Optimal disparity estimation in natural stereo images

Johannes Burge
Center for Perceptual Systems and Department of Psychology, University of Texas at Austin, Austin, TX, USA

Wilson S. Geisler
Center for Perceptual Systems and Department of Psychology, University of Texas at Austin, Austin, TX, USA

A great challenge of systems neuroscience is to understand the computations that underlie perceptual constancies, the ability to represent behaviorally relevant stimulus properties as constant even when irrelevant stimulus properties vary. As signals proceed through the visual system, neural states become more selective for properties of the environment, and more invariant to irrelevant features of the retinal images. Here, we describe a method for determining the computations that perform these transformations optimally, and apply it to the specific computational task of estimating a powerful depth cue: binocular disparity. We simultaneously determine the optimal receptive field population for encoding natural stereo images of locally planar surfaces and the optimal nonlinear units for decoding the population responses into estimates of disparity. The optimal processing predicts well-established properties of neurons in cortex. Estimation performance parallels important aspects of human performance. Thus, by analyzing the photoreceptor responses to natural images, we provide a normative account of the neurophysiology and psychophysics of absolute disparity processing. Critically, the optimal processing rules are not arbitrarily chosen to match the properties of neurophysiological processing, nor are they fit to match behavioral performance. Rather, they are dictated by the task-relevant statistical properties of complex natural stimuli. Our approach reveals how selective invariant tuning—especially for properties not trivially available in the retinal images—could be implemented in neural systems to maximize performance in particular tasks.

Introduction

Front-facing eyes evolved, at least in part, to support binocular depth perception (Figure 1a). Stereopsis—perceiving depth from disparity—is a perceptual constancy that pervades the animal kingdom. In the binocular zone, each eye's view yields a slightly different image of the scene. The local differences between the retinal images—the binocular disparities—are powerful signals for fixating the eyes and computing the depth structure of the scene. The critical step in enabling the use of disparity in service of these tasks is to estimate disparity itself. Once the disparities are estimated, metric depth can be computed by triangulation given the fixation of the eyes. There have been many computational studies of disparity estimation (Banks, Gepshtein, & Landy, 2004; Cormack, Stevenson, & Schor, 1991; Marr & T. Poggio, 1976; Qian, 1997; Qian & Zhu, 1997; Read & Cumming, 2007; Tyler & Julesz, 1978) and of the behavioral limits in humans (Banks et al., 2004; Cormack et al., 1991; Marr & T. Poggio, 1976; Ogle, 1952; Panum, 1858; Tyler & Julesz, 1978). The underlying neural mechanisms of disparity processing have also been extensively researched (Cumming & DeAngelis, 2001; DeAngelis, Ohzawa, & Freeman, 1991; Nienborg, Bridge, Parker, & Cumming, 2004; Ohzawa, DeAngelis, & Freeman, 1990). However, there is no widely accepted ideal observer theory of disparity estimation in natural images. Such a theory would be a useful tool for evaluating performance in disparity-related tasks, providing principled hypotheses for neural mechanisms, and developing practical applications.

Deriving the ideal observer for disparity estimation is a hierarchical, multistep process (see Figure 2). The first step is to model the photoreceptor responses to stereo images of natural scenes. The second step is to learn the optimal set of binocular filters for disparity estimation from a large collection of natural images. The third step is to determine how the optimal filter responses should be combined to obtain units that are selective for particular disparities and maximally invariant to stimulus dimensions other than disparity (e.g., texture). The fourth and final step is to read out the population response to obtain the optimal disparity estimates. In addition to carrying out these steps, we

Citation: Burge, J., & Geisler, W. S. (2014). Optimal disparity estimation in natural stereo images. *Journal of Vision*, *14*(2):1, 1–18, http://www.journalofvision.org/content/14/2/1, doi:10.1167/14.2.1.

doi: 10.1167/14.2.1 Received May 28, 2013; published February 3, 2014

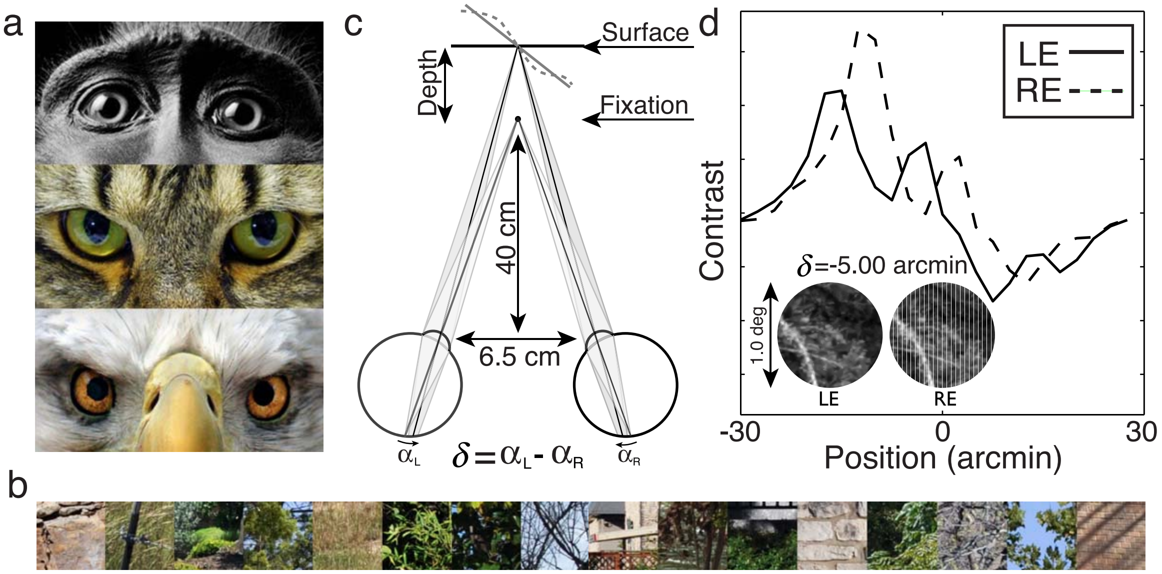

Figure 1. Natural scene inputs, disparity geometry, and example left and right eye signals. (a) Animals with front facing eyes. (b) Example natural images used in the analysis. (c) Stereo geometry. The eyes are fixated and focused at a point straight ahead at 40 cm. We considered retinal disparity patterns corresponding to fronto-parallel and slanted surfaces. Non-planar surfaces were also considered (see Discussion). (d) Photographs of natural scenes are texture mapped onto planar fronto-parallel or slanted surfaces. Here, the left and right eye retinal images are perspective projections (inset) of a fronto-parallel surface with 5 arcmin of uncrossed disparity. Left and right eye signals are obtained by vertically averaging each image; these are the signals available to neurons with vertically oriented receptive fields. The signals are not identical shifted copies of each other because of perspective projection, added noise, and cosine windowing (see text). We note that across image patches there is considerable signal variation due to stimulus properties (e.g., texture) unrelated to disparity. A selective, invariant neural population must be insensitive to this variation.

show how the optimal computations can be implemented with well-understood cortical mechanisms. Interestingly, the binocular receptive field properties of cortical neurons, and human behavioral performance in disparity-related tasks, parallel those of the optimal estimator. The results help explain the exquisite ability of the human visual system to recover the 3-D structure of natural scenes from binocular disparity.

The computational approach employed here has implications not just for understanding disparity estimation, but for the more general problem of understanding neural encoding and decoding in specific tasks. Our approach merges and extends the two major existing computational approaches for understanding the neural encoding-decoding problem. The first existing approach—efficient coding—focuses on how to efficiently (compactly) represent natural sensory signals in neural populations (Hoyer & Hyvärinen, 2000; Lewicki, 2002; Li & Atick, 1994; Olshausen & Field, 1996; Simoncelli & Olshausen, 2001). While this approach has provided general insights into retinal and cortical encoding, its goal is to represent the sensory signals without loss of information. Because it has a different goal, it does not elucidate the encoding or the decoding necessary for specific perceptual tasks. In a certain sense, it provides no more insight into the computations underlying specific perceptual abilities than the photoreceptor responses themselves. The second existing approach focuses on how to decode populations of neurons tuned to the stimulus variable associated with a specific perceptual task (Girshick, Landy, & Simoncelli, 2011; Jazayeri & Movshon, 2006; Ma, Beck, Latham, & Pouget, 2006; Read & Cumming, 2007). However, this approach focuses on response variability that is intrinsic to the neural system rather than response variability that is due to the natural variation in the stimuli themselves. Specifically, this approach assumes neurons that have perfectly invariant tuning functions. These neurons give the same response (except for neural noise) to all stimuli having the same value of the variable of interest. In general, this cannot occur under natural viewing conditions. Variation in natural signals along irrelevant dimensions inevitably causes variation in the tuning function. Thus, this approach does not address how encoding affects the neural encoding-decoding problem. It is critical to consider natural sensory signals and task-specific encoding and decoding simultaneously, because encoded signals relevant for one task may be irrelevant for another, and because task-specific decoding must discount irrelevant variation in the encoded signals.

