Systems/Circuits

Optimal Estimation of Local Motion-in-Depth with Naturalistic Stimuli

Daniel Herrera-Esposito[1] and Johannes Burge[1,2,3]

[1]Department of Psychology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, [2]Neuroscience Graduate Group, University of Pennsylvania, Philadelphia, Pennsylvania 19104, and [3]Bioengineering Graduate Group, University of Pennsylvania, Philadelphia, Pennsylvania 19104

Estimating the motion of objects in depth is important for behavior and is strongly supported by binocular visual cues. To understand both how the brain should estimate motion in depth and how natural constraints shape and limit performance in two local 3D motion tasks, we develop image-computable ideal observers from a large number of binocular video clips created from a dataset of natural images. The observers spatiotemporally filter the videos and nonlinearly decode 3D motion from the filter responses. The optimal filters and decoder are dictated by the task-relevant image statistics and are specific to each task. Multiple findings emerge. First, two distinct filter subpopulations are spontaneously learned for each task. For 3D speed estimation, filters emerge for processing either changing disparities over time or interocular velocity differences, cues that are used by humans. For 3D direction estimation, filters emerge for discriminating either left–right or toward–away motion. Second, the filter responses, conditioned on the latent variable, are well-described as jointly Gaussian, and the covariance of the filter responses carries the information about the task-relevant latent variable. Quadratic combination is thus necessary for optimal decoding, which can be implemented by biologically plausible neural computations. Finally, the ideal observer yields nonobvious—and in some cases counterintuitive—patterns of performance like those exhibited by humans. Important characteristics of human 3D motion processing and estimation may therefore result from optimal information processing in the early visual system.

Key words: **Bayesian perception; binocular vision; ideal observer; motion in depth; natural scene statistics**

Significance Statement

Humans and other animals extract and process features of natural images that are useful for estimating motion in depth, an ability that is crucial for successful interaction with the environment. However, the enormous diversity of natural visual inputs that are consistent with a given 3D motion—natural stimulus variability—presents a challenging computational problem. The neural populations that support the estimation of motion in depth are under active investigation. Here, we study how to optimally estimate local 3D motion with naturalistic stimulus variability. We show that the optimal computations are biologically plausible and that they reproduce sometimes counterintuitive performance patterns independently reported in the human psychophysical literature. Novel testable hypotheses for future neurophysiological and psychophysical research are discussed.

Introduction

Accurate estimation of motion in depth from image information is important for successful interaction with the environment. To estimate 3D motion, animals with binocular vision combine information extracted from the two two-dimensional (2D) images formed by the eyes (Fig. 1*A*). However, many questions remain about how the brain does, and should, estimate 3D motion from binocular images of natural scenes (Cormack et al., 2017; Rosenberg et al., 2023). Ideal observer analysis is a useful tool for providing normative accounts of how perceptual tasks should be solved and for establishing theoretical bounds on sensory-perceptual performance. Ideal observers must be defined with respect to a specific task, a stimulus ensemble, and a set of constraints. When the relevant biological constraints are accurately modeled, ideal observer analysis can aid experimentalists in developing principled hypotheses about real biological systems that can be tested empirically (Geisler, 1989; Burge, 2020). When the stimulus ensemble is representative—or captures important properties—of natural stimuli, ideal observer analysis can provide insight into the design principles that optimize performance of the task of interest in natural conditions.

Received March 12, 2024; revised Oct. 30, 2024; accepted Nov. 6, 2024.

Author contributions: D.H.-E. and J.B. designed research; D.H.-E. performed research; D.H.-E. analyzed data; D.H.-E. and J.B. wrote the paper.

This work was supported by the National Eye Institute and the Office of Behavioral and Social Sciences Research–National Institutes of Health Grant R01-EY028571 to J.B.

The authors declare no competing financial interests.

Correspondence should be addressed to Daniel Herrera-Esposito at dherresp@sas.upenn.edu or Johannes Burge at jburge@psych.upenn.edu.

https://doi.org/10.1523/JNEUROSCI.0490-24.2024

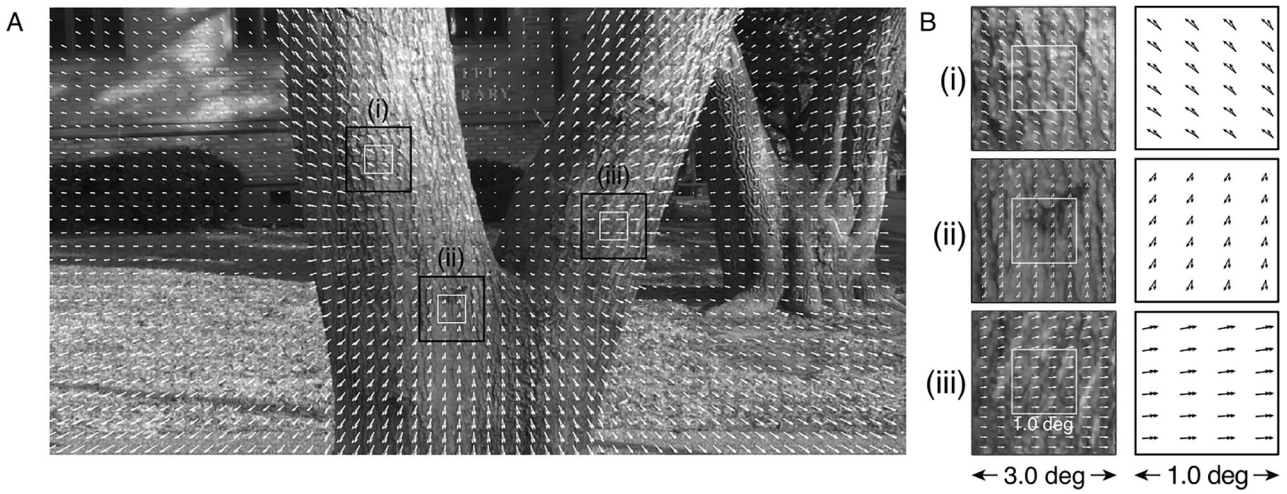

Figure 1. Binocular velocity fields for self-motion in depth in a natural scene. ***A***, Retinal velocities for the left and right eye produced by forward self-motion. For a grid of points in the scene, two arrows are shown, representing the local left- and right-eye retinal velocities. Instantaneous velocities were computed using projective geometry from coregistered range data at each pixel. ***B***, High-detail patches of 3° (left column) and 1° (right column). The 1° patches are the size of the local regions used to estimate motion in depth. Many 1° patches depicting the scene are dominated by near-constant velocity fields.

Previous studies have used Bayesian estimators to model 3D motion estimation (Lages, 2006; Welchman et al., 2008; Rokers et al., 2018). However, previous models are typically not image-computable (but see Peng and Shi, 2014; Wu et al., 2020) and do not account for nuisance image variability arising from the diversity of surfaces in natural scenes (Fig. 2*A*). As a consequence, these models must make assumptions about the information in—and the statistical properties of—the stimulus encodings. However, because these models do not specify what computations (e.g., receptive fields) extract the relevant information or verify the assumed statistical properties of the stimulus encodings, reasonable concerns can arise about the applicability of the models to real biological systems. Such concerns can be alleviated by developing image-computable ideal observer models that are grounded in the statistical properties of natural images. [However, image-computable ideal observers are not without their own assumptions. For ideal observers to aid in understanding how biological systems operate in natural conditions, the analyzed stimulus ensembles should capture important features of the natural visual diet and the modeled biological constraints should approximate those of the system under study (Burge, 2020).]

In this article, we develop image-computable Bayesian ideal observer models for the task of local 3D motion estimation. Although 3D motion estimation in complex scenes can be aided by global stimulus features and global computations, information extraction at a more local scale almost certainly occurs before more global computations (Hubel and Wiesel, 1962). Neurons that are relevant for 3D motion estimation with large receptive fields have been identified in areas MT, MST, and FST (Czuba et al., 2014; Sanada and DeAngelis, 2014; L. W. Thompson et al., 2023; Rosenberg et al., 2023), but these large receptive fields are formed by pooling over smaller, more local receptive fields from earlier areas like V1. Given the hierarchical organization of the visual cortex, to develop an understanding of how the selectivities of neurons in later areas emerge and of the computations that best serve 3D motion estimation, it is useful to determine the stimulus features and local computations that best support performance at the local scale of V1 receptive fields (Fig. 1*B*).

We develop two image-computable Bayesian ideal observers for the tasks of local 3D speed estimation or local 3D direction estimation. Both ideal observers are constrained by the front-end of the human visual system and by the statistics of naturalistic binocular video clips (Fig. 2*B*). We use the term "naturalistic" to indicate that the videos share many but not all statistical properties of natural visual stimulation (see below). We use projective geometry to generate binocular video datasets of naturally textured flat frontoparallel surfaces moving in depth. [Although these videos contain the statistical properties of natural images and the local 3D motion cues most studied in the literature, they do not contain the dynamic (dis)occlusions that occur at depth boundaries nor the motion parallax that results from local within-surface depth variability in natural scenes (see Discussion).] For each task, the respective ideal observer first applies a small set of task-optimized spatiotemporal filters to the binocular videos. The filters are learned via Accuracy Maximization Analysis (AMA), a Bayesian method for finding receptive fields that select for the stimulus features that provide the most useful information for a given task (Geisler et al., 2009; Burge and Jaini, 2017). The ideal observer performs optimal nonlinear decoding of the filter responses, using the filter response statistics, and probabilistic inference, to yield optimal estimates of the task-relevant latent variable (i.e., 3D speed or 3D direction; Fig. 2*A*,*B*).

The most important results are as follows. For each task, two filter types with distinct functional specializations are learned. The response statistics of the learned filters to videos with naturalistic variability dictate that optimal estimation of local 3D speed and 3D direction requires computations that can be well approximated by an extension of the energy model of cortical neural responses (Adelson and Bergen, 1985; Ohzawa et al., 1990; DeAngelis et al., 1991). Finally, for both 3D speed and direction estimation, the models provide a normative account of many previously reported aspects of human psychophysical performance.

Materials and Methods

3D motion estimation tasks. We developed ideal observer models for two different local 3D motion estimation tasks—3D speed estimation and 3D direction estimation—from datasets of naturalistic binocular videos. To learn the optimal filters for each task, a labeled set of input binocular videos was generated, where each video was consistent with a differently textured planar surface moving in 3D space in a given speed

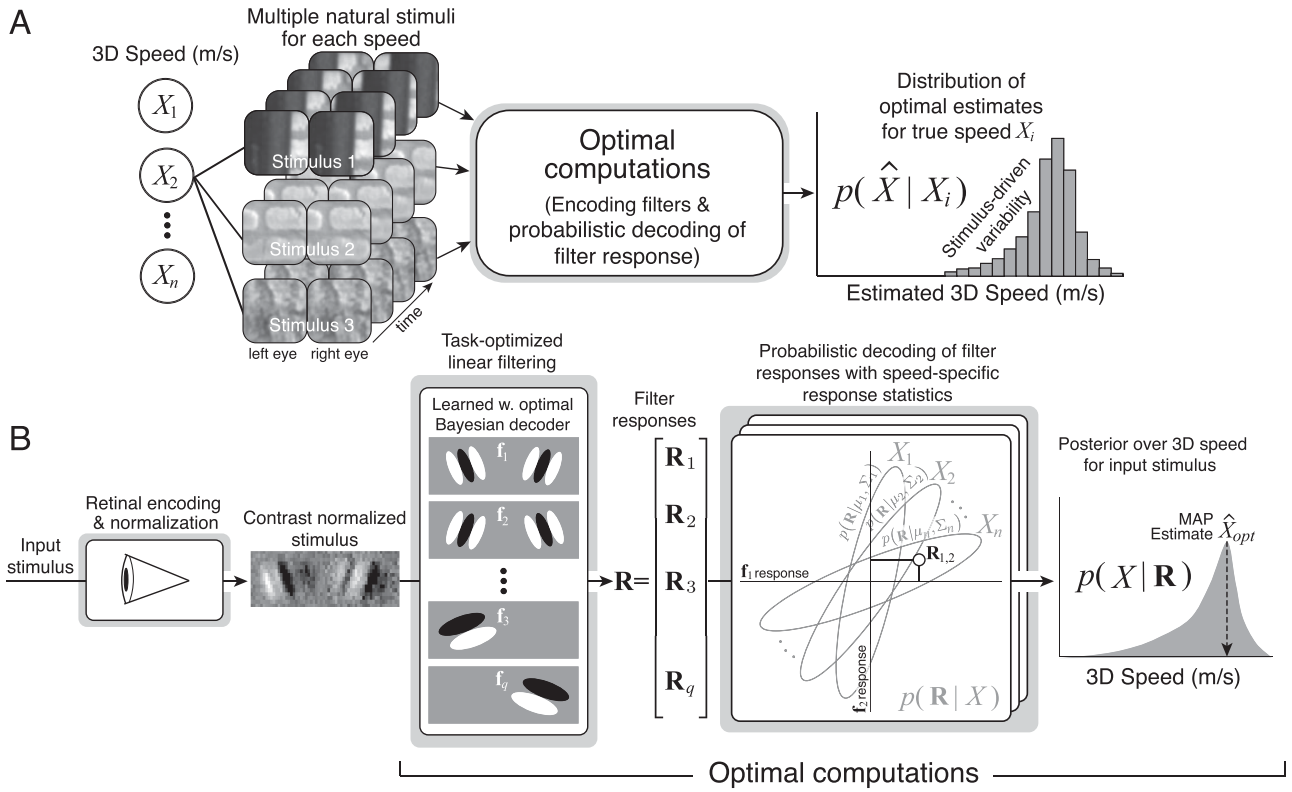

Figure 2. Ideal observer model for 3D speed estimation with naturalistic stimuli. ***A***, In natural viewing, many different retinal stimuli are associated with the same value of a latent variable (X). The naturalistic dataset used here mimics this natural ("nuisance") variability: many different binocular videos correspond to the same 3D motion trajectory in depth. For each binocular image video having a certain 3D motion (here, 3D speed), the ideal observer outputs an estimate ($\hat{X}_{opt}$) of that speed that is optimal for a given cost function. Across all binocular image videos having that same 3D speed (X_i), the corresponding distribution of optimal estimates indicates the effect of nuisance variability on performance. The variance of the estimates is largely stimulus-driven and attributable to naturalistic image variability. ***B***, Information flow through the ideal observer. First, an input stimulus is encoded by the retina and normalized to yield a contrast-normalized retinal stimulus. Then, task-optimized receptive fields are applied to the stimulus, yielding a set of filter responses (R). Next, each set of filter responses is decoded using the sufficient statistics (Σ_i) that characterize the conditional probability distributions of response for all speeds. Finally, the resulting posterior probability distribution over 3D speed is then used to obtain an optimal estimate.

and direction, behind a windowed circular aperture (Fig. 3*A*). The objective was to estimate, from the input videos, the 3D motion of the surface.

