Mechanisms of visual motion detection

Paul R. Schrater[1], David C. Knill[2] and Eero P. Simoncelli[3]

[1] *Department of Psychology, University of Minnesota, N218 Elliott Hall, 75 E. River Dr., Minneapolis, Minnesota 55455, USA*

[2] *Department of Psychology, University of Pennsylvania, 3815 Walnut St., Philadelphia, Pennsylvania 19104, USA*

[3] *Center for Neural Science, New York University, 4 Washington Place, New York, New York 10003, USA*

Correspondence should be addressed to P.R.S. (schrater@eye.psych.umn.edu)

Visual motion is processed by neurons in primary visual cortex that are sensitive to spatial orientation and speed. Many models of local velocity computation are based on a second stage that pools the outputs of first-stage neurons selective for different orientations, but the nature of this pooling remains controversial. In a human psychophysical detection experiment, we found near-perfect summation of image energy when it was distributed uniformly across all orientations, but poor summation when it was concentrated in specific orientation bands. The data are consistent with a model that integrates uniformly over all orientations, even when this strategy is sub-optimal.

When a viewer moves relative to the environment, the visual image projected onto the retina changes accordingly. Within small regions of the retina and for short durations, this motion is commonly approximated as a two-dimensional translation. The field of velocity vectors associated with each such region is referred to as 'optic flow'. Physiological and psychophysical experiments demonstrate that mammalian cortex uses mechanisms sensitive to local image motion[1]. These mechanisms are generally considered to form the neural substrate for representing optic flow.

Neurons in primary visual cortex perform the first stage of cortical motion processing. They are selective for both the spatial orientation and speed of the visual input within a spatially localized region. Because of their orientation specificity, however, these neurons are impaired by an ambiguity commonly known as the 'aperture problem': each neuron can signal the speed only of motion perpendicular to the orientation to which it is tuned[2]; thus, it is insensitive to the velocity component parallel to this orientation. It is suggested that a second stage of processing, commonly associated with visual area MT, computes an unambiguous representation of local pattern velocity by selectively combining the outputs of the first-stage detectors[2–5]. Although such an integration stage seems largely consistent with psychophysical and physiological data, the precise form of the combination rule remains unclear. Here we describe psychophysical experiments designed to test the predictions of several general combination rules.

The simplest combination strategy is to integrate responses of all those first-stage mechanisms consistent with a particular two-dimensional velocity[2]. It is convenient to describe this construction in the three-dimensional spatiotemporal frequency domain, where the spectral energy of a rigidly translating image is concentrated on an oblique plane (**Fig. 1a** and **b**). The orientation of this plane uniquely specifies the translational velocity (both speed and direction) of the luminance pattern[6]. First-stage motion detectors in the mammalian visual system can be described as computing the spectral energy within limited bands of spatiotemporal frequency[7]. Thus, a pattern-velocity detector may be constructed by summing the weighted outputs of first-stage mechanisms tuned to spatiotemporal frequency bands that lie on a common plane (**Fig. 1c**)[8–10]. The population response of a family of detectors built in this way can determine the presence and velocity of local translating patterns, and the responses of individual detectors are well matched to the behaviors of a subset of neurons in visual area MT[10].

Alternatively, the visual system could use an adaptive integration rule, selectively combining only those first-stage detectors that are tuned to the spatial structure of the image. For example, one type of model makes initial robust estimates of the components of a pattern by selecting those detectors responding to the stimulus but ignoring those responding to noise[11]. These one-dimensional estimates are then combined to produce an estimate of the two-dimensional pattern velocity that is most consistent with the measured one-dimensional velocity components. In general, such adaptive pooling rules produce more efficient motion detectors than fixed pooling rules because the detector is better matched to the signal. Human observers adapt their spatial pooling to improve the detection of static images[12]. Thus, it is plausible that they may do the same for moving images.