Arguably, the most significant advances in behavioral and systems neuroscience have resulted from the study of neural populations associated with particular

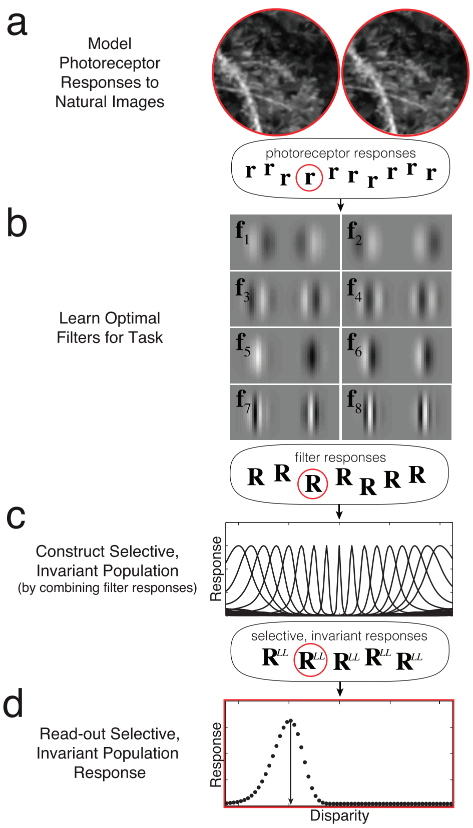
a
Model Photoreceptor Responses to Natural Images
photoreceptor responses
r r r r r r r r r r
b
Learn Optimal Filters for Task
f1
f2
f3
f4
f5
f6
f7
f8
filter responses
R R R R R R R
c
Construct Selective, Invariant Population
(by combining filter responses)
Response
selective, invariant responses
R^LL R^LL R^LL R^LL R^LL
d
Read-out Selective, Invariant Population Response
Response
Disparity

Figure 2. Hierarchical processing steps in optimal disparity estimation. (a) The photoreceptor responses are computed for each of many natural images, for each of many different disparities. (b) The optimal filters for disparity estimation are learned from this collection of photoreceptor responses to natural stimuli. These are the eight most useful vertically oriented filters (receptive fields) (see also Figure 3a). Left and right eye filter weights are shown in gray. Filter responses are given by the dot product between the photoreceptor responses and filter weights. (c) The optimal selective, invariant units are constructed from the filter responses. Each unit in the population is tuned to a particular disparity. These units result from a unique combination of the optimal filter responses (see also Figure 7). (d) The optimal readout of the selective, invariant population response is determined. Each black dot shows the response of one of the disparity-tuned units in (c) to the particular image shown in (a). The peak of the population response is the optimal (MAP) estimate. Note that **r**, **R**, and $\mathbf{R}^{LL}$ are vectors representing the photoreceptor, filter, and disparity-tuned-unit population responses to particular stimuli. Red outlines represent the responses to the particular stereo-image patch in (a). The steps in this procedure are general, and will be useful for developing ideal observers for other estimation tasks.

sensory and perceptual tasks. Here, we describe an approach for determining the encoding and decoding of natural sensory stimuli required for optimal performance in specific tasks, and show that well-established cortical mechanisms can implement the optimal computations. Our results provide principled, testable hypotheses about the processing stages that give rise to the selective, invariant neural populations that are thought to underlie sensory and perceptual constancies.

Methods and results

Natural disparity signals

The first step in deriving the ideal observer for disparity estimation is to simulate the photoreceptor responses to natural scenes (see Figure 2a). The vector of photoreceptor responses, **r** (see Figure 2a), is determined by the luminance (Figure 1b) and depth structure of natural scenes, projective viewing geometry, and the optics, sensors, and noise in the vision system. We model each of these factors for the human visual system (see Methods details). We generate a large number of noisy, sampled, stereo-images of small (1° × 1°), fronto-parallel surface patches for each of a large number of disparity levels within Panum's fusional range (Panum, 1858) (−16.875 to 16.875 arcmin in 1.875 arcmin steps). This range covers approximately 80% of disparities that occur at or near the fovea in natural viewing (Liu, Bovik, & Cormack, 2008), and these image patches represent the information available to the vision system for processing. Although we focus first on fronto-parallel patches, we later show that the results are robust to surface slant and are likely to be robust to other depth variations occurring in natural scenes. We emphasize that in this paper we simulate retinal images by projecting natural image patches onto planar surfaces (Figure 1c), rather than simulating retinal images from stereo photographs. Later, we evaluate the effects of this simplification (see Discussion).

The binocular disparity in the images entering the two eyes is given by

$$\delta = \alpha_L - \alpha_R, \tag{1}$$

where α_L and α_R are the angles between the retinal projections of a target and fixation point in the left and right eyes. This is the definition of absolute retinal disparity. The specific pattern of binocular disparities depends on the distance between the eyes, the distance and direction that the eyes are fixated, and the distance and depth structure of the surfaces in the scene. We consider a viewing situation in which the eyes are separated by 6.5 cm (a typical interocular separation) and are fixated on a point 40 cm straight ahead (a typical arm's length) (Figure 1c).

Optimal linear binocular filters

Estimating disparity (solving the correspondence problem) is difficult because there often are multiple points in one eye's image that match a single point in the other (Marr & T. Poggio, 1976). Therefore, accurate estimation requires using the local image region around each point to eliminate false matches. In mammals, neurons with binocular receptive fields that weight and sum over local image regions first appear in primary visual cortex. Much is known about these neurons. It is not known, however, why they have the specific receptive fields they do, nor is it known how their responses are combined for the task of estimating disparity.

To discover the linear binocular filters that are optimal this task (see Figure 2b), we use a Bayesian statistical technique for dimensionality reduction called accuracy maximization analysis (AMA) (Geisler, Najemnik, & Ing 2009). The method depends both on the statistical structure of the sensory signals and on the task to be solved. This feature of the method is critically important because information relevant for one task may be irrelevant for another. For any fixed number of filters (feature dimensions), AMA returns the linear filters that maximize accuracy for the specific computational task at hand (see Methods details). (Note that feature dimensions identified with other dimensionality-reduction techniques like PCA and ICA may be irrelevant for a given task.) Importantly, AMA makes no a priori assumptions about the shapes of the filters (e.g., there no requirement that they be orthogonal). We applied AMA to the task of identifying the disparity level, from a discrete set of levels, in a random collection of retinal stereo images of natural scenes. The number of disparity levels was sufficient to allow continuous disparity estimation from -15 to 15 arcmin (see Methods details).

Before applying AMA, we perform a few additional transformations consistent with the image processing known to occur early in the primate visual system (Burge & Geisler, 2011). First, each eye's noisy sampled image patch is converted from a luminance image to a windowed contrast image $c(\mathbf{x})$ by subtracting off and dividing by the mean and then multiplying by a raised cosine of 0.5° at half height. The window limits the maximum possible size of the binocular filters; it places no restriction on minimum size. The size of the cosine window approximately matches the largest V1 binocular receptive field sizes near the fovea (Nienborg et al., 2004). Next, each eye's sampled image patch is averaged vertically to obtain what are henceforth referred to as left and right eye signals. Vertical averaging is tantamount to considering only vertically oriented filters, for this is the operation that vertically oriented filters perform on images. All orientations can provide information about binocular disparity (Chen & Qian, 2004; DeAngelis et al., 1991); however, because canonical disparity receptive fields are vertically oriented, we focus our analysis on them. An example stereo pair and corresponding left and right eye signals are shown in Figure 1d. Finally, the signals are contrast normalized to a vector magnitude of 1.0: $c_{norm}(\mathbf{x}) = c(\mathbf{x})/||c(\mathbf{x})||$. This normalization is a simplified version of the contrast normalization seen in cortical neurons (Albrecht & Geisler, 1991; Albrecht & Hamilton, 1982; Heeger, 1992) (see Supplement). Seventy-six hundred normalized left and right eye signals (400 Natural Inputs × 19 Disparity Levels) constituted the training set for AMA.

The eight most useful linear binocular filters (in rank order) are shown in Figure 3a. These filters specify the subspace that a population of neurophysiological receptive fields *should* cover for maximally accurate disparity estimation. Some filters are excited by nonzero disparities, some are excited by zero (or near-zero) disparities, and still others are suppressed by zero disparity. The left and right eye filter components are approximately log-Gabor (Gaussian on a log spatial frequency axis). The spatial frequency selectivity of each filter's left and right eye components are similar, but differ between filters (Figure 3b). Filter tuning and bandwidth range between 1.2–4.7 c/° and 0.9–4.2 c/°, respectively, with an average bandwidth of approximately 1.5 octaves. Thus, the spatial extent of the filters (see Figure 3a) is inversely related to its spatial frequency tuning: As the tuned frequency increases, the spatial extent of the filter decreases. The filters also exhibit a mixture of phase and position coding (Figure 3c), suggesting that a mixture of phase and position coding is optimal. Similar filters result (Figure S1a–c) when the training set contains surfaces having a distribution of different slants (see Discussion). (Note that the cosine windows bias the filters more toward phase than position encoding. Additional analyses have nevertheless shown that windows having a nonzero position offset, i.e., position disparity, do not qualitatively change the filters.)

Interestingly, some filters provide the most information about disparity when they are not responding. For example, the disparity is overwhelmingly likely to be zero when filter $\mathbf{f}_5$ (see Figure 3, Supplementary Figure S5) does not respond because anticorrelated intensity patterns in the left- and right-eye images are very unlikely in natural images. A complex cell with this binocular receptive field would produce a disparity tuning curve similar to the tuning curve produced by a classic tuned-inhibitory cell (Figure S5; Poggio, Gonzalez, Krause, 1988).

The properties of our linear filters are similar to those of binocular simple cells in early visual cortex (Cumming & DeAngelis, 2001; De Valois, Albrecht, &

Thorell, 1982; DeAngelis et al., 1991; Poggio et al., 1988). Specifically, binocular simple cells have a similar range of shapes, a similar range of spatial frequency tuning, similar octave bandwidths of 1.5, and similar distributions of phase and position coding (Figure 3c). Despite the similarities between the optimal filters and receptive fields in primary visual cortex, we emphasize that our explicit aim is not to account for the properties of neurophysiological receptive fields (Olshausen & Field, 1996). Rather, our aim is to determine the filters that encode the retinal information most useful for estimating disparity in natural images. Thus, neurophysiological receptive fields with the properties described here would be ideally suited for disparity estimation.

The fact that many neurophysiological properties of binocular neurons are predicted by a task-driven analysis of natural signals, with appropriate biological constraints and zero free parameters, suggests that similar ideal observer analyses may be useful for understanding and evaluating the neural underpinnings of other fundamental sensory and perceptual abilities.

Optimal disparity estimation

The optimal linear binocular filters encode the most relevant information in the retinal images for disparity estimation. However, optimal estimation performance can only be reached if the joint responses of the filters are appropriately combined and decoded (Figure 2c, d). Here, we determine the Bayes optimal (nonlinear) decoder by analyzing the responses of the filters in Figure 3 to natural images with a wide range of disparities. Later, we show how this decoder could be implemented with linear and nonlinear mechanisms commonly observed in early visual cortex. (We note that the decoder that was used to learn the AMA filters from the training stimuli cannot be used for arbitrary stimuli; see Methods details.)

