Behavioral/Systems/Cognitive

A Three-Dimensional Spatiotemporal Receptive Field Model Explains Responses of Area MT Neurons to Naturalistic Movies

Shinji Nishimoto[1] **and Jack L. Gallant**[1,2]

[1]Helen Wills Neuroscience Institute and [2]Department of Psychology, University of California, Berkeley, California 94720

Area MT has been an important target for studies of motion processing. However, previous neurophysiological studies of MT have used simple stimuli that do not contain many of the motion signals that occur during natural vision. In this study we sought to determine whether views of area MT neurons developed using simple stimuli can account for MT responses under more naturalistic conditions. We recorded responses from macaque area MT neurons during stimulation with naturalistic movies. We then used a quantitative modeling framework to discover which specific mechanisms best predict neuronal responses under these challenging conditions. We find that the simplest model that accurately predicts responses of MT neurons consists of a bank of V1-like filters, each followed by a compressive nonlinearity, a divisive nonlinearity, and linear pooling. Inspection of the fit models shows that the excitatory receptive fields of MT neurons tend to lie on a single plane within the three-dimensional spatiotemporal frequency domain, and suppressive receptive fields lie off this plane. However, most excitatory receptive fields form a partial ring in the plane and avoid low temporal frequencies. This receptive field organization ensures that most MT neurons are tuned for velocity but do not tend to respond to ambiguous static textures that are aligned with the direction of motion. In sum, MT responses to naturalistic movies are largely consistent with predictions based on simple stimuli. However, models fit using naturalistic stimuli reveal several novel properties of MT receptive fields that had not been shown in prior experiments.

Introduction

Area MT is an important site of motion processing that lies downstream from areas V1 and V2 (Felleman and Van Essen, 1991; Born and Bradley, 2005). Many studies have examined how MT neurons represent motion information, using synthetic stimuli such as bars (Albright, 1984; Okamoto et al., 1999), gratings (Movshon et al., 1985; Pack and Born, 2001; Perrone and Thiele, 2001), dots (Britten et al., 1993), and noise (Livingstone et al., 2001). Several influential models have been proposed to account for these neurophysiological findings (Simoncelli and Heeger, 1998; Rust et al., 2006; Bradley and Goyal, 2008).

The ultimate goal of visual neuroscience is to understand the neural mechanisms mediating normal vision. For this reason, it is generally agreed that models of visual processing should ultimately predict responses observed during natural vision (Rust and Movshon, 2005; Wu et al., 2006; Stanley, 2008). Do the neuronal models of area MT developed from experiments that used synthetic stimuli predict responses under more natural viewing conditions? The answer to this question is not known, because no neurophysiology study has yet reported data that reflect the full range of stimulus–response relationships that can occur during natural vision. Natural moving stimuli occupy a three-dimensional spatiotemporal frequency domain: two dimensions of space and one of time. Previous neurophysiological studies of MT have only focused on a subspace within the three-dimensional frequency domain: a one-dimensional ring (i.e., direction) (Movshon et al., 1985; Pack and Born, 2001; Smith et al., 2005; Rust et al., 2006; Majaj et al., 2007), a two-dimensional slice (Perrone and Thiele, 2001; Priebe et al., 2003), or a cylinder (Okamoto et al., 1999). When a model is constructed based on data in a restricted stimulus subspace, generalizing the model to naturalistic stimuli is an ill-posed problem that will inevitably involve untested assumptions.

Recent studies raise another concern: the way that neurons represent visual information might be different if measured using synthetic (e.g., white noise or grating) versus more naturalistic stimuli. Several groups have addressed this issue in area V1 (David and Gallant, 2005; Felsen et al., 2005; Sharpee et al., 2006). These studies found that while models developed using synthetic stimuli generally explained responses evoked by naturalistic stimuli, receptive fields observed using synthetic stimuli deviated systematically from those observed using naturalistic stimuli. These deviations suggest that neurons possess nonlinear mechanisms that depend on stimulus statistics. Given the stimulus-dependent deviations found in V1, it is possible that some

Received Dec. 30, 2010; revised Aug. 10, 2011; accepted Aug. 13, 2011.

Author contributions: S.N. and J.L.G. designed research; S.N. performed research; S.N. contributed unpublished reagents/analytic tools; S.N. analyzed data; S.N. and J.L.G. wrote the paper.

This work was supported by grants from the National Eye Institute and the National Institute of Mental Health (J.L.G.). James Mazer wrote the neurophysiology software suite, and Stephen David wrote the database software. We thank Kendrick Kay, Thomas Naselaris, An Vu, and Liberty Hamilton for comments on this manuscript. We also thank Michael Oliver, Ryan Prenger, Michael Wu, and Ben Willmore for their help and for fruitful discussions.

The authors declare no competing financial interests.

Correspondence should be addressed to Jack L. Gallant, University of California at Berkeley, 3210 Tolman Hall, #1650, Berkeley, CA 94720. E-mail: gallant@berkeley.edu.

DOI:10.1523/JNEUROSCI.6801-10.2011

 0270-6474/11/3114551-14$15.00/0

properties of MT neurons might differ depending on whether they are measured using synthetic stimuli versus under more naturalistic conditions.

Here we addressed this issue by using naturalistic motion-enhanced movies to characterize receptive properties of macaque area MT neurons. These movies allowed us to probe the full three-dimensional frequency domain within the time constraints of neurophysiological experiments. We used a quantitative modeling approach to characterize responses of single MT neurons to these movies. We compared recovered receptive fields with theoretical predictions and with the results of previous studies that used synthetic stimuli.

Materials and Methods

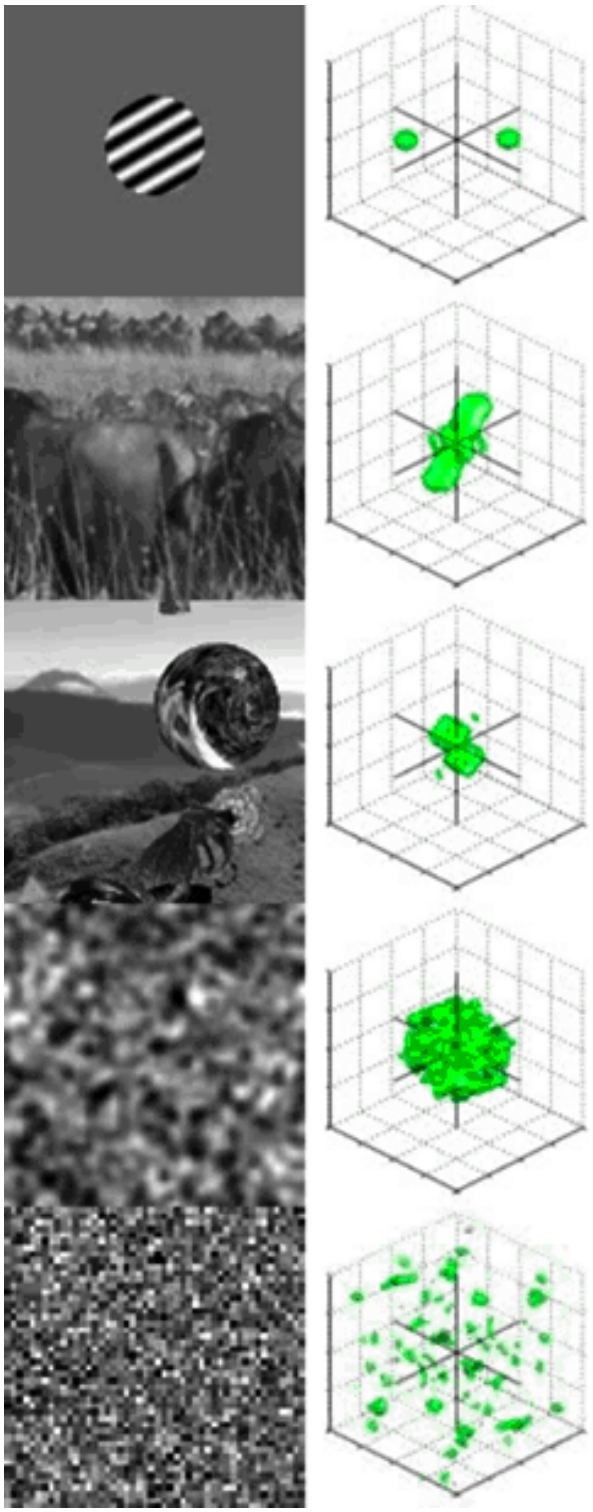
Physiology and behavioral tasks. Extracellular single-unit recordings were made from two adult male macaques (*Macaca mulatta*), prepared for recording as described previously (Mazer and Gallant, 2003). Recordings were made with epoxy-coated tungsten electrodes (FHC). Signals were amplified, bandpass filtered, and sorted (Plexon Instruments) to identify single units. Area MT was located by exterior cranial landmarks, anatomical images from magnetic resonance imaging (MRI), and/or physiological properties. During recordings, subjects performed a fixation task for liquid reward. Eye movements were monitored by an infrared eye tracker at either 250 or 500 Hz (EyeLink II; SR Research). All animal procedures were approved by the Animal Care and Use Committees at the University of California, Berkeley, and met or exceeded all NIH and U.S. Department of Agriculture regulations.

Visual stimuli. The primary stimuli consisted of motion-enhanced natural movies (see Fig. 1; Movie 1) constructed by combining full-screen natural movies (background) with an overlay of textured, moving three-dimensional objects (foreground). The movies were obtained from high-definition natural movie libraries provided by the Cornell Laboratory of Ornithology or the BBC. The moving objects consisted of cubes, spheres, and animal shapes, synthesized using a three-dimensional rendering library (Panda3D by Disney and Carnegie Melon University). The object textures were static natural images obtained from the McGill Calibrated Color Image Database (Olmos and Kingdom, 2004). The objects moved around a virtual three-dimensional space, and their accelerations for spatial position and rotation angles were updated by random walk. Movies were converted to grayscale by taking the average luminance across the three color channels. After the movies were constructed and concatenated into a single sequence, simulated saccadic eye movements were introduced by cutting the movies into sequences of 350 ± 50 ms and shuffling the order of the segments (David et al., 2004). (Note that this scheme provides a synchronous scene cut of both foreground and background.)

To ensure that the motion-enhanced natural movies did not produce biased receptive field estimates, some recordings were made with natural movies that contained no motion enhancement. These movies were constructed exactly as described above, but without the addition of rendered foreground objects.

The display device was a Sony Trinitron CRT monitor with a refresh rate of 83 Hz and a display resolution of 640×480 pixels (36×27 degrees of visual field at the viewing distance of 57 cm). Stimuli were presented while subjects fixated on a small spot ($<0.1°$) for 3–5 s per trial (1° diameter fixation window). After each successful fixation, there was a 200 ms delay (neutral gray background, 60 cd/m^2) followed in turn by the movie. The first 220 ms of the movie shown on each trial overlapped with the final frames of the movies shown on the previous trial. This permitted us to reduce the effects of the initial transient response during receptive field estimation by removing the associated responses before analysis (David et al., 2004).