We designed a set of psychophysical detection experiments to rigorously test the predictions of three models for pooling of spectral energy: (1) a planar power detector that sums the signal energy over all orientations in a fixed planar region of spatiotemporal frequency, (2) an adaptive planar power detector that sums energy only over those regions of a plane containing the signal and (3) an adaptive unrestricted power detector that can sum energy over arbitrary regions of spatiotemporal frequency. Experimental stimuli were constructed by summing dynamic random noise patterns that were band-limited using spatiotemporally oriented filters analogous to those used to describe first-stage motion mechanisms. Each of the three models makes distinct subthreshold summation predictions for the detection of such stochastic signals embedded in white noise. The planar power detector predicts optimal summation performance when signal energy is distributed uniformly over all spatial orientations in a plane. However, this model predicts suboptimal performance when the signal energy is concentrated in a subset of planar orientation

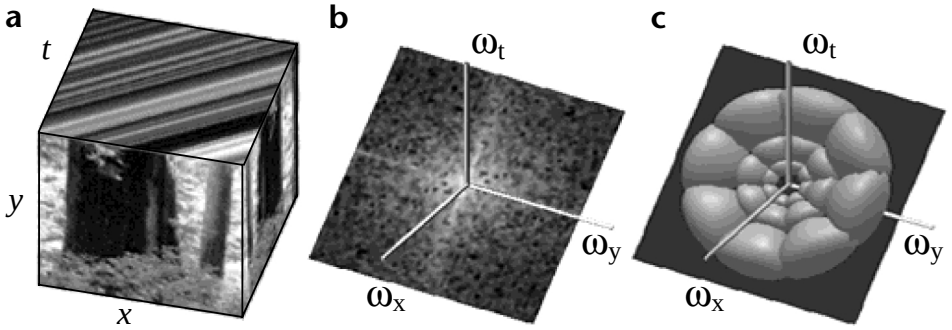

Fig. 1. A translational motion detector. (**a**) Space-time luminance pattern of an image translating to the right. This is a representation of the intensity information in the retinal image (the *x*–*y* plane) over time (*t*). The rightward motion can be inferred from the oriented pattern on the *x*–*t* face. (**b**) The Fourier amplitude spectrum of the luminance pattern, represented by the intensity of points in a three-dimensional spatiotemporal frequency domain. Non-zero Fourier amplitudes are constrained to lie on a plane through the origin. The orientation of this plane uniquely specifies the direction and speed of translation. (**c**) Construction of a translation detector[10], illustrated in the Fourier domain. Pairs of balls symmetric about the origin indicate the Fourier amplitude spectra of band-pass filters whose peak frequencies lie in the plane. A translation detector can be constructed by summing the squared outputs of such filters.

bands or distributed over non-coplanar regions of spatiotemporal frequency, because the model sums both signal and noise in these cases. Similarly, the adaptive planar power detector predicts optimal performance for any signal whose energy is concentrated in a plane, regardless of orientation content, but sub-optimal performance for stimuli composed of non-coplanar distributions of signal energy. Finally, the adaptive unrestricted power detector predicts optimal performance for any distribution of signal energy, with no improvement in performance for planar distributions.

Results

In the first experiment, we tested these predictions by measuring detection thresholds for three types of spatially localized stochastic signals embedded in white noise. Signals were generated by filtering spatiotemporally white noise with different sets of band-pass filters. 'Component' stimuli (**Fig. 2a**) had spectral energy confined to a single pass-band. 'Plaid' stimuli (**Fig. 2b**) had two component bands on the same plane. 'Planar' stimuli (**Fig. 2c**) were constructed using a set of component filters arranged in a ring on an oblique plane (so that signal energy was uniformly distributed over an annulus in the plane). A two-alternative forced choice task was used to measure detection thresholds for each of the three stimuli. Stimuli were presented in two consecutive intervals: one contained the signal added to broadband white noise and the other only white noise. Subjects were asked to judge which interval contained the signal. Thresholds were estimated by fitting a psychometric function to the data and selecting the contrast energy producing 81.1% detectability. Different stimulus types were presented in distinct experimental sessions, and subjects were told which stimulus type was being used. Before data collection, subjects were given several hours practice for each stimulus type over the course of several days, until performance saturated (see Methods).