First, we examine the joint distribution of filter responses to the training stimuli, conditioned on each disparity level δ_k. (The dot product between each filter and a contrast normalized left- and right-eye signal gives each filter's response. The filter responses are represented by the vector $\mathbf{R}$; see Figure 2b.) The conditional response distributions $p(\mathbf{R}|\delta_k)$ are approximately Gaussian (83% of the marginal distributions conditioned on disparity are indistinguishable from Gaussian; K-S test, $p > 0.01$, see Figure 4a). (The approximately Gaussian form is largely due to the contrast normalization of the training stimuli; see Discussion.) Each of these distributions was fit with a multidimensional Gaussian $gauss(\mathbf{R};\, \mathbf{u}_k,\, \mathbf{C}_k)$ estimated from the sample mean vector $\hat{\mathbf{u}}_k$ and the sample covariance matrix $\hat{\mathbf{C}}_k$. Figure 4a shows sample filter response distributions (conditioned on several disparity levels) for the first two AMA filters. Much of the information about disparity is contained in the covariance between the filter responses, indicating that a nonlinear decoder is required for optimal performance.

The posterior probability of each disparity given an observed response vector of AMA filter responses $\mathbf{R}$, can be obtained via Bayes' rule:

$$p(\delta_k|\mathbf{R}) = \frac{gauss(\mathbf{R}; \mathbf{u}_k, \mathbf{C}_k)p(\delta_k)}{\sum_{l=1}^{N} gauss(\mathbf{R}; \mathbf{u}_l, \mathbf{C}_l)p(\delta_l)}. \tag{2}$$

Here, we assume that all disparities are equally likely. Hence, the prior probabilities $p(\delta)$ cancel out. (This assumption yields a lower bound on performance; performance will increase somewhat in natural conditions when the prior probabilities are not flat, a scenario that is considered later.)

The solid curves in Figure 4b show posterior probability distributions averaged across all natural stimuli having the same disparity. Each posterior can be conceptualized as the expected response of a population of (exponentiated) log-likelihood neurons (see below), ordered by their preferred disparities. Colored areas show variation in the posterior probabilities due to irrelevant image variation. When the goal is to obtain the most probable disparity, the optimal estimate is given by the maximum a posteriori (MAP) read-out rule

$$\hat{\delta} = \underset{\delta}{\operatorname{argmax}}\, p(\delta|\mathbf{R}). \tag{3}$$

The posterior distributions given by Equation 2 could of course also be read out with other decoding rules, which would be optimal given different goals (cost functions).

The accuracy of disparity estimates from a large collection of natural stereo-image test patches (29,600 Test Patches: 800 Natural Inputs × 37 Disparity Levels) is shown in Figure 5a. (Note: None of the test patches were in the training set, and only half the test disparity levels were in the training set.) Disparity estimates are unbiased over a wide range. At zero disparity, estimate precision corresponds to a detection threshold of $\sim$ 6 arcsec. As disparity increases (i.e., as the surface patch is displaced from the point of fixation) precision decreases approximately exponentially with disparity (Figure 5b). Sign confusions are rare (3.7%, Figure 5c). Similar, but somewhat poorer, performance is obtained with surfaces having a cosine distribution of slants (Figures 5d, e, Figure S1f, see Discussion). In summary, optimal encoding (filters in Figure 3) and decoding (Equations 2 and 3) of task-

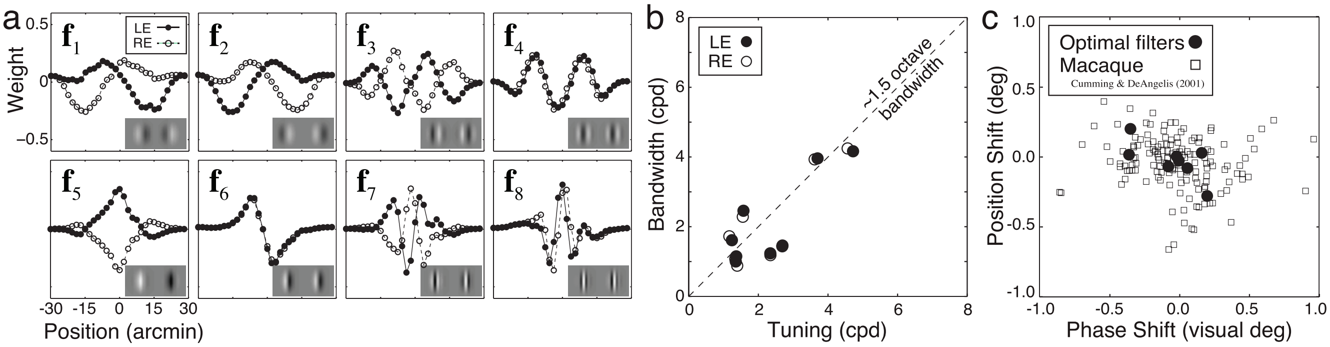

Figure 3. Optimal linear binocular receptive fields for disparity estimation. (a) Spatial receptive fields. Solid lines with closed symbols indicate the left eye filter components. Dashed lines with open symbols indicate right eye filter components. Insets show 2-D versions of the 1-D filters. (b) Luminance spatial frequency tuning versus spatial frequency bandwidth. The filters have bandwidths of approximately 1.5 octaves. (c) Phase and position shift coding for optimal binocular filters (circles) and binocular cells in macaque (squares) (Cumming & DeAngelis, 2001). Phase shifts are expressed in equivalent position shifts. Note that the filters were optimized for the fovea, whereas macaque cells were recorded from a range of different eccentricities.

relevant information in natural images yields excellent disparity estimation performance.

This pattern of performance is consistent with human performance: Human disparity detection thresholds are exquisite, a few arcsec on average (Blakemore, 1970; Cormack et al., 1991); discrimination thresholds decrease exponentially with disparity (Badcock & Schor, 1985; Blakemore, 1970; McKee, Levi, & Bowne, 1990; Stevenson, Cormack, Schor, & Tyler, 1992); and sign confusions occur with a similar pattern and a similar proportion of the time (Landers & Cormack, 1997).

Most psychophysical and neurophysiological data has been collected with artificial stimuli, usually random-dot stereograms (RDSs). We asked how well our optimal estimator performs on RDS stimuli. Given that our estimator was trained on natural stimuli, it is interesting to note that performance is very similar (but slightly poorer) with RDSs (Figure S2a–e). Fortunately, this finding suggests that under many circumstances RDS stimuli are reasonable surrogates for natural stimuli, when measuring disparity processing in biological vision systems.

Our optimal estimator also performs poorly on anticorrelated stimuli (e.g., stimuli in which one eye's image is contrast reversed), just like humans (Cumming, Shapiro, & Parker, 1998). Relevant information exists in the optimal filter responses to anticorrelated stimuli, although the conditional response distributions are more poorly segregated and the responses are generally slightly weaker (Figure S2f, g). The primary reason for poor performance is that an optimal estimator trained on natural images cannot accurately decode the filter responses to anticorrelated stimuli (Figure S2h, i). This fact may help explain why binocular neurons often respond strongly to anticorrelated stereograms, and why anticorrelated stereograms appear like ill-specified volumes of points in depth.

Optimal disparity encoding and decoding with neural populations

Although the AMA filters appear similar to binocular simple cells (Figure 3, Figure S1), it may not be obvious how the optimal Bayesian rule for combining their responses is related to processing in visual cortex. Here, we show that the optimal computations can be implemented with neurally plausible operations—linear excitation, linear inhibition, and simple static non-linearities (thresholding and squaring). Appropriate weighted summation of binocular simple and complex cell population responses can result in a new population of neurons having tightly tuned, unimodal disparity tuning curves that are largely invariant (see Figure 2c).

The key step in implementing a Bayes-optimal estimation rule is to compute the likelihood—or equivalently the log likelihood—of the optimal filter responses, conditioned on each disparity level. For the present case of a uniform prior, the optimal MAP estimate is simply the disparity with the greatest log likelihood. Given that the likelihoods are Gaussian, the log likelihoods are quadratic:

$$\ln[gauss(\mathbf{R}|\delta_k)] = -0.5(\mathbf{R}-\mathbf{u}_k)^T\mathbf{C}_k^{-1}(\mathbf{R}-\mathbf{u}_k) + const, \quad (4)$$

where $\mathbf{u}_k$ and $\mathbf{C}_k$, which are the mean and covariance matrices of the AMA filter responses to a large collection of natural stereo-image patches having disparity δ_k (see Figure 4a). By multiplying through and collecting terms, Equation 4 can be expressed as

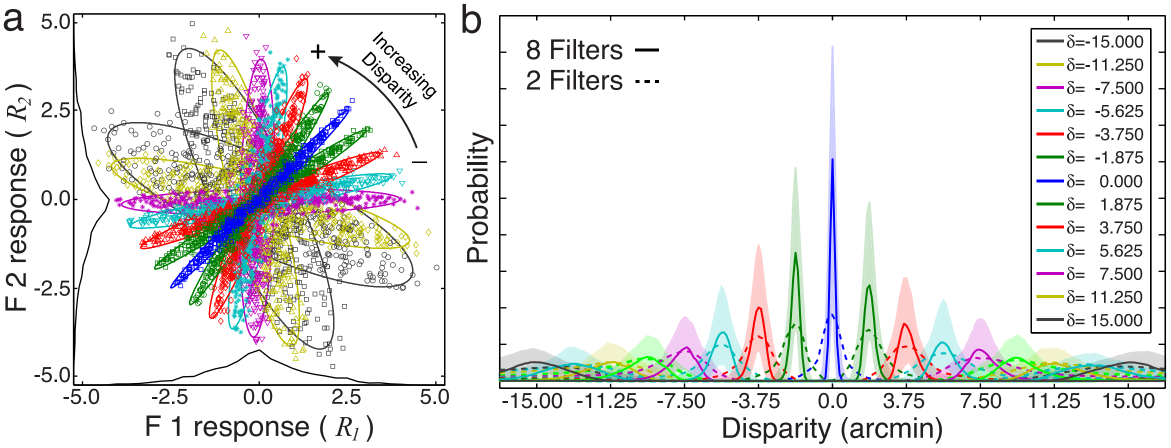

Figure 4. Joint filter response distributions conditioned on disparity for filters F1 and F2 (see Figure 3a). (a) Joint filter responses to each of the 7,600 image patches in the training set. Different colors and symbols denote different disparity levels. Contours show Gaussian fits to the conditional filter response distributions. The black curves on the *x* and *y* axes represent the marginal response distributions, $p(R_1)$ and $p(R_2)$. (b) Posterior probability distributions, averaged across all stimuli at each disparity level if only filters F1 and F2 are used (dotted curves), and if all eight filter responses are used (solid curves). Using eight AMA filters instead of two increases disparity selectivity. Shaded areas represent 68% confidence intervals on the posterior probabilities; this variation is due to natural stimulus variation that is irrelevant for estimating disparity. Natural stimulus variation thus creates response variability even in hypothetical populations of noiseless neurons.

