanism (Harris, Nefs, & Grafton, 2008). Consistent with this expectation, our results with the IOVD energy model and URDS stimuli shown in Fig. 7(b) show no significant degradation in the RMSE of MID estimates. However, our results with the CD energy model and URDS stimuli shown in Fig. 7(a) are inconsistent with this expectation, since we see no significant degradation in the RMSE of MID estimates. This result can be explained as follows. As indicated by Eq. (11), the population response in the SDU population will always have a single peak, whether or not the input stimulus has a well defined disparity. Since there is no well defined disparity between the left and right eye stimulus for a URDS, the absolute location of the peak is not meaningful. Nonetheless, the relative motion of the peak over time is still indicative of the relative velocity between the left and right eye stimuli, since the location of the peak, $\Phi(x, y, t)$, depends upon the phases of the complex valued spatial Gabor RF outputs for the left and right eyes, $V_1(x, y, t)$ and $V_2(x, y, t)$ in (10) and these phases vary approximately linearly with shifts in the left and right inputs. Since the phases of the left and right eye Gabor outputs are well defined whether or not there are similar patterns in the left and right eye, this phenomenon does not depend upon the existence of any possible spurious matches between dots over time in the URDS stimulus, which has been proposed as one possible mechanism for MID perception in URDS stimuli for a CD mechanism (Allison, Howard, & Howard, 1998).

Our finding that the RMSE of MID velocity estimates for anticorrelated and correlated RDS stimuli are identical for the IOVD energy model is expected, as monocular velocity cues are preserved. Since binocularly anticorrelated stimuli do not lead to the perception of depth, it has been argued that psychophysical experiments showing similar perception of MID on ARDS and