Recordings were made from 52 single area MT neurons while showing a total of 20,000–40,000 frames of motion-enhanced natural movies (mean number of frames, 27,120). These training data were used to estimate the spectral receptive field for each neuron. Each neuron was also probed with a different motion-enhanced natural movie, 2000 frames in length, and repeated 5–20 times. These validation data were



Movie 1. Comparison of the three-dimensional amplitude spectra for several different stimulus classes. The left column shows several different classes of stimuli that have been used in visual neurophysiology experiments. From the top to the bottom, these are grating sequences, natural movies, motion-enhanced natural movies, pink noise, and white noise. The right column shows the three-dimensional spectral amplitudes of these stimuli. The green blobs delineate the isosurfaces for the spectral amplitudes. There are clear differences in the amplitude spectra of these stimuli. The amplitude spectrum of gratings is sparse, while the spectrum of white noise is dense. Natural movies and motion-enhanced natural movies have very similar $1/f$ amplitude spectra, but because the motion-enhanced natural movies are biased toward higher temporal frequencies, they span the frequency space more efficiently. The amplitude spectrum of pink noise is also $1/f$, but in other respects the spectrum of pink noise is quite different from that of natural movies.

used to evaluate prediction accuracy. The number of spikes obtained from a single neuron in the training data was 8413 on average, and the number of spikes in the validation data was 6108 on average. To estimate a single response time course from the repeated validation data, we first averaged the responses across repeats and then applied a 12 ms temporal Gaussian filter.

As a control, on a subset of 15 MT neurons we collected additional validation data using 2000 frames of natural movies without motion enhancement. Each movie was repeated 5–20 times. The number of spikes in the natural movie data set was 5266.

Notes on stimulus design. The stimuli used in this study were designed with three important constraints in mind. First, because motion information in natural stimuli can only be defined by three-dimensional changes of luminance patterns, the stimulus set should cover the full three-dimensional frequency domain (two for space, one for time). Second, because neurophysiological recordings from single neurons tend to be data limited, the stimuli should facilitate efficient estimation of receptive field properties. Third, because stimulus statistics might affect receptive field properties, the stimuli should be naturalistic. Note that "natural stimuli" do not belong to a discrete category. Rather, naturalness is a continuum. A very natural stimulus would possess a natural dynamic luminance range, natural color distribution, binocular disparity, and so on. It would also reflect the influence of the observer's eye movements and motion through the environment. A very unnatural stimulus would be white noise or gratings. The stimuli used in any neurophysiological experiment lies somewhere on this continuum.

The motion-enhanced natural movies used here satisfy all three of these constraints. They span the full three-dimensional frequency domain. They contain naturalistic texture and naturally structured motion (i.e., rotation, expansion, contraction, and translations). They tend to reduce the spatial and temporal correlations found in natural movies, thereby making it easier to correct estimated spectral receptive fields for stimulus bias.

Movie 1 shows the three-dimensional spectrum of the drifting gratings, natural stimuli, motion-enhanced movies, $1/f$ noise, and white noise. Drifting gratings are very unnatural and do not sample the three-dimensional frequency domain efficiently. Natural movies have natural three-dimensional spectrum, but they do not contain much motion information and so are likely to be an inefficient stimulus for characterizing area MT neurons. Motion-enhanced movies retain the second-order $1/f$ amplitude spectrum characteristic of natural movies, but contain substantially more motion information. White noise has a very different spectrum from natural movies, and $1/f$ noise has no statistical structure beyond second order.

Model estimation. It is difficult in principle to model the nonlinear relationship between stimulus and response in visual neurons (Wu et al., 2006). Estimation of receptive field properties for these nonlinear neurons requires an optimization method that can search through a complex error surface without becoming stuck in a local minimum. One way to solve this problem is to nonlinearly transform the stimulus into a new space in which the relationship between the transformed stimulus and the response is linear. In this case, linear optimization methods can be used to find the optimal weights that map between the nonlinearly transformed stimulus and measured responses. In this study we used a V1 filter bank (see below, The V1 filter bank and Tests for nonlinearities) to perform the nonlinear transformation, and we used regularized linear regression to find the optimal weights (see below, Regression by boosting). Note that a similar approach (i.e., using linear combinations of nonlinear local spectral measurements) was used in previous studies to characterize receptive fields of neurons in early visual (Nishimoto et al., 2006) and auditory areas (Theunissen et al., 2000).

The V1 filter bank. The bank of V1 filters chosen to represent each MT neuron were selected from a Gabor basis. The Gabor basis consisted of several thousand individual Gabor filters (see below), each defined as follows:

$$G_{i,p}(x,y,t) = \exp\left(-\frac{(x-cx_i)^2+(y-cy_i)^2}{2ws_i^2}-\frac{(t-ct_i)^2}{2wt_i^2}\right) * \sin((x-cx_i)*fx_i+(y-cy_i)*fy_i+(t-ct_i)ft_i+p), \quad (1)$$

where fx_i, fy_i, and ft_i represent the spatial and temporal frequency; cx_i, cy_i, and ct_i give the center of each Gabor filter in each dimension of the space-time domain; ws_i and wt_i give the width of the Gaussian envelope in space and time; and p gives phase.

The process of filtering each movie $I(x,y,t)$ with these Gabor filters was modeled as linear multiplication:

$$L_{i,p}(t) = \sum_x \sum_y \sum_\tau G_{i,p}(x,y,\tau) I(x, y, t-\tau). \quad (2)$$

The V1 simple cell inputs were modeled as follows:

$$S_{i,p}(t) = HR[L_{i,p}(t)], \quad (3)$$

where $HR[*]$ is half-wave rectification, and $p = 0°$, 90°, 180°, and 270°.

The V1 complex cell inputs were modeled as follows:

$$C_i(t) = \sqrt{L_{i,0}^2 + L_{i,90}^2}. \quad (4)$$

Note that $S(t)$ and $C(t)$ are time series. In this study we call these (and their variants, described below) V1 filter outputs, and denote them as $X(t)$. A previous neurophysiological study showed that V1 afferents to area MT are predominantly direction-selective complex cells, not simple cells (Movshon and Newsome, 1996). Our modeling results confirm this finding: most of the response variance is captured by the model complex cells, and the model simple cells have only small effect on prediction performance (data not shown). However, to be sure that we would obtain the most accurate model possible for each neuron, we included both $S(t)$ and $C(t)$ in this study.

The entire bank of V1 filters used here consisted of 5956 basis functions spanning 12 different directions, five different spatial frequencies, and six different velocities. The spatial frequency of the filters was log distributed from zero to six cycles per classical receptive field (cRF). The temporal frequency was log distributed from 0 to 30 Hz. The filters were spatially tiled on to a two-dimensional Cartesian grid. Grid spacing was set separately at each scale to ensure that the grid width was proportional to the spatial width of the Gaussian envelope in Equation 1. Each adjacent pair of filters was separated by 2.2 σ of the Gaussian envelope. The size of Gaussian envelope was set proportional to the spatial frequency such that one cycle of the sine wave was two σ of the envelope. The overall spatial analysis window was set to two times the size of the classical receptive field. Our preliminary analysis showed that predictions were not improved when the spatial frequency of the simple cell filters increased beyond two cycles per cRF (data not shown). Therefore, to reduce the computational burden, these filters were limited to be no higher than two cycles per cRF.

Regression by boosting. Boosting with early stopping procedure (Friedman, 2001; David et al., 2007; Willmore et al., 2010) was used to model the relationship between V1 filter outputs and responses of each MT neuron. The procedure has the effect of shrinking the total sum of absolute weights (compared with the ordinary least squares regression). This suppresses small weights that cannot be estimated accurately with the data available. The procedure produces a robust fit even when the number of model parameters to be estimated is much larger than the number of data samples. Note that only the training data were used for fitting the model weights; the validation data were preserved for estimating model predictions.

The regression model was defined as follows (see Fig. 2):

$$Y(t) = \sum_i \sum_\tau X_i(t-\tau) W_i(\tau), \quad (5)$$

where $X(t)$ represents the V1 filter outputs given some input (i.e., a segment of a movie), $Y(t)$ is the predicted response, and $W(t)$ is a weight matrix containing linear weights between $X(t)$ and $Y(t)$. (Note that the weight matrix contains weights for correlation delays, τ, up to 10 frames or 130 ms.) According to this definition, fitting the model to the responses of each neuron is simply a matter of estimating the optimal weight matrix.

An iterative procedure was used to estimate the weight matrix as follows: (1) Set all the elements of the weight matrix $W(t)$ to 0. (2) Calculate gradient of the square error E between model and neural responses:

$$E = (Y(t) - r(t))^2, \quad (6)$$

$$\frac{\partial E}{\partial W} = \sum_t (Y(t) - r(t)) X_i(t - \tau). \quad (7)$$

(3) Identify the element with the steepest gradient:

$$(i_m, \tau_m) = \text{argmax}\left(\left| \frac{\partial E}{\partial W} \right| \right). \quad (8)$$

(4) Update the element in the weight matrix by a small step size ε:

$$W_{i_m}(\tau_m) \leftarrow W_{i_m}(\tau_m) - \text{sgn}\left(\frac{\partial E}{\partial W_{i_m,\tau_m}} \right) \varepsilon. \quad (9)$$

(5) Loop back to Step 2 until the termination criterion is met.

Early stopping with cross-validation was used to determine when to terminate boosting. On each iteration of the fitting procedure, 80% of the data from the training set were used to fit the model, and the remaining 20% of the training data were used to evaluate predictions. The boosting procedure was terminated when prediction errors on this training subset began to increase. To estimate parameters optimally this entire procedure was repeated five times, each time reserving a different 20% of the training data as a prediction subset. The final weight estimates were obtained by averaging across these five repetitions. Note that the validation data were never used in any aspect of the fitting procedure, but were only used to estimate prediction accuracy of the final model.

Tests for nonlinearities. To identify additional nonlinearities that might be critical for the model, we developed a switching framework that allowed us to compare several different nonlinear mechanisms directly (see Fig. 2). Each stage could be switched in or out of the circuit, and we exhaustively explored all possible models by parametrically varying which elements were included or excluded from the model. (The V1 filters were present in all cases and were never switched out of the circuit.) The switching framework included three kinds of nonlinearity: (1) luminance and contrast normalization placed before the V1 filters, (2) a static nonlinearity, and (3) divisive normalization.