$$\begin{aligned}\ln[gauss(\mathbf{R}|\delta_k)] = &\sum_{i=1}^{n} w_{ik}R_i + \sum_{i=1}^{n} w_{iik}R_i^2 \\ &+ \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_{ijk}(R_i + R_j)^2 \\ &+ const', \end{aligned} \tag{5}$$

where R_i is the response of the ith AMA filter, and w_{ik}, w_{iik}, and w_{ijk} are the weights for disparity δ_k. The weights are simple functions of the mean and covariance matrices from Equation 4 (see Supplement). Thus, a neuron which responds according to the log likelihood of a given disparity (an LL neuron)—that is, $R_k^{LL} = \ln[gauss(\mathbf{R}|\delta_k)]$—can be obtained by weighted summation of the linear (first term), squared (second term), and pairwise-sum-squared (third term) AMA filter responses (Equation 5). The implication is that a large collection of LL neurons, each with a different preferred disparity, can be constructed from a small, fixed set of linear filters simply by changing the weights on the linear filter responses, squared-filter responses, and pairwise-sum-squared filter responses.

A potential concern is that these computations could not be implemented in cortex. AMA filters are strictly linear and produce positive and negative responses (Figure 6a), whereas real cortical neurons produce only positive responses. However, the response of each AMA filter could be obtained by subtracting the outputs of two half-wave rectified simple cells that are "on" and "off" versions of each AMA filter (see Supplement, Figure S3a). The squared and pairwise-sum-squared responses are modeled as resulting from a linear filter followed by a static squaring nonlinearity (Figure 6b); these responses could be obtained from cortical binocular complex cells (see Supplement). (Note that our "complex cell" differs from the definition given by some prominent expositors of the disparity energy model, [Figure 6d], Cumming & DeAngelis, 2001; Ohzawa, 1998; Qian, 1997.) The squared responses could be obtained by summing and squaring the responses of on and off simple cells (see Supplement, Figure S3a, b). Finally, the LL neuron response could be obtained via a weighted sum of the AMA filter and model complex cell responses (Figure 6c). Thus, all the optimal computations are biologically plausible.

Figure 6 shows processing schematics, disparity tuning curves, and response variability for the three filter types implied by our analysis: an AMA filter, a model complex cell, a model LL neuron, and, for comparison, a standard disparity energy neuron (Cumming & DeAngelis, 2001). (The model complex cells are labeled "complex" because in general they exhibit temporal frequency doubling and a fundamental to mean response ratio that is less than 1.0, Skottun et al., 1991.) Response variability due to retinal image features that are irrelevant for estimating disparity is indicated by the gray area; the smaller the gray area, the more invariant the neural response to irrelevant image features. The AMA filter is poorly tuned to disparity and gives highly variable responses to different stereo-image patches with the same disparity

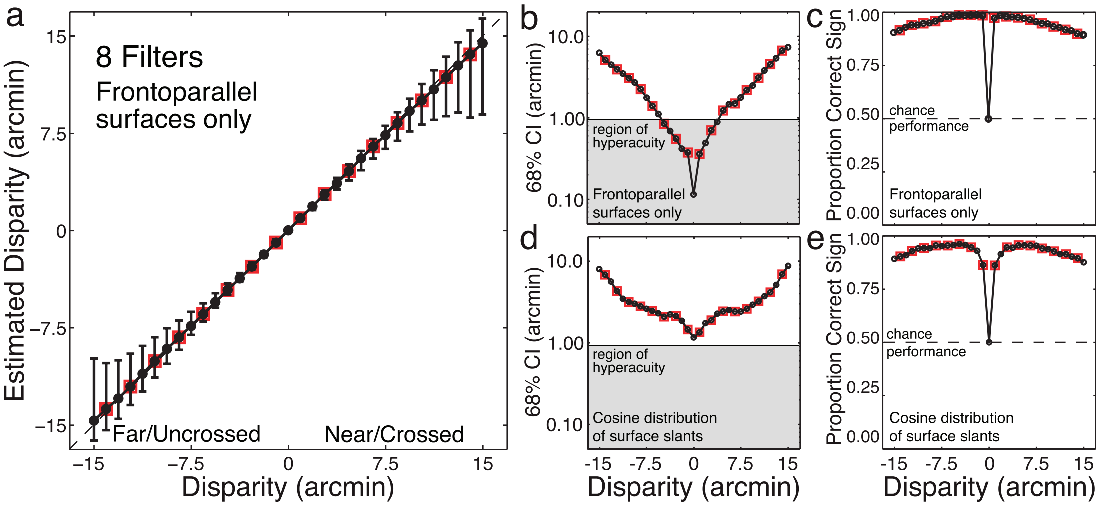

Figure 5. Accuracy and precision of disparity estimates on test patches. (a) Disparity estimates of fronto-parallel surfaces displaced from fixation using the filters in Figure 3. Symbols represent the median MAP readout of posterior probability distributions (see Figure 4b). Error bars represent 68% confidence intervals on the estimates. Red boxes mark disparity levels not in the training set. Error bars at untrained levels are no larger than at the trained levels, indicating that the algorithm makes continuous estimates. (b) Precision of disparity estimates on a semilog axis. Symbols represent 68% confidence intervals (same data as error bars in Figure 5a). Human discrimination thresholds also rise exponentially as stereo stimuli are moved off the plane of fixation. The gray area shows the hyperacuity region. (c) Sign identification performance as a function of disparity. (d), (e) Same as in (b), (c), except that data is for surfaces with a cosine distribution of slants.

(Figure 6a). The model complex cell is better tuned, but it is not unimodal and its responses also vary severely with irrelevant image information (Figure 6b). (The disparity tuning curves for all model complex cells are shown in Figure S5.) In contrast, the LL neuron is sharply tuned, is effectively unimodal, and is strongly response invariant (Figure 6c). That is, it responds similarly to all natural image patches of a given disparity. The disparity tuning curves for a range of LL neurons are shown in Figure 7a. When slant varies, similar but broader LL neuron tuning curves result (Figure S1d, e).

These results show that it is potentially misleading to refer to canonical V1 binocular simple and complex cells as disparity tuned because their responses are typically as strongly modulated by variations in contrast pattern as they are by variations in disparity (gray area, Figure 6a, b). The LL neurons, on the other hand, are tuned to a narrow range of disparities, and respond largely independent of the spatial frequency content and contrast.

The LL neurons have several interesting properties. First, their responses are determined almost exclusively by the model complex cell inputs because the weights on the linear responses (Equation 5, Figure 6a, c) are generally near zero (see Supplement). In this regard, the LL neurons are consistent with the predictions of the standard disparity energy model (Cumming & DeAngelis, 2001; Ohzawa, 1998). However, standard disparity energy neurons are not as narrowly tuned or as invariant (Figure 6d).

Second, each LL neuron receives strong inputs from multiple complex cells (Figure 7b). In this regard, the LL neurons are inconsistent with the disparity energy model, which proposes that disparity-tuned cells are constructed from two binocular subunits. The potential value of more than two subunits has been previously demonstrated (Qian & Zhu, 1997). Recently, it has been shown that some disparity-tuned V1 cells are often modulated by a greater number of binocular subunits than two. Indeed, as many as 14 subunits can drive the activity of a single disparity selective cell (Tanabe, Haefner, & Cumming, 2011).

Third, the weights on the model complex cells (Figure 7b)—which are determined by the conditional response distributions (Figure 4a)—specify how information at different spatial frequencies should be combined.

Fourth, as preferred disparity increases, the number of strong weights on the complex cell inputs decreases (Figure 7b). This occurs because high spatial frequencies are less useful for encoding large disparities (see below). Thus, if cells exist in cortex that behave similarly to LL neurons, the number and spatial frequency tuning of binocular subunits driving their response should decrease as a function of their preferred disparity.

Fifth, for all preferred disparities, the excitatory (positive) and inhibitory (negative) weights are in a