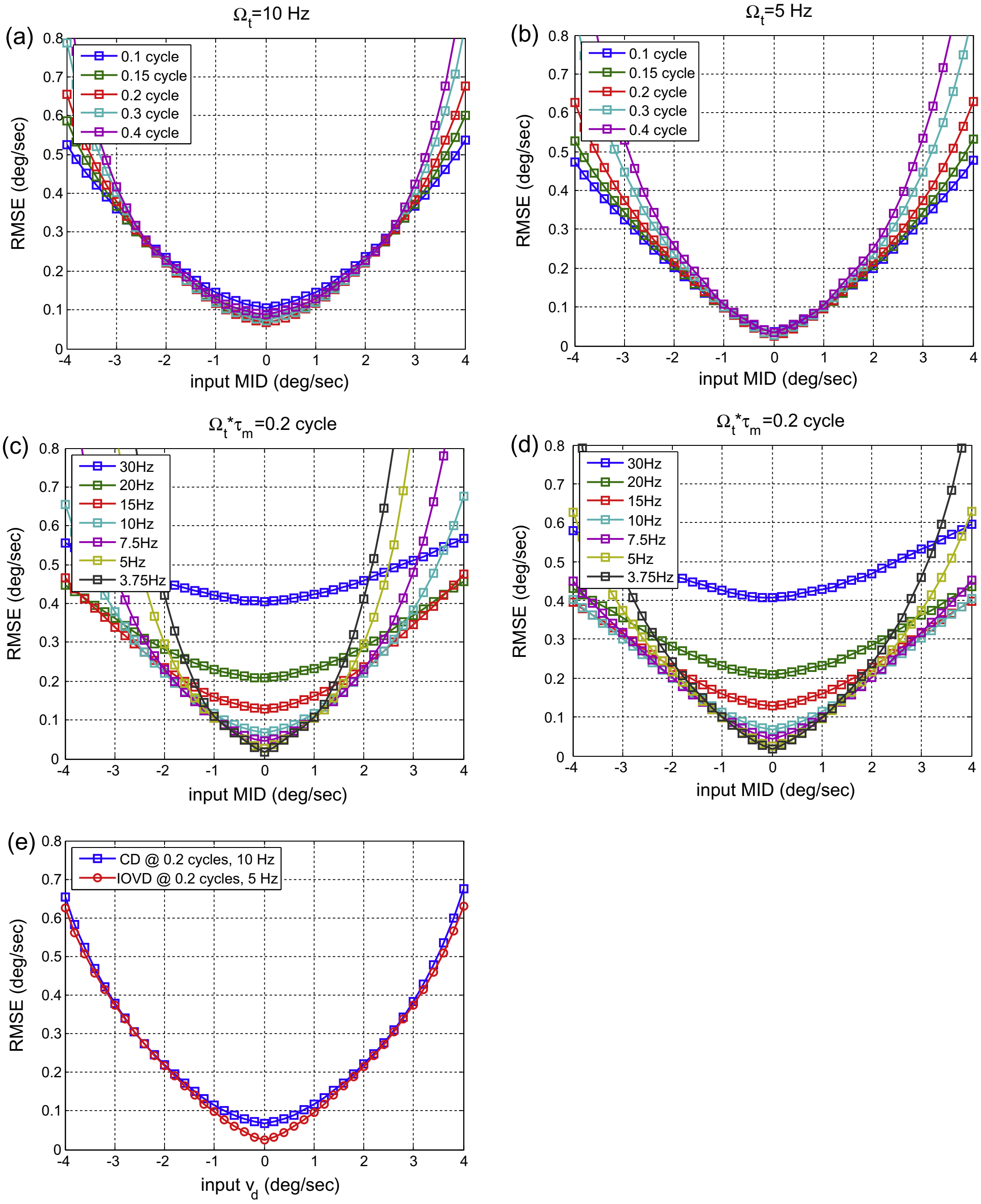

Fig. 16. The effect of changing the temporal filter parameters on the RMSE of MID velocity estimation. (a and b) The RMSE as a function of input MID velocity for different values of relative time constant for the CD (a) and IOVD models (b). (c and d) The RMSE as a function of input MID velocity for different center frequencies for the CD (a) and IOVD models (b). (e) Comparison of the RMSE for the CD and IOVD energy models for parameters chosen so that the RMSE is comparable for MID velocities between ±4 deg/s.

RDS provide evidence that IOVDs are used as a cue to MID, and that binocular anti-correlation is a useful tool for distinguishing the contributions of the CD and IOVD mechanisms. This conclusion is contradicted by our finding that the CD energy model does not exhibit degraded MID perception for ARDS stimuli. The reason that perception by our models is similar for ARDS and RDS is similar to that given in the previous paragraph. For anticorrelated stimuli, the disparity tuning curves of the SDU are inverted (Cumming & Parker, 1997). This leads to an inversion of the population response, so that the trough, rather than the peak, in the response encodes the stimulus disparity. This inversion disrupts depth perception, but the motion of the trough (and peak) in the population response still encodes the changing disparity.

On balance, our results suggest that many psychophysical results on MID perception may be equally well accounted for by neurally plausible models of either the CD or IOVD mechanisms, and that many stimuli commonly used to distinguish between the two mechanisms may be insufficient. Our models as presented do not include the effects of adaptation or after effects. Thus, we cannot model the results of experiments based on these effects (Brooks, 2002a, 2002b; Czuba et al., 2012; Fernandez & Farell, 2005; Sakano, Allison, & Howard, 2012; Shioiri et al., 2003). Elaboration of these models in order to model these effects would be a natural follow up to this study. In addition, it would be interesting to investigate the model responses to new stimuli for probing 3D motion perception, such as the dichoptic pseudoplaid stimulus (Rokers et al., 2011).

The learning mechanisms described here for developing the spatial patterns of interconnectivity required to support the CD and IOVD energy models suggest another approach to studying the potential contributions of CD and IOVD mechanisms toward MID perception. Our results here suggest that the connectivity required by the CD mechanism may be easier to develop in response to natural image input. The first stages of both the CD and IOVD developmental models can learn the connectivity to support the disparity or motion selectivity hard-wired into the CD and IOVD energy models. This is consistent with other recent work studying developmental models of the neuronal selectivity observed in primary visual cortex (Hyvärinen & Hoyer, 2001; Olshausen & Field, 1996; Tosic & Olshausen, 2011). However, we find that units in the second stage of the CD developmental model generally exhibit the broad invariance to stimulus trajectory observed in the CD energy model, whereas the units in the second stage of the IOVD developmental model generally do not. These developmental models predict that neurons in higher cortical areas tuned to MID will more likely be constructed according to a CD mechanism, rather than an IOVD mechanism.

Acknowledgments

This work was supported in part by the General Research Fund of the Hong Kong Research Grants Council (Project #619111).

Appendix A.

We describe here the factors considered in the choice of the temporal filter parameters of the TCPF and TSPF units, $\Omega_{t,\mathrm{IOVD1}}$, $\Omega_{t,\mathrm{CD2}}$, τ_{IOVD1} and τ_{CD2}.