Binocular video synthesis. In the 3D speed estimation task, each video shows a planar surface moving with a given speed, from 1.0 m straight-ahead, directly toward or away from the observer (Fig. 3*B*, top). Three-dimensional speeds ranged from −2.5 m/s (receding) to 2.5 m/s (approaching) and were sampled in 0.1 m/s increments (51 total speeds). These 3D speeds correspond to monocular retinal speeds ranging from −4.6 to 4.6 °/s. In the 3D direction estimation task, each video shows a surface moving at a fixed 3D speed of 0.15 m/s, from 1.0 m straight-ahead, in a given direction in the XZ plane (Fig. 3*B* bottom). Three-dimensional directions in the XZ plane were sampled in 7.5° increments, for a total of 48 (i.e., 360/7.5) different directions.

Eight hundred naturalistic binocular videos were sampled for each 3D motion: 500 for training and 300 for testing. For each naturalistic video, we textured the planar surface with a unique image patch sampled from a natural image dataset (Burge et al., 2016) and then moved the surface in depth (Fig. 3*A*–*C*). The variation across videos at each 3D motion constitutes naturalistic "nuisance" stimulus variability in the dataset for the task. The left- and right-eye retinal images were computed frame-by-frame using projective geometry (Fig. 3*C*,*D*).

To do so, we consider a set of rays that are equally spaced in the visual angle emanating from the nodal point of each eye and that intersect the 3D surface. As the 3D surface moves in depth, the locations of the intersection points between the rays and the 3D surface change with time. The coordinates of these intersection points are (x_{T_L}, y_{T_L}, z_T) and (x_{T_R}, y_{T_R}, z_T) for the left and right eyes, respectively, and are straightforwardly determined using similar triangles with reference to a projection plane located at a fixed arbitrary distance from the observer (Fig. 3*C*). Specifically,

$$x_{\mathrm{T_L}} = (x_\mathrm{P} + I/2) \cdot (z_\mathrm{T}/z_\mathrm{P}) + I/2, \quad (1a)$$

$$x_{\mathrm{T_R}} = (x_\mathrm{P} - I/2) \cdot (z_\mathrm{T}/z_\mathrm{P}) - I/2, \quad (1b)$$

$$y_{\mathrm{T_L}} = y_\mathrm{P} \cdot (z_\mathrm{T}/z_\mathrm{P}), \quad (1c)$$

$$y_{\mathrm{T_R}} = y_\mathrm{P} \cdot (z_\mathrm{T}/z_\mathrm{P}), \quad (1d)$$

where x_P and y_P are x- and y-locations at which the set of rays intersects the projection plane, z_P is the distance of the projection plane from the observer (e.g., 1.0 m), z_T is the distance of the surface on a given time step, and I is the interocular distance (e.g., 60 mm). [Note that because the projection plane is always at a fixed distance, the coordinates (x_P, y_P, z_P) of the intersection points of the rays with the projection plane do not change with time.] When the intersection points with the target surface occur at interpixel texture locations, the surface texture intensities were linearly interpolated. This procedure was carried out for each eye with a set of rays that intersected the projection plane in a grid of 60×60 points equally spaced in the visual angle, such that they cover a square $1 \times 1°$ area.

This stimulus generation procedure creates geometrically correct monocular and binocular cues to 3D motion in the left- and right-eye

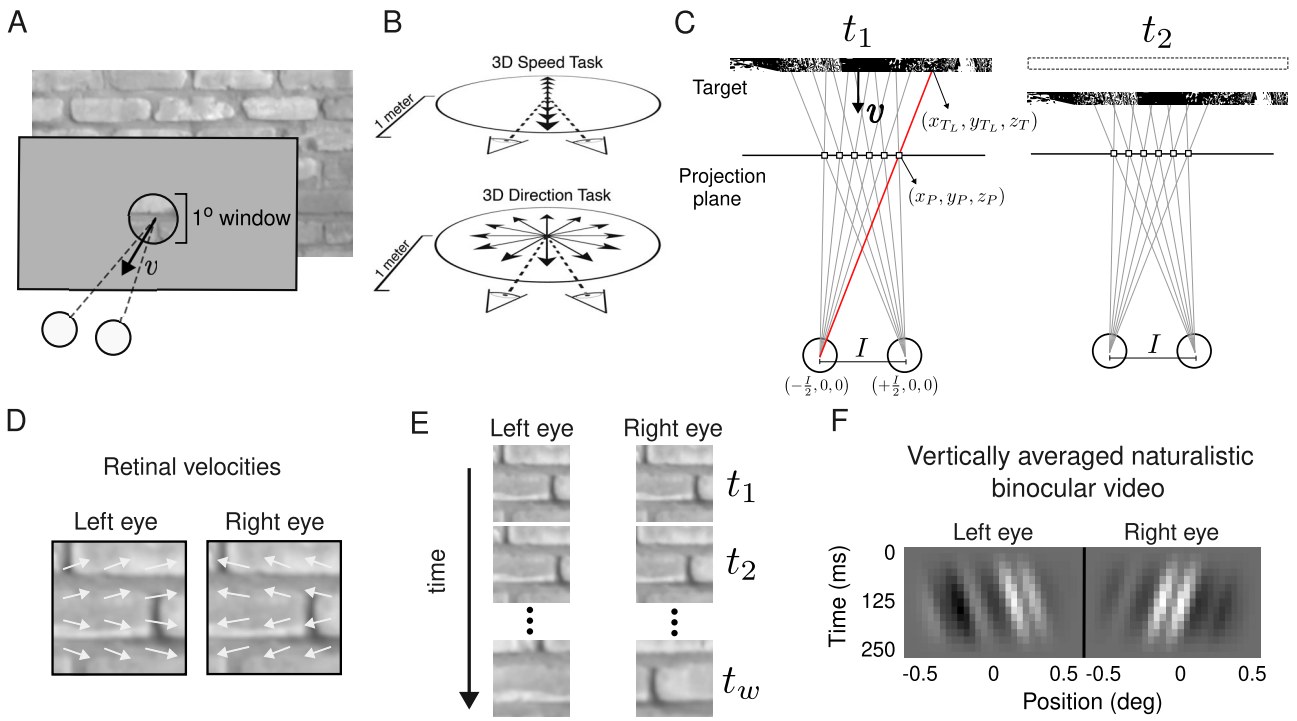

Figure 3. 3D motion tasks and stimuli. ***A***, A flat frontoparallel surface moves in 3D with motion vector *v* (black arrow). The observer views the surface through a fixed 1° window. (Note that the actual window had fuzzy cosine-attenuated, not sharp, edges.) ***B***, Motions through depth (black arrows) for the 3D speed estimation (top) and 3D direction estimation (bottom) tasks. ***C***, Projective geometry for moving binocular retinal videos (plan view). The target surface changes position from time-step t_1 to time-step t_2. Rays, equally spaced in visual angle, project from each eye and intersect with both an arbitrary projection plane (x_P, y_P, z_P; white squares) and the target surface (x_T, y_T, z_T). Note that the intersection points with the projection plane do not change with time and are the same for both eyes, whereas the intersection points with the target surface do change with time and are different for the two eyes. Example intersection points for an arbitrary ray (red) are shown. ***D***, Retinal velocity fields on a single frame of a movie. The binocular videos contain both monocular looming cues and stereo cues to motion in depth. However, in local regions, looming cues (i.e., the expansion or reduction of image features with time) are often more subtle than those depicted here (Fig. 1). ***E***, An example binocular video. Note the change in size of image features across time due to looming (top vs bottom). Also, with sufficient elapsed time, few if any corresponding points remain in the left- and right-eye images. ***F***, A vertically averaged naturalistic binocular video, moving toward the observer at 1.4 m/s. Each horizontal slice of the stimulus corresponds to a vertically averaged frame of a binocular video.

image movies. These cues include looming, changing disparities over time (CDOT), and interocular velocity differences (IOVD). However, the retinal videos in the dataset do not include nonuniformities in the retinal flow field—motion parallax and dynamic (dis)occlusions—that are due to local depth variation within and between surfaces in natural scenes. Nonetheless, there are good empirical reasons to expect that, at the local (i.e., 1°) scale of the analyses, the current results should be representative of the general case (see Discussion).

Binocular video preprocessing. To ensure that stimuli included constraints similar to those imposed by the front-end processing of the human visual system, we incorporated the effects of physiological optics, the cone photoreceptors, and response normalization. The effects of physiological optics were incorporated by convolving the frames of each video with a radially symmetric point-spread function approximating the optical blur introduced by the foveal optics of the human eye with a 4 mm pupil. The point-spread function was obtained by taking the inverse Fourier transform of the modulation transfer function (MTF). The human MTF was modeled as the weighted sum of two exponential functions as follows:

$$\mathrm{MTF}(f) = (1 - w) \cdot \exp(-k_1 f) + w \cdot \exp(-k_2 f),$$

where f is spatial frequency in cycles per degree, w is the mixing weight, and k_1 and k_2 are the shape parameters. The parameter values, which were fit to empirical measurements, were $w = 0.22$, $k_1 = 0.172$, and $k_2 = 0.037$ (Navarro et al., 1993).

The effects of photoreceptor physiology were incorporated by convolving the frames of the video with a temporal impulse response (TIR) function approximating that of the cone photoreceptors. The TIR was modeled using a sum of two gamma-shaped functions. Specifically,

$$\mathrm{TIR}(t) = w \cdot \mathrm{Gamma}(t - d_1; k_1, \tau_1) + (1 - w) \cdot \mathrm{Gamma}(t - d_2; k_2, \tau_2),$$

where $\mathrm{Gamma}(x; k, \tau)$ is the gamma distribution density function, k and τ are the shape and scale parameters, d is a response delay, and t is time in milliseconds. The parameter values, which were fit to physiological data, were $w = 0.353$, $d_1 = 2.2$, $k_1 = 4.16$, $\tau_1 = 4.79$, $d_2 = 1.9$, $k_2 = 1.64$, and $\tau_2 = 23.73$ (Schneeweis and Schnapf, 1995). The resulting TIR peaked at 16 ms with a full-width at half-height of 30 ms. Note that to prevent temporal aliasing, before the retinal videos were convolved with the TIR, retinal videos were computed at a high temporal framerate (i.e., 480 Hz).

The retinal videos, which had a duration of 250 ms, were then downsampled in space and in time such that the resulting (binocular) retinal videos had a spatial resolution of 30 pix/deg and a temporal resolution 60 Hz. (We have verified that, because the most useful information for the task resides in lower spatial frequencies, this downsampling has little practical effect on performance.) A raised cosine spatiotemporal window was then applied to the videos. Each frame of the video was then vertically averaged to produce a 1D binocular video (Fig. 3*F*). Vertical averaging the stimuli does not change the response of vertically oriented filters because such filters themselves vertically average the stimuli (Burge and Geisler, 2015). The ideal observers described in the article can thus be conceived of as operating on the outputs of receptive fields within vertically oriented orientation columns. It is known, however, that human observers use perspective cues to discriminate toward and

away motion (L. Thompson et al., 2019; Fulvio et al., 2020), the usefulness of which are attenuated by vertical averaging. A clear direction for future work is to generalize the analysis to videos that are not vertically averaged. Technical improvements in the computational efficiency of the procedures for receptive-field learning will facilitate these efforts (Burge and Jaini, 2017; Jaini and Burge, 2017). Such improvements are well underway. Finally, the binocular intensity videos were then converted to binocular contrast videos by subtracting and dividing by the mean intensity of each video. A given binocular video obtained with this process (Fig. 3*F*) is represented (for training or testing) as a vector with 900 elements (2 eyes × 30 pixels × 15 frames).

Additional stimulus sets. The main stimulus sets for 3D speed and 3D direction estimation were each supplemented with an additional stimulus set that included disparity variability. Disparity variability was used to simulate the small fixational errors that are associated with vergence jitter and the increased disparity variability in the peripheral visual field. Disparity was introduced across videos by changing the vergence posture of the eyes by an amount dictated by a random sample from a mean-zero normal distribution. (Note that this change in eye posture causes the rays from the two eyes to intersect the projection plane in different, as opposed to the same, locations; Fig. 3*C*.) The stimulus set for 3D speed estimation was also supplemented by cue-isolating datasets. The IOVD-isolating stimulus set contained stimuli in which CDOT cue had been eliminated, and the CDOT-isolating stimulus set contained stimuli in which the IOVD cue had been eliminated. In IOVD-isolating binocular videos, each eye had coherent motion signals consistent with the same 3D speed, but contrast patterns in the two eyes were mismatched and uncorrelated (Fig. 8*A*). IOVD-isolating stimuli were generated from the original dataset by pairing left- and right-eye videos from differently textured surfaces moving with the same 3D speed. In CDOT-isolating videos, each frame had a disparity consistent with a surface moving in 3D, but no coherent motion within each eye across frames. CDOT-isolating stimuli were generated by interlacing frames from different videos with the same 3D speed (Fig. 8*D*).

Ideal observer model. In this section, the 3D motion associated with each video is referred to as the latent variable X, which can take on values of $X_1, \ldots, X_n$. In the 3D speed estimation task, e.g., X is the 3D speed of the surface, and X_i is a particular value of 3D speed. As described above, for each motion X_i, the dataset contained a large set of binocular contrast videos $\mathbf{s}_{ij}$, where i indexes the true 3D motion and j indexes one of the 800 videos corresponding to motion i (Fig. 2*A*). To solve the task of estimating the true 3D motion depicted by stimulus $\mathbf{s}_{ij}$, we train an ideal observer model using AMA. The ideal observer model has three different stages: preprocessing, encoding, and decoding (Fig. 2*B*).