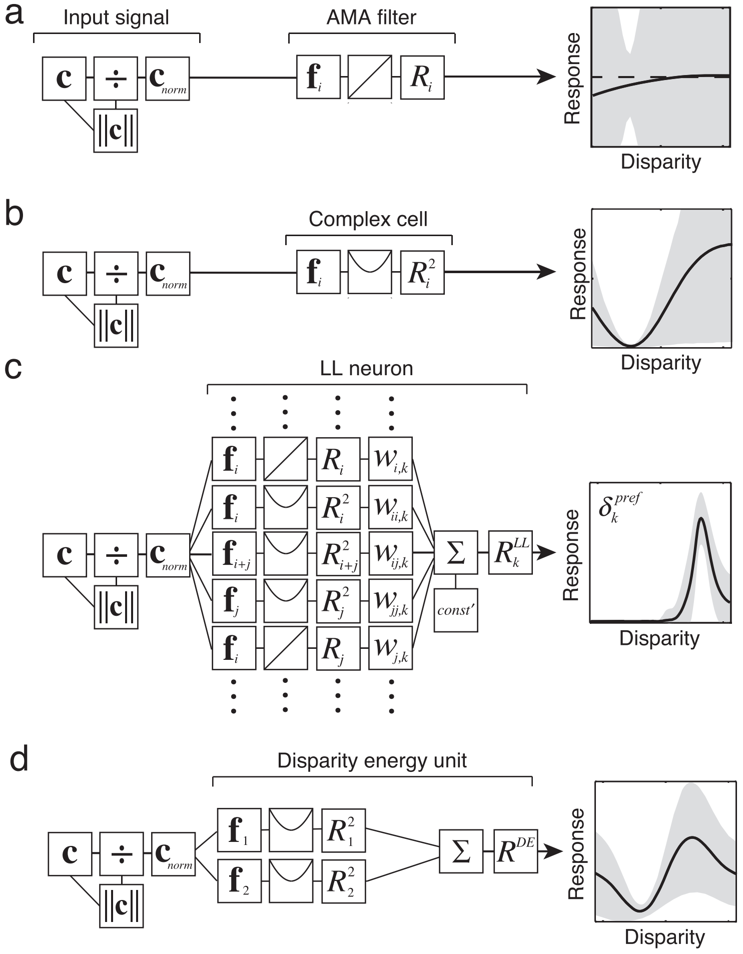

Figure 6. Biologically plausible implementation of selective, invariant tuning: processing schematics and disparity tuning curves for an AMA filter, a model complex cell, and a model log-likelihood (LL) neuron. For comparison, a disparity energy unit is also presented. In all cases, the inputs are contrast normalized photoreceptor responses. Disparity tuning curves show the mean response of each filter type across many natural image patches having the same disparity, for many different disparities. Shaded areas show response variation due to variation in irrelevant features in the natural patches (not neural noise). Selectivity for disparity and invariance to irrelevant features (external variation) increase as processing proceeds. (a) The filter response is obtained by linearly filtering the contrast normalized input signal with the AMA filter. (b) The model complex cell response is obtained by squaring the linear AMA filter response. (c) The response of an LL neuron, with preferred disparity δ_k, is obtained by a weighted sum of linear and squared filter responses. The weights can be positive/excitatory or negative/inhibitory (see Figure 7). The weights for an LL neuron with a particular preferred disparity are specified by the filter response distribution to natural images having that disparity (Figure 4a, Equations 4, 5, S1– 4). In disparity estimation, the filter responses specify that the weights on the linear filter responses are near zero (see Supplement). (d) A standard disparity energy unit is obtained by simply summing the squared responses of two binocular linear filters that are in quadrature (90° out of phase with each other). Here, we show the tuning curve of a disparity energy unit having binocular linear filters (subunits) with left and right-eye components that are also 90° out of phase with each other (i.e., each binocular subunit is selective for a nonzero disparity).

classic push-pull relationship (Ferster & Miller, 2000) (Figure 7b), consistent with the fact that disparity selective neurons in visual cortex contain both excitatory and suppressive subunits (Tanabe et al., 2011).

Sixth, the LL neuron disparity tuning curves are approximately log-Gaussian in shape (i.e., Gaussian on a log-disparity axis). Because their standard deviations are approximately constant on a log disparity axis, their disparity bandwidths increase linearly with the preferred disparity (Figure 7c). In this respect, many cortical neurons behave like LL neurons; the disparity tuning functions of cortical neurons typically have more low frequency power than is predicted from the standard energy model (Ohzawa, DeAngelis, & Freeman, 1997; Read & Cumming, 2003). Additionally, psychophysically estimated disparity channels in humans exhibit a similar relationship between bandwidth and preferred disparity (Stevenson et al., 1992). Human disparity channels, however, differ in that they have inhibitory side-lobes (Stevenson et al., 1992); that is,

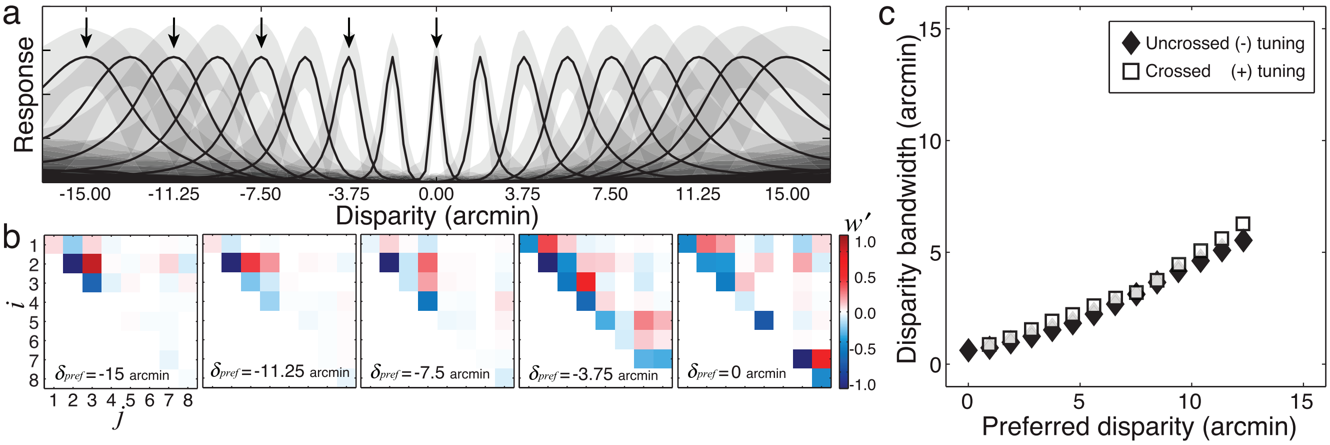

Figure 7. Constructing selective, invariant disparity-tuned units (LL neurons). (a) Tuning curves for several LL neurons, each with a different preferred disparity. Each point on the tuning curve represents the average response across a collection of natural stereo-images having the same disparity. Gray areas indicate ±1 *SD* of response due to stimulus-induced response variability. (b) Normalized weights on model complex cell responses (see Supplement, Equation 5, Figure 6c) for constructing the five LL neurons marked with arrows in (a). Positive weights are excitatory (red). Negative weights are inhibitory (blue). On-diagonal weights correspond to model complex cells having linear receptive fields like the filters in Figure 3a. Off-diagonal weights correspond to model complex cells having linear receptive fields like the scaled pairwise sums of the AMA filters (see Figures S3–S5). High spatial frequencies are not useful for estimating large disparities (Figure S7). Thus, the number of strongly weighted complex cells decreases as the magnitude of the preferred disparity increases from zero. (c) LL neuron bandwidth (i.e., full-width at half-height of disparity tuning curves) as a function of preferred disparity. Bandwidth increases approximately linearly with tuning.

they have the center-surround organization that is a hallmark of retinal ganglion cell, LGN, and V1 receptive fields. Understanding the basis of this center-surround organization is an important direction for future work.

V1 binocular neurons are unlikely to have receptive fields exactly matching those of the simple and complex cells implied by the optimal AMA filters, but V1 neurons are likely to span the subspace spanned by the optimal filters. It is thus plausible that excitatory and inhibitory synaptic weights could develop so that a subset of the neurons in V1, or in other cortical areas, signal the log likelihood of different specific disparities (see Figure 7b). Indeed, some cells in cortical areas V1, V2, and V3/V3a exhibit sharp tuning to the disparity of random dot stimuli (Cumming, 2002; Ohzawa et al., 1997; G. F. Poggio et al., 1988; Read & Cumming, 2003).

Computational models of estimation from neural populations often rely on the assumption that each neuron is invariant and unimodally tuned to the stimulus property of interest (Girshick et al., 2011; Jazayeri & Movshon, 2006; Lehky & Sejnowski, 1990; Ma et al., 2006). However, it is often not discussed how invariant unimodal tuning arises. For example, the binocular neurons (i.e., complex cells) predicted by the standard disparity energy model do not generally exhibit invariant unimodal tuning (Figure 6d). Our analysis shows that neurons with unimodal tuning to stimulus properties not trivially available in the retinal images (e.g., disparity) can result from appropriate linear combination of nonlinear filter responses.

To obtain optimal disparity estimates, the LL neuron population response (represented by the vector $\mathbf{R}^{LL}$ in Figure 2c) must be read out (see Figure 2d). The optimal read-out rule depends on the observer's goal (the cost function). A common goal is to pick the disparity having the maximum a posteriori probability. If the prior probability of the different possible disparities is uniform, then the optimal MAP decoding rule reduces to finding the LL neuron with the maximum response. Nonuniform prior probabilities can be taken into account by adding a disparity-dependent constant to each LL neuron response before finding the peak response. There are elegant proposals for how the peak of a population response can be computed in noisy neural systems (Jazayeri & Movshon, 2006; Ma et al., 2006). Other commonly assumed cost functions (e.g., MMSE) yield similar performance.

In sum, our analysis has several implications. First, it suggests that optimal disparity estimation is best understood in the context of a population code. Second, it shows how to linearly sum nonlinear neural responses to construct cells with invariant unimodal tuning curves. Third, it suggests that the eclectic mixture of binocular receptive field properties in cortex may play a functional role in disparity estimation. Fourth, it provides a principled hypothesis for how neurons may compute the posterior probabilities of stimulus properties (e.g., disparity, defocus, motion) not trivially available in the retinal image(s). Thus, our analysis provides a recipe for how to increase selectivity and invariance in the visual processing stream: invariance to image content variability and selectivity for the stimulus property of interest.

Discussion

Using principles of Bayesian statistical decision theory, we showed how to optimally estimate binocular disparity in natural stereo images, given the optical, sensor, and noise properties of the human visual system. First, we determined the linear binocular filters that encode the retinal image information most useful for estimating disparity in natural stereo-images. The optimal linear filters share many properties with binocular simple cells in the primary visual cortex of monkeys and cats (Figure 3, Figure S1). Then, we determined how the optimal linear filter responses should be nonlinearly decoded to maximize the accuracy of disparity estimation. The optimal decoder was not based on prior assumptions, but rather was dictated by the statistical pattern of the joint filter responses to natural stereo images. Overall performance was excellent and matched important aspects of human performance (Figure 5). Finally, we showed that the optimal encoder and decoder can be implemented with well-established neural operations. The operations show how selectivity and invariance can emerge along the visual processing stream (Figures 6 & 7).

External variability and neural noise

A visual system's performance is limited by both external (stimulus-induced) variability and intrinsic neural response variability (i.e., neural noise). Many theoretical studies have focused on the impact of neural noise on neural computation (Berens, Ecker, Gerwinn, Tolias, & Bethge, 2011; Cumming & DeAngelis, 2001; DeAngelis et al., 1991; Jazayeri & Movshon, 2006; Ma et al., 2006; Nienborg et al., 2004; Ohzawa, DeAngelis, & Freeman, 1990). However, in many real-world situations, stimulus-induced response variability (i.e., variability due to nuisance stimulus properties) is a major or the dominant source of variability. Our analysis demonstrates that the external variation in natural image content—that is, variation in image features irrelevant to the task—constitutes an important source of response variability in disparity estimation (see Figures 4, 5, 7). Thus, even in a hypothetical visual system composed of noiseless neurons, significant response variability will occur when estimating disparity under natural conditions (see Figures 4 and 5).