To compare detection performance across stimuli, we assumed that performance for the component stimulus reflects the detection efficiency of a band-limited, first-stage motion-energy detector mechanism. (The component signal has a bandwidth similar to psychophysically optimal bandwidths[13]). Specifically, we assumed a model in which subjects made detection judgments for the component stimulus based on output of a filter matched to the component signal; this output was then corrupted by additive internal noise. The internal noise represented the added uncertainty created by neural noise, filter/signal mismatch and central-decision noise. We estimated internal noise by comparing subjects' thresholds with that of an ideal observer for the component stimulus. Assuming approximately constant internal noise across all stimulus conditions, we derived optimal summation predictions for the plaid and planar stimuli. These predictions reflected the performance of ideal observers that sum energy over only the spatiotemporal frequency bands containing the signal in a stimulus and were corrupted by a constant level of internal noise. (Such a model gives the commonly cited 'square-root law' for contrast summation.)

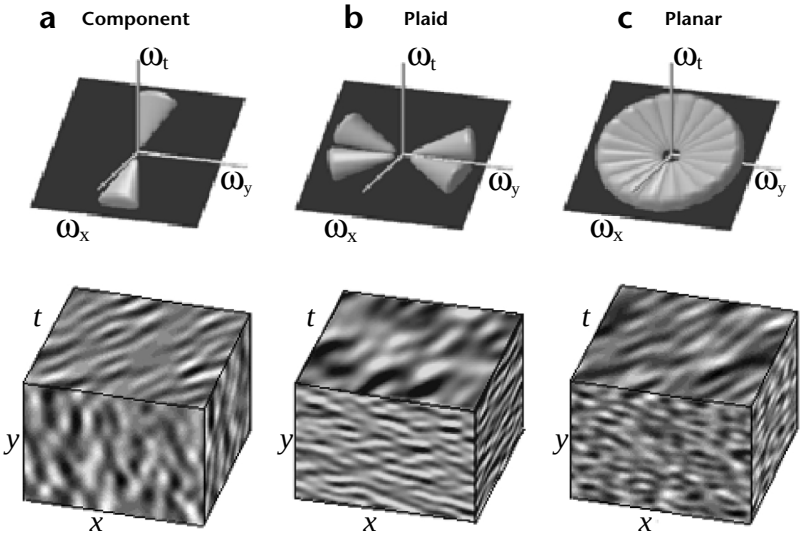

Fig. 2. Filter sets and examples of their corresponding signals. Top row, level surfaces (65% of peak response) of the three different filter sets used to generate stimuli. Bottom row, space–time luminance patterns of signals produced by passing spatiotemporal Gaussian white noise through the corresponding filter sets. (**a**) The 'component' stimulus, constructed from a spatially and temporally band-pass filter. The *x*–*y* face of the stimulus shows structures that are spatially band-pass and oriented along the *x*-axis. The orientation of structures on the *x*–*t* face indicates rightward motion. (**b**) The 'plaid' stimulus, constructed from two 'component' filters lying in a common plane. The *x*–*y* face of the stimulus shows a mixture of spatial structures with dominant orientations close to the *y*-axis. (**c**) The 'planar' stimulus, constructed from a set of ten 'component' filters lying in a common plane. The stimulus is spatially band-passed and isotropic (*x*-*y* face) and moves rightward (*x*–*t* face).