The preprocessing stage involves biologically realistic steps similar to those occurring in the retina and early visual cortex. First, we add a random sample of spatiotemporal white noise to each pixel to mask spatial detail that is undetectable by the human visual system. Then, contrast normalization is implemented as follows:

$$\mathbf{c} = \frac{\mathbf{s} + \gamma}{\|\mathbf{s} + \gamma\|}, \tag{2}$$

where $\mathbf{c}$ is the contrast-normalized stimulus and $\gamma \sim N(0, \sigma_p)$ is the noise sample vector, which is computationally equivalent to divisive normalization that is included in standard models of neural response (Albrecht and Geisler, 1991; Heeger, 1992; A. Iyer and Burge, 2019; Burg et al., 2021). The noise level ($\sigma_p = 0.008$) is set to be consistent with maximum human contrast sensitivity (Campbell and Robson, 1968). Normalization is performed separately for each eye's component of the binocular stimulus, consistent with recent neurophysiological reports, but the results presented here are largely robust to whether two monocular normalization factors (Mitchell et al., 2023) or a single binocular normalization factor (Hou et al., 2020) is used.

During the encoding stage, a small set of linear spatiotemporal filters is applied to the noisy normalized stimuli $\mathbf{c}$, and a sample of Gaussian response noise is added to the response of each filter to obtain a scalar noisy filter response. Each filter is constrained to have unit magnitude. The noisy response of each filter $\mathbf{f}_l$ is given as follows:

$$R_l = \mathbf{f}_l^T \mathbf{c} + \eta, \tag{3}$$

where $\eta \sim N(0, \sigma_r^2)$ is a sample of Gaussian response noise. The response noise ($\sigma_r = 0.1$) was set to be consistent with the dynamic range of real neurons in the cortex (Geisler and Albrecht, 1997), though results are robust to the specific value within a reasonable range (Burge and Jaini, 2017). Denoting the set of filters $\mathbf{f} = [\mathbf{f}_1, \ldots, \mathbf{f}_q]$ and the population filter response vector $\mathbf{R} = [R_1, \ldots, R_q] \in \mathbb{R}^q$, the noisy filter responses, conditional on a particular stimulus $p(\mathbf{R}|\mathbf{c}) \sim N(\mathbf{f}^T\mathbf{c}, \mathrm{I}\sigma_r^2)$, are normally distributed. The noisy responses of the linear filters can be conceptualized as the stimulus drives to simple cells in the early visual cortex.

In the decoding stage, the value of the latent variable X is inferred from the vector of filter responses $\mathbf{R}$ that a stimulus caused (Fig. 2*B*). Although neuroscience has provided detailed characterizations of how individual stimuli drive the responses of neurons, an individual stimulus-conditioned response distribution $p(\mathbf{R}|\mathbf{c})$ is not sufficient to decode the 3D motion. For this task, the 3D motion-conditioned response distributions $p(\mathbf{R}|X_i)$ are required, which can be obtained by marginalizing out individual binocular videos corresponding to a given value of the latent variable, $p(\mathbf{R}|X_i) = \sum_c p(\mathbf{R}|\mathbf{c})p(\mathbf{c}|X_i)$.

The posterior probability $p(X_i|\mathbf{R})$ of a latent-variable value X_i given a set of filter responses is related to the product of the likelihood $L(X_i; \mathbf{R}) = p(\mathbf{R}|X_i)$ and the prior via Bayes rule $p(X_i|\mathbf{R}) = p(\mathbf{R}|X_i)p(X_i)/p(\mathbf{R})$, where $p(X_i)$ is the prior of latent-variable value and $p(\mathbf{R}) = \sum_i p(\mathbf{R}|X_i)p(X_i)$ is the marginal probability of the response. We use a flat prior throughout the analysis.

When the conditional response distributions $p(\mathbf{R}|X_i) \sim N(0, \Sigma_i)$ are mean-zero Gaussian-distributed, as filter response distributions tend to be with appropriate normalization (Wainwright and Simoncelli, 1999; A. Iyer and Burge, 2019; Ni and Burge, 2024), the likelihood that the population response $\mathbf{R}$ was elicited by a stimulus having arbitrary latent-variable value X_i is computed by evaluating the response in the equation for the corresponding Gaussian. Specifically, the likelihood is given as follows:

$$L(X_i; \mathbf{R}) = \frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} \exp\left(\frac{-1}{2}\mathbf{R}^T\Sigma_i^{-1}\mathbf{R}\right), \tag{4}$$

where the conditional covariance Σ_i is estimated from the responses to all stimuli in the training set corresponding to latent-variable value X_i. Note that the term inside the exponential is a quadratic function of $\mathbf{R}$. When the 3D motion-conditioned responses are Gaussian-distributed, as they approximately are here, energy model-like computations are therefore required for optimal decoding (see Results). To obtain an optimal estimate $\hat{X}_{opt}$, the posterior probability for each value of the latent variable $p(X = X_i|\mathbf{R})$ should be obtained. From the posterior probability distribution $p(X|\mathbf{R})$, we obtain the maximum a posteriori (MAP) estimate of the value of the latent variable (i.e., the level of the latent variable with the highest posterior probability). In the case of a flat prior probability distribution, MAP estimates and maximum likelihood estimates are equal to one another.

Filter learning. The optimal filters (receptive fields) were learned using AMA-Gauss, a computationally efficient version of AMA (Jaini and Burge, 2017). AMA-Gauss learns filters given the (verifiable) assumption that, when conditioned on each value of the latent variable, the filter responses are Gaussian-distributed (see Results). Previous work on related tasks—binocular disparity estimation and 2D motion estimation—shows that the filters learned using the assumption that the conditional response distributions are Gaussian are near-identical to those learned without the assumption (Burge and Jaini, 2017; Jaini and Burge, 2017). The loss function was chosen to be the Kullback–Leibler divergence between the computed posterior probability distribution and an idealized posterior probability distribution with all its mass at

the correct value of the latent variable. In this case, the cost associated with each stimulus is given by the negative logarithm of the posterior probability at the correct value of the latent variable as follows:

$$C_{ij} = -\ln[\, p_{ij}(X = X_i|\mathbf{R})], \tag{5}$$

where $\mathbf{R}$ are the stochastic filter responses to the ijth stimulus, with latent-variable value of $X = X_i$. The total cost is the stimulus-specific cost averaged across all stimuli in the dataset as follows:

$$\bar{C} = \frac{1}{N}\sum_{ij} C_{ij}. \tag{6}$$

Stochastic gradient descent was used during optimization, with batches of 2,048 stimuli, out of a total of 25,500 and 24,000 stimuli for the 3D speed and 3D direction tasks, respectively. The step-size decreased by 20% every 10 epochs. One hundred epochs were used. The loss always plateaued. Given the assumption that the conditional response distributions are Gaussian, the sufficient response statistics—i.e., the mean and covariance μ_j and Σ_j of the filter responses—were recomputed after each update of the filters. The whole training set was used to update the statistics at each step.

Filters were learned in rank-ordered pairs. The first pair of filters (i.e., Filters 1 and 2) was learned first. This should be the most useful pair in isolation. Then, the second pair (i.e., Filters 3 and 4) was learned, while the first pair was held fixed. The second pair should be the most useful pair to be used in conjunction with the first pair. Then, the third pair (i.e., Filters 5 and 6) was learned while the first two pairs were held fixed and so on. The filters in the figures reflect this ordering. Ten filters were learned for each task. Learning additional filters returns copies of previously learned filters; when AMA returns copies of previously learned filters, it signifies that the reduction of internal noise is more beneficial to performance than the extraction of new stimulus features (Burge and Jaini, 2017). Performance also began to asymptote with >10 filters. We note, however, that although 10 filters extract most of the relevant information for our stimulus set, more filters will almost certainly be needed for stimulus sets that include all spatial orientations (i.e., no vertical averaging). To reduce the chance of local minima, we trained each pair of filters seven times with different starting seeds; the pair of filters with the lowest final loss was used, while the rest were discarded. This procedure resulted in highly consistent sets of filters for different starting conditions.

Interpolating response statistics. To obtain MAP estimates that were not limited by the spacing of the latent variable in the training set, we interpolated between the statistics of the latent-variable values in the training set. Specifically, we fitted a cubic spline to each element of the sample mean and sample covariance functions of the filter responses [$\mu(X)$ and $\Sigma(X)$, respectively]. Then, using the cubic spline, we interpolated the response statistics, for values of the latent variable that were not in the dataset, upsampling by a factor of 10. This procedure works because the response statistics change smoothly with X and because the values of X were finely spaced in the training dataset. We verified the accuracy of the interpolated results by comparing interpolated statistics to left-out empirical statistics. (Similar results are obtained by interpolating between the sampled values of the posterior probability distribution using a cubic spline.) This procedure ensured that the posterior probability distributions over the latent variable, and hence the MAP estimates, were not quantized by the latent-variable spacing in the training set. Results are qualitatively the same without interpolation, although some quantization effects are apparent.

Generality of findings. There are two important caveats to note before proceeding further. First, the binocular videos described above are compatible with many different 3D motions, depending on the starting location of the target (Longuet-Higgins, 1984; Lages and Heron, 2010). Information about 3D location must be available to obtain unambiguous estimates of 3D motion from such binocular videos. Just as solutions to the stereo-correspondence problem do not address how 3D location is estimated, the current computations do not address how 3D location is estimated; we assume such information is known to the observer. However, the resulting estimates of 3D motion from the ideal observer trained on videos from one distance (1 m away, straight-ahead), can be adapted for a different 3D location by appropriate geometrical transformation.

The second caveat is that, given the design of the current analyses, the 3D speed and 3D direction tasks both could, in principle, be performed by analyzing 2D speed from one eye only. This same issue has been faced by human psychophysical studies with closely related designs. In such studies, control experiments have examined whether humans can, when tasked to do so, reliably report 2D motion in one or the other eye. Such studies have come to the justified conclusion that human 3D motion estimation and discrimination performance is not accounted for by explicit reliance on 2D motion signals in each eye alone. Stereomotion suppression (Tyler, 1971; Brooks and Stone, 2006), a general difficulty with utrocular (i.e., the eye of origin) discrimination and identification (Ono and Barbeito, 1985), and various other controls contribute to this consensus (Harris and Watamaniuk, 1995; Rokers et al., 2008; Czuba et al., 2010). The fact that humans cannot explicitly report 2D motion of course does not mean that the image-computable ideal observers developed here do not make exclusive use of 2D motion signals in each eye, but there are at least two reasons to think otherwise and to instead conclude that binocular comparisons carry distinct information for the task. First, for each respective task, an optimization procedure that finds the most useful stimulus features returns binocular receptive fields. Second, binocular comparisons of monocular filter responses clearly carry information over and above that provided by the responses of each monocular filter in isolation. Hence, binocular comparisons substantively underlie ideal observer performance in both the 3D speed and 3D direction tasks. Nevertheless, the development of expanded observer models that jointly estimate 3D speed and 3D direction is important future work, as they will enable yet stronger conclusions to be drawn regarding the binocular computations that should underlie 3D motion estimation (see Discussion for more).

Code. Stimuli were generated using a database of natural images (Burge et al., 2016) and custom-written software in MATLAB. Filter-learning routines and ideal observer analyses were implemented in Pytorch (https://github.com/dherrera1911/3D_motion_ideal_observer).

Results

We developed image-computable ideal observers for two different local 3D motion estimation tasks with naturalistic stimuli: 3D speed estimation (Fig. 3*B*, top) and 3D direction estimation (Fig. 3*B*, bottom). The training and testing datasets of naturalistic stimuli were obtained by applying the laws of projective geometry to approximate the retinal stimulation that would be caused by flat frontoparallel surfaces, textured with natural scenes, moving in depth relative to the observer (Fig. 3*C*,*D*). Although dynamic (dis)occlusions and other effects of inter- and intrasurface depth variability are not in the stimulus set, there are good empirical reasons to expect that, at the local scale of the analysis, the results will be representative of the general case (see Discussion). The front-end processing of the stimuli is matched to that of the human visual system. Specifically, the information available for subsequent processing is constrained by physiological optics (Wyszecki et al., 1968; Thibos et al., 1992; Navarro et al., 1993), the photopic (L + M) sensitivity and temporal dynamics of the foveal cone photoreceptors (Schneeweis and Schnapf, 1995; Stockman and Sharpe, 2000), and luminance contrast normalization (Albrecht and Geisler, 1991; Heeger, 1992; A. Iyer and Burge, 2019; Burg et al., 2021; Ni and Burge, 2024; see Materials and Methods for details). The ideal observers are defined by a set of task-optimized encoding receptive fields and the computations that implement probabilistic decoding of the

receptive-field responses. The ideal observer outputs optimal estimates of local 3D motion. The decoding computations are dictated by the response statistics, which are themselves dictated by the task-relevant properties of the stimulus ensemble. We analyze the computations that support observer performance with naturalistic stimuli and compare ideal observer performance to previously reported patterns of human performance. The current analysis takes important steps toward a better understanding of the neural computations that support optimal 3D speed and 3D direction estimation in natural scenes and of the patterns that characterize human estimation and discrimination of 3D motion.

3D speed estimation

We developed an ideal observer for the estimation of local 3D speed from naturalistic binocular videos. The task-optimized filters, the filter responses, and ideal observer performance are shown in Figure 4. The filters are Gabor-like and localized in spatiotemporal frequency, like typical receptive fields in the early visual cortex (Fig. 4*A*). The filter responses, conditional on different 3D speeds, are shown in Figure 4, *B* and *C*. Responses to stimuli of the same 3D speed are approximately Gaussian-distributed. The 3D speed estimates, which are decoded by the ideal observer from the filter responses, are accurate and precise (Fig. 4*D*). Also, the estimation error increases systematically with speed (Fig. 4*E*), similar to how human 2D speed discrimination thresholds increase as a function of speed (McKee et al., 1986; Chin and Burge, 2020).

The filters that optimize 3D speed estimation performance select for interesting stimulus properties. Two distinct filter types were spontaneously learned for the task (Fig. 4*A*). The first four filters are monocular; they assign strong weights to one eye and weights near zero to the other eye (Fig. 4*A*, top). The fact that monocular filters are learned may be surprising given that the inputs to all filters were binocular videos. (The emergence of monocular filters is also nontrivial. Performing PCA on the dataset, e.g., does not return monocular filters.) The remaining six filters are binocular; they weight information from each of the two eyes approximately equally, with left- and right-eye components that select for the same spatiotemporal frequencies (Fig. 4*A*, bottom). The variation in ocular dominance across the filters may well have functional importance for this task (see below). Variation in ocular dominance is a well known property of cortical neurons (Hubel and Wiesel, 1962; Katz and Crowley, 2002). Also, the binocular receptive fields have the interesting property that the position of each eye's preferred feature changes smoothly in opposite directions across time, entailing that the preferred binocular disparity changes across time. This property is the signature of receptive fields that jointly encode motion and disparity. The presence of such receptive fields in the early visual cortex and their potential functional importance have been written about extensively (see Discussion).