The luminance and contrast normalization were implemented as a stimulus preprocessing stage interposed between the movie and the V1 filters:

$$I'(x,y,t) = \frac{I(x,y,t) - \text{Lum}(t)}{\text{Con}(t)}, \quad (10)$$

where the $I'(x,y,t)$ is the normalized luminance. The time course of luminance, $\text{Lum}(t)$ was defined as follows:

$$\text{Lum}(t) = \sum_x \sum_y I(x,y,t). \quad (11)$$

$\text{Con}(t)$, the time course of contrast, was defined as follows:

$$\text{Con}(t) = \sqrt{\sum_x \sum_y (I(x,y,t) - \text{Lum}(t))^2} \quad (12)$$

The static output nonlinearity was implemented as a half-wave rectification followed by a power function:

$$X_i'(t) = [X_i(t)]^\alpha \quad (13)$$

where Xi represents the output of the linearized Gabor filters, and Xi' represents the nonlinearly transformed output of the filter bank. Three values for α were used: half-wave rectification given by $\alpha = 1.0$, a compressive nonlinearity given by $\alpha = 0.5$, and an expansive nonlinearity given by $\alpha = 2.0$. Note that contrast responses for V1 neurons are often described using the Naka–Rushton equation (Albrecht and Hamilton, 1982). The Naka–Rushton equation can be linear, compressive, or expansive, depending on the range of stimulus contrast. Which form the contrast response of area MT neurons will take under naturalistic conditions is an open question that can be addressed using our modeling framework.

Divisive normalization was implemented as follows:

$$X_i'(t) = \frac{X_i(t)}{\sum_n X_n^{\text{norm}}(t) + \beta} \quad (14)$$

where $\sum_n X_n^{\text{norm}}(t)$ represents pooled responses of the V1 filters. The $X_n^{\text{norm}}(t)$ were prenormalized so that each of the nth filters had unit SD over the time course of output. (The second term in the denominator, β, is the semisaturation constant for normalization.)

Note that divisive normalization is not selective for any particular range of spatial positions or spatial or temporal frequencies. The suppressive effect is global, both spatially and spectrally. In contrast, the suppressive spectral receptive field (see Figs. 4, 5, blue blobs) is spatially and spectrally localized.

Relationship to other models. The MT neuron model developed here offers both a generalization and a simplification of models proposed previously. Our model is similar in many respects to those proposed in previous neurophysiological studies (Perrone and Thiele, 2001; Rust et al., 2006). However, those studies only probed receptive field properties within a one- or two-dimensional subspace of the full three-dimensional frequency domain [i.e., a two-dimensional slice in the study by Perrone and Thiele (2001) and a one-dimensional ring in the study by Rust et al. (2006)]. Because our model describes receptive field properties within the full three-dimensional spectral domain, it is more general than those proposed previously.

Our model produces receptive fields that are in many respects consistent with those proposed in other studies (Simoncelli and Heeger, 1998; Rust et al., 2006), even though the model requires fewer nonlinear mechanisms than those proposed previously. Simoncelli and Heeger (1998) proposed that rectification and normalization occur within area MT. Rust et al. (2006) proposed a directionally dependent normalization. In our preliminary modeling work (data not shown), we explored models with a second output nonlinearity located within MT (Simoncelli and Heeger, 1998; Rust et al., 2006). However, we found that the second output nonlinearity had no significant effect on predictions, so we discarded it to simplify the model. We also did not include a divisive normalization component within MT, because that component would require strong assumptions regarding the specific form of MT receptive fields (Simoncelli and Heeger, 1998).

A recent study suggested that there are local nonlinear interactions between the receptive subfields of area MT neurons (Majaj et al., 2007). We did not include such interactions in our modeling effort for two reasons. First, we were most interested in the organization of receptive field profiles within the three-dimensional frequency domain, and any interaction effects would not materially affect these estimates. Second, including nonlinear interaction terms between the constituent V1 filters would dramatically increase the number of parameters of the model and so make estimation much more difficult. (Note that the MT model used here contains ~6000 regression channels. Including all possible two-way interactions would require regressions of $\sim 6000^2$, or ~36,000,000 channels).

Our model is also limited in temporal respects. The movies used in this experiment were shown at a frame rate of 83 Hz. Therefore, we did not attempt to model responses at a time scale finer than 12 ms. For this reason, the model does not address subframe nonlinear responses (e.g., transients and bursts).

Optimal velocity plane. To compare estimated spectral receptive fields across the sample of area MT neurons, we computed two indices for each neuron: an on-plane index and a horizontal–vertical ratio index. Both these measures required estimating the optimal velocity plane, that is, the plane within the three-dimensional frequency domain that best fits the spectral amplitude distribution of the excitatory receptive field. The optimal velocity plane crosses the zero point and can be defined by its azimuth (direction) and elevation (speed). Note that

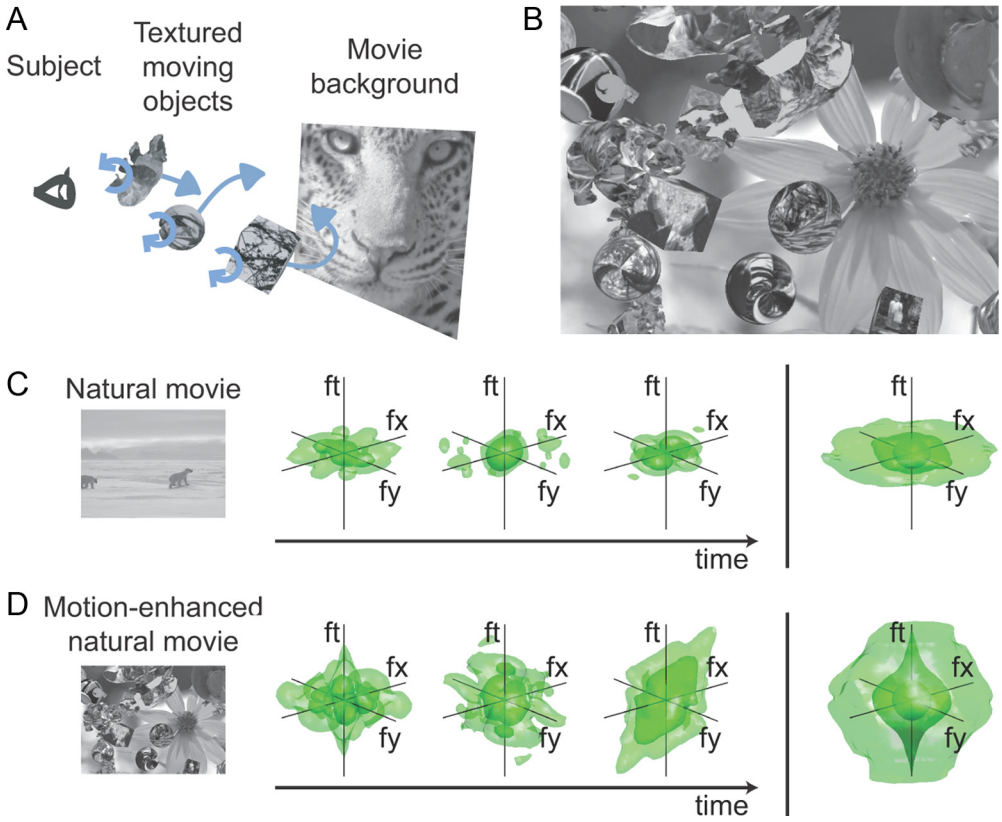


Figure 1. Spatial and spectral structure of motion-enhanced natural movies. ***A***, Schematic diagram of motion-enhanced natural movies. The movies are constructed by combining two distinct components: the background is a natural movie and the foreground contains several textured objects that move along random trajectories. The entire display is updated approximately three times per second to simulate the visual stimulation that would occur due to natural saccadic eye movements. The addition of these foreground objects increases high temporal frequency energy and decorrelates the stimulus, thereby increasing efficiency of receptive field estimation. ***B***, One typical frame of a motion-enhanced natural movie. ***C***, Three-dimensional frequency spectrum of natural movies. The leftmost column shows one frame from a natural movie. The middle columns show three snapshots of the three-dimensional amplitude spectrum of three specific frames from the movie, plotted in the three-dimensional spatiotemporal frequency domain. The rightmost column shows the average spectrum for a long movie. Three isospectral surfaces are plotted (1, 4, and 16% of the maximum) to facilitate visualization. The amplitude spectrum follows a $1/f$ distribution in both space and time. ***D***, Three-dimensional amplitude spectrum for motion-enhanced natural movies. The format is the same as in ***C***. The amplitude spectrum of motion-enhanced natural movies has relatively more energy at high temporal frequencies than is found in natural movies but is otherwise similar to the amplitude spectrum found in natural movies.

the spectrum of any image that translates in a fixed direction and at constant speed will lie on a plane (Watson and Ahumada, 1985; Simoncelli and Heeger, 1998).

To find the azimuth and elevation of the optimal plane for each neuron, we introduced two constraints. The maximum coverage constraint identified the plane that had the maximal coverage of excitatory components on and near the plane. This was found by summing V1 filter weights whose temporal frequency was within ± 1 octave or ± 5 Hz from the plane (whichever was largest). The symmetry constraint identified the plane at the optimal direction. This was found by summing the V1 filter weights separately on the two sides of the azimuth of the plane, subtracting these quantities and taking the negative. [The symmetry constraint was important for neurons such as the one shown in Fig. 4*B*, where the maximal coverage constraint alone would not produce a unique optimal velocity plane (see also Fig. 10 and Discussion).] We manually balanced the importance of the two constraints to produce the most stable estimates of the optimal plane across the neurons in our sample.

Null hypotheses test for on-plane ratio. To determine statistical significance of the on-plane ratio index, we used Monte Carlo simulation to obtain the null distribution. First, 150 model area MT neurons were constructed by randomly assigning weights from a normal distribution with mean 0 and SD 1 (arbitrary units). Second, these model neurons were used to filter the motion-enhanced natural movies that had been used as stimuli in the experiment. Poisson noise was added to the model responses at this stage. Third, spectral receptive fields were estimated for each of the model neurons, using the same techniques described above. Finally, the on-plane ratio index was calculated for each model neuron. These ratios constituted the null comparison distribution. A Wilcoxon rank-sum test was used to assess statistical significance.

Results

We recorded from 52 area MT neurons in two animals while they performed a simple fixation task. During recording, each MT neuron was stimulated with 22,000 to 42,000 frames of motion-enhanced natural movies (Fig. 1; Movie 1). To ensure that the motion-enhanced natural movies did not bias estimated receptive field models, additional recordings were made from a subset of 15 MT neurons using 2000 frames of natural movies without motion enhancement. The data acquired from each neuron were split into two parts: the first part was used to fit receptive field models, and the second was used to evaluate model predictions. A modeling framework based on nonlinear system identification (David et al., 2004; Nishimoto et al., 2006; Wu et al., 2006; Willmore et al., 2010) was used to fit several quantitative computational receptive field models to the data recorded from each MT neuron. To facilitate interpretation and comparison of tuning properties, all receptive fields were visualized in the three-dimensional spatiotemporal frequency domain. Here we refer to these as spectral receptive fields.