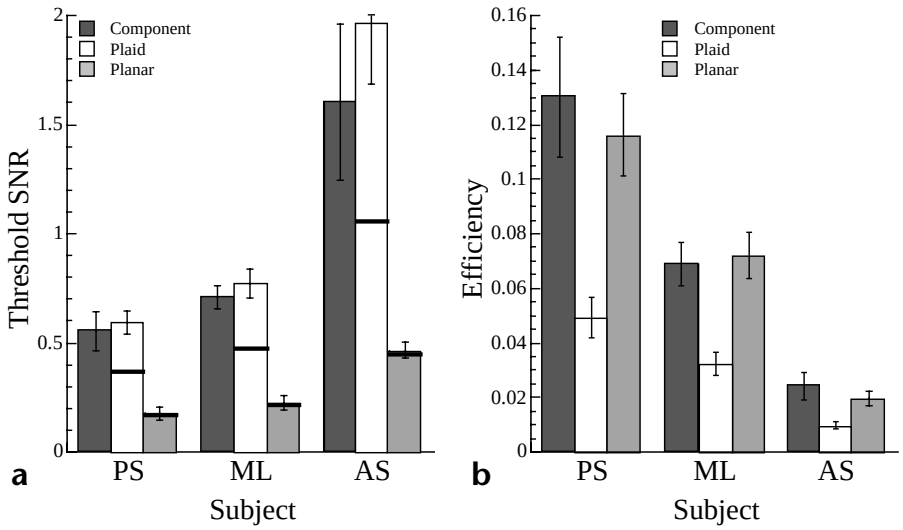

Fig. 3. Detection thresholds. (**a**) Performance of three subjects for the three stochastic signals of Fig. 2. Threshold signal-to-noise ratio (SNR) for 81.1% detectability. SNR is calculated as the ratio of the signal power to the background-noise power. Heavy black lines indicate predictions for ideal summation, derived from the component condition thresholds. (**b**) Detection efficiencies for the three stimulus types. Efficiencies are plotted in proportions, with 1.0 reflecting perfect performance (matching that of an ideal observer tuned to the signal structure for that particular stimulus). The differences between the efficiencies of the pattern stimuli (plaid and planar stimuli) and the component stimulus provide a quantitative measure of summation of the pattern's components.

Figure 3a shows subjects' detection thresholds for the component, plaid and planar stimuli. The heavy black lines on the plaid and planar plots represent predictions of ideal summation derived from the component data. The plaid data show little or no summation, whereas thresholds for the planar stimulus fell well within the range predicted by perfect summation. Figure 3b shows an alternative characterization of the results. We plotted subjects' detection efficiencies for each of the three types of stimuli. Detection efficiencies were computed by comparing subjects' detection thresholds with those of ideal observers optimally tuned to the signals contained in the stimuli. (Note that the ideal observers are different for each of the three stimulus types.) Efficiency provides a measure of the fraction of stimulus information effectively used by subjects in performing the detection task. Because perfect summation is ideal for these experiments, efficiency is also a measure of summation. The significantly lower detection efficiencies for the plaid stimulus than for the component stimulus reflect poor summation for the plaid. Conversely, equal detection efficiencies for the planar stimulus and the component stimulus reflect nearly optimal summation for the planar stimulus.

Of the three detection models described above, the data are clearly most consistent with the planar power-detector model. The experiment, however, only compared summation performance for two types of 'pattern' stimuli. Motivated by the observation that the planar power-detector model predicts progressively improved summation when signal energy is distributed more broadly across spatial orientations in a plane, but not when the energy is distributed out of a plane, we ran a second experiment. We generated a new set of stochastic signals to test this prediction. Signals for the stimuli were created by passing spatiotemporal white noise through three configurations of filters (**Fig. 4a**). The first was a plaid signal, similar to the plaid used in experiment 1. The second was a 'planar triplet', created by adding a component band to the plaid, in the same plane as the plaid. The third was a non-planar triplet, created by adding a component band out of the plane of the plaid. Detection thresholds were measured using the same method described for experiment 1. The planar power-detector model predicts improved summation for the planar triplet relative to the plaid, but no improved summation for the non-planar triplet.

Detection thresholds for the planar triplet were, as predicted, lower than for the plaid, whereas there was no significant difference between those for plaid and non-planar-triplet stimuli (**Fig. 4b**). Figure 4c shows subjects' efficiencies for detecting each of the three signals used in the experiment. Detection efficiency was significantly better for the planar triplet than for the plaid, implying improved summation for the triplet stimulus (**Fig. 4c**). Whereas detection efficiencies for the planar triplet were substantially better than for the plaid stimulus, subjects' efficiencies for the planar triplet remained lower than for the planar stimulus in experiment 1 (**Fig. 3**). Again, this is consistent with the planar power-detector model, which predicts progressively better summation as signal energy is distributed more broadly over a plane in spatiotemporal frequency, attaining ideal summation only when the energy is distributed uniformly over the plane—a stimulus that best matches the putative detector.

Discussion

We can summarize the qualitative results of experiments 1 and 2 as follows: subthreshold summation of signal-contrast energy improved as the energy was distributed more and more broadly around different orientations in a plane in spatiotemporal frequency, but did not improve when the energy was distributed into non-planar regions of spatiotemporal frequency. The lack of summation for non-planar regions agrees with previous studies suggesting a lack of generic summation for components moving in opposite directions[14–16]. Stated more plainly, detection efficiencies improve as more orientations are added to a moving pattern, as long as the motion of those orientation components are consistent with the velocity of the pattern. Of the three detection models proposed in the introduction, these results are most consistent with the planar power-detector model.