To understand why two filter types are learned, we examined how each type supports 3D speed estimation. We examined estimation performance for versions of the ideal observer that used only the monocular filters, or only the binocular filters, and

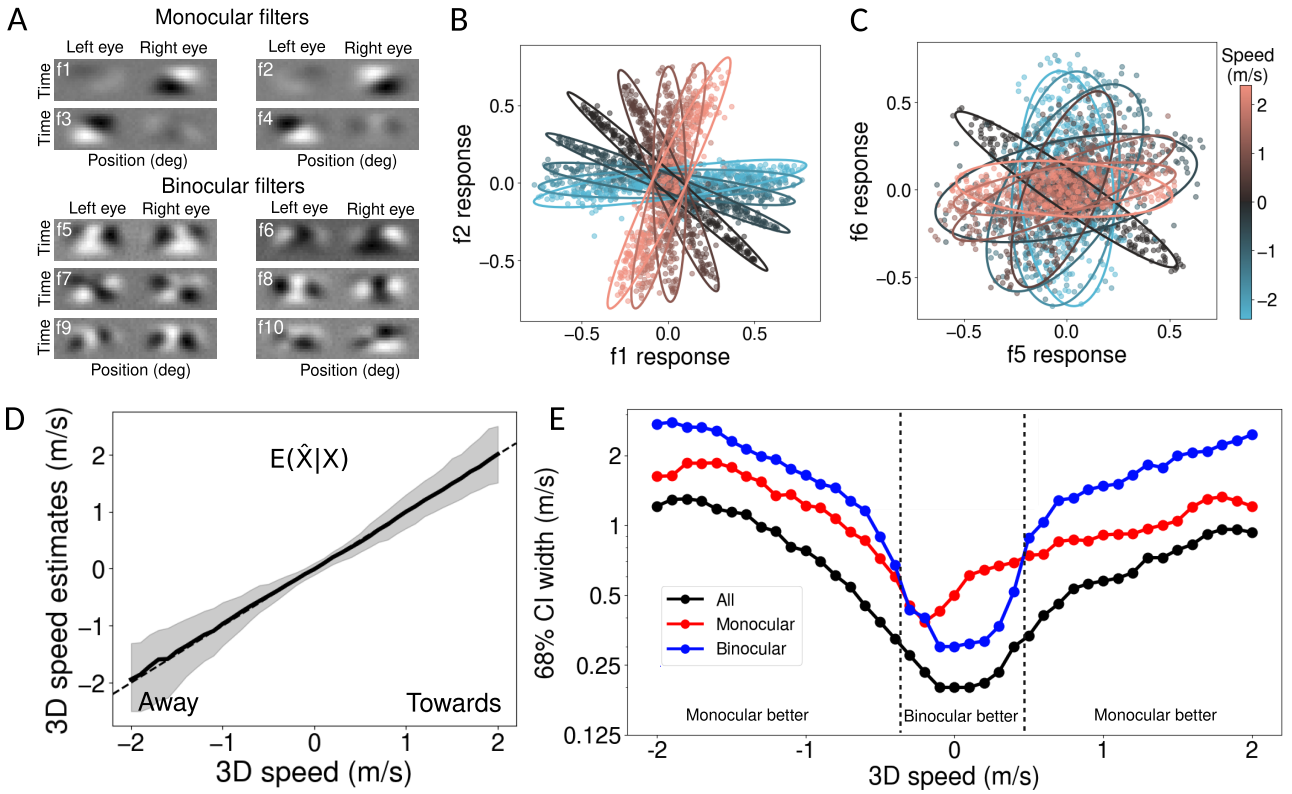

Figure 4. Task-optimized filters, responses, and 3D speed estimation performance. ***A***, Optimal binocular spatiotemporal filters for 3D speed estimation. Two specialized filter subpopulations spontaneously emerged. Monocular filters (top) support the extraction of IOVD cues (see the main text). Binocular filters (bottom) support the extraction of CDOT cues (also see the main text). ***B***, Conditional filter responses of the first pair of monocular filters in ***A*** for a subset of the 3D speeds in the dataset (colors). Each point is the expected filter response to an individual stimulus (i.e., without added filter noise). Ellipses show the best-fitting Gaussians. ***C***, Same as ***B*** but for a pair of binocular filters. ***D***, Median estimated 3D speeds as a function of true 3D speed. The gray area shows 68% confidence intervals of the response distribution. ***E***, Confidence intervals on estimates derived from monocular filters alone (red), binocular filters alone (blue), or all filters together (black). Dashed lines indicate speed ranges where monocular filters and binocular filters are better.

compared the precision of the estimates obtained with each model across speeds. Each filter type has a domain of specialization. At slow 3D speeds, estimates derived from the six binocular filters alone have smaller confidence intervals (i.e., are more precise) than the estimates derived from the four monocular filters alone. At fast speeds, the opposite is true (Fig. 4*E*). (Similar results are obtained when the number of filters are matched between the two groups, e.g., four binocular filters vs four monocular filters.) When all filters are used together, the confidence intervals are more similar to those of the monocular filters at fast speeds and are more similar to those of the binocular filters at slow speeds. Thus, the two filter types are engaged in a division of labor. The filters extract complementary task-relevant information from the stimuli.

Note that there is an asymmetry in the precision of 3D speed estimates for toward versus away motions (positive and negative speeds, respectively), which is most pronounced for estimates derived from monocular filters only. The asymmetries in precision are not present in a control stimulus set lacking looming cues, indicating that looming cues are responsible for the toward–away asymmetries. Both filter subpopulations on their own support largely accurate estimates that are weakly biased toward slower speeds, and a weak bias toward negative speeds for monocular filters. (The latter is eliminated when looming cues are eliminated.) These biases are small in magnitude compared with the estimate's variability (data not shown).

Interestingly, the two distinct filter types support extraction of two different stereo cues that are used by humans and nonhuman primates to estimate 3D motion: the IOVD cue and the CDOT cue (Cormack et al., 2017). The IOVD cue is the difference in the speeds of the retinal images of the moving object in the two eyes, produced by motion in depth. Computing this cue involves estimating the velocity of each retinal image first and then determining the velocity differences between the eyes. The monocular filters are well-suited to this computation. The CDOT cue is the change in binocular disparity over time as an object moves in depth. Computing this cue involves computing the binocular disparity at each time point and then determining how the disparity changes with time. The binocular filters are well-suited to this computation (see below, Cue-isolating stimuli link filter subpopulations to stereomotion cues). Note that, as 3D speeds increase, corresponding left- and right-eye image features move through the receptive fields of the binocular filters in ever-smaller fractions of the temporal integration period. This helps explain why the binocular filters—which support extraction of the CDOT cue—are less useful at fast 3D speeds than monocular filters. Binocular disparity, and how it changes over time (i.e., the CDOT cue), cannot be computed when corresponding image features are absent.

Psychophysical and neurophysiological investigations have provided strong evidence that IOVD and CDOT cues are both used by the visual system (Cormack et al., 2017). Although in natural circumstances, moving objects tend to produce IOVD and CDOT cues that are consistent with one another, each cue must be processed with different neural computations. The IOVD cue is computed by first taking time derivatives of the monocular images and then performing a binocular comparison. The CDOT cue is computed by first performing a binocular comparison—which depends on corresponding image features (see above)—and then taking a time derivative. Specialized IOVD- and CDOT-isolating stimuli have been used to show that the two cues underlie different 3D motion sensitivity profiles under different stimulus conditions. Humans rely on IOVD cues more heavily at high speeds and in the peripheral visual field and on CDOT cues more heavily at low speeds and in the fovea (Czuba et al., 2010; Cormack et al., 2017). The fact that the ideal observer is supported by distinct filter subpopulations in a manner that is consistent with human psychophysical results suggests that the adaptive manner in which humans use these motion cues across conditions may reflect near-optimal processing of the available visual information.

Likelihood neurons for local 3D speed estimation and the energy model

The filter responses, conditioned on each 3D speed, are well approximated by Gaussian distributions (Fig. 4*B*,*C*). The information about 3D speed in the filter responses is carried near-exclusively by their covariance, because the mean filter responses are very nearly equal to zero for all speeds. These results justify the use of AMA-Gauss, which approximates each conditional response distribution $p(\mathbf{R}|X_i)$ as Gaussian during the filter-learning process (Jaini and Burge, 2017). These results also indicate that quadratic combination of the filter responses is required for computing the likelihood of the different speeds $L(X_i; R)$ and thus for optimal inference of local 3D speed. So energy model-like computations (see below) are normative for 3D speed estimation from naturalistic signals. The log-likelihood that a particular 3D speed elicited the observed response is given by a weighted quadratic combination of the filter responses as follows:

$$\ln[L(X_i; \mathbf{R})] = \frac{-1}{2}\mathbf{R}^T\mathbf{Q}_i\mathbf{R} + K, \tag{7}$$

where $Q_i = \Sigma_i^{-1}$ is a fixed set of weights that is equal to the inverse covariance matrix and K is a constant that depends on the determinant of the covariance matrix (Eq. 4). Thus, the likelihood of a given 3D speed is given by a quadratic combination of the filter responses with a fixed set of weights. The inverse covariance matrices are, in turn, determined by the statistical properties of filter responses to naturalistic image movies. Quadratic combinations can be implemented with biologically realistic computations (Adelson and Bergen, 1985; Fleet et al., 1996; Burge and Geisler, 2014; Jaini and Burge, 2017), and optimal combination weights can be learned through Hebbian-like learning (Pagan et al., 2016).

From the above, we can conceive an energy-like neuron that computes the likelihood of a given 3D speed. Its tuning curve can be obtained by computing its mean response across many stimuli at each of many different 3D speeds. 3D speed tuning curves for a population of such neurons (where each neuron computes the likelihood of a different 3D speed) are approximately log-Gaussian in shape (Fig. 5*A*). Neurons with log-Gaussian–shaped tuning curves for 2D speed are widely observed in area MT (Nover et al., 2005). However, although the responses of MT neurons to stimuli having different 3D speeds toward and away an observer have been recorded (Sanada and DeAngelis, 2014), 3D speed tuning curves of neurons in cortical areas thought to be functionally involved in 3D speed estimation (V1, MT, MST, FST) have, to our knowledge, not been reported in the literature (Cormack et al., 2017; Rosenberg et al., 2023). Such reports would be of interest.

The quadratic computations underlying construction of each 3D speed-tuned "likelihood neuron" (Fig. 5*C*) and/or of the likelihood function over speed for a response elicited by a particular stimulus (Eq. 7) bear obvious similarities to the computations specified by the energy model, a popular descriptive model of

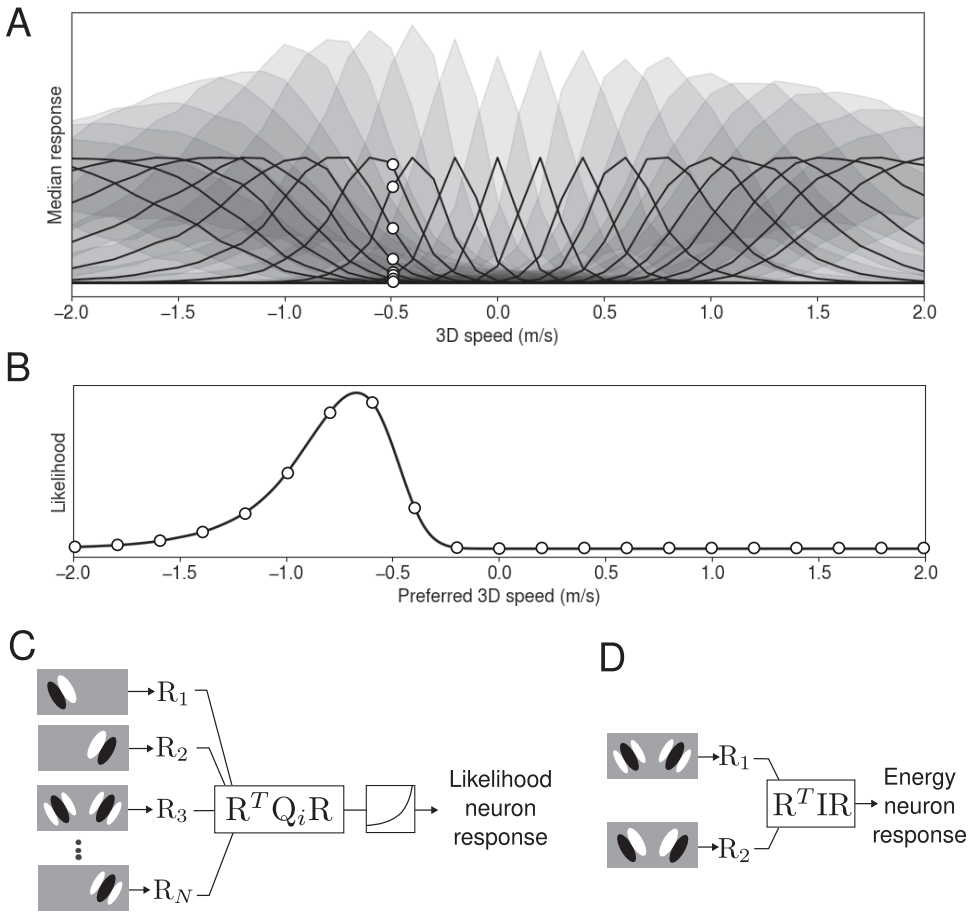

Figure 5. Likelihood-neuron 3D speed tuning curves and supporting computations. ***A***, 3D speed tuning curves for a set of (population-normalized) likelihood neurons. Each solid curve represents the tuning curve of a likelihood neuron as a function of 3D speed. Each tuning curve was obtained by computing the median response of each likelihood neuron to 300 binocular movies at each 3D speed. The max response of each neuron has been scaled to 1.0. White circles show the responses across likelihood neurons to a median stimulus having a particular 3D speed. Gray regions indicate ±1 standard deviation of the response due to naturalistic stimulus variation. ***B***, Population (normalized) response from a set of likelihood neurons to a single video. The black line shows the (normalized) response of likelihood neurons with different selectivities to a stimulus with speed of −0.5 m/s. The population response is the likelihood function of 3D speed for this stimulus. White circles show the preferred 3D speeds of the cells with tuning curves shown in ***A***. ***C***, Each likelihood neuron is constructed by a weighted quadratic combination of the responses of the optimal receptive fields (Fig. 2*B*), followed by a static exponential output nonlinearity. The weights $Q_i = \Sigma_i^{-1}$ are given by the inverse covariance of the optimal receptive-field responses to stimuli having the preferred speed of the likelihood neuron. The weights are thus dictated by the task-relevant naturalistic image statistics. ***D***, The energy model posits an unweighted sum of the squared responses of two receptive fields. This corresponds to a quadratic combination with the identity matrix *I*.

complex cells. The energy model computes the stimulus energy—the phase-invariant content—that is present in a spatial or spatiotemporal frequency band within a local stimulus region (Adelson and Bergen, 1985). This phase-invariance is achieved via quadratic combination: squaring and summing the outputs from a pair of orthogonal quadrature filters (Fig. 5*D*). Energy-like computations underlie models of neural activity recovered from responses to arbitrary contrast patterns (Rust et al., 2005; Park et al., 2013) and models of neural selectivity for a multitude of different latent variables, including disparity (DeAngelis et al., 1991; Ohzawa, 1998; Burge and Geisler, 2014), 2D motion (Adelson and Bergen, 1985; Burge and Geisler, 2015), and motion in depth (Peng and Shi, 2010, 2014; Wu et al., 2020). Such computations have also been shown to support the optimal estimation of these behaviorally relevant variables from naturalistic stimuli, including focus error (Burge and Geisler, 2011, 2012), binocular disparity (Burge and Geisler, 2014; Jaini and Burge, 2017), and 2D motion (Burge and Geisler, 2015; Chin and Burge, 2020). Here, from a task-specific analysis of naturalistic signals, we have shown that they support the optimal estimation of local 3D speed. The current results therefore provide a normative explanation, grounded in naturalistic scene statistics, for the success of a common descriptive model of neural response.