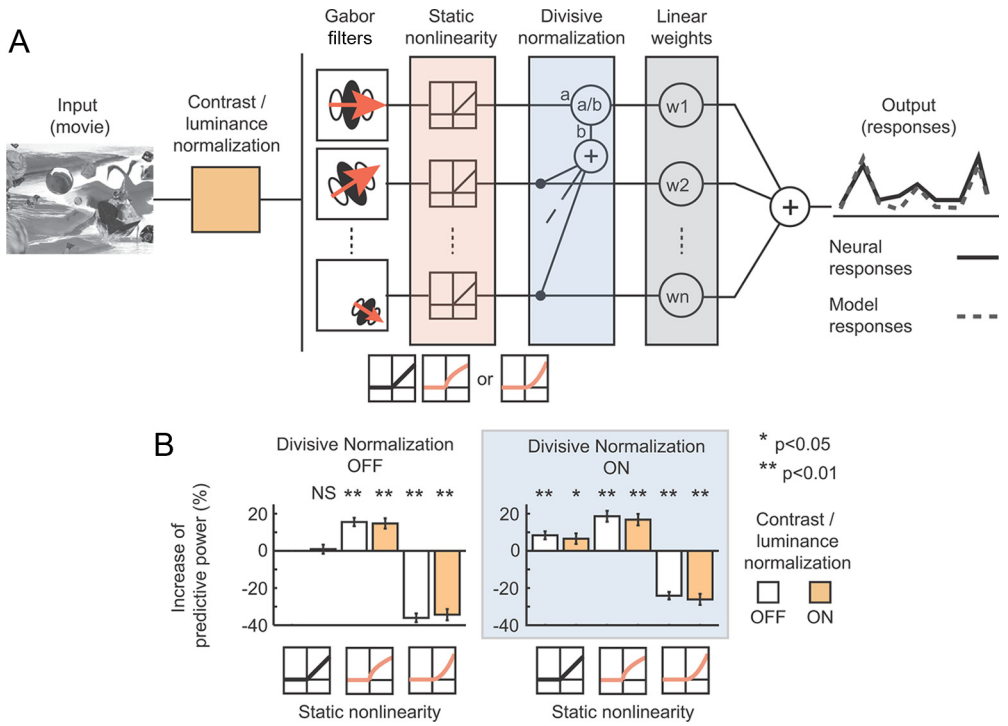


Figure 2. Analysis of MT neurons using the switched model. ***A***, The switched model framework used to describe each MT neuron. The complete model consists of several linear and nonlinear filtering stages. Incoming images are first transformed by a nonlinear contrast and luminance normalization stage (orange square). These signals are fed into a bank of simple and complex type V1 filters (schematic Gabor filters). The output of each V1 filter is rectified with a static linear, compressive, or expansive nonlinearity with half-wave rectification (pink rectangle). These signals are fed in turn into a divisive normalization stage (blue rectangle). Finally, the results are summed linearly according to weights estimated by regularized linear regression (gray rectangle). ***B***, To determine which nonlinear filtering components improve model predictions, we systematically switched each of the nonlinear processing stages shown in ***A*** in and out of the model and compared the predictions of each model. A total of 12 different models [i.e., 2 (with or without luminance and contrast normalization) × 3 (three types of static nonlinearity) × 2 (with or without divisive normalization) different switching conditions] were examined, and results were averaged across all 52 MT neurons in the sample. Each bar represents the average increase in predictive power relative to the simplest model examined. (The simplest model contains linear half-wave rectification, no luminance normalization, and no divisive normalization.) Results from models with or without the luminance and contrast normalization stage are represented as open or filled bars, respectively (see legend, right). The shapes of the three static nonlinearities tested here are shown below each bar. (The three static nonlinearities were linear, compressive, and expansive, all with half-wave rectification.) The left and right panels compare results from models without or with the divisive normalization stage, respectively. Note that the leftmost bar shows the comparison to itself (the simplest model) and thus is guaranteed to be zero. Error bars show bootstrap estimates of the SE ($n = 52$). Significance of prediction power relative to the simplest model is shown above each bar (*$p < 0.05$; **$p < 0.01$; Wilcoxon signed-rank test with Bonferroni correction; NS, not significant).

Linear and nonlinear mechanisms that predict natural visual responses in area MT

The most common framework for modeling single neurons in area MT is to describe each neuron in terms of its inputs from previous stages of visual processing (Simoncelli and Heeger, 1998; Rust et al., 2006; Bradley and Goyal, 2008). The simplest plausible model consists of two stages: a spatiotemporal Gabor filter that represents a pool of area V1 simple and complex neurons (Adelson and Bergen, 1985; Jones and Palmer, 1987), and a linear pooling mechanism that selectively integrates information from a specific subset of putative V1 inputs. This two-stage model provides a simple framework for describing MT neurons, but a complete functional model that can account for responses to naturalistic stimuli will likely require additional nonlinear mechanisms like those that have been reported at more peripheral stages of processing (Kaplan et al., 1987; Heeger, 1992a,b; Carandini et al., 1997; Mante et al., 2005; Bonin et al., 2006). To identify these critical nonlinearities efficiently, we developed a switched model framework that encompassed several different nonlinear mechanisms identified previously in area MT or at more peripheral stages of processing: luminance and contrast normalization (Kaplan et al., 1987; Bonin et al., 2006), a static nonlinearity (Heeger, 1992a), and divisive normalization (Heeger, 1992b; Carandini et al., 1997).

The prediction accuracy of each of the nonlinear models is summarized in Figure 2*B* (see Materials and Methods for details). Each bar in the histograms shows how well one specific model predicts responses in the validation data set, relative to the simplest model that includes only the V1 filtering stage with static rectification and no additional nonlinearities. Contrast normalization does not have a significant effect on predictions compared with the simplest model ($p > 0.10$; Wilcoxon signed-rank test with Bonferroni correction). However, the static output nonlinearity does have a significant effect. The compressive static output nonlinearity consistently increases predictions ($p < 0.01$), while the expansive nonlinearity always decreases predictions ($p < 0.01$). Divisive normalization also improves predictions significantly ($p < 0.01$). The model that contains both a compressive nonlinearity and divisive normalization has significantly more predictive power than a model that contains only a compressive nonlinearity ($p < 0.01$). Note, however, that the effects of the compressive and expansive nonlinearities do not depend on whether divisive normalization is present or not. Based on these results, we conclude that the simplest model of MT neurons that gives accurate predictions of responses to motion-enhanced natural movies consists of a bank of V1 filters, each followed by a compressive nonlinearity, a divisive nonlinearity, and a linear

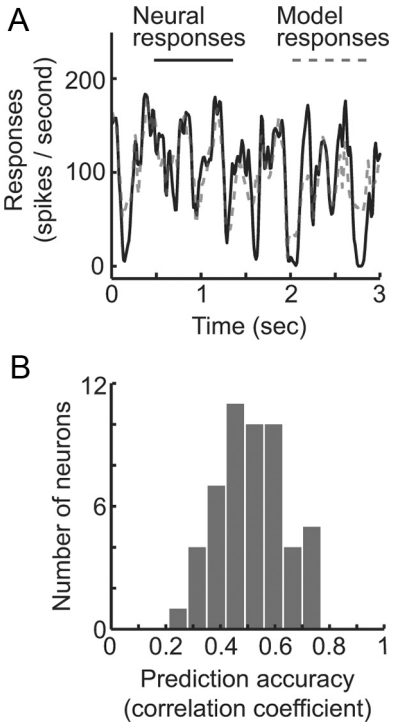


Figure 3. Prediction accuracy of the neuronal model used in this study. ***A***, Model predictions for one MT neuron. The horizontal axis indicates time, and the vertical axis the response rate. The solid curve gives the mean response observed over time (10 repetitions), and the dotted curve indicates predicted responses. For this neuron, the correlation between predicted and observed responses is $r = 0.72$. ***B***, Summary of prediction performance across the entire sample of 52 neurons. The average correlation is $r = 0.52$. The proportion of variance explained by the model is 35%, which is approximately comparable to the predictions obtained in previous studies of areas V1 and V2 from our laboratory (V1, 40%; V2, 30%) (David and Gallant 2005; Willmore et al., 2010).

pooling stage whose weights are determined uniquely for each neuron.

Figure 3*A* illustrates the predictions of the compressive and divisive V1 filter model fit to responses of one MT neuron. The correlation between model predictions and observed responses for this neuron is quite good ($r = 0.72$) considering the inherent variability of neural responses. Figure 3*B* summarizes prediction performance of the compressive/divisive V1 filter model across the entire sample of 52 MT neurons. The average correlation between model predictions and observed responses is $r = 0.52$, and the model accounts for 35% of the explainable response variance (David and Gallant, 2005). This is somewhat lower than the response variance that can be explained in LGN neurons (~60%) (Mante et al., 2008), but comparable to the response variance explained in V1 (~40%) (David and Gallant, 2005; Willmore et al., 2010) and in V2 (~30%) (Willmore et al., 2010). The prediction accuracy achieved here is remarkable given that predictions were estimated for time-varying responses, evoked by movies that were not used to fit the model.

Spectral receptive fields of area MT neurons

In a seminal theoretical paper, Simoncelli and Heeger (1998) hypothesized that the receptive fields of MT neurons should lie on a single plane within the three-dimensional frequency domain. Their reasoning was based on the fact that the three-dimensional power spectrum of an image translating at a fixed velocity will lie on a plane whose azimuth and elevation reflect the speed and direction of image motion (Watson and Ahumada, 1985; Simoncelli and Heeger, 1998; Bradley and Goyal, 2008). Thus, any neuron that has a planar receptive field in the three-dimensional frequency domain will be optimally tuned for one specific image velocity. Due to the spatiotemporal bandpass nature of the V1 inputs to MT, Simoncelli and Heeger (1998) predicted that spectral receptive fields of MT neurons would form a ring in the three-dimensional frequency domain. In the same paper, Simoncelli and Heeger (1998) also postulated that MT neurons might possess suppressive receptive fields that lie off the optimal excitatory velocity plane. These suppressive components would tend to sharpen velocity tuning.

The velocity plane tuning model proposed by Simoncelli and Heeger (1998) is consistent with many previous neurophysiological studies in area MT (Movshon et al., 1985; Rodman and Albright, 1987; Snowden et al., 1991; Britten et al., 1993; Perrone and Thiele, 2001), and with human psycophysical studies (Schrater and Simoncelli, 1998; Schrater et al., 2000). However, previous neurophysiological studies examined only a one- or a two-dimensional subspace within the full three-dimensional frequency domain (Okamoto et al., 1999; Perrone and Thiele, 2001; Priebe et al., 2003). Therefore, none of them provided direct evidence of tuning along the three-dimensional velocity plane (see also Discussion, Relationship to previous reports of speed-tuned neurons). Furthermore, no previous study has investigated three-dimensional suppressive tuning in area MT.