We first fix the center frequencies to $\Omega_{t,\mathrm{IOVD1}} = 5$ Hz and $\Omega_{t,\mathrm{CD2}} = 10$ Hz, and examine the effect of varying the time constants, which we express in terms of the relative time constant, $\Omega_t\tau$, which has units of cycles of the center frequency. Fig. 16(a) and (b) show the RMSE of MID velocity estimation for direct trajectories as the relative time constant varies from 0.1 to 0.4 cycles (10–40 ms for the CD model and 20–80 ms for the IOVD model). Both models exhibit similar trends. For large MID velocities (near ±4 deg/s), the estimation error increases with the relative time constant. For MID velocities near zero, the error decreases first, then increases. For both models, the lowest error at zero MID occurs for a relative time constant of 0.2 cycles.

We then fix the relative time constant at 0.2 cycles, and vary the center frequency Ω_t from 3.75 to 30 Hz. Fig. 16(c) and (d) show that increasing the center frequency reduces the estimation RMSE for larger MID velocities, but increases RMSE for MID velocities near zero. We also observe that in order to achieve comparable estimation error at larger MID velocities, the temporal frequency of the CD model should be chosen twice as large as the temporal frequency in the IOVD model.

Fig. 16(e) compares the estimation error of the CD and IOVD models for the parameters used in the main text, $\Omega_{t,\mathrm{IOVD1}} = 5$ Hz, $\Omega_{t,\mathrm{CD2}} = 10$ Hz, and relative time constant $\Omega_t\tau = 0.2$ cycles. The estimation error for the two models is quite similar for MID velocities larger than 1 deg/s. However, the estimation error for the IOVD model is lower than that of the CD model for smaller MID velocities.

References

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

Allison, R., Howard, I., & Howard, A. (1998). Motion in depth can be elicited by dichoptically uncorrelated textures. *Perception, 27*, 46.

Barlow, H. B., Blakemore, C., & Pettigrew, J. D. (1967). The neural mechanism of binocular depth discrimination. *The Journal of Physiology, 193*(2), 327–342.

Brooks, K. (2001). Stereomotion speed perception is contrast dependent. *Perception, 30*(6), 725–731.

Brooks, K. R. (2002a). Interocular velocity difference contributes to stereomotion speed perception. *Journal of Vision, 2*(3), 218–231.

Brooks, K. R. (2002b). Monocular motion adaptation affects the perceived trajectory of stereomotion. *Journal of Experimental Psychology: Human Perception and Performance, 28*(6), 1470.

Brooks, K. R., & Stone, L. S. (2004). Stereomotion speed perception: Contributions from both changing disparity and interocular velocity difference over a range of relative disparities. *Journal of Vision, 4*(12), 1061–1079.

Brooks, K. R., & Stone, L. S. (2006). Stereomotion suppression and the perception of speed: Accuracy and precision as a function of 3D trajectory. *Journal of Vision, 6*(11), 1214–1223.

Chen, G., Lu, H. D., & Roe, A. W. (2008). A map for horizontal disparity in monkey V2. *Neuron, 58*(3), 442–450.

Chen, Y., & Qian, N. (2004). A coarse-to-fine disparity energy model with both phase-shift and position-shift receptive field mechanisms. *Neural Computation, 16*(8), 1545–1577.

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

Cumming, B. G., & DeAngelis, G. C. (2001). The physiology of stereopsis. *Annual Review of Neuroscience, 24*, 203–238.

Cumming, B. G., & Parker, A. J. (1994). Binocular mechanisms for detecting motion-in-depth. *Vision Research, 34*(4), 483–495.

Cumming, B. G., & Parker, A. J. (1997). Responses of primary visual cortical neurons to binocular disparity without depth perception. *Nature, 389*(6648), 280–283.

Cynader, M., & Regan, D. (1978). Neurones in cat parastriate cortex sensitive to the direction of motion in three-dimensional space. *Journal of Physiology, 274*, 549–569.

Czuba, T. B., Rokers, B., Guillet, K., Huk, A. C., & Cormack, L. K. (2011). Three-dimensional motion aftereffects reveal distinct direction-selective mechanisms for binocular processing of motion through depth. *Journal of Vision, 11*(10).

Czuba, T. B., Rokers, B., Huk, A. C., & Cormack, L. K. (2010). Speed and eccentricity tuning reveal a central role for the velocity-based cue to 3D visual motion. *Journal of Neurophysiology, 104*(5), 2886–2899.

Czuba, T. B., Rokers, B., Huk, A. C., & Cormack, L. K. (2012). To CD or not to CD: Is there a 3D motion aftereffect based on changing disparities? *Journal of Vision, 12*(4).

Dayan, P., & Abbott, L. F. (2001). Theoretical neuroscience: Computational and mathematical modeling of neural systems. In *Computational neuroscience* (pp. 60–74). Cambridge, Mass: MIT Press.