We extended the analysis one step further by comparing performance of a particular planar power detector on the five different 'pattern' stimuli used in the experiments with that of the subjects. The detector optimally summed energy over the band of frequencies contained in the planar stimulus from experiment 1 (using a matched 'power' filter—see Methods). We assumed this detector's output to be corrupted by levels of internal noise estimated from subjects' detection thresholds for the component stimulus in experiment 1. Figure 5 shows the model predictions for the five pattern stimuli used. Given the assumptions built into the model concerning the exact spatiotemporal frequency band covered by the planar power detector, the match is sur-

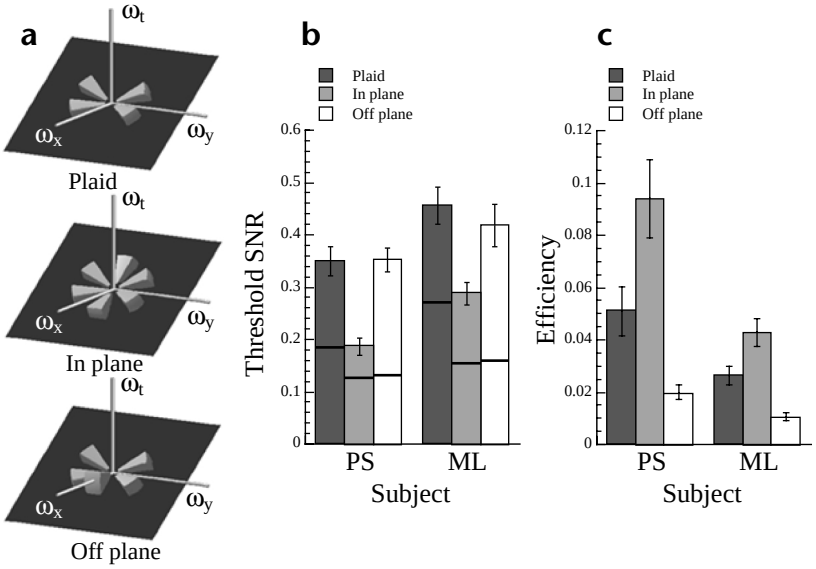

Fig. 4. Experiment 2. (**a**) Filters used to generate stimuli. (**b**) Threshold SNRs for detecting the three types of stimuli in. Black bands indicate the predictions of ideal summation, based on subjects' detection thresholds for the component stimulus used in experiment 1. (**c**) Detection efficiencies for the three stimulus types. The results show a large increase in efficiency for the in-plane triplet over the plaid and a decrease for the off-plane triplet, indicating increased summation with additional power on a common plane.

prisingly good. That is, not only do the qualitative results follow the predictions of the planar power-detector model, but the quantitative results are well fit by a pre-defined instantiation of the model (without fitting the parameters of the model to the data).

Our experiments critically test the predictions of a simple model for velocity-tuned motion detectors in the mammalian visual system. The model detectors sum outputs of early motion-energy detectors with different preferred orientations and spatiotemporal frequency bands that intersect a common plane. The result is a family of velocity-tuned detectors that measure the stimulus energy in fixed, orientationally broadband, planar regions of spatiotemporal frequency. Such mechanisms have been plausibly linked to the subclass of 'pattern' neurons in the middle temporal (MT) visual area[10]. The model may be contrasted with adaptive velocity-estimation models dynamically tuned to the spatial-orientation content of the image. Our subthreshold summation results favor the fixed pooling model over adaptive models.

As with all psychophysical studies, we had to make certain assumptions to draw conclusions about the underlying neural mechanisms. In particular, we assumed that detection of single component signals is mediated by responses of first-stage energy mechanisms, but that detection of compound signals is mediated by second-stage responses. In addition, we assume that observers can learn (through training) to attend to the most sensitive mechanism for each stimulus.

Although our experiments strongly argue for second-stage mechanisms that pool over orientation, further processing is required to account for the full range of motion-perception phenomena. For example, although planar pooling models are capable of representing more than one motion in a given local region[10], they are unable to provide a consistent account of whether stimuli appear to move coherently. Additional computations are necessary to account for such phenomena. Athough our experimental results suggest fixed pooling over orientation, they do not preclude the possibility of adaptive integration over spatial position, time or spatial frequency (scale). Adaptive integration over spatial position, for example, would be a useful strategy for preventing combination of information across occlusion boundaries. The experiments described herein may be naturally extended to examine the integration of first-stage mechanisms over these other domains.