In this section, we present data that resolve both of these long-standing issues. To directly examine excitatory and suppressive tuning, we visualize the receptive field of each neuron in the full three-dimensional frequency domain. Spectral receptive fields were obtained by first multiplying the three-dimensional amplitude spectrum of each V1 filter by its fitted weight (Fig. 2*A*) and then summing across filters and correlation delays. Excitatory and suppressive receptive fields were obtained by summing spectra for either positive or negative weights separately.

The receptive fields of some of the MT neurons in our sample are consistent with the predictions of Simoncelli and Heeger (1998). One such neuron is shown in Figure 4*A*. The first three columns show the spectral receptive field of this neuron, viewed from three different angles. (For clarity, the plot has been rotated so that the preferred direction of motion is aligned with the *x*-axis.) The three rows show the excitatory components (top, positive fit weights), the suppressive components (middle, negative fit weights), and the combined receptive field (bottom). The transparent red and blue surfaces in each panel delineate the excitatory and suppressive isospectral surfaces. These surfaces were obtained by thresholding the aggregated amplitude spectrum of the Gabor filters at 25, 50, and 75% of the spectral peak. For this neuron, the excitatory receptive field forms a ring that lies on a single velocity plane in the frequency domain, and the suppressive receptive field lies off the excitatory plane.

The receptive fields of some of the other MT neurons in our sample are not rings, but rather encompass a narrow range of spatial and temporal frequencies. One such neuron is shown in Figure 4*B* (format same as Fig. 4*A*). The excitatory receptive field of this neuron is confined to a single point in the three-dimensional frequency domain, and there is little evidence of any substantial suppressive receptive field.

The neurons shown in Figure 4 represent the most extreme examples in our sample. In fact, most of the MT neurons lie between these two extremes. Two examples that are more typical of the sample as a whole are shown in Figure 5 (format same as Fig. 4). The excitatory receptive fields of these neurons lie predominantly on a single velocity plane, and they are elongated

along the frequency axis perpendicular to the optimal direction. However, they form only a partial ring in the plane. Thus, these neurons are insensitive to frequencies near the zero temporal frequency axis (compare Figs. 4*A*, 5). As far as we know, this pattern of tuning in area MT has not been described previously.

Excitatory spectral receptive fields lie on a plane

Simoncelli and Heeger (1998) predicted that the excitatory receptive fields of MT neurons should tend to lie on a single plane in the three-dimensional frequency domain. To address this issue quantitatively, we created an index that describes the proportion of the spectral receptive field of each MT neuron that lies on versus off of the optimal velocity plane. If the excitatory receptive field of an MT neuron lies on or near the optimal plane, then this on-plane ratio index will be near 1, and if it lies off of this plane, the index will be near 0. Figure 6*A* summarizes the on-plane ratio for the excitatory receptive fields of all 52 area MT neurons in our sample. The average ratio is 0.59.

Because our definition of the optimal velocity plane relies on maximizing the coverage of excitatory receptive fields on and near this plane (see Materials and Methods), by definition the on-plane ratio will be biased toward positive values. Therefore, to determine which index values were significantly greater than chance, we ran a Monte Carlo simulation to estimate the null distribution (see Materials and Methods). We created 150 model compressive/divisive neurons, each model seeded with random weights. We then estimated the on-plane ratio for each of these random model neurons. The average on-plane ratio for this null model is 0.36 (Fig. 6*A*, dashed line). This value is significantly lower than the value we observed across the real sample of neurons ($p < 0.01$, Wilcoxon rank-sum test). These data confirm that the excitatory receptive fields of MT neurons tend to lie on a single plane in the three-dimensional frequency domain, consistent with the predictions of Simoncelli and Heeger (1998).

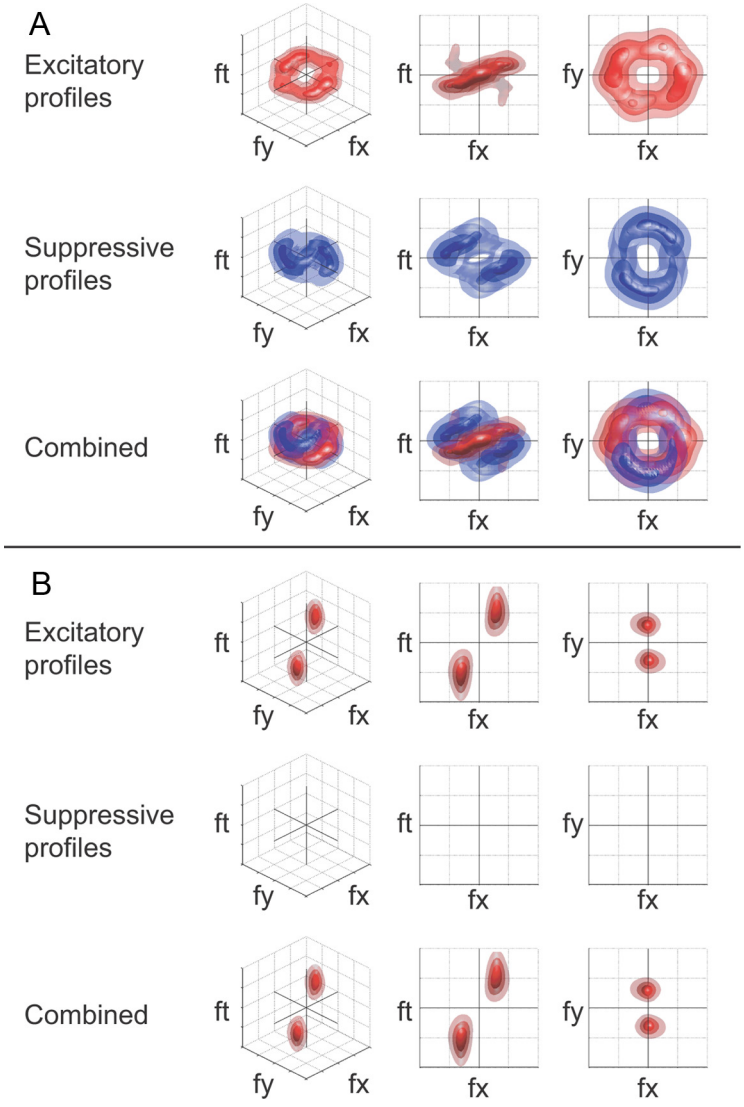


Figure 4. Estimated excitatory spectral receptive fields for two MT neurons. ***A***, Excitatory spectral receptive field for an MT neuron that is consistent with the predictions of Simoncelli and Heeger (1998). The three columns represent different views of the three-dimensional frequency domain. The red shells in the top row indicate excitatory isospectral contours (25, 50, and 75% of the spectral peak), the blue shells in the middle row indicate suppressive isospectral contours, and the bottom row shows both excitatory and suppressive shells in the same plots. Ticks for each axis show five cycles per receptive field for spatial frequency and 25 Hz for temporal frequency. To facilitate visualization, the spectral receptive field has been rotated so that the preferred direction of motion is aligned with the *x*-axis in the frequency domain. The excitatory receptive field for this neuron forms a ring in the three-dimensional frequency domain, and the suppressive receptive field encompasses a wide band of off-plane frequencies. Thus, this neuron is tuned for a single velocity. ***B***, Spectral receptive field for an MT neuron that encompass a narrow range of spatial and temporal frequencies. The format is the same as in ***A***.

Suppressive spectral receptive fields lie off the optimal excitatory plane

Simoncelli and Heeger (1998) also speculated that the suppressive receptive fields of MT neurons tend to lie off the optimal excitatory plane. To address this issue we simply applied the on-plane ratio to the suppressive (rather than the excitatory) spectral receptive field of each neuron. If the suppressive receptive field of an MT neuron tends to avoid the optimal velocity plane, then this ratio will be near 0, and if it lies on the optimal plane, then the ratio will be near 1. Figure 6*B* summarizes the suppressive on-plane ratio obtained across our sample. The average ratio is 0.18, which is significantly smaller than the value for the null model described above ($p < 0.01$, Wilcoxon rank-sum test). These data confirm that the suppressive receptive fields of MT neurons tend to avoid the optimal excitatory plane, as proposed by Simoncelli and Heeger (1998).

Excitatory spectral receptive fields of most MT neurons form a partial ring in the plane

Inspection of the spectral receptive fields estimated for individual MT neurons suggests that the population might differ along one simple dimension: the degree to which their excitatory receptive fields fill the optimal velocity plane within the three dimensional frequency domain. To characterize this, we created a separate

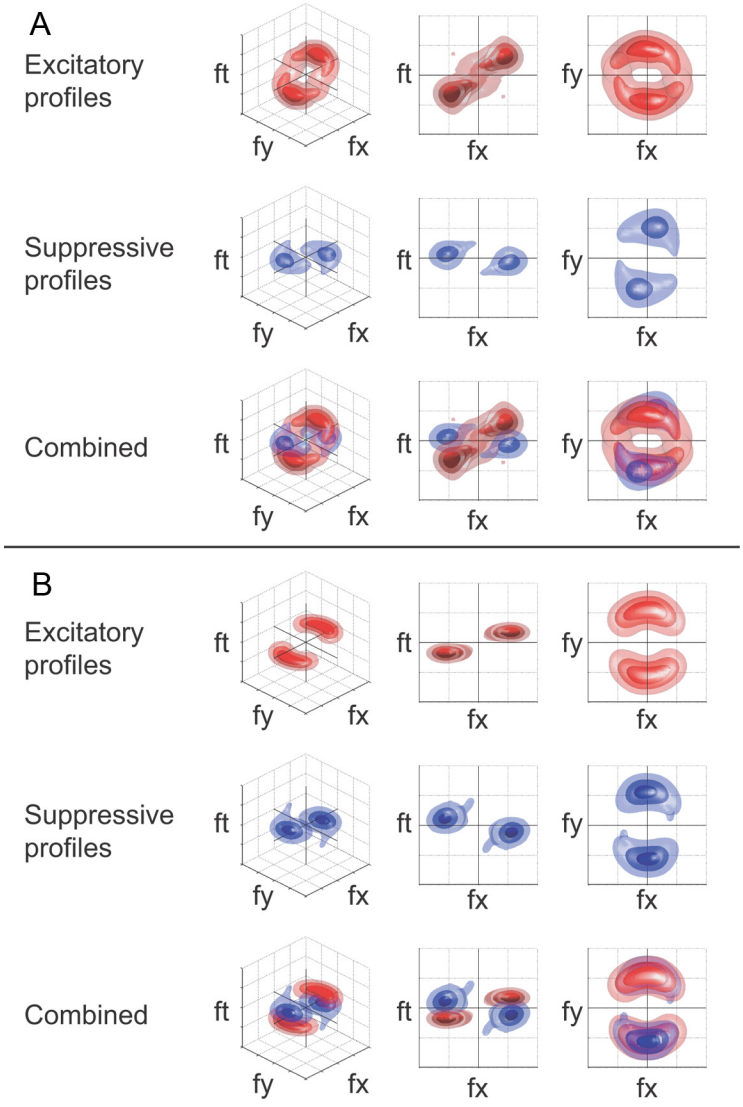


Figure 5. Estimated spectral receptive fields for two MT neurons typical of the sample as a whole. The format is the same as in Figure 4. ***A***, An MT neuron whose excitatory receptive field forms a partial ring in the optimal velocity plane. This neuron is relatively insensitive to temporal frequencies near zero. Thus, this neuron is tuned for a specific velocity but is not very sensitive to static texture oriented along the optimal direction of motion. ***B***, A second MT neuron whose excitatory receptive field forms a partial ring in the three-dimensional frequency domain. This neuron is tuned for a specific velocity, but will not respond to static texture oriented along the optimal direction of motion.