However, there are some notable differences between the classic version of the energy model, and the quadratic computations that support the performance of the ideal observer. First, unlike in the energy model, likelihood neurons combine the responses of far more than only two receptive fields. Complex cells in the cortex are now widely reported to be driven by far more than two subunit receptive fields (Rust et al., 2005; Tanabe et al., 2011; McFarland et al., 2013; Park et al., 2013). Such complex cells may be the neurophysiological analogs of likelihood neurons. Second, also unlike in the energy model—which combines subunit receptive-field responses in a straight quadratic sum—likelihood neurons use a weighted quadratic combination of receptive-field responses, where the weights are dictated by the response statistics to naturalistic stimuli (Eq. 7; Figs. 4*B*,*C*, 5*C*,*D*). Third, this statistics-based approach to neural computation automatically—and in a principled fashion—determines how information across different spatiotemporal frequency channels should be combined. Energy model-based frameworks have

previously used heuristics to combine information across frequencies (Read and Cumming, 2007). Finally, the filters in the ideal observer model are not constrained to be a quadrature pair. Because of how the stimulus-feature selectivity of the receptive-field population interacts with the effects of internal noise, filter pairs that span the same subspace as a quadrature pair but are nonorthogonal can, under certain biotypical circumstances, produce a higher quality encoding and support better latent-variable estimation performance than orthogonal filter pairs (Burge and Jaini, 2017). All these differences between the computations underlying likelihood neurons and those underlying the classic version of the energy model are consistent with available neurophysiological data. The demonstration that energy model-like neural computations are optimal for specific tasks with natural(istic) stimuli provides a normative explanation for why evidence of quadratic computations has been widely observed in neural systems: quadratic computations optimize task performance with the stimuli that visual systems evolved to process.

3D direction estimation

Next, we developed an ideal observer for local 3D direction estimation. The results for the 3D direction estimation task are similar in many ways to the 3D speed estimation results. The filters that optimize 3D direction estimation, their responses to naturalistic stimuli, and the ideal observer estimates of 3D direction are shown in Figure 6.

Some filters have almost identical patterns of weights in the two eyes (Fig. 6*A*, top). Others have distinct patterns of weights in the two eyes (Fig. 6*A*, bottom). These differences distinguish two filter subpopulations that, again, spontaneously emerge for the task (see below). The filter response distributions are reasonably approximated as mean-zero Gaussian—as they were for the 3D speed task—entailing that the task-relevant information is carried by the covariance of the filter responses (Fig. 6*B*,*C*; although some filter responses are not perfectly Gaussian (Fig. 6*B*, **inset**), the Gaussian approximation captures all task-relevant information up to the second order (i.e., the covariance of filter responses), and, because the Gaussian approximation is not unreasonable in most cases, the performance differences between quadratic decoding and optimal decoding should be minor. To verify, we simulated a new set of responses from Gaussian distributions with covariances equal to the filter response covariances. All main results were reproduced. Still, to extract all task-relevant information, more sophisticated decoding computations than the quadratic computations described here (Fig. 5*C*) would be required.) and that quadratic (energy model-like) computations are necessary for optimal decoding (see above).

Interestingly, unlike the filters that are optimized for 3D speed estimation (Fig. 4), no individual pair of filters clearly discriminates the full range of 3D directions. Rather, different subpopulations specialize in discriminating different cardinal 3D directions (left/right vs toward/away).

One subpopulation—the frontoparallel filter subpopulation—is populated by filters having the same weights in the two eyes and yields discriminable conditional response distributions for 3D directions with different frontoparallel components (Fig. 6*B*). For a fixed frontoparallel speed, however, these filters yield response distributions that perfectly overlap for different toward and away directions (Fig. 6*B*).

The other subpopulation—the toward–away filter subpopulation—is populated by filters having different weights in the two

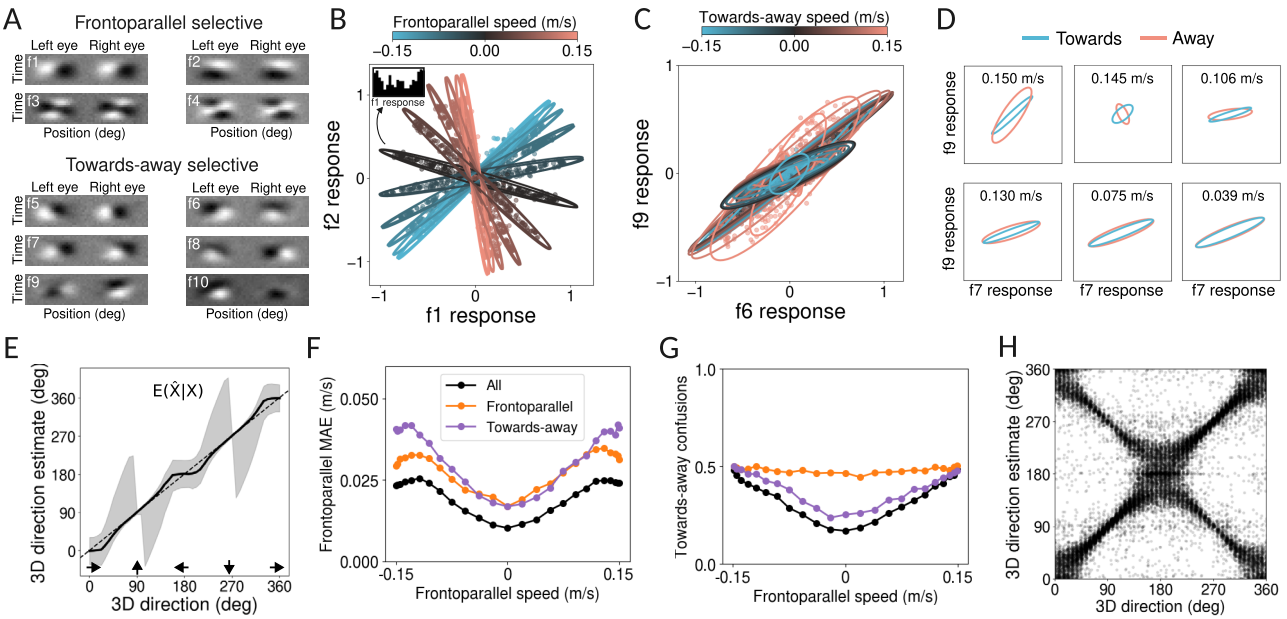

Figure 6. 3D direction estimation. ***A***, Filters learned for 3D direction estimation. Top, Frontoparallel selective. Bottom, Toward–away selective. ***B***, Response distributions for the pair of frontoparallel selective filters, conditional on 3D direction, and color coded by the frontoparallel motion component (only a subset of the directions). Ellipses show the best-fitting Gaussians. Filter responses without added noise are shown. Three-dimensional directions with identical frontoparallel components but opposite toward–away components produce perfectly overlapping response distributions for this pair of filters. Inset shows a response histogram for filter f_1, for the 90° direction. ***C***, Same as ***B***, but for toward–away filters, color coded by their toward–away motion component. ***D***, Same as ***C***, each panel shows only a direction pair that have the same frontoparallel component but opposite toward–away component. Panel headers indicate the magnitude of the toward–away speed for each pair. The color of the ellipse indicates whether the frontoparallel component is toward or away from the observer. ***E***, Estimated 3D directions. The gray area shows 68% confidence intervals. Three-dimensional directions, as seen from above, are indicated by the inset arrows. ***F***, Mean absolute error of the frontoparallel component of estimated motion as a function of true 3D motion direction (collapsed for directions with the same frontoparallel component). ***G***, Proportion of toward–away sign confusions obtained from the different types of filters as a function of true 3D motion direction. ***H***, Model estimates across the dataset. Each point shows the estimate from a unique stimulus.

eyes. The responses of these filters do not clearly segregate according to the frontoparallel component of 3D motion, but they do support discrimination of the toward and away motion component (Fig. 6*C*). In particular, when conditioned on the frontoparallel component, responses segregate according to the toward and away directions, and the segregation decreases with the magnitude of the toward–away component (Fig. 6*D*). We call the two filter subpopulations frontoparallel filters and toward–away filters.

When 3D direction is estimated only with the frontoparallel filter subpopulation, the frontoparallel component of the 3D direction estimate is largely accurate for all speeds. In contrast, when 3D direction is estimated only with the toward–away filter subpopulation, the frontoparallel component of the estimates become inaccurate for large frontoparallel speeds (Fig. 6*F*). Three-dimensional direction estimates that are derived exclusively from the frontoparallel filter subpopulation confuse the sign of toward–away motion ∼50% of the time (Fig. 6*G*, orange), whereas 3D direction estimates derived from the toward–away filter subpopulation confuse the sign a much smaller percentage of the time (Fig. 6*G*, purple). The two types of filters have clear functional specializations.

Although one may expect the toward–away filters learned for this task (Fig. 6*A*, bottom) to be similar to the 3D speed filters that discriminate different speeds in the sagittal plane (Fig. 4*A*), the two subpopulations are different. Notably, while the binocular 3D speed filters are selective for opposite motions in the two eyes, the toward–away 3D direction filters are not. This is not unexpected, considering the binocular geometry of 3D direction perception (see below) and the many differences between the two tasks (e.g., the range of toward–away speeds, the added frontoparallel components in the direction task). It is important to note, however, that much neurophysiological work probes neural selectivity for 3D motion by presenting opposite motion in the two eyes (Czuba et al., 2014; Sanada and DeAngelis, 2014; L. W. Thompson et al., 2023). The current computational results suggest that neurons involved in discriminating 3D motion may not necessarily prefer opposite monocular motion directions but rather subtly different monocular speeds in the same motion direction.

Ideal observer performance in the 3D direction estimation task is similar to key aspects of human psychophysical performance, as it is with the 3D speed estimation task. For a non-negligible proportion of the stimuli, the ideal observer confuses toward and away motion directions, yielding a characteristic X-shaped pattern of estimates (Fig. 6*H*). This pattern of responses is characteristic of human performance in laboratory-based 3D motion tasks (Fulvio et al., 2015; Rokers et al., 2018; Bonnen et al., 2020), indicating that the counterintuitive estimation errors may be a consequence of optimal decoding of 3D direction from stereo-based cues to 3D motion. Also, for some directions of motion, model estimates are biased toward frontoparallel directions (Fig. 6*E*,*H*; note the response cluster ∼180°). Frontoparallel bias is a known feature of human binocular 3D direction discrimination; the bias has been previously attributed to slow-speed priors (Welchman et al., 2008; Rokers et al., 2018). Here, however, we show that the same behavior is exhibited by an ideal observer that does not include a prior for slow speeds (also see Rideaux and Welchman, 2020). Understanding why biased estimation occurs in the absence of a slow-speed prior is an important topic for future work (also see below).

Likelihood neurons for local 3D direction estimation and the energy model

Tuning curves for 3D direction-selective likelihood neurons can be obtained using the same approach that was used to obtain tuning curves for 3D speed-selective likelihood neurons. Before discussing their properties, however, it is useful to consider how 3D viewing geometry, and viewing distance in particular, impacts the relationship between 3D motion and motion on the two retinas (Fig. 7*A*). Differently shaded regions of the plot correspond to 3D target directions that produce qualitatively different patterns of binocular retinal motion.

Targets moving along 3D trajectories that pass between the eyes (Fig. 7*A*, dark gray regions) create opposite retinal motion directions in the two eyes. A straight-ahead target moving toward the nose, e.g., produces rightward motion in the left-eye retinal image and leftward motion in the right-eye retinal image. However, at typical natural viewing distances (i.e., 0.3 m and beyond; Sprague et al., 2015), the vast majority of 3D directions do not pass between the eyes and therefore generate retinal motions with the same direction in the two eyes (i.e., left or right). At 0.3 m, e.g., the target must be moving within ±6° of straight toward or away from the observer, for an observer with a typical interocular eye separation. This range of directions shrinks as the target moves farther away (e.g., at 1 m, the range is ±2° from straight toward or away; Fig. 7).