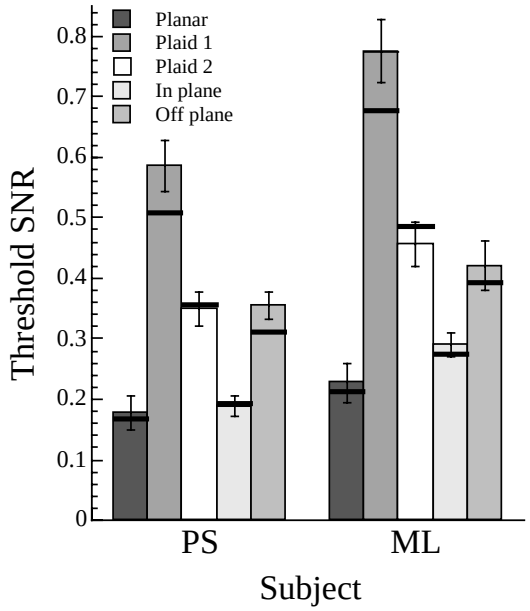

Fig. 5. Threshold SNRs for detecting the five types of pattern stimuli replotted from experiments 1 and 2, where Plaid 1 in the legend denotes the plaid from the first experiment and Plaid 2 from the second. Plaid 1 differs from Plaid 2 in that its energy is more diffusely spread over frequency. Black bands indicate the predictions of a planar filter, based on subjects' detection thresholds for the component stimulus used in experiment 1.

Methods

Stimuli in the experiment subtended 2.2° of visual angle, had a duration of 0.43 s and were displayed in 12-bit precision on a calibrated gamma-corrected gray-scale monitor. Observers fixated a small black dot 0.2° above the stimulus and were given several hours practice for each stimulus type over the course of several days until performance saturated. Observers were given auditory feedback after every trial in all experimental sessions. Only one signal type was used in a given 1-h experimental session. Average signal energy was varied using the method of constant stimuli. The average background noise energy was fixed at 0.26 degrees$^2 \cdot$ s.

Thresholds were estimated from the maximum-likelihood fits of two-parameter Weibull functions, for which the threshold parameter α is a measure of signal-to-noise power at 81.1% correct for a 2AFC task. Standard errors were computed both from

the Hessian of the likelihood function of the fit and using a Monte Carlo sampling method.

The component filter has an amplitude spectrum given by $C(\omega_r, \omega_\theta, \omega_\phi) = R(\omega_r)\,|\cos(\omega_\theta)|^9\,|\cos(\omega_\phi - \omega_{\phi_0})|^9$ in spherical frequency coordinates. The frequency radius ω_r is given by $\omega_r = \sqrt{\omega_x^2 + \omega_y^2 + (\omega_t/2.1)^2}$ (cycles per degree, cycles per degree, cycles per s). The elevation shift $\omega_{\phi_0} = 36.9°$ acts to rotate the filter out of the spatial frequency plane into a plane corresponding to a 1.93 degrees per s translation downward. The radial frequency function R is a smooth box function whose transitions from 0 to 1 are given by 1/2 cycle of a cosine function. This function has low/high cutoffs of (0.49, 7.6) cycles per degree and the cosine transition regions have widths of 1.45 cycles per degree. The plaid signal was constructed by adding two component filters, rotated by +70° and −70° within the common plane. The planar signal was constructed using a sum of 10 component filters, rotated by multiples of 18° and constrained to lie in a common plane. For the second experiment, the filters were modified to decrease the spectral spread. They have the form $BP(\omega_r, \omega_\theta, \omega_\phi) = W_r(\omega_r)W_\theta(\omega_\theta)W_\phi(\omega_\phi)$, where W_x is a smooth box function on the variable x. W_r had a transition region width of 1.45 and low/high frequency cutoffs of (0.49, 7.6) cycles per degree visual angle. W_θ and W_ϕ had transition widths of 8° and high/low cutoffs that spanned 36°. These filters were rotated to lie in the same positions as the component and plaid filters for the plaid and planar triplet. The off-plane non-planar triplet component, however, had an elevation shift corresponding to a slower 0.26 degrees per s translation downward.