index that describes how well the excitatory spectral receptive field of each neuron fills the optimal velocity plane. First, we divided the optimal plane into four quadrants, two centered around the vertical axis (along the optimal direction) and two centered around the horizontal axis (the axis embedded in the zero temporal frequency plane). Then we integrated the excitatory spectral receptive field amplitudes in the vertical versus horizontal quadrants and took the ratio of these values. According to this horizontal–vertical index, a ratio of 1 indicates that an MT neuron forms a perfect ring in the optimal velocity plane, while a ratio of 0 indicates that the neuron is tuned to a single spatial and temporal frequency.

Figure 7*A* summarizes the horizontal–vertical ratios estimated for all 52 area MT neurons in our sample. The distribution is clearly continuous. At one end of the distribution lie MT neurons that have spectral receptive fields that form a ring in the optimal frequency plane, consistent with predictions of Simoncelli and Heeger (1998). At the opposite end of the distribution lie neurons that have spectral receptive fields confined to a unique spatial and temporal frequency. However, the receptive fields of the large majority of MT neurons lie between these extremes, forming a partial ring in the optimal velocity plane and avoiding the region near zero temporal frequency. Thus, the vast majority of MT neurons are relatively insensitive to static patterns aligned with the optimal direction of motion as opposed to what would be predicted according to the proposal of Simoncelli and Heeger (1998). [Note that there was no significant correlation between prediction performance and the horizontal–vertical ratio ($p > 0.10$).]

The horizontal–vertical ratio is correlated with simulated pattern selectivity

Many previous studies have used plaid stimuli to assess motion selectivity of area MT neurons (Movshon et al., 1985; Pack and Born, 2001; Smith et al., 2005; Rust et al., 2006; Majaj et al., 2007). A plaid consists of two superimposed gratings that move in different directions. MT neurons vary substantially in their responses to plaids. Some MT neurons are selective to the direction of the component gratings, whereas others are selective for the aggregate direction. These are called component-selective and pattern-selective neurons, respectively. Recent studies have reported that MT neurons do not form discrete component- and pattern-selective classes, but rather that selectivity is distributed continuously between these two extremes (Smith et al., 2005; Rust et al., 2006).

We performed a simulation to determine how the continuum of tuning in the optimal plane that we report here is related to the continuum of component and pattern selectivity found in previous studies (Smith et al., 2005; Rust et al., 2006). We first simulated responses of all MT neurons in our sample to both single grating and plaids. Then we used the pattern index developed in previous studies (Smith et al., 2005) to characterize plaid selectivity. The pattern index quantifies directional selectivity for plaid stimuli: it is positive if a neuron is selective for the aggregate direction of a plaid, and it is negative if a neuron is selective for the direction of the component gratings. Thus, the results of this simulation can be interpreted as a prediction about the plaid selectivity that we would expect to obtain for each of the neurons in our sample had we probed them with plaids.

Figure 7*B* summarizes the pattern index distribution predicted for all 52 area MT neurons in our sample. The distribution forms a clear continuum that captures the major distributions reported in previous studies (e.g., Rust et al., 2006, their Fig. 2*b*). The pattern index for most neurons ranges from −6 to 2, and highly pattern-selective neurons are relatively rare. To determine how the pattern index used in plaid studies is related to the horizontal–vertical ratio developed here, we compared these two in-

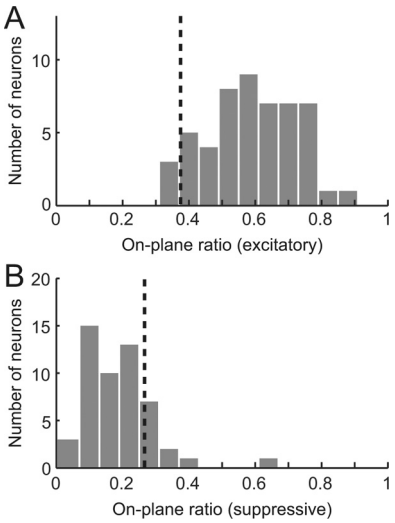


Figure 6. Spectral tuning to frequencies on versus off of the optimal plane within the three-dimensional frequency domain. ***A***, For each neuron, we estimated the plane that captures the largest amount of positive (excitatory) weight within the three-dimensional frequency domain. We then calculated the on-plane ratio, the ratio of positive weights near the plane over the total positive weights. Across the sample of 52 neurons, this ratio is significantly larger than chance ($p < 0.01$, Wilcoxon rank-sum test), indicating that the excitatory spectral receptive fields of MT neurons tend to lie along the optimal velocity plane. ***B***, On-plane ratios as in ***A***, but calculated using the negative (suppressive) weights for each neuron, and the same optimal plane as described in ***A***. Across the sample, the ratio is significantly smaller than chance ($p < 0.01$, Wilcoxon rank-sum test), indicating that the suppressive spectral receptive fields of MT neurons tend to lie off of the optimal velocity plane.

dices for each MT neuron in our sample (Fig. 7*C*). The correlation between the two indices is significant ($r = 0.46$, $p < 0.01$; *t* test for correlation coefficients). Thus, responses to plaids can be partly described in terms of the horizontal–vertical tuning ratio. There was no significant correlation between prediction performance and the pattern index ($p > 0.10$).

Control to identify any bias arising from the use of motion-enhanced natural movies

One potential concern with our results is that the stimuli used in our experiments were motion-enhanced natural movies whose spatial and temporal frequency spectra differ somewhat from those of natural movies (for details, see Materials and Methods; Fig. 1). To ensure that our use of motion-enhanced natural movies did not bias receptive field estimates we recorded from a subset of 15 area MT neurons using both motion-enhanced natural movies and simple natural movies as stimuli. We then compared predictions obtained when neuronal models were fit using motion-enhanced natural movies and tested using a separate set of motion-enhanced natural movies versus when those same models were tested using simple natural movies without motion enhancement. If simple natural movies evoke nonlinear responses that cannot be described by receptive fields estimated using motion-enhanced natural movies, then models estimated using motion-enhanced movies will fail to predict responses to movies without motion enhancement (David et al., 2004).

Receptive field models of area MT neurons estimated using motion-enhanced natural movies predicted responses to novel simple natural movies just as well as they predicted responses to novel motion-enhanced natural movies (average $r = 0.533$ for natural movies, average $r = 0.537$ for motion-enhanced movies; $p > 0.10$, Wilcoxon signed-rank test). This important control demonstrates that models estimated using motion-enhanced natural movies accurately describe and predict responses to natural movies. Based on this result, the predictions reported throughout this manuscript reflect the average prediction for both natural and motion-enhanced and natural movies (when the latter were available).

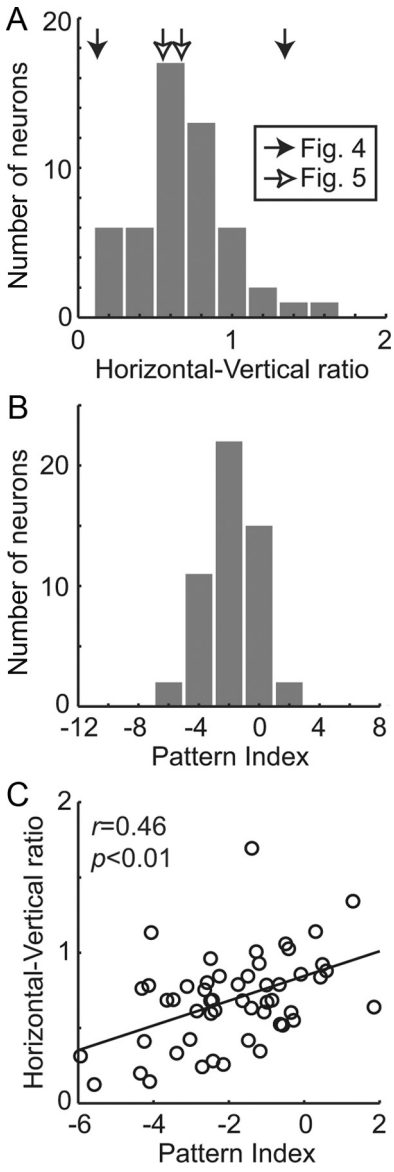


Figure 7. MT neurons differ in their sensitivity to low temporal frequencies. ***A***, Distribution of the horizontal–vertical ratio across the entire sample of 52 area MT neurons. The four neurons shown in Figures 4 and 5 are indicated by arrows. Only a small fraction of MT neurons have profiles consistent with the Simoncelli and Heeger (1998) model (ratio of ~1; Fig. 4*A*) or the unique energy model (ratio of ~0; Fig. 4*B*). The majority of MT neurons have profiles that are midway between these two extremes (Fig. 5). These neurons are insensitive to temporal frequencies near zero, so they do not respond to static texture patterns aligned with the optimal direction of motion. ***B***, Distribution of the pattern index derived from simulated responses to plaid and grating stimuli across the entire sample of MT neurons. The distribution generally agrees with previous studies (Rust et al. 2006, their Fig. 2*b*). ***C***, Joint scatter plot of horizontal–vertical ratio and pattern index. There is a significant correlation ($p < 0.01$) between these two indices.