Straight-ahead targets moving along 3D trajectories that do not pass between the eyes create retinal motions with the same directions in both eyes (Fig. 7*A*, light gray regions). So, for the vast majority of 3D directions, the frontoparallel component of the 3D target motion dominates the retinal motion such that the difference between the retinal motions in the two eyes will typically be much smaller than the shared retinal motion

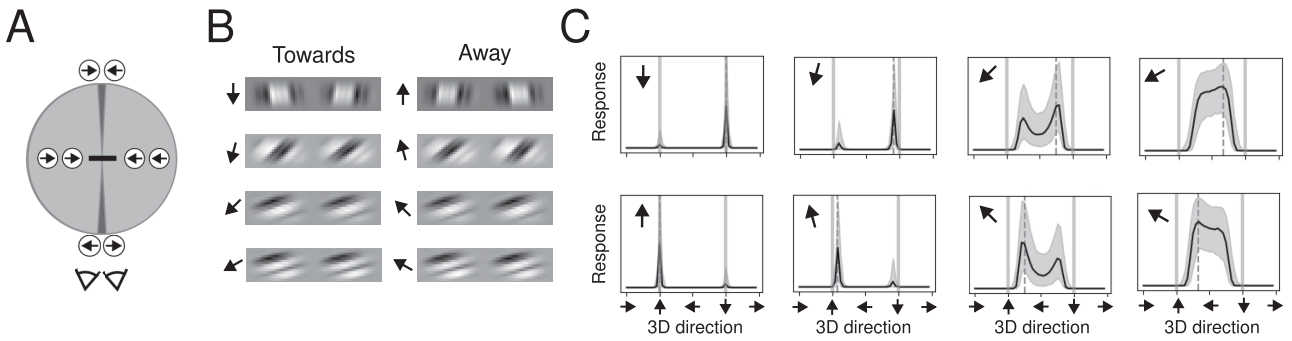

Figure 7. ***A***, Viewing geometry with a fixated target at a distance of 1.0 m, the simulated viewing distance throughout the article. ***B***, Example binocular retinal stimuli for 3D directions moving toward (left column) and away (right column) from the observer, with identical frontoparallel motion components but differently signed toward–away motion components (rows). ***C***, Likelihood-neuron tuning curves (1 m viewing distance). Inset arrows in each subplot indicate the preferred 3D direction of the likelihood neuron associated with each tuning curve.

(Fig. 7*B*). This property of binocular retinal videos under typical viewing conditions helps explain why the receptive fields for 3D direction estimation in our model tend to select for stimulus features that drift in the same direction in the two eyes with subtly different speeds in each eye (Fig. 6*A*). The same viewing geometry has been invoked in arguments about why there is a relatively low proportion of neurons in area MT that prefer opposite directions of motion in the two eyes (Czuba et al., 2014).

Likelihood-neuron 3D direction tuning curves are shown in Figure 7*C*. Many of these tuning curves are bimodal, with peaks that are associated with 3D directions having identical frontoparallel motion components but oppositely signed toward–away motion components. Despite the fact that the toward–away component of the 3D motion vector is large—many times, it is larger than the frontoparallel component—3D motions with opposite toward–away components often elicit similar likelihood-neuron responses because the retinal stimuli associated with these very different 3D directions are distinguished only by subtle differences in left- and right-eye retinal motion speeds (Fig. 7*B*). The tuning curves of other likelihood neurons—those that prefer 3D directions that are more nearly in the frontoparallel plane—are unimodal with distinctive plateaus (Fig. 7*C*, right column). The projective geometry entails that differences between the retinal motion speeds in the two eyes become even smaller as the true 3D motion direction approaches frontoparallel. This leads to a considerable range of directions close to frontoparallel that produce high activations of the likelihood neurons. However, unlike tuning curves for 3D speed (Fig. 5*A*), none of the tuning curves for 3D direction have the classic bell shapes that are depicted in textbooks.

The bimodal tuning curves help account for the toward–away confusions in the 3D direction estimates (Fig. 6*G*,*H*). Although the likelihood function—i.e., the population activity elicited by a single stimulus across the set of likelihood neurons (Fig. 5*B*)—most often peaks at the correct 3D direction, the double peaks in the tuning curves imply that the likelihood function will, for a non-negligible subset of stimuli, peak at a 3D direction with the correct frontoparallel component and an incorrect toward–away sign. Maximum likelihood estimators report, as the estimate, the latent-variable value that corresponds to the peak of the likelihood function. (Maximum likelihood and MAP estimators produce identical estimates when the prior over the latent variable is flat, as it is by design in our stimulus sets.) The likelihood functions which are dictated by the stimulus statistics therefore underlie the sign confusions in 3D direction estimation (Fig. 6*H*).

Cue-isolating stimuli link filter subpopulations to stereo-motion cues

Three-dimensional motion processing and perception have previously been probed in psychophysical experiments with stimuli that isolate the two primary binocular cues to motion in depth (Cormack et al., 2017). To test how each filter subpopulation for 3D speed estimation performs with such stimuli, we generated two additional stimulus sets: IOVD-isolating stimuli, in which disparity signals were removed (Fig. 8*A*), and CDOT-isolating stimuli, in which monocular velocity signals were removed (Fig. 8*D*; see Materials and Methods). We show that performance based on these filter subpopulations are differentially affected by the cue-isolating stimuli: monocular filters are better-suited for processing IOVD cues, and binocular filters are better-suited for processing CDOT cues.

With IOVD-isolating stimuli, binocular filter-based performance is dramatically reduced compared with the original stimuli (especially at slow speeds), whereas monocular filter-based performance is only slightly affected (Fig. 8*B*,*C*). CDOT-isolating stimuli (Fig. 8*D*) reduce the performance of both filter subpopulations compared with the original stimuli (Fig. 8*E*,*F*) but harm the performance of the monocular filters most. Therefore, monocular filters are better-suited than binocular filters to process IOVD cues, and binocular filters are better-suited than monocular filters to process CDOT cues. [Note that we report the ability of the model to correctly identify the toward–away direction of motion, similar to Czuba et al. (2010); analogous conclusions are reached when other measures of performance are used.]

Despite the fact that the monocular and binocular filter subpopulations have clear functional specializations, the IOVD- and CDOT-isolating stimulus sets do not perfectly isolate their functions. Although IOVD-isolating stimuli drastically affect binocular filter-based performance, as expected, they also harm monocular filter-based performance to some degree. A similar statement can be made about CDOT-isolating stimuli and the binocular filters (Peng and Shi, 2010, 2014). These findings raise the prospect that new stimulus sets could be constructed that maximally disassociate the activity of each subpopulation and the perceptual performance that they support. Doing so has the potential to strengthen conclusions drawn from neurophysiological and imaging studies that seek to determine whether neurons in a given area (e.g., MT) carry segregated signals from separate subcircuits versus signals that have been merged or combined (Joo et al., 2016; Cormack et al., 2017). In a cross-cue adaptation study, precombination circuits should not show adaptation transfer, whereas postcombination circuits should. But, when adaptation transfer is observed, the strength of the conclusions that can be drawn depends on the degree to which the stimulus sets actually isolate the subcircuits.

Disparity variability affects optimal 3D motion processing

We next analyzed the effect of adding an additional source of nuisance variability to the task: disparity variability. In all the results presented up until now, the first frame of each binocular retinal video started with zero disparity, which is tantamount to assuming the target surface was perfectly fixated on the first time step. However, small convergent or divergent eye movements cause fixation errors, which are known as vergence noise or vergence jitter. Under typical conditions, the standard deviation of vergence noise ranges between 2 arcmin and 10 arcmin and is known to harm stereo-based depth discrimination performance (Ukwade et al., 2003a,b).

We examined how 3D speed and direction estimation is affected by disparity variability within and exceeding this range (Fig. 9*A*). Note that, in all cases, the decoder—i.e., the computation of the likelihood—used response statistics that incorporated the simulated level of disparity variability.

For 3D speed estimation, disparity variability reduces estimation precision for low speeds (Fig. 9*B*) but not for fast speeds (Fig. 9*C*). The reduction in precision at low speeds occurs because disparity variability strongly affects the binocular filters. For obvious reasons, it does not affect the monocular filters. When disparity variability is sufficiently high, it eliminates the advantage of binocular filters at low speeds. For 3D direction estimation, disparity variability increases toward–away confusions (Fig. 9*D*), but it has only a small effect on the frontoparallel component of the direction estimates (Fig. 9*E*).

These effects are relevant for understanding two striking aspects of human psychophysical performance. First, the human ability to accurately report 3D direction from stereo-3D cues is substantially

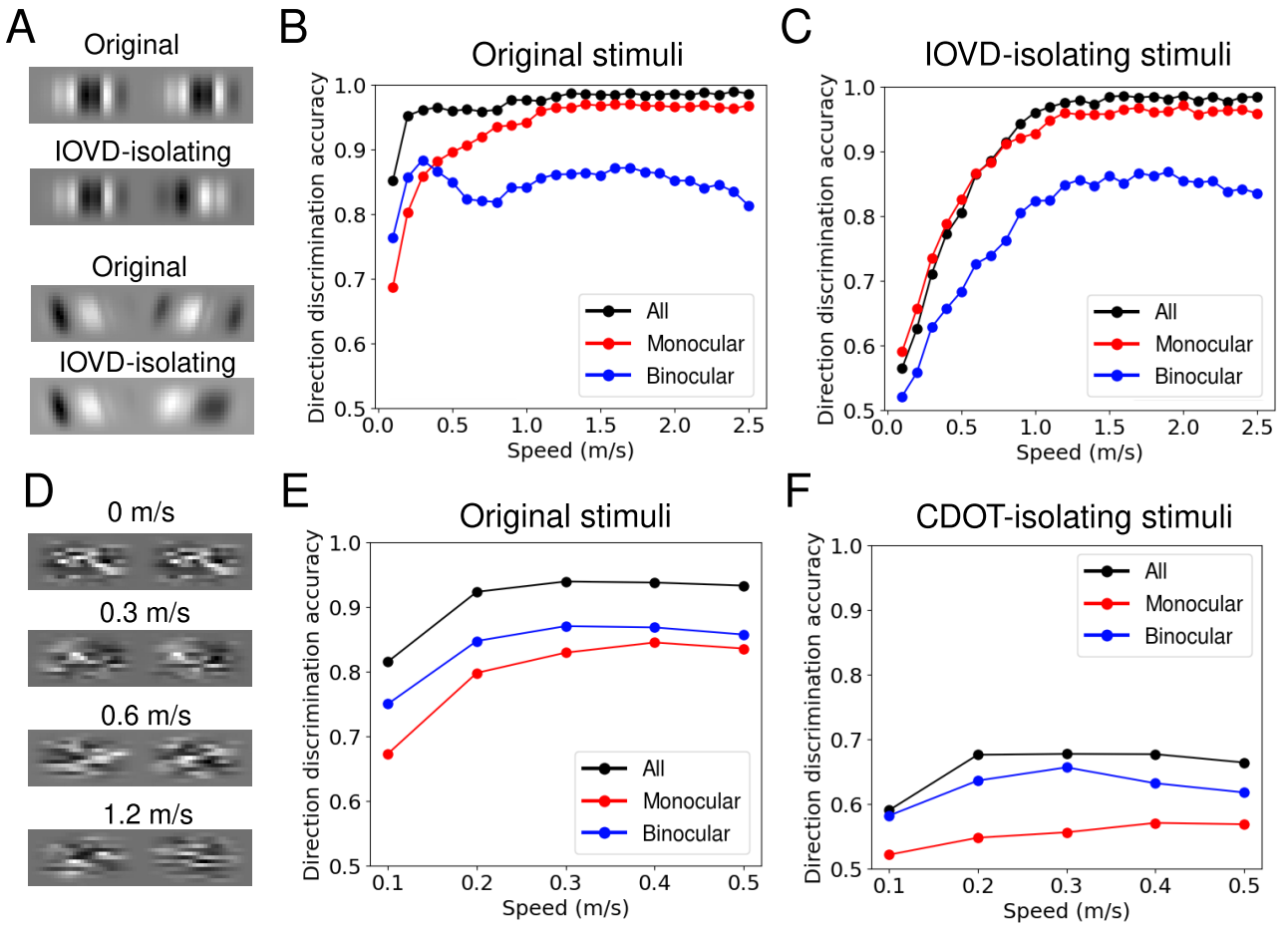

Figure 8. Three-dimensional speed estimation with cue-isolating stimuli. ***A***, Example stimuli from the original dataset and corresponding IOVD-isolating stimuli: one binocular video is stationary (0 m/s; top); another is moving toward the observer (0.3 m/s; bottom). ***B***, Accuracy of direction (i.e., toward vs away) discrimination in 3D speed estimation with original stimuli when different filter subpopulations are used (colors). ***C***, Direction accuracy with IOVD-isolating stimuli. ***D***, Examples CDOT-isolating stimuli across a range of slow speeds. ***E***, Accuracy of direction (i.e., toward vs away) estimation with slow-moving original stimuli. Only slow speeds were analyzed because, at fast 3D speeds, corresponding stimulus features quickly exit the spatial receptive field, and CDOT cues become unavailable. ***F***, Accuracy with CDOT-isolating stimuli.

affected by variability in fixation disparity: when disparity variability is present, human toward–away confusions in 3D motion tasks increase substantially (Fulvio et al., 2015). Second, the visual periphery relies more strongly on IOVD cues for toward versus away motion processing than the fovea does (Czuba et al., 2010; Cormack et al., 2017), and the periphery is also exposed to greater disparity variability than the fovea in natural viewing (Sprague et al., 2015). The results in Figure 9 therefore suggest that the periphery preferentially relies on IOVD cues in part because of higher disparity variability. Of course, there are other relevant differences between the fovea and the peripheral visual field—e.g., fast retinal motion speeds are more common in the periphery than at the fovea (Blakemore, 1970; Rogers and Bradshaw, 1993). So a dedicated modeling effort would have to be undertaken to understand their relative importance. Still, the results are intriguing and may help explain functional differences between foveal and peripheral vision.

Discussion

Despite the importance of 3D motion estimation and discrimination for behavior, relatively little is known about the computations that the brain uses or what computations are optimal given the constraints of the perceptual system, for estimating 3D motion from the retinal input. To gain insight into the normative computations that the brain should use, we developed image-computable ideal observers grounded in the task-relevant natural scene statistics for each of the two tasks: local 3D speed estimation and local 3D direction estimation. For each task, two subpopulations of filters spontaneously emerge from a Bayesian filter-learning routine, each with its own functional specialization. For local 3D speed estimation, each subpopulation is specialized for processing one of the two well characterized binocular cues to 3D motion (IOVD and CDOT cues), providing an normative account of the use of these cues by humans. For 3D direction estimation, one subpopulation strongly selects for image features supporting estimation of left–right 3D motion components; another provides information that is useful for discriminating toward–away 3D motion components. A generalization of the energy model description of complex cells closely approximates the computations underlying optimal inference with naturalistic signals for both the 3D speed and 3D direction estimation tasks, and the performance of the ideal observer in each task mirrors intriguing—and sometimes counterintuitive—patterns of human performance, suggesting that human performance patterns are accounted for by the optimal computations.