Ideal observers for the task are matched filter power detectors[17]. Let $F(\omega_x, \omega_y, \omega_t)$ denote the spectrum of the signal normalized to one, and $N(\omega_x, \omega_y, \omega_t)$ the (complex) spectrum of the noise. Let $|X(\omega_x, \omega_y, \omega_t)|^2 = X(\omega_x, \omega_y, \omega_t) \cdot X(\omega_x, \omega_y, \omega_t)^*$ denote the inner product of the complex function with its complex conjugate. Then the spectrum of the received signal $|S(\omega_x, \omega_y, \omega_t)|^2$ is $s^2\,|F(\omega_x, \omega_y, \omega_t)|^2 + |N(\omega_x, \omega_y, \omega_t)|^2$ on the signal plus noise interval and $|N(\omega_x, \omega_y, \omega_t)|^2 = N^2$ on the noise-alone interval. The best possible performance can be achieved by computing the energy $E = \int |F(\omega_x, \omega_y, \omega_t)|^2\,|S(\omega_x, \omega_y, \omega_t)|^2\,d\omega_x\,d\omega_y\,d\omega_t$, over both signal plus noise and noise intervals and using the difference of these values as a decision variable[12]. In practice, the set of frequencies are quantized, and the integral is computed as a sum over 1.3×10^5 frequency samples.

Efficiencies are approximately given by the ratio of ideal to human squared energy thresholds. This approximation was corrected through simulations of the ideal, and the simulation values are shown in the figures. Performance of the matched filter depends on the probability, $p(E_s - E_n > 0)$, that the difference in energies measured by the filter on two intervals is greater than zero. To compute this probability, we derive an approximate expression for the distribution of $E_s - E_n$. For the stimuli used, the amplitude spectrum of the signal $S(\omega_x, \omega_y, \omega_t)$ consists of independent (real and imaginary) Gaussian-distributed random variables at each frequency with zero mean and variance given by either $s^2\,|F(\omega_x, \omega_y, \omega_t)|^2 + |N(\omega_x, \omega_y, \omega_t)|^2$ on the signal interval or $|N(\omega_x, \omega_y, \omega_t)|^2$ on the noise-alone interval. The power spectrum, as the product of the signal amplitude with its complex conjugate, consists of Chi-squared random variables with two degrees of freedom $\chi^2(2)$ at each frequency, which are scaled by the squared amplitude spectrum of the signal. Because the energy is the sum of an extremely large number of $\chi^2(2)$ random variables (one for each frequency), weighted by the product of the filter and the signal power spectrum, we can use the law of large numbers to generate an extremely good bound on performance (<< 1% error from true performance). By the law of large numbers, $E_s - E_n$ is Gaussian distributed, and hence can be characterized by its mean and variance. The mean and variance of $E_s - E_n$ are simply the sums of the means and variances of the $\chi^2(2)$ variables at each frequency. Computing the means and variances of the $\chi^2(2)$ variables that have been scaled by $|F(\omega_x, \omega_y, \omega_t)|^2 \cdot (s^2\,|F(\omega_x, \omega_y, \omega_t)|^2 + |N(\omega_x, \omega_y, \omega_t)|^2)$ on the signal interval and $|F(\omega_x, \omega_y, \omega_t)|^2 \cdot (|N(\omega_x, \omega_y, \omega_t)|^2)$ on the noise interval, and letting $F^m S^n = \Sigma\,|F(\omega_x, \omega_y, \omega_t)|^m\,|N(\omega_x, \omega_y, \omega_t)|^n$, then with a little algebra, the mean and variance $E_s - E_n$ are given by $\mu_{id} = 2s^2F^2F^2$ and $\sigma^2_{id} = 8(s^4F^4F^4 + 2\,s^2F^4F^2\,N + 2\,F^4N^2)$, respectively. Then $p(E_s - E_n > 0) = 1 - \Phi(0, \mu_{id}, \sigma^2_{id})$, where Φ is the cumulative Gaussian distribution.