Control to ensure that the model estimation procedure can recover spectral receptive fields of any shape

The model-fitting algorithms used here are closely related to those used in previous papers from our laboratory (David and

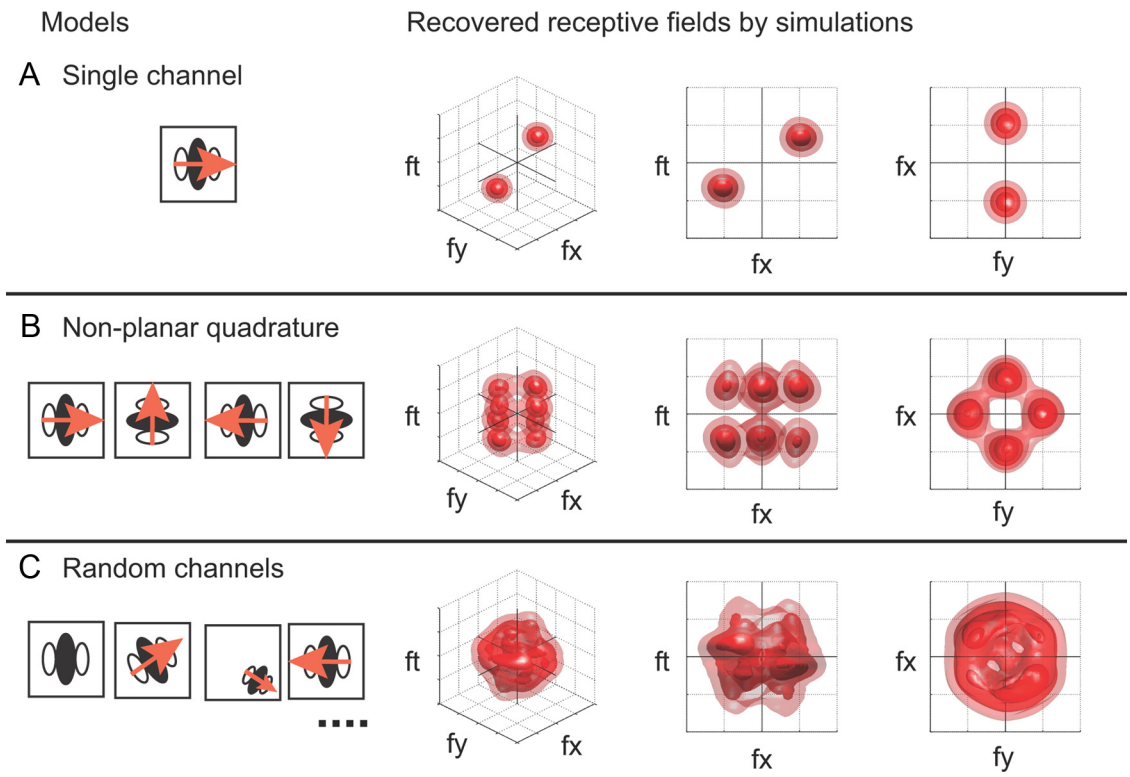


Figure 8. Demonstration that our receptive field estimation procedures can recover spectral receptive fields of various shapes. ***A***, A simulated MT neuron that receives input from a set of V1 neurons that are all tuned to a narrow range of spatial and temporal frequencies. These input neurons are all consistent with the compressive/divisive Gabor model (but only excitatory weights were included here). The left column shows a schematic description of the simulated MT neuron. The right column shows the spectral receptive field for the simulated neuron estimated using the same procedures used to characterize the real MT neurons in this study. The format is the same as in Figure 4. Our procedure correctly recovers the spectral receptive field. ***B***, A simulated MT neuron that receives input from four sets of V1 neurons each tuned to a different direction, but where all inputs are tuned for the same spatial and temporal frequency. Our procedure correctly recovers the spectral receptive field. ***C***, A simulated MT neuron that receives input from many V1 neurons, each tuned to a random direction and spatial and temporal frequency. Our procedure correctly recovers the spectral receptive field.

Gallant, 2005; Willmore et al., 2010) and from other groups (David et al., 2007). These powerful algorithms are rather complicated, and some readers might be concerned that the fitting procedures might have biased the results so as to produce the spectral receptive fields reported here. For example, given that translational motion in natural movies always will produce a planar three-dimensional frequency spectrum, would it ever be possible to recover nonplanar receptive fields? To address this concern, we used the same receptive field estimation procedures described previously in this paper to estimate receptive fields of three simulated neurons whose receptive field organization was quite different from that observed in our sample of real MT neurons.

Figure 8 shows spectral receptive fields for three simulated MT neurons (the format is the same as in Fig. 4; note that the model neurons did not have suppressive receptive fields). The first simulated MT neuron receives input from a set of V1 neurons that are all tuned to a narrow range of spatial and temporal frequencies. Our procedure correctly recovers the spectral receptive field for this neuron. The second simulated MT neuron receives input from four sets of V1 neurons each tuned to a different direction, but where all inputs are tuned for the same spatial and temporal frequency. This is an interesting test case because this simulated MT neuron does not have a planar spectral tuning profile. However, our procedure still recovers the correct spectral receptive field. Finally, the third simulated MT neuron receives input from many V1 neurons, each tuned to a random direction and spatial and temporal frequency. This is another interesting test case because this model MT neuron does not have a planar spectral tuning profile, and it does not avoid low temporal frequencies. Once again our procedure correctly recovers the spectral receptive field. These results demonstrate that the receptive field estimation procedure used in this study can characterize arbitrary spectral receptive fields, regardless of their organization within the three-dimensional frequency domain.

Discussion

Area MT has been the target of intensive neurophysiological investigation over the last 25 years (for review, see Born and Bradley, 2005). However, most previous studies of MT have used simple, parameterized stimuli that spanned only a subspace within the full three-dimensional frequency domain, and it is unclear how the mechanisms revealed under those conditions will generalize to natural vision. We investigated this issue by recording responses of MT neurons evoked by naturalistic movies. We found that the simplest model of MT neurons that accurately predicts responses to these movies consists of a bank of Gabor filters, each followed by either a half-wave rectification or motion-energy computation, a compressive nonlinearity, a divisive nonlinearity, and a linear pooling stage whose weights are determined uniquely for each neuron. This result confirms that concepts of motion coding in MT developed using synthetic stimuli (Simoncelli and Heeger, 1998) are generally valid under more naturalistic conditions.

Our study provides the first reconstructions of spectral receptive fields of MT neurons within the full three-dimensional fre-

quency domain, and it demonstrates that these neurons have planar receptive fields within this domain. The excitatory receptive fields of a few of these neurons form a ring in the optimal velocity plane, consistent with predictions in Simoncelli and Heeger (1998). Another small group of MT neurons have excitatory receptive fields tuned for one unique spatial and temporal frequency. However, the receptive fields of most MT neurons form a partial ring in the optimal velocity plane, avoiding very low temporal frequencies. In sum, the entire population of MT neurons can be characterized along three dimensions: the orientation and the elevation of the optimal velocity plane, and the extent to which the excitatory receptive field forms a ring in the optimal plane (Fig. 9).

Simoncelli and Heeger (1998) also predicted that the receptive fields of some MT neurons might form a partial ring in the optimal velocity plane. However, they predicted that this partial ring would be elongated toward the origin (i.e., zero spatial and temporal frequency). In contrast, we find that these partial ring receptive fields are elongate horizontally along the optimal direction of motion (Fig. 5). This has important functional implications: elongation toward the origin does not preserve velocity tuning, while horizontal elongation along the optimal direction of motion does preserve velocity tuning (see also below and Fig. 10).

We used a computational simulation to show that the MT receptive fields estimated here explain general aspects of plaid responses reported previously (Movshon et al., 1985; Pack and Born, 2001; Smith et al., 2005; Rust et al., 2006; Majaj et al., 2007). However, while several previous studies have reported that a small minority of MT neurons are extremely pattern selective (Smith et al., 2005; Rust et al., 2006), our simulations did not identify any such neurons. One possibility is that extreme pattern selectivity depends on a tuned normalization mechanism (Rust et al., 2006) that was not included explicitly in our model. However, our model does include a compressive nonlinearity on each channel, and this might perform a function similar to the tuned normalization of Rust et al. (2006). Furthermore, Rust et al. (2006) reported no significant relationship between tuned normalization and the pattern index, suggesting that the mechanism does not play a major role in explaining the pattern index quantitatively. Another possibility is that extreme pattern selectivity is only observed when MT neurons are probed with simple plaids,

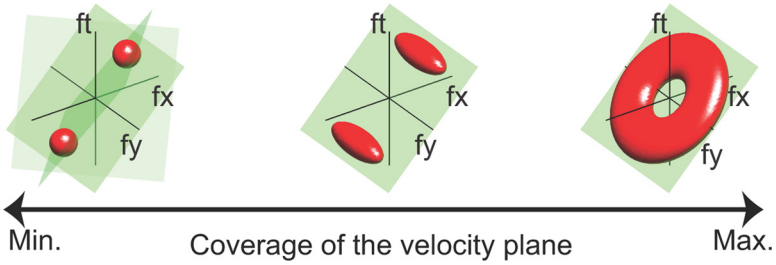


Figure 9. Area MT neurons vary in the degree to which their excitatory spectral receptive fields form a ring within the optimal velocity plane. On one extreme lie MT neurons whose spectral receptive fields are tuned for one unique spatial and temporal frequency. These neurons are not tuned for speed or velocity because there are infinite combinations of speed and direction that are consistent with the receptive field. On the other extreme lie neurons whose spectral receptive fields form a ring on the optimal velocity plane. These neurons are tuned for velocity as originally proposed by Simoncelli and Heeger (1998). However, they also respond to static texture that is aligned with the optimal direction of motion. The majority of MT neurons lie between these extremes. These neurons have excitatory spectral receptive fields that are elongated parallel to the optimal velocity plane, forming a partial ring in the plane and avoiding low temporal frequencies. These neurons are also tuned for velocity, but they do not respond to static texture.