Determining the best performance possible without neural noise is essential for understanding the effect of neural noise (Haefner & Bethge, 2010; Lewicki, 2002; Li & Atick, 1994; Olshausen & Field, 1996; Simoncelli & Olshausen, 2001), because it sets bounds on the amount of neural noise that a system can tolerate, before performance significantly deteriorates. Indeed, the impact of neural noise on encoding and decoding in natural viewing cannot be evaluated without considering the impact of stimulus-induced response variability on the population response (see Supplement).

Seventy years ago, Hecht, Shlaer, and Pirenne (1942) showed in a landmark study that dark-adapted humans can detect light in the periphery from the absorption of only five to eight quanta, and that external variability (in the form of photon noise) was the primary source of variability in human performance. This result contradicted the notion (widespread at the time) that response variability is due primarily to variability within the organism (e.g., intrinsic noise). Similarly, our results underscore the importance of analyzing the variability of natural signals, when analyzing the response variability of neural circuits that underlie performance in critical behavioral tasks.

Comparison with the disparity energy model

The most common neurophysiological model of disparity processing is the disparity energy model, which aims to account for the response properties of binocular complex cells (Cumming & DeAngelis, 2001; DeAngelis et al., 1991; Ohzawa, 1998; Ohzawa et al., 1990). The model had remarkable initial success. However, an increasing number of discrepancies have emerged between the neurophysiology and the model's predictions. The standard disparity energy model, for example, does not predict the reduced response to anticorrelated stereograms observed in cortex. It also does not predict the dependence of disparity selective cells on more than two binocular subunits. Fixes to the standard model have been proposed to account for the observed binocular response properties (Haefner & Cumming, 2008; Qian & Zhu, 1997; Read, Parker, & Cumming, 2002; Tanabe & Cumming, 2008). We have shown, however, that a number of these properties are a natural consequence of an optimal algorithm for disparity estimation.

The most common computational model of disparity processing is the local cross-correlation model (Banks et al., 2004; Cormack et al., 1991; Tyler & Julesz, 1978). In this model, the estimated disparity is the one that produces the maximum local cross-correlation between the images in the two eyes. Local cross-correlation is the computational equivalent of appropriately processing large populations of disparity-energy complex cells having a wide range of tuning characteristics (Anzai, Ohzawa, & Freeman, 1999; Fleet, Wagner, & Heeger, 1996). Cross-correlation is optimal when the disparities across a patch are small and constant, because it is then akin to the simple template matching of noise-limited ideal observers. Cross-correlation is not optimal for larger or varying disparities, because then the left- and right-eye signals falling within a binocular receptive field often differ substantially.

How does our ideal observer for disparity estimation compare with the standard disparity-energy and cross-correlation models? To enable a meaningful comparison, both methods were given access to the same image information. Doing so requires a slight modification to the standard cross correlation algorithm (see Supplement). The eight AMA filters in Figure 3 outperform cross-correlation in both accuracy and precision, when the disparities exceed approximately 7.5 arcmin (Figure S6). This performance increase is impressive given that local cross-correlation has (in effect) access to a large bank of filters whereas our method uses only eight. Furthermore, increasing the number of AMA filters to 10–12 gives essentially equivalent performance to cross-correlation for smaller disparities.

Spatial frequency tuning of optimal linear binocular filters

The optimal linear binocular filters (the AMA filters) are selective for low to mid spatial frequencies (1.2–4.7 c/°) and not for higher frequencies (see Figure 3, Figure S1). To develop an intuition for why, we examined the binocular contrast signals as a function of frequency that result from a binocularly viewed, high-contrast edge (Figure S7). The binocular contrast signals barely differ above ~6 c/°, indicating that higher spatial frequencies carry little information for disparity estimation.

This analysis of binocular contrast signals provides insight into another puzzling aspect of disparity processing. Human sensitivity to disparity modulation (i.e., sinusoidal modulations in disparity-defined depth) cuts off at very low frequencies (~4 c/°) (Banks et al., 2004; Tyler, 1975). V1 binocular receptive fields tuned to the highest useful (luminance) spatial frequency (~6 c/°) have widths of ~8 arcmin, assuming the typical octave bandwidth for cortical neurons. Given that receptive fields cannot signal variation finer than their own size, the fact that humans cannot see disparity modulations higher than ~4 c/° is nicely predicted by the ~8 arcmin width suggested by the present analysis. This size matches previous, psychophysically based estimates of the smallest useful mechanism in disparity estimation (Banks et al., 2004; Harris, McKee, & Smallman, 1997).

Eye movement jitter has previously been proposed as an explanation for why useful binocular information is restricted to low spatial frequencies (Vlaskamp, Yoon, & Banks, 2011). The argument is that jitter and the relatively long integration time of the stereo system could “smear out” the disparity signals. Eye movement jitter certainly can degrade stereo information enough to render high frequency signals unmeasurable (Figure S7c, d). However, the analysis in the supplement (Figure S7) suggests that the low frequency at which humans lose the ability to detect disparity modulation may also be explained by the statistics of natural images.

Generality of findings

Although our analysis is based on calibrated natural stereo-image patches and realistic characterizations of human optics, sensors, and neural noise, there are some simplifying assumptions that could potentially limit the generality of our conclusions. Here we examine the effect of these assumptions.

The effect of surface slant and nonplanar depth structure

The stereo-image patches were projections of fronto-parallel surfaces. In natural scenes, surfaces are often slanted, which causes the disparity to change across a patch. We evaluated the effect of surface slant by repeating our analysis with a training set having a distribution of slants (see Figure 5d, e and Supplement). This distribution of slants produced a distribution of disparity gradients that are comparable to those that occur when viewing natural scenes (Hibbard, 2008) (Figure S1g). The optimal binocular filters are only very slightly affected by surface slant (c.f., Figure 3 and Figure S1a). These differences would be difficult to detect in neurophysiological experiments. This may help explain why there has been little success in finding neurons in early visual cortex that are tuned to nonzero slants (Nienborg et al., 2004). Disparity estimation performance is also quite similar, although estimate precision is somewhat reduced (Figure 5d, e).

Even after including the effect of surface slant, our training set lacks the non-planar depth structure—within-patch depth variation and depth discontinuities (occlusions)—that is present in natural scenes (Figure S8a–c). To determine how the magnitude of these other sources depth variation compare to that due to slant, we analyzed a set of range images obtained with a high-precision Riegl VZ400 range scanner (see Supplement). We find that most of the depth variation in the range images is captured by variation in slant (Figure S8d). Given that variation in slant has little effect on the optimal receptive fields, and given that locations of fine depth structure and monocular occlusion zones ($\sim$9% of points in our dataset) are most likely random with respect to the receptive fields, then it is likely that the other sources of variation will have little effect.

Nevertheless, occlusions create monocular zones that carry information supporting da Vinci stereopsis in humans (Kaye, 1978; Nakayama & Shimojo, 1990). Neurophysiological evidence suggests that binocular mechanisms exist for locating occlusion boundaries (Tsao, Conway, & Livingstone, 2003). In the future, when databases become available of high-quality natural stereo images with coregistered RGB and distance information, the general approach described here may prove useful for determining the filters that are optimal for detecting occlusion boundaries.

The effect of a realistic disparity prior

Our analysis assumed a flat prior probability distribution over disparity. Two groups have recently published estimates of the prior distribution of disparities, based on range measurements in natural scenes and human eye movement statistics (Hibbard, 2008; Liu et al., 2008). Does the prior distribution of disparity signals have a significant effect on the optimal filters and estimation performance? To check, we modified the frequency of different disparities in our training set to match the prior distribution of disparities encountered in natural viewing (Liu et al., 2008). Linear filter shapes, LL neuron tuning curve shapes, and performance levels were robust to differences between a flat and realistic disparity prior.

It has been hypothesized that the prior over the stimulus dimension of interest (e.g., disparity) may determine the optimal tuning curve shapes and the optimal distribution of peak tunings (i.e., how the tuning curves should tile the stimulus dimension) (Ganguli & Simoncelli, 2010). In evaluating this hypothesis, one must consider the results presented in the present paper. Our results show that natural signals and the linear receptive fields that filter those signals place strong constraints on the shapes that tuning curves can have. Specifically, the shapes of the LL neuron disparity tuning curve (approximately log-Gaussian) are robust to changes in the prior. Thus, although the prior may influence the optimal distribution of peak tuning (our analysis does not address this issue), it is unlikely to be the sole (or even the primary) determinant of tuning curve shapes.

The effect of contrast normalization

Contrast normalization significantly contributes to the Gaussian form of the filter response distributions. One potential advantage of Gaussian response distributions is that pair-wise (and lower order) statistics fully characterize the joint responses from an arbitrary number of filters, making possible the decoding of large filter population responses. In the analysis presented here, all binocular signals were normalized to a mean of zero and a vector magnitude of 1.0 before being projected onto the binocular filters. This is a simplified form of the contrast normalization that is a ubiquitous feature of retinal and early cortical processing (Carandini & Heeger, 2012). The standard model of contrast normalization in cortical neurons is given by $c_{norm}(\mathbf{x}) = c(\mathbf{x})/\sqrt{||c(\mathbf{x})||^2 + nc_{50}^2}$ where n is the dimensionality of the vector and c^{50} is the half-saturation constant (Albrecht & Geisler, 1991; Albrecht & Hamilton, 1982; Heeger, 1992). (The half-saturation constant is so-called because, in a neuron with an output squaring nonlinearity, the response rate will equal half its maximum when the contrast of the stimulus equals the value of c_{50}.)

We examined the effect of different half-saturation constants. The optimal filters are robust to different values of c_{50}. The conditional response distributions are not. For large values of c_{50} the response distributions have tails much heavier than Gaussians. When $c_{50} = 0.0$ (as it did throughout the paper), the distributions are well approximated but somewhat lighter-tailed than Gaussians. The distributions (e.g., see Figure 4a) are most Gaussian on average when $c_{50} = 0.1$ (Figure S9). On the basis of this finding, we hypothesize a new function for cortical contrast normalization in addition to the many already proposed (Carandini & Heeger, 2012): Contrast normalization may help create conditional filter response distributions that are Gaussian, thereby making simpler the encoding and decoding of high-dimensional subspaces of retinal image information.