De Valois, R. L., Albrecht, D. G., & Thorell, L. G. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research, 22*(5), 545–559.

DeAngelis, G. C., Ghose, G. M., Ohzawa, I., & Freeman, R. D. (1999). Functional micro-organization of primary visual cortex: Receptive field analysis of nearby neurons. *Journal of Neuroscience, 19*(10), 4046–4064.

DeAngelis, G. C., & Newsome, W. T. (1999). Organization of disparity-selective neurons in macaque area MT. *The Journal of Neuroscience, 19*(4), 1398–1415.

DeAngelis, G. C., Ohzawa, I., & Freeman, R. D. (1993). Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development. *Journal of Neurophysiology, 69*(4), 1091–1117.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*(1), 1–47.

Fernandez, J. M., & Farell, B. (2005). Seeing motion in depth using inter-ocular velocity differences. *Vision Research, 45*(21), 2786–2798.

Fleet, D. J., Wagner, H., & Heeger, D. J. (1996). Neural encoding of binocular disparity: Energy models, position shifts and phase shifts. *Vision Research, 36*(12), 1839–1857.

Foster, K. H., Gaska, J. P., Nagler, M., & Pollen, D. A. (1985). Spatial and temporal frequency selectivity of neurones in visual cortical areas V1 and V2 of the macaque monkey. *Journal of Physiology, 365*, 331–363.

Geisler, W. S. (2003). Ideal observer analysis. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (pp. 825–837). MIT Press.

Harris, J. M., & Dean, P. J. (2003). Accuracy and precision of binocular 3-D motion perception. *Journal of Experimental Psychology: Human Perception and Performance, 29*(5), 869–881.

Harris, J. M., McKee, S. P., & Watamaniuk, S. N. (1998). Visual search for motion-in-depth: Stereomotion does not 'pop out' from disparity noise. *Nature Neuroscience, 1*(2), 165–168.

Harris, J. M., Nefs, H. T., & Grafton, C. E. (2008). Binocular vision and motion-in-depth. *Spatial Vision, 21*(6), 531–547.

Harris, J. M., & Rushton, S. K. (2003). Poor visibility of motion in depth is due to early motion averaging. *Vision Research, 43*(4), 385–392.

Heeger, D. J. (1987). Model for the extraction of image flow. *Journal of the Optical Society of America A: Optics, Image Science, and Vision, 4*(8), 1455–1471.

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

Hubel, D. H., & Wiesel, T. N. (1963). Shape and arrangement of columns in cat's striate cortex. *The Journal of Physiology, 165*(3), 559–568.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology, 195*(1), 215.

Hyvärinen, A., & Hoyer, P. O. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research, 41*(18), 2413–2423.

Kara, P., & Boyd, J. D. (2009). A micro-architecture for binocular disparity and ocular dominance in visual cortex. *Nature, 458*(7238), 627–631.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE, 78*(9), 1464–1480.

Kohonen, T., Kaski, S., & Lappalainen, H. (1997). Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM. *Neural Computation, 9*(6), 1321–1344.

Lamberty, K., Gobbelé, R., Schoth, F., Buchner, H., & Waberski, T. D. (2008). The temporal pattern of motion in depth perception derived from ERPs in humans. *Neuroscience Letters, 439*(2), 198–202.

Likova, L. T., & Tyler, C. W. (2007). Stereomotion processing in the human occipital cortex. *Neuroimage, 38*(2), 293–305.

López-Rubio, E., Munoz-Pérez, J., & Gómez-Ruiz, J. A. (2004). A principal components analysis self-organizing map. *Neural Networks, 17*(2), 261–270.

Maldonado, P. E., Gödecke, I., Gray, C. M., & Bonhoeffer, T. (1997). Orientation selectivity in pinwheel centers in cat striate cortex. *Science, 276*(5318), 1551–1555.

Meng, Y., & Shi, B. E., 2009. Normalized phase shift motion energy neuron populations for image velocity estimation. In *International joint conference on neural networks*, Atlanta, Georgia.

Norcia, A. M., & Tyler, C. W. (1984). Temporal frequency limits for stereoscopic apparent motion processes. *Vision Research, 24*(5), 395–401.

Ohki, K., Chung, S., Kara, P., Hubener, M., Bonhoeffer, T., & Reid, R. C. (2006). Highly ordered arrangement of single neurons in orientation pinwheels. *Nature, 442*(7105), 925–928.