Limitations

The current analyses of 3D motion estimation are limited in some respects. Chiefly, the current results analyze the estimation

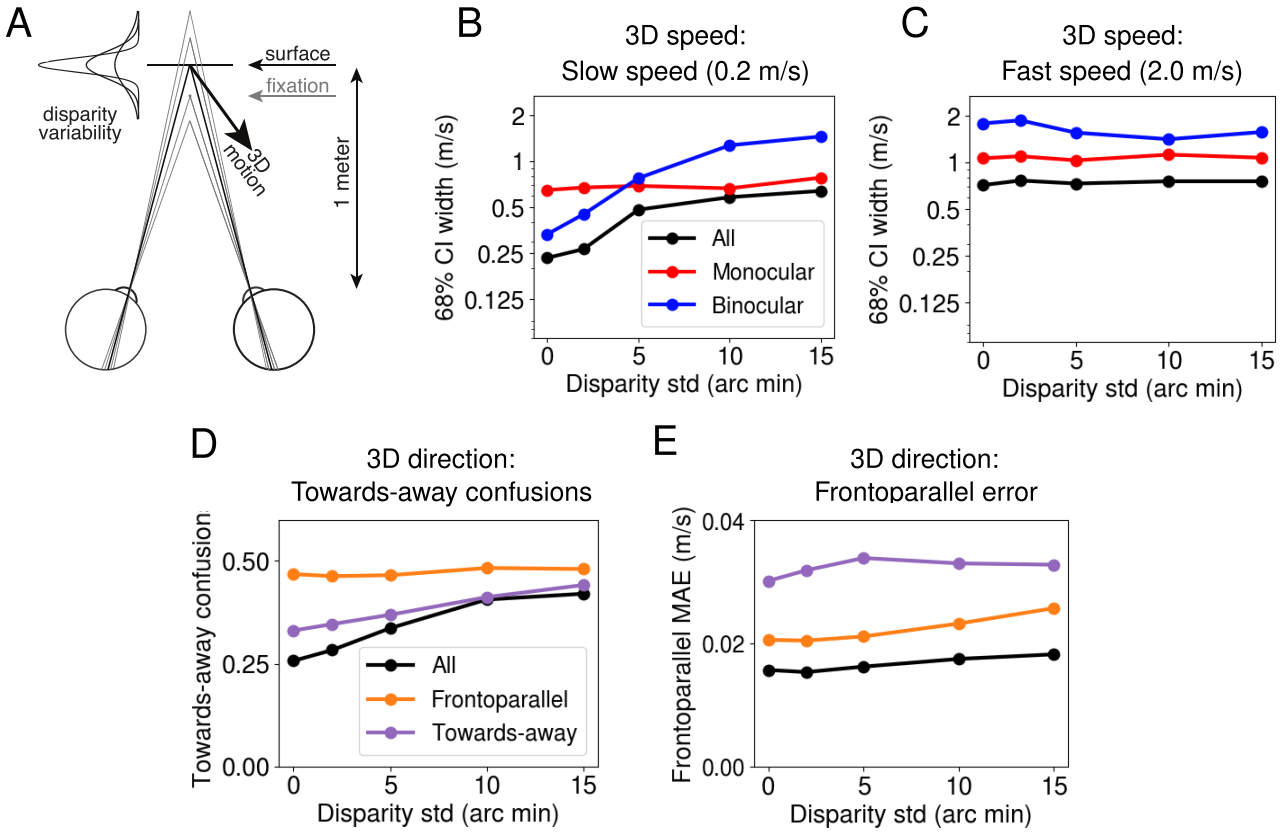

Figure 9. The effects of disparity variability (i.e., vergence noise) are condition-specific. ***A***, Disparity variability geometry. ***B***, Confidence intervals of the 3D speed estimates, as a function of disparity variability for a low target speed (0.2 m/s) for different filter subpopulations (colors). ***C***, Same as ***A*** for a fast target speed (2.0 m/s). ***D***, Proportion of toward–away sign confusions as a function of disparity variability, for an example target direction of 22.5° and speed of 0.15 m/s, for different filter subpopulations (colors) (similar results are obtained for other directions). ***E***, Mean absolute error of the estimated frontoparallel motion component as a function of disparity variability, for a target direction of 22.5°.

of motion in depth from local (1°) retinal image patches. Neurons in cortical areas that are most often implicated in 3D motion processing and representation—MT, MST, and FST—have receptive fields that are substantially larger than 1° (Cormack et al., 2017; Rosenberg et al., 2023). These large receptive fields result from computations that pool local monocular and binocular motion signals across space. The current work does not address how to pool information across multiple spatial locations. Rather, it focuses on how to combine information at individual spatial locations for local 3D motion estimation (Simoncelli, 2003; Burge, 2020). Cortical computations at this local spatial scale occur in Area V1. Neurons in Area V1 have receptive fields that are approximately the same size as the analyzed image patches (Priebe et al., 2006). If the available information for the task is made best-possible use of in V1, it will enable neurons in later areas—which process more global information—to maximize performance (see Joint encoding of motion and disparity). Understanding optimal extraction of local information should help identify neural performance patterns that signify involvement in the task and inform investigations of how neural selectivity changes along the cortical processing hierarchy (Hubel and Wiesel, 1962; L. W. Thompson et al., 2023).

Next, as already discussed, the binocular videos in the main stimulus set depict the motion of flat frontoparallel surfaces, textured with natural images. The stimulus set therefore does not contain the nonuniform motion signals associated with two sources of depth variation: local depth variation within surfaces (e.g., slant, tilt, and/or nonplanarity) and depth discontinuities across surface boundaries. It is an open empirical question whether these factors qualitatively change the ideal observer results. However, there are empirical reasons to expect the current results to be representative for the current local tasks. First, formal analyses of closely related tasks—binocular disparity estimation (Burge and Geisler, 2014) and 2D speed estimation (Burge and Geisler, 2015)—have shown that the task-optimized receptive fields, computations, and estimation performance are robust to nonuniformities that arise from (naturalistic) local depth variation within surfaces (Burge and Geisler, 2014; Extended Data Fig. 8-1). Second, patches including depth boundaries—which cause motion parallax and dynamic (dis)occlusions—are relatively rare (~15%; A. V. Iyer and Burge, 2018). Future work should address how the identification of depth boundaries can contribute to scene segregation and global estimates of the 3D motion structure across the entirety of a real-world scene (Fig. 1).

Another potential limitation is that 3D speed estimation and 3D direction estimation were analyzed as separate tasks. Because of the strong tradition of studying 3D speed and 3D direction estimation and discrimination separately, there is value in understanding the optimal solutions to the isolated tasks. But estimating 3D speed and direction simultaneously may require somewhat different computations than those described here. On the other hand, specialization of function is common in the visual brain, so the current results may be representative of the general solution. To adjudicate these possibilities, future work should examine 3D motion estimation as a single task in which both speed and direction can vary. The current results will

provide a useful benchmark against which to compare future findings.

Finally, the current efforts to model 3D motion estimation do not address how the 3D location of the target is estimated by the visual system. This is important because estimating how something is moving depends on determining from where it is moving, i.e., information about a target's 3D location (i.e., distance and direction) is necessary to obtain an accurate estimate of 3D motion from left- and right-eye retinal image motions (Lages and Heron, 2010). In cue-impoverished circumstances, in which 3D location information is poor, humans cannot accurately estimate 3D motion (Rushton and Duke, 2009). In typical conditions, however, location information can be extracted from a combination of image-based and extraretinal cues (Backus et al., 1999; Watt et al., 2005), and when 3D location is accurately estimated, stereo-3D motion estimation is limited by the accuracy with which binocular retinal motion is estimated. Just as the core computation underlying stereo-based depth estimation —solving the stereo-correspondence problem (Tyler and Julesz, 1978; Read and Cumming, 2007; Burge and Geisler, 2014)— does not depend on the estimation of target location (e.g., the location of one of two surfaces in depth), the core computations underlying local 3D motion estimation do not depend on it. Hence, just as solving the correspondence problem is a critical subtask for disparity-based depth estimation, the computations described solve a critical subtask for estimating motion-in-depth estimation (also see Materials and Methods).

Although the current work is not without limitations, it marks an important step forward, in showing how natural scene statistics dictate which local measurements visual systems should make—i.e., what stimulus features neuronal receptive fields should select for—if their function is to support the estimation of motion in depth in natural scenes.

Joint encoding of motion and disparity

The neurophysiological underpinnings of 3D motion perception are a topic of current interest (Rosenberg et al., 2023). The results here indicate that binocular receptive fields having disparity preferences that change continuously throughout the temporal integration period—i.e., that jointly encode motion and disparity—optimally support the estimation of 3D speed from local retinal image information (Fig. 4*A*, bottom), especially at slow speeds (Fig. 4*E*). Computational models based on V1 complex cells with such subunit receptive fields have been proposed to account for neural activity underlying aspects of 3D motion estimation performance (Qian and Andersen, 1997; Anzai et al., 1999; Pack et al., 2003). However, other modeling efforts show that such neurons are not required to account for the data (Read and Cumming, 2005a), and targeted neurophysiological investigations have produced scant evidence that such neurons exist in V1 (Read and Cumming, 2005b).

Interestingly, units with opposite direction preferences in the two eyes have been reported in Area MT (Zeki, 1974; Albright, 1984; Czuba et al., 2014; Sanada and DeAngelis, 2014). What is unclear is how the motion selectivity of these MT units emerges or how such selectivity is—or should be—related to neural selectivity for, and perceptual estimation of, motion in depth (Sanada and DeAngelis, 2014). It could be inherited directly from binocular V1 units having space-time receptive fields with opposite direction preferences in the two eyes, as the current analyses suggest would be optimal (Fig. 4*A*, bottom), although, as noted, Read and Cumming (2005b) searched specifically for such V1 units and found no evidence that they exist. Another possibility is that the selectivity emerges from pooling outputs from two binocular units with the same direction and spatiotemporal frequency preferences in the two (i.e., left- and right-eye) components of each unit but opposite direction preferences across the two units. Whatever the process by which the selectivity emerges, because the ideal detectors appear not to be present in the cortex, sensory-perceptual deficits are likely to be involved (see below).

There are loosely analogous results in stereo-surface-orientation perception. There is no strong evidence that, in the early visual system, binocular receptive fields exist that are optimized for extracting image information about 3D surface orientations. One can show that receptive fields that are optimized for surfaces slanted around a vertical axis have different spatial frequency preferences in the two eyes (Banks et al., 2004; Oluk et al., 2022) and that for surfaces slanted around a horizontal axis, optimal receptive fields have different orientation preferences in the two eyes (Greenwald and Knill, 2009). However, despite careful efforts to uncover evidence for such receptive fields, using both psychophysical (Banks et al., 2004; Vlaskamp et al., 2009; Oluk et al., 2022) and neurophysiological methods (Bridge et al., 2001; Bridge and Cumming, 2001; Nienborg et al., 2004), no evidence has been accrued. Of course, whether or not such binocular receptive fields exist in the early visual cortex is not determinative of whether humans can estimate 3D surface orientation. Humans can clearly do so (Stevens, 1983; Knill, 1998; Hillis et al., 2004; Watt et al., 2005; Kim and Burge, 2018, 2020), and there are cortical areas that contain neurons that select for 3D surface orientation (i.e., slant and tilt), including areas MT, V3A, and CIP (Welchman et al., 2005; Rosenberg et al., 2013; Rosenberg and Angelaki, 2014; Elmore et al., 2019). However, the lack of optimal early detectors entails that the visual system will be limited by the information loss at the early (nonoptimal) detectors (see Banks et al., 2004 for a nice discussion of this issue).

What is the utility of this discussion for an image-computable ideal observer analysis of local 3D motion estimation? The apparent absence of neurons with joint motion-disparity selectivity in V1 suggests that humans may be subject to deficits in 3D motion perception that the ideal observer is not, as is the case with stereo-based 3D surface orientation perception. Targeted experiments could be developed to test whether those deficits in 3D motion perception in fact occur. Such experiments should be conducted at slow 3D speeds, because it is at these speeds that the deficits imposed by an apparent lack of optimal receptive fields for the task place the most dramatic limits on performance (Fig. 4*A*,*E*).

Priors and perceptual estimation

We have developed ideal observer models for the estimation of 3D motion from local regions of the retinal images. These models are grounded in natural image statistics, incorporate constraints imposed by front-end properties of the visual system, and make use of the tools of probabilistic inference. However, unlike many probabilistic models of motion perception (Weiss et al., 2002; Stocker and Simoncelli, 2006; Welchman et al., 2008; Rokers et al., 2018), we imposed a flat motion prior—which is consistent with the relative number of stimuli at each level of the latent variable in the stimulus set—rather than a zero-motion prior. Estimation biases that are typically attributed to the action of a zero-motion prior, nevertheless emerged.

Two examples are as follows. First, estimates of 3D direction were biased toward the frontoparallel direction. Second, estimates of 3D speed from low-contrast stimuli in the dataset tend to be more biased toward slow speeds (data not shown). Both the first finding (Rokers et al., 2018), and an analog of the

second finding in 2D motion estimation (Weiss et al., 2002), have been appealed to as evidence for the action of a slow-motion prior in perceptual estimation.

The current results indicate that the same patterns of estimation errors can emerge from an accuracy-maximizing ideal observer with a flat prior. The current results suggest that patterns of estimation errors that have been attributed to priors can also be caused by properties of natural stimuli and/or constraints imposed by the front-end of visual systems that have nothing to do with priors. The current findings thus invite reevaluation of common claims in the literature that perceptual biases constitute evidence for nonflat priors (Rideaux and Welchman, 2020).

Conclusion

A spate of recent research has investigated the psychophysical limits and neurophysiological underpinnings of 3D motion estimation (Cormack et al., 2017; Rosenberg et al., 2023). The development of image-computable ideal observers for 3D motion estimation—which are grounded in natural scene statistics—can help interpret existing results by providing principled benchmarks against which neural response properties and psychophysical performance patterns can be evaluated. They also suggest new hypotheses that can be tested experimentally or provide reasons to question consensus views about how to best understand widely observed neural and perceptual phenomena (Burge, 2020). The ideal observers developed here, for the tasks of estimating 3D speed and 3D direction from local regions of the retinal images, show that many—sometimes counterintuitive—aspects of neural activity and human performance may be a consequence of optimal information processing in the visual system. Increasing the realism of the stimulus sets and the generality of tasks should provide deeper insights still into the computations and performance patterns that characterize 3D motion perception in natural viewing.

References

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

Albrecht DG, Geisler WS (1991) Motion selectivity and the contrast-response function of simple cells in the visual cortex. Vis Neurosci 7:531–546.

Albright TD (1984) Direction and orientation selectivity of neurons in visual area MT of the macaque. J Neurophysiol 52:1106–1130.