The effect of different optics in the two eyes

We modeled the optics of the two eyes as being identical, whereas refractive power and monochromatic aberrations often differ somewhat between the eyes (Marcos & Burns, 2000). To evaluate the effect of normal optical variations, we convolved the retinal projections with point-spread functions previously measured in the first author's left and right eyes (Burge & Geisler, 2011). (The first author has normal optical variation.) Then, we determined optimal filters and performance. Filters and performance levels are similar to those in Figures 3 and 4. Thus, our results are robust to typical optical differences between the eyes. However, if the optics in one eye is severely degraded relative to the other, then the optimal filters are quite different, and disparity estimation performance is significantly poorer.

Interestingly, a persistent > 4 diopter difference in refractive power between the two eyes is the most important risk factor for the development of amblyopia (Levi, McKee, & Movshon, 2011), a condition characterized by extremely poor or absent stereopsis (among other deficits). A four diopter difference between the left and right eye optics drastically reduces (i.e., nearly eliminates) computable interocular contrast signals that covary systematically with disparity (Equation S6, Figure S7e–h).

Conclusions

The method presented here provides a prescription for how to optimally estimate local depth cues from images captured by the photosensors in any vision system. Our analysis of the depth cue of binocular disparity provides insight into the computations and neural populations required for accurate estimation of binocular disparity in animals viewing natural scenes. Specifically, by analyzing the information available at the photoreceptors, we improve upon existing computational methods for solving the stereo correspondence problem and provide a normative account of a range of established neurophysiological and psychophysical findings. This study demonstrates the power of characterizing the properties of natural signals that are relevant for performing specific natural tasks. A similar recent analysis provided insight into the neural computations required for accurate estimation of the focus error (defocus) in local regions of retinal images of natural scenes (Burge & Geisler, 2011; Stevenson et al., 1992). The same approach seems poised to provide insight into the neural encoding and decoding that underlie many other fundamental sensory and perceptual tasks.

Methods details

Natural disparity signals

To determine the luminance structure of natural scenes, we photographed natural scenes on and around the University of Texas at Austin campus with a tripod-mounted Nikon D700 14-bit SLR camera (4256 × 2836 pixels) fitted with a Sigma 50 mm prime lens (Burge & Geisler, 2011). To ensure sharp photographs, the camera lens was focused on optical infinity, and all imaged objects were at least 16 m from the camera. RAW photographs were converted to luminance values using the measured sensitivities of the camera (Burge & Geisler, 2011; 2012) and the human photopic sensitivity function (Stockman & Sharpe, 2000). Twelve hundred 256 × 256 pixel patches were randomly selected from 80 photographs (15 Patches × 80 Images; Figure 1b). Four hundred were used for training; the other 800 were used for testing.

To approximate the depth structure of natural scenes, we texture map the luminance patches onto planar fronto-parallel or slanted surface patches that were straight ahead at eye height. After perspective projection, this procedure yields the same pattern of retinal stimulation that would be created by viewing photographs pasted on surfaces (e.g., walls) differently slanted in depth. The disparity of the surface patch was defined as the disparity of the surface's central point

$$\delta = 2\left[\tan^{-1}\left(\frac{IPD/2}{d_{fixation} + \Delta}\right) - \tan^{-1}\left(\frac{IPD/2}{(d_{fixation})}\right)\right], \quad (6)$$

where $d_{fixation}$ is the fixation distance to the surface center, IPD is the interpupillary distance, and Δ is the depth. Depth is given by $\Delta = d_{surface} - d_{fixation}$ where $d_{surface}$ is the distance to the surface. The disparity of other points on the surface patch varies slightly across the surface patch. The change in disparity across the surface is greater when the surface patches are slanted.

Each surface patch was either coincident with or displaced from the point of fixation (Figure 1c); that is, the surface patches were positioned at one of 37 depths corresponding to 37 disparity levels within Panum's fusional range (Panum, 1858) (–16.875 to 16.875 arcmin in equally spaced steps). Each surface patch subtended 2° of visual angle from the cyclopean eye. Surfaces were then slanted (or not) and left- and right-eye projections were determined via perspective projection, ensuring that horizontal and vertical disparities were correct (Figure 1c). Next, left and right projections were resampled at 128 samples/°, approximately the sampling rate of the human foveal cones (Curcio, Sloan, Kalina, & Hendrickson, 1990). Finally, the images were cropped such that they subtended 1° from each eye (± 0.5° about the left- and right-eye foveae).

Optics

Patches were defocused with a polychromatic point-spread function based on a wave-optics model of the human visual system. The model assumed a 2 mm pupil (a size typical for a bright, sunny day) (Wyszecki & Stiles, 1982), human chromatic aberrations (Thibos, Ye, Zhang, & Bradley, 1992), a single refracting surface, and the Fraunhoffer approximation, which implies that at or near the focal plane the optical transfer function (OTF) is given by the cross-correlation of the generalized pupil function with its complex conjugate. In humans and nonhuman primates, accommodative and vergence eye-movement systems are tightly coupled. In other words, the focus distance usually equals the fixation distance (Fincham & Walton, 1957). We set the focus distance equal to the fixation distance of 40 cm and defocused the images appropriate for each disparity. See Supplement for further details of how the optics were simulated.

Sensor array responses

The information for disparity estimation depends on the depth and luminance structure of natural scenes, projective geometry, the optics of the two eyes, the sensor arrays, and the sensor noise. These factors together determine the pattern of noisy sensor responses

$$r_e(\mathbf{x}) = [I_e(\mathbf{x}) * psf_e(\mathbf{x})]samp(\mathbf{x}) + \eta \qquad (7)$$

for each eye e. $I_e(\mathbf{x})$ represents an eye-specific luminance image of the light striking the sensor array at each location $\mathbf{x} = (x, y)$ in a hypothetical optical system that causes zero degradation in image quality. Each eye's optics degrade the idealized image; the optics are represented by a polychromatic point-spread function $psf_e(\mathbf{x})$ that contains the effects of defocus, chromatic aberration, and photopic wavelength sensitivity. The sensor arrays are represented by a spatial sampling functions $samp(\mathbf{x})$. Finally, the sensor responses are corrupted by noise η; the noise level was set just high enough to remove retinal image detail that is undetectable by the human visual system (Williams, 1985). (In pilot studies, we found that the falloff in contrast sensitivity at low frequencies had a negligible effect on results; for simplicity we did not model its effects.) Note that Equation 7 represents a luminance (photopic) approximation of the images that would be captured by the photoreceptors. This approximation is sufficiently accurate for the present purposes (Burge & Geisler, 2011).

Accuracy maximization analysis (AMA)

AMA is a method for dimensionality reduction that finds the low-dimensional set of features in sensory signals that are most useful for performing a specific task. The dependence of AMA on task renders it distinct from many other popular methods for dimensionality reduction. Principal components analysis (PCA), for example, finds the dimensions that account for the most variation in the sensory signals. There is, of course, no guarantee that factors causing the most variation in the sensory signals will be useful for the task at hand. Color, lighting, and the multitude of textures in natural images, for example, cause significant variation in the retinal images. Yet all of this variation is irrelevant for the task of estimating disparity. Thus, PCA returns features that are not necessarily useful to the task, while AMA is specifically designed to ignore the irrelevant variation.

The logic of AMA is as follows. Consider encoding each training stimulus with a small population of filters with known (internal) noise characteristics. The filter responses are given by the dot product of the contrast-normalized stimulus, scaled by the maximum response rate, plus response noise (here, a small amount of Gaussian noise). With a known noise model, it is straightforward to compute the mean and variance of each filter's response to each training stimulus. Then, a closed-form expression is derived for the approximate accuracy of the Bayesian optimal decoder (Geisler, Najemnik, & Ing, 2009). Finally, this closed-form expression can be used to search the space of linear filters to find those that give the most accurate performance. If the algorithm does not settle in local minima, it finds the Bayes-optimal filters for maximizing performance in a given task (Geisler et al., 2009). Here, different random initializations yielded the same final estimated filters. A Matlab implementation of AMA and a short discussion of how to apply it are available at http://jburge.cps.utexas.edu/research/Code.html.

It is important to note that the Bayesian optimal decoder used in AMA requires knowing the means and variances for each possible stimulus, and hence it can only be used to decode the training stimuli. In other words, AMA can find the optimal linear filters given a large enough training set, but the decoder it uses to learn those filters cannot be applied to arbitrary stimuli. A separate analysis (like the one described in the body of the present paper) is required to determine how to optimally decode the AMA filter responses for arbitrary test stimuli.

Estimating disparity

Increasing the number of disparity levels in the training set increases the accuracy of disparity estimation. However, increasing the number of disparity levels in the training set also increases the training set size and the computational complexity of learning filters via AMA. A balance must be struck. Excellent continuous estimates are obtained using 1.875 arcmin steps for training followed by interpolation of 577 filter response distributions (i.e., 31 interpolated response distribution between each training step, corresponding to one interpolated distribution every ~3.5 arcsec). Interpolated distributions were obtained by fitting cubic splines through the response distribution mean vectors and covariance matrices. This procedure resulted in 577 LL neurons, which resulted in posterior probability distributions that were defined by 577 discrete points. MAP estimates were obtained by selecting the disparity with the highest posterior probability. Increasing the number of interpolated distributions (i.e., LL neurons) had no effect on performance.

Keywords: natural scene statistics, perceptual constancy, ideal observer, Bayesian statistics, population code, encoding, decoding, selectivity, invariance, depth perception, stereopsis, hierarchical model, disparity energy model, simple cells, complex cells

Acknowledgments

The authors thank Larry Cormack, Bruno Olshausen, and Jonathan Pillow for comments on an earlier draft. JB was supported by NSF grant IIS-1111328, NIH grant 2R01-EY011747, and NIH Training Grant 1T32-EY021462-01A1. WSG was supported by NIH grant 2R01-EY011747.

Commercial relationships: none.

Corresponding author: Johannes Burge.
Email: jburge@mail.cps.utexas.edu.
Address: Center for Perceptual Systems, University of Texas at Austin, Austin, TX, USA.

References

Albrecht, D. G., & Geisler, W. S. (1991). Motion selectivity and the contrast-response function of simple cells in the visual cortex. *Visual Neuroscience, 7*(6), 531–546.