Ohzawa, I., DeAngelis, G. C., & Freeman, R. D. (1990). Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science, 249*(4972), 1037–1041.

Ohzawa, I., DeAngelis, G. C., & Freeman, R. D. (1996). Encoding of binocular disparity by simple cells in the cat's visual cortex. *Journal of Neurophysiology, 75*(5), 1779–1805.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*(6583), 607–609.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research, 37*(23), 3311–3325.

Peng, Q., & Shi, B. E. (2010). The changing disparity energy model. *Vision Research, 50*(2), 181–192.

Pettigrew, J. D., Nikara, T., & Bishop, P. O. (1968). Binocular interaction on single units in cat striate cortex: Simultaneous stimulation by single moving slit with receptive fields in correspondence. *Experimental Brain Research, 6*(4), 391–410.

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

Qian, N., & Andersen, R. A. (1997). A physiological model for motion-stereo integration and a unified explanation of Pulfrich-like phenomena. *Vision Research, 37*(12), 1683–1698.

Rashbass, C., & Westheimer, G. (1961a). Disjunctive eye movements. *Journal of Physiology, 159*, 339–360.

Rashbass, C., & Westheimer, G. (1961b). Independence of conjugate and disjunctive eye movements. *Journal of Physiology, 159*, 361–364.

Regan, D. (1993). Binocular correlates of the direction of motion in depth. *Vision Research, 33*(16), 2359–2360.

Regan, D., & Beverley, K. I. (1973). Some dynamic features of depth perception. *Vision Research, 13*(12), 2369–2379.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*, 1019–1025.

Rokers, B., Cormack, L. K., & Huk, A. C. (2008). Strong percepts of motion through depth without strong percepts of position in depth. *Journal of Vision, 8*(4), 1–10.

Rokers, B., Cormack, L. K., & Huk, A. C. (2009). Disparity- and velocity-based signals for three-dimensional motion perception in human MT+. *Nature Neuroscience, 12*(8), 1050–1055.

Rokers, B., Czuba, T. B., Cormack, L. K., & Huk, A. C. (2011). Motion processing with two eyes in three dimensions. *Journal of Vision, 11*(2).

Sabatini, S. P., Solari, F., Andreani, G., Bartolozzi, C., & Bisio, G. M. (2001). A hierarchical model of complex cells in visual cortex for the binocular perception of motion-in-depth. In *Proc. neural information processing systems* (pp. 1271–1278), Vancouver, British Columbia, Canada.

Sabatini, S. P., & Solari, F. (2004). Emergence of motion-in-depth selectivity in the visual cortex through linear combination of binocular energy complex cells with different ocular dominance. *Neurocomputing, 58–60*, 865–872.

Sabatini, S. P., Solari, F., Cavalleri, P., & Bisio, G. M. (2003). Phase-based binocular perception of motion in depth: Cortical-like operators and analog VLSI architectures. *EURASIP Journal on Applied Signal Processing, 2003*(1), 690–702.

Sakano, Y., Allison, R. S., & Howard, I. P. (2012). Motion aftereffect in depth based on binocular information. *Journal of Vision, 12*(1).

Shioiri, S., Kakehi, D., Tashiro, T., & Yaguchi, H. (2009). Integration of monocular motion signals and the analysis of interocular velocity differences for the perception of motion-in-depth. *Journal of Vision, 9*(13), 10, 11–17.

Shioiri, S., Kakehi, D., Tashiro, T., & Yaguchi, H. (2003). Investigating perception of motion in depth using monocular motion aftereffect. *Journal of Vision, 3*(9). 856–856.

Shioiri, S., Nakajima, T., Kakehi, D., & Yaguchi, H. (2008). Differences in temporal frequency tuning between the two binocular mechanisms for seeing motion in depth. *Journal of Optical Society of America A—Optics Image Science and Vision, 25*(7), 1574–1585.

Shioiri, S., Saisho, H., & Yaguchi, H. (2000). Motion in depth based on inter-ocular velocity differences. *Vision Research, 40*(19), 2565–2572.

Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research, 38*(5), 743–761.

Tosic, I., & Olshausen, B. A. (2011). Hierarchical inference of disparity. *Nature Precedings*. http://dx.doi.org/10.1038/npre.2011.5824.1.

Tyler, C. W. (1971). Stereoscopic depth movement: Two eyes less sensitive than one. *Science, 174*(12), 958–961.

Watson, A. B., & Ahumada, A. J. Jr., (1985). Model of human visual-motion sensing. *Journal of the Optical Society of America A: Optics, Image Science, and Vision, 2*(2), 322–341.

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