To compute summation predictions, we modeled component performance as ideal but with the variance given by $\sigma^2_{id} + \sigma^2_{internal}$, so that $p(E_s - E_n > 0) = 1 - \Phi(0, \mu_{id}, \sigma^2_{id} + \sigma^2_{internal})$. Setting $p(E_s - E_n > 0) = 81.1$ and plugging in the subject's threshold into μ_{id} and σ^2_{id}, we solved for $\sigma^2_{internal}$. To generate predictions, the $\sigma^2_{internal}$ was then added to the ideal variance for each of the other signal types, and a predicted threshold for 81.1% was computed. Predictions for the planar model were generated in a similar manner. If we represent the planar filter amplitude spectrum as $P(\omega_x, \omega_y, \omega_t)$ then the planar model computes the energies $E_{pl} = \Sigma\,|P(\omega_x, \omega_y, \omega_t)|^2 \cdot |S(\omega_x, \omega_y, \omega_t)|^2$ on both intervals. The decision variable mean and variance are then given by $\mu_{id} = 2s^2P^2F^2$ and $\sigma^2_{id} = 8(s^4P^4F^4 + 2s^2P^4F^2\,\boldsymbol{N} + 2P^4\boldsymbol{N}^2)$. Threshold predictions were generated from these expressions using the internal noise estimated from the observer's component stimulus performance.

Acknowledgements

P.S. was supported by an NIH training grant, D.C.K. was supported by a grant from the NIH and E.P.S. was supported by a Sloan Fellowship, an NSF CAREER grant and the Sloan Program in Theoretical Neurobiology at New York University.

Received 11 August; accepted 11 November 1999

1. Nakayama, K. Biological image motion processing: a review. *Vision Res.* **25**, 625–660 (1985).
2. Adelson, E. H. & Movshon, J. A. Phenomenal coherence of moving visual patterns. *Nature* **300**, 523–525 (1982).
3. Albright, T. D. Direction & orientation selectivity of neurons in visual area MT of the macaque. *J. Neurophysiol.* **52**, 1106–1130 (1984).
4. Movshon, J. A., Adelson, E. H., Gizzi, M. S. & Newsome, W. T. in *Experimental Brain Research Supplementum II: Pattern Recognition Mechanisms* (eds. Chagas, C., Gattass, R. & Gross, C.) 117–151 (Springer, New York, 1986).
5. Morgan, M. J. Spatial filtering precedes motion detection. *Nature* **355**, 344–346 (1992).
6. Watson, A. B. & Ahumada, A. J. in *Motion: Perception & Representation* (ed. Tsotsos, J. K.) 1–10 (Association for Computing Machinery, New York, 1983).
7. Adelson, E. H. & Bergen, J. R. Spatio-temporal energy models for the perception of motion. *J. Opt. Soc. Am. A* **2**, 284–299 (1985).
8. Heeger, D. J. Model for the extraction of image flow. *J. Opt. Soc. Am. A* **4**, 1455–1471 (1987).
9. Grzywacz, N. M. & Yuille, A. L. A model for the estimate of local image velocity by cells in the visual cortex. *Proc. R. Soc. Lond. B Biol. Sci.* **239**, 129–161 (1990).
10. Simoncelli, E. P. & Heeger, D. A model of neuronal responses in visual area MT. *Vision Res.* **38**, 743–761 (1998).
11. Derrington, A. M. & Suero, M. Motion of complex patterns is computed from the perceived motions of their components. *Vision Res.* **31**, 139–149 (1991).
12. Kersten, D. Statistical efficiency for the detection of visual noise. *Vision Res.* **27**, 1029–1043 (1987).
13. Watson, A. B. & Turano, K. The optimal motion stimulus. *Vision Res.* **35**, 325–336 (1994).
14. Levinson, E. & Sekuler, R. The independence of channels in human vision selective for direction of movement. *J. Physiol. (Lond.)* **250**, 347–366 (1975).
15. Watson, A. B., Thompson, P. G., Murphy, B. J. & Nachmias, J. Summation and discrimination of gratings moving in opposite directions. *Vision Res.* **20**, 341–347 (1980).
16. Dobkins, K. R. & Teller, D. Y. Infant contrast detectors are selective for direction of motion. *Vision Res.* **36**, 281–294 (1996).
17. van Trees, H. L. *Detection, Estimation, and Modulation Theory* Part III (Wiley, New York, 1971).