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

Backus BT, Banks MS, van Ee R, Crowell JA (1999) Horizontal and vertical disparity, eye position, and stereoscopic slant perception. Vision Res 39: 1143–1170.

Banks MS, Gepshtein S, Landy MS (2004) Why is spatial stereoresolution so low? J Neurosci 24:2077–2089.

Blakemore C (1970) The range and scope of binocular depth discrimination in man. J Physiol 211:599–622.

Bonnen K, Czuba TB, Whritner JA, Kohn A, Huk AC, Cormack LK (2020) Binocular viewing geometry shapes the neural representation of the dynamic three-dimensional environment. Nat Neurosci 23:113–121.

Bridge H, Cumming BG (2001) Responses of macaque V1 neurons to binocular orientation differences. J Neurosci 21:7293–7302.

Bridge H, Cumming BG, Parker AJ (2001) Modeling V1 neuronal responses to orientation disparity. Vis Neurosci 18:879–891.

Brooks KR, Stone LS (2006) Stereomotion suppression and the perception of speed: accuracy and precision as a function of 3D trajectory. J Vis 6:1214–1223.

Burg MF, Cadena SA, Denfield GH, Walker EY, Tolias AS, Bethge M, Ecker AS (2021) Learning divisive normalization in primary visual cortex. PLoS Comput Biol 17:e1009028.

Burge J (2020) Image-computable ideal observers for tasks with natural stimuli. Annu Rev Vis Sci 6:491–517.

Burge J, Geisler WS (2011) Optimal defocus estimation in individual natural images. Proc Natl Acad Sci U S A 108:16849–16854.

Burge J, Geisler WS (2012) Optimal defocus estimates from individual images for autofocusing a digital camera. In: Digital photography VIII 8299:124–135.

Burge J, Geisler WS (2014) Optimal disparity estimation in natural stereo images. J Vis 14:1.

Burge J, Geisler WS (2015) Optimal speed estimation in natural image movies predicts human performance. Nat Commun 6:7900.

Burge J, Jaini P (2017) Accuracy maximization analysis for sensory-perceptual tasks: computational improvements, filter robustness, and coding advantages for scaled additive noise. PLoS Comput Biol 13: e1005281.

Burge J, McCann BC, Geisler WS (2016) Estimating 3D tilt from local image cues in natural scenes. J Vis 16:2.

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

Chin BM, Burge J (2020) Predicting the partition of behavioral variability in speed perception with naturalistic stimuli. J Neurosci 40:864–879.

Cormack LK, Czuba TB, Knöll J, Huk AC (2017) Binocular mechanisms of 3D motion processing. Annu Rev Vis Sci 3:297–318.

Czuba TB, Huk AC, Cormack LK, Kohn A (2014) Area MT encodes three-dimensional motion. J Neurosci 34:15522–15533.

Czuba TB, Rokers B, Huk AC, Cormack LK (2010) Speed and eccentricity tuning reveal a central role for the velocity-based cue to 3D visual motion. J Neurophysiol 104:2886–2899.

DeAngelis GC, Ohzawa I, Freeman RD (1991) Depth is encoded in the visual cortex by a specialized receptive field structure. Nature 352:156–159.

Elmore LC, Rosenberg A, DeAngelis GC, Angelaki DE (2019) Choice-related activity during visual slant discrimination in macaque CIP but not V3A. eNeuro 6:ENEURO.0248-18.2019.

Fleet DJ, Wagner H, Heeger DJ (1996) Neural encoding of binocular disparity: energy models, position shifts and phase shifts. Vision Res 36:1839–1857.

Fulvio JM, Ji M, Thompson L, Rosenberg A, Rokers B (2020) Cue-dependent effects of VR experience on motion-in-depth sensitivity. PLoS One 15: e0229929.

Fulvio JM, Rosen ML, Rokers B (2015) Sensory uncertainty leads to systematic misperception of the direction of motion in depth. Atten Percept Psychophys 77:1685–1696.

Geisler WS (1989) Sequential ideal-observer analysis of visual discriminations. Psychol Rev 96:267–314.

Geisler WS, Albrecht DG (1997) Visual cortex neurons in monkeys and cats: detection, discrimination, and identification. Vis Neurosci 14: 897–919.

Geisler WS, Najemnik J, Ing AD (2009) Optimal stimulus encoders for natural tasks. J Vis 9:17.

Greenwald HS, Knill DC (2009) Orientation disparity: a cue for 3D orientation? Neural Comput 21:2581–2604.

Harris JM, Watamaniuk SNJ (1995) Speed discrimination of motion-in-depth using binocular cues. Vision Res 35:885–896.

Heeger DJ (1992) Normalization of cell responses in cat striate cortex. Vis Neurosci 9:181–197.

Hillis JM, Watt SJ, Landy MS, Banks MS (2004) Slant from texture and disparity cues: optimal cue combination. J Vis 4:967–992.

Hou C, Nicholas SC, Verghese P (2020) Contrast normalization accounts for binocular interactions in human striate and extra-striate visual cortex. J Neurosci 40:2753–2763.

Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160:106–154.

Iyer AV, Burge J (2018) Depth variation and stereo processing tasks in natural scenes. J Vis 18:4.

Iyer A, Burge J (2019) The statistics of how natural images drive the responses of neurons. J Vis 19:4.

Jaini P, Burge J (2017) Linking normative models of natural tasks to descriptive models of neural response. J Vis 17:16.

Joo SJ, Czuba TB, Cormack LK, Huk AC (2016) Separate perceptual and neural processing of velocity- and disparity-based 3D motion signals. J Neurosci 36:10791–10802.

Katz LC, Crowley JC (2002) Development of cortical circuits: lessons from ocular dominance columns. Nat Rev Neurosci 3:34–42.

Kim S, Burge J (2018) The lawful imprecision of human surface tilt estimation in natural scenes. Elife 7:e31448.

Kim S, Burge J (2020) Natural scene statistics predict how humans pool information across space in surface tilt estimation. PLoS Comput Biol 16: e1007947.
Knill DC (1998) Discrimination of planar surface slant from texture: human and ideal observers compared. Vision Res 38:1683–1711.
Lages M (2006) Bayesian models of binocular 3-D motion perception. J Vis 6:14.
Lages M, Heron S (2010) On the inverse problem of binocular 3D motion perception. PLoS Comput Biol 6:e1000999.
Longuet-Higgins HC (1984) The visual ambiguity of a moving plane. Proc R Soc Lond B Biol Sci 223:165–175.
McFarland JM, Cui Y, Butts DA (2013) Inferring nonlinear neuronal computation based on physiologically plausible inputs. PLoS Comput Biol 9: e1003143.
McKee SP, Silverman GH, Nakayama K (1986) Precise velocity discrimination despite random variations in temporal frequency and contrast. Vision Res 26:609–619.
Mitchell BA, Carlson BM, Westerberg JA, Cox MA, Maier A (2023) A role for ocular dominance in binocular integration. Curr Biol 33:3884–3895.e5.
Navarro R, Artal P, Williams DR (1993) Modulation transfer of the human eye as a function of retinal eccentricity. J Opt Soc Am A 10:201–212.
Ni L, Burge J (2024) Feature-specific divisive normalization improves natural image encoding for depth perception.
Nienborg H, Bridge H, Parker AJ, Cumming BG (2004) Receptive field size in V1 neurons limits acuity for perceiving disparity modulation. J Neurosci 24:2065–2076.
Nover H, Anderson CH, DeAngelis GC (2005) A logarithmic, scale-invariant representation of speed in macaque middle temporal area accounts for speed discrimination performance. J Neurosci 25:10049–10060.
Ohzawa I (1998) Mechanisms of stereoscopic vision: the disparity energy model. Curr Opin Neurobiol 8:509–515.
Ohzawa I, DeAngelis GC, Freeman RD (1990) Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. Science 249:1037–1041.
Oluk C, Bonnen K, Burge J, Cormack LK, Geisler WS (2022) Stereo slant discrimination of planar 3D surfaces: frontoparallel versus planar matching. J Vis 22:6.
Ono H, Barbeito R (1985) Utrocular discrimination is not sufficient for utrocular identification. Vision Res 25:289–299.
Pack CC, Born RT, Livingstone MS (2003) Two-dimensional substructure of stereo and motion interactions in macaque visual cortex. Neuron 37:525–535.
Pagan M, Simoncelli EP, Rust NC (2016) Neural quadratic discriminant analysis: nonlinear decoding with V1-like computation. Neural Comput 28: 2291–2319.
Park IM, Archer EW, Priebe N, Pillow JW (2013) Spectral methods for neural characterization using generalized quadratic models. In: *Advances in neural information processing systems* (Burges CJ, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds), Vol 26. pp 2454–2462. Curran Associates, Inc.
Peng Q, Shi BE (2010) The changing disparity energy model. Vision Res 50: 181–192.
Peng Q, Shi BE (2014) Neural population models for perception of motion in depth. Vision Res 101:11–31.
Priebe NJ, Lisberger SG, Movshon JA (2006) Tuning for spatiotemporal frequency and speed in directionally selective neurons of macaque striate cortex. J Neurosci 26:2941–2950.
Qian N, Andersen RA (1997) A physiological model for motion-stereo integration and a unified explanation of Pulfrich-like phenomena. Vision Res 37:1683–1698.
Read JCA, Cumming BG (2005a) All Pulfrich-like illusions can be explained without joint encoding of motion and disparity. J Vis 5:901–927.
Read JCA, Cumming BG (2005b) Effect of interocular delay on disparity-selective V1 neurons: relationship to stereoacuity and the Pulfrich effect. J Neurophysiol 94:1541–1553.
Read JCA, Cumming BG (2007) Sensors for impossible stimuli may solve the stereo correspondence problem. Nat Neurosci 10:1322–1328.
Rideaux R, Welchman AE (2020) But still it moves: static image statistics underlie how we see motion. J Neurosci 40:2538–2552.
Rogers BJ, Bradshaw MF (1993) Vertical disparities, differential perspective and binocular stereopsis. Nature 361:253–255.
Rokers B, Cormack LK, Huk AC (2008) Strong percepts of motion through depth without strong percepts of position in depth. J Vis 8:6.
Rokers B, Fulvio JM, Pillow JW, Cooper EA (2018) Systematic misperceptions of 3-D motion explained by Bayesian inference. J Vis 18:23.
Rosenberg A, Angelaki DE (2014) Gravity influences the visual representation of object tilt in parietal cortex. J Neurosci 34:14170–14180.
Rosenberg A, Cowan NJ, Angelaki DE (2013) The visual representation of 3D object orientation in parietal cortex. J Neurosci 33:19352–19361.
Rosenberg A, Thompson LW, Doudlah R, Chang T-Y (2023) Neuronal representations supporting three-dimensional vision in nonhuman primates. Annu Rev Vis Sci 9:337–359.
Rushton SK, Duke PA (2009) Observers cannot accurately estimate the speed of an approaching object in flight. Vision Res 49:1919–1928.
Rust NC, Schwartz O, Movshon JA, Simoncelli EP (2005) Spatiotemporal elements of macaque V1 receptive fields. Neuron 46:945–956.
Sanada TM, DeAngelis GC (2014) Neural representation of motion-in-depth in area MT. J Neurosci 34:15508–15521.
Schneeweis DM, Schnapf JL (1995) Photovoltage of rods and cones in the macaque retina. Science 268:1053.
Simoncelli EP (2003) Local analysis of visual motion. In: *The visual neurosciences* (Chalupa LM, Werner JS, eds), Vol 2, pp 1616–1623. The MIT Press.
Sprague WW, Cooper EA, Tošić I, Banks MS (2015) Stereopsis is adaptive for the natural environment. Sci Adv 1:e1400254.
Stevens KA (1983) Slant-tilt: the visual encoding of surface orientation. Biol Cybern 46:183–195.
Stocker AA, Simoncelli EP (2006) Noise characteristics and prior expectations in human visual speed perception. Nat Neurosci 9:578–585.
Stockman A, Sharpe LT (2000) The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype. Vision Res 40:1711–1737.
Tanabe S, Haefner RM, Cumming BG (2011) Suppressive mechanisms in monkey V1 help to solve the stereo correspondence problem. J Neurosci 31:8295–8305.
Thibos LN, Ye M, Zhang X, Bradley A (1992) The chromatic eye: a new reduced-eye model of ocular chromatic aberration in humans. Appl Opt 31:3594–3600.
Thompson L, Ji M, Rokers B, Rosenberg A (2019) Contributions of binocular and monocular cues to motion-in-depth perception. J Vis 19:2.
Thompson LW, Kim B, Rokers B, Rosenberg A (2023) Hierarchical computation of 3D motion across macaque areas MT and FST. Cell Rep 42: 113524.
Tyler CW (1971) Stereoscopic depth movement: two eyes less sensitive than one. Science 174:958–961.
Tyler CW, Julesz B (1978) Binocular cross-correlation in time and space. Vision Res 18:101–105.
Ukwade MT, Bedell HE, Harwerth RS (2003a) Stereopsis is perturbed by vergence error. Vision Res 43:181–193.
Ukwade MT, Bedell HE, Harwerth RS (2003b) Stereothresholds with simulated vergence variability and constant error. Vision Res 43:195–204.
Vlaskamp BNS, Filippini HR, Banks MS (2009) Image-size differences worsen stereopsis independent of eye position. J Vis 9:17.
Wainwright MJ, Simoncelli E (1999) Scale mixtures of Gaussians and the statistics of natural images. Adv Neural Inf Process Syst 12:855–861.
Watt SJ, Akeley K, Ernst MO, Banks MS (2005) Focus cues affect perceived depth. J Vis 5:834–862.
Weiss Y, Simoncelli EP, Adelson EH (2002) Motion illusions as optimal percepts. Nat Neurosci 5:598–604.
Welchman AE, Deubelius A, Conrad V, Bülthoff HH, Kourtzi Z (2005) 3D shape perception from combined depth cues in human visual cortex. Nat Neurosci 8:820–827.
Welchman AE, Lam JM, Bülthoff HH (2008) Bayesian motion estimation accounts for a surprising bias in 3D vision. Proc Natl Acad Sci U S A 105:12087–12092.
Wu W, Hatori Y, Tseng C, Matsumiya K, Kuriki I, Shioiri S (2020) A motion-in-depth model based on inter-ocular velocity to estimate direction in depth. Vision Res 172:11–26.
Wyszecki G, Stiles VS, Kelly KL (1968) Color science: concepts and methods, quantitative data and formulas. Phys Today 21:83–84.
Zeki SM (1974) Cells responding to changing image size and disparity in the cortex of the rhesus monkey. J Physiol 242:827–841.