Albrecht, D. G., & Hamilton, D. B. (1982). Striate cortex of monkey and cat: Contrast response function. *Journal of Neurophysiology, 48*(1), 217–237.

Anzai, A., Ohzawa, I., & Freeman, R. D. (1999). Neural mechanisms for processing binocular information I. Simple cells. *Journal of Neurophysiology*, *82*(2), 891–908.

Badcock, D. R., & Schor, C. M. (1985). Depth-increment detection function for individual spatial channels. *Journal of the Optical Society of America A*, *2*(7), 1211–1215.

Banks, M. S., Gepshtein, S., & Landy, M. S. (2004). Why is spatial stereoresolution so low? *Journal of Neuroscience*, *24*(9), 2077–2089.

Berens, P., Ecker, A. S., Gerwinn, S., Tolias, A. S., & Bethge, M. (2011). Reassessing optimal neural population codes with neurometric functions. *Proceedings of the National Academy of Sciences, USA, 108*(11), 4423.

Blakemore, C. (1970). The range and scope of binocular depth discrimination in man. *The Journal of Physiology, 211*(3), 599–622.

Burge, J., & Geisler, W. S. (2011). Optimal defocus estimation in individual natural images. *Proceedings of the National Academy of Sciences, USA, 108*(40), 16849–16854.

Burge, J., & Geisler, W. S. (2012). Optimal defocus estimates from individual images for autofocusing a digital camera. *Proceedings of the IS&T/SPIE 47th Annual Meeting,* January, 2012, Burlingame, CA.

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience, 13*(1), 51–62.

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

Cormack, L. K., Stevenson, S. B., & Schor, C. M. (1991). Interocular correlation, luminance contrast and cyclopean processing. *Vision Research*, *31*(12), 2195–2207.

Cumming, B. G. (2002). An unexpected specialization for horizontal disparity in primate primary visual cortex. *Nature*, *418*(6898), 633–636.

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

Cumming, B. G., Shapiro, S. E., & Parker, A. J. (1998). Disparity detection in anticorrelated stereograms. *Perception*, *27*(11), 1367–1377.

Curcio, C. A., Sloan, K. R., Kalina, R. E., & Hendrickson, A. E. (1990). Human photoreceptor topography. *The Journal of Comparative Neurology, 292*(4), 497–523.

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

DeAngelis, G. C., Ohzawa, I., & Freeman, R. D. (1991). Depth is encoded in the visual cortex by a specialized receptive field structure. *Nature*, *352*(6331), 156–159.

Ferster, D., & Miller, K. D. (2000). Neural mechanisms of orientation selectivity in the visual cortex. *Annual Review of Neuroscience, 23*(1), 441–471.

Fincham, E., & Walton, J. (1957). The reciprocal actions of accommodation and convergence. *The Journal of Physiology, 137*(3), 488–508.

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

Ganguli, D., & Simoncelli, E. P. (2010). Implicit encoding of prior probabilities in optimal neural populations. *Advances in Neural Information Processing Systems (NIPS), 23,* 658–666.

Geisler, W. S., Najemnik, J., & Ing, A. D. (2009). Optimal stimulus encoders for natural tasks. *Journal of Vision, 9*(13):17, 1–16, http://www.journalofvision.org/content/9/13/17, 10.1167/9.13.17. [PubMed] [Article]

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932.

Haefner, R., & Bethge, M. (2010). Evaluating neuronal codes for inference using Fisher information. *Advances in Neural Information Processing Systems, 23,* 1–9.

Haefner, R. M., & Cumming, B. G. (2008). Adaptation to natural binocular disparities in primate V1 explained by a generalized energy model. *Neuron*, *57*(1), 147–158.

Harris, J. M., McKee, S. P., & Smallman, H. S. (1997). Fine-scale processing in human binocular stereopsis. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 14*(8), 1673–1683.

Hecht, S., Shlaer, S., & Pirenne, M. H. (1942). Energy, quanta, and vision. *The Journal of General Physiology, 25*(6), 819–840.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*(2), 181–197.

Hibbard, P. B. (2008). Binocular energy responses to natural images. *Vision Research*, *48*(12), 1427–1439.

Hoyer, P. O., & Hyvärinen, A. (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, *11*(3), 191–210.

Jazayeri, M., & Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience*, *9*(5), 690–696.

Kaye, M. (1978). Stereopsis without binocular correlation. *Vision Research*, *18*(8), 1013–1022.

Landers, D. D., & Cormack, L. K. (1997). Asymmetries and errors in perception of depth from disparity suggest a multicomponent model of disparity processing. *Perception & Psychophysics*, *59*(2), 219–231.

Lehky, S. R., & Sejnowski, T. J. (1990). Neural model of stereoacuity and depth interpolation based on a distributed representation of stereo disparity. *Journal of Neuroscience*, *10*(7), 2281–2299.

Levi, D. M., McKee, S. P., & Movshon, J. A. (2011). Visual deficits in anisometropia. *Vision Research*, *51*(1), 48–57.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, *5*(4), 356–363.

Li, Z., & Atick, J. J. (1994). Efficient stereo coding in the multiscale representation. *Network: Computation in Neural Systems, 5,* 157–174.

Liu, Y., Bovik, A. C., & Cormack, L. K. (2008). Disparity statistics in natural scenes. *Journal of Vision*, *8*(11):19, 1–14, http://www.journalofvision.org/content/8/11/19, doi:10.1167/8.11.19 [PubMed] [Article].

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

Marcos, S., & Burns, S. A. (2000). On the symmetry between eyes of wavefront aberration and cone directionality. *Vision Research*, *40*(18), 2437–2447.

Marr, D., & Poggio, T. (1976). Cooperative computation of stereo disparity. *Science*, *194*(4262), 283–287.

McKee, S. P., Levi, D. M., & Bowne, S. F. (1990). The imprecision of stereopsis. *Vision Research*, *30*(11), 1763–1779.

Nakayama, K., & Shimojo, S. (1990). Da Vinci stereopsis: Depth and subjective occluding contours from unpaired image points. *Vision Research, 30*(11), 1811–1825.

Nienborg, H., Bridge, H., Parker, A. J., & Cumming, B. G. (2004). Receptive field size in V1 neurons limits acuity for perceiving disparity modulation. *Journal of Neuroscience*, *24*(9), 2065–2076.

Ogle, K. N. (1952). On the limits of stereoscopic vision. *Journal of Experimental Psychology*, *44*(4), 253–259.

Ohzawa, I. (1998). Mechanisms of stereoscopic vision: The disparity energy model. *Current Opinion in Neurobiology*, *8*(4), 509–515.

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

Ohzawa, I., DeAngelis, G. C., & Freeman, R. D. (1997). Encoding of binocular disparity by complex cells in the cat's visual cortex. *Journal of Neurophysiology*, *77*(6), 2879–2909.

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

Panum, P. (1858). *Physiologische Untersuchungen über das Sehen mit zwei Augen* [Translation: *Psychological investigations on seeing with two eyes*]. Kiel: Schwerssche Buchandlung.

Poggio, G. F., Gonzalez, F., & Krause, F. (1988). Stereoscopic mechanisms in monkey visual cortex: Binocular correlation and disparity selectivity. *Journal of Neuroscience*, *8*(12), 4531–4550.

Qian, N. (1997). Binocular disparity and the perception of depth. *Neuron*, *18*(3), 359–368.

Qian, N., & Zhu, Y. (1997). Physiological computation of binocular disparity. *Vision Research*, *37*(13), 1811–1827.

Read, J. C. A., & Cumming, B. G. (2003). Testing quantitative models of binocular disparity selectivity in primary visual cortex. *Journal of Neurophysiology*, *90*(5), 2795–2817.

Read, J. C. A., & Cumming, B. G. (2007). Sensors for impossible stimuli may solve the stereo correspondence problem. *Nature Neuroscience*, *10*(10), 1322–1328.

Read, J. C. A., Parker, A. J., & Cumming, B. G. (2002). A simple model accounts for the response of disparity-tuned V1 neurons to anticorrelated images. *Visual Neuroscience*, *19*(6), 735–753.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24,* 1193–1216.

Skottun, B. C., De Valois, R. L., Grosof, D. H., Movshon, J. A., Albrecht, D. G., & Bonds, A. B. (1991). Classifying simple and complex cells on the basis of response modulation. *Vision Research*, *31*(7-8), 1079–1086.

Stevenson, S. B., Cormack, L. K., Schor, C. M., & Tyler, C. W. (1992). Disparity tuning in mechanisms of human stereopsis. *Vision Research*, *32*(9), 1685–1694.

Stockman, A., & Sharpe, L. T. (2000). The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision Research*, *40*(13), 1711–1737.

Tanabe, S., & Cumming, B. G. (2008). Mechanisms underlying the transformation of disparity signals from V1 to V2 in the macaque. *Journal of Neuroscience*, *28*(44), 11304–11314.

Tanabe, S., Haefner, R. M., & Cumming, B. G. (2011). Suppressive mechanisms in monkey V1 help to solve the stereo correspondence problem. *Journal of Neuroscience*, *31*(22), 8295–8305.

Thibos, L. N., Ye, M., Zhang, X., & Bradley, A. (1992). The chromatic eye: A new reduced-eye model of ocular chromatic aberration in humans. *Applied Optics, 31*(19), 3594–3600.

Tsao, D. Y., Conway, B. R., & Livingstone, M. S. (2003). Receptive fields of disparity-tuned simple cells in macaque V1. *Neuron*, *38*(1), 103–114.

Tyler, C. W. (1975). Spatial organization of binocular disparity sensitivity. *Vision Research*, *15*(5), 583–590.

Tyler, C. W., & Julesz, B. (1978). Binocular cross-correlation in time and space. *Vision Research*, *18*(1), 101–105.

Vlaskamp, B. N. S., Yoon, G., & Banks, M. S. (2011). Human stereopsis is not limited by the optics of the well-focused eye. *Journal of Neuroscience*, *31*(27), 9814–9818.

Williams, D. R. (1985). Visibility of interference fringes near the resolution limit. *Journal of the Optical Society of America A*, *2*(7), 1087–1093.

Wyszecki, G., & Stiles, W. (1982). *Color science: Concepts and methods, quantitative data and formulas*. New York: John Wiley & Sons.