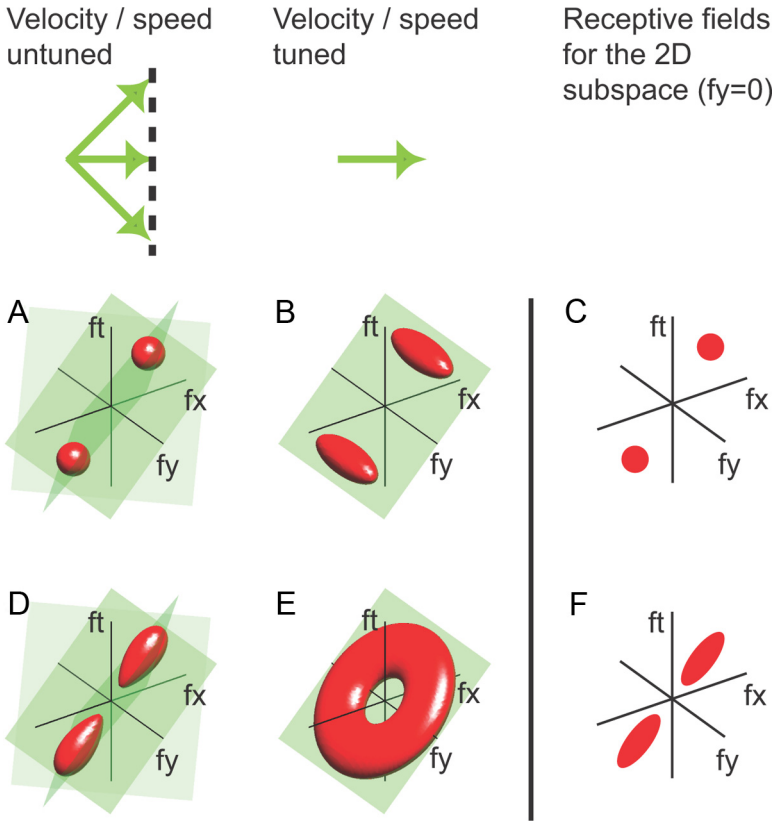


Figure 10. Spectral receptive fields estimated within a two-dimensional frequency subspace cannot unambiguously reveal tuning for speed and velocity. ***A***, Spectral receptive field of a hypothetical neuron that is tuned for a unique spatial and temporal frequency. The format is the same as in Figure 4, except that the only excitatory receptive fields are shown. This neuron is not tuned for speed and velocity because there are many velocity planes (green) that pass through the receptive field. ***B***, Spectral receptive field of a hypothetical neuron that is similar to those typically observed in area MT. This neuron is tuned for both speed and velocity, because one specific velocity plane maximizes overlap with the receptive field. ***C***, Two-dimensional slice ($fy = 0$) (Perrone and Thiele 2001; Priebe et al. 2003) through the spectral receptive field of the hypothetical neurons shown in ***A*** and ***B***. A researcher who only had access to the receptive fields within this subspace might conclude that neither of the neurons shown in ***A*** and ***B*** are tuned for speed, although the neuron shown in ***B*** is tuned for both speed and velocity. ***D***, Hypothetical neuron whose spectral receptive field is elongated toward the origin. This neuron is also not tuned for speed or velocity, because there are many velocity planes that pass through the receptive field. ***E***, Spectral receptive field of a hypothetical neuron that forms a ring as proposed by Simoncelli and Heeger (1998). The neuron is tuned for velocity, because only one velocity plane maximizes overlap with the receptive field. ***F***, Two-dimensional slice through the spectral receptive field of the hypothetical neurons shown in ***D*** and ***E***. A researcher who had access to only the receptive fields within this subspace might conclude that both of the neurons shown in ***D*** and ***E*** are tuned for speed, although the neuron shown in ***D*** is not tuned for speed or velocity. These examples illustrate that it is impossible to definitively determine whether a neuron is tuned for speed or velocity by examining only a two-dimensional subspace within the full three-dimensional frequency domain.

and that responses change when these neurons are probed using more naturalistic stimuli. Analogous stimulus-dependent effects have been reported in studies of area V1 (David et al., 2004; Felsen et al., 2005; Sharpee et al., 2006).

We found that luminance and contrast normalization does not appear to improve model predictions beyond what can be achieved using a simpler model without these mechanisms (Fig. 2). However, although our stimuli span the range of luminance commonly used in neurophysiology experiments (8 to 130 cd/m^2), it is conceivable that these luminance and contrast normalization might be important under more natural conditions containing a wider luminance range (Lewen et al., 2001).

Relationship to previous reports of speed-tuned neurons

Several neurophysiological studies have used drifting gratings to measure speed tuning in area MT (Perrone and Thiele, 2001; Priebe et al., 2003). These studies optimized the direction of grating drift for each neuron individually while systematically varying spatial and temporal frequency. Thus, each study probed a two-dimensional slice of the full three-dimensional frequency domain. However, speed (and velocity) tuning cannot be established unequivocally with stimuli that are confined to a two-dimensional slice. To see why this is so, consider the four hypothetical neurons shown in Figure 10. Figure 10*A* shows the spectral receptive field for a hypothetical neuron tuned for a unique combination of spatial and temporal frequencies. This neuron is not tuned for speed or velocity because many different velocity planes (i.e., many different combinations of speeds and directions; shown in green) pass through the receptive field (Movshon et al., 1985; Simoncelli and Heeger, 1998; Bradley and Goyal, 2008). Figure 10*B* shows the spectral receptive field of a hypothetical neuron similar to those typically found in area MT. This neuron is tuned for speed and velocity because only one velocity plane maximizes overlap with the receptive field. Figure 10*C* shows a two-dimensional slice through the spectral receptive field of the hypothetical neurons shown in *A* and *B*. In the three-dimensional space, the two-dimensional subspace examined in previous studies (Perrone and Thiele, 2001; Priebe et al., 2003) forms the slice defined by $f_y = 0$. A researcher who had access only to receptive fields within this subspace might conclude that neither of the neurons shown in Figure 10, *A* and *B*, are tuned for speed, though the neuron shown in *B* is tuned for speed (and velocity).

Figure 10*D* shows a hypothetical neuron whose spectral receptive field is elongated toward the origin, as predicted by Simoncelli and Heeger (1998). This neuron is not tuned for speed or velocity because many different velocity planes pass through the receptive field. Figure 10*E* shows the spectral receptive field of a hypothetical neuron that forms a ring, as predicted by Simoncelli and Heeger (1998). This neuron is tuned for speed and velocity because only one velocity plane maximizes overlap with the receptive field. Figure 10*F* shows a two-dimensional slice through the spectral receptive fields of the hypothetical neurons shown in Figure 10, *D* and *E*. A researcher who only had access to receptive fields within this subspace might conclude that both of the neurons shown in Figure 10, *D* and *E*, are tuned for speed, though the neuron shown in *D* is not tuned for speed (or velocity). These examples demonstrate that speed and velocity tuning cannot be established by measuring the receptive field within a two-dimensional subspace. This paper represents the first attempt to examine three-dimensional spectral selectivity, which allows direct assessment of speed and velocity tuning.

Functional implications of partial ring structures in the frequency domain

An MT neuron that forms a ring in the optimal three-dimensional velocity plane (Simoncelli and Heeger, 1998) will be tuned for one particular velocity, depending on the slant and tilt of the optimal plane. However, such a neuron will also respond to static stimuli oriented parallel to the optimal direction of motion. Our results show that most area MT neurons avoid this problem. These neurons form a partial ring in the optimal velocity plane, systematically avoiding the region around zero temporal frequency. They respond to moving patterns at the optimal velocity, but they do not tend to respond to static patterns that are oriented parallel to the optimal direction. This scheme provides a representation of image motion that is less ambiguous than that proposed by the Simoncelli and Heeger (1998) model. Consistent with this, Albright (1984) reported that although some MT neurons do respond to a static bar oriented parallel to the optimal direction, these responses are much less vigorous than those elicited by moving stimuli. Our study provides a clear explanation for this phenomenon, and confirms that area MT is optimized to process moving patterns.

It is currently unclear how area MT neurons develop their very precise receptive fields and why most of them are insensitive to low temporal frequencies. Each MT neuron receives input from a specific population of direction-selective V1 neurons (Movshon and Newsome, 1996). Direction-selective neurons in V1 are almost exclusively bandpass for temporal frequency (Hawken et al., 1996) and so do not respond to static stimuli. The fact that most MT neurons do not respond to zero temporal frequency could therefore merely reflect a bias that already exists in the V1 neurons that project to MT. This bias could be genetic, or it could be caused by the learning algorithm that governs the development of receptive fields in MT. If this feature is the product of a learning rule, this finding could provide a new constraint on how corticocortical circuits are optimized to represent information during natural vision.

References

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

Albrecht DG, Hamilton DB (1982) Striate cortex of monkey and cat: contrast response function. J Neurophysiol 48:217–237.

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

Bonin V, Mante V, Carandini M (2006) The statistical computation underlying contrast gain control. J Neurosci 26:6346–6353.

Born RT, Bradley DC (2005) Structure and function of visual area MT. Annu Rev Neurosci 28:157–189.

Bradley DC, Goyal MS (2008) Velocity computation in the primate visual system. Nat Rev Neurosci 9:686–695.

Britten KH, Shadlen MN, Newsome WT, Movshon JA (1993) Responses of neurons in macaque MT to stochastic motion signals. Vis Neurosci 10:1157–1169.

Carandini M, Heeger DJ, Movshon JA (1997) Linearity and normalization in simple cells of the macaque primary visual cortex. J Neurosci 17:8621–8644.

David SV, Gallant JL (2005) Predicting neuronal responses during natural vision. Network 16:239–260.

David SV, Vinje WE, Gallant JL (2004) Natural stimulus statistics alter the receptive field structure of V1 neurons. J Neurosci 24:6991–7006.

David SV, Mesgarani N, Shamma SA (2007) Estimating sparse spectro-temporal receptive fields with natural stimuli. Network 18:191–212.

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1:1–47.

Felsen G, Touryan J, Han F, Dan Y (2005) Cortical sensitivity to visual features in natural scenes. PLoS Biol 3:e342.

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232.
Hawken MJ, Shapley RM, Grosof DH (1996) Temporal-frequency selectivity in monkey visual cortex. Vis Neurosci 13:477–492.
Heeger DJ (1992a) Half-squaring in responses of cat striate cells. Vis Neurosci 9:427–443.
Heeger DJ (1992b) Normalization of cell responses in cat striate cortex. Vis Neurosci 9:181–197.
Jones JP, Palmer LA (1987) An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J Neurophysiol 58:1233–1258.
Kaplan E, Purpura K, Shapley RM (1987) Contrast affects the transmission of visual information through the mammalian lateral geniculate nucleus. J Physiol 391:267–288.
Lewen GD, Bialek W, de Ruyter van Steveninck RR (2001) Neural coding of naturalistic motion stimuli. Network 12:317–329.
Livingstone MS, Pack CC, Born RT (2001) Two-dimensional substructure of MT receptive fields. Neuron 30:781–793.
Majaj NJ, Carandini M, Movshon JA (2007) Motion integration by neurons in macaque MT is local, not global. J Neurosci 27:366–370.
Mante V, Frazor RA, Bonin V, Geisler WS, Carandini M (2005) Independence of luminance and contrast in natural scenes and in the early visual system. Nat Neurosci 8:1690–1697.
Mante V, Bonin V, Carandini M (2008) Functional mechanisms shaping lateral geniculate responses to artificial and natural stimuli. Neuron 58:625–638.
Mazer JA, Gallant JL (2003) Goal related activity in area V4 during free viewing visual search: evidence for a ventral stream salience map. Neuron 40:1241–1250.
Movshon JA, Newsome WT (1996) Visual response properties of striate cortical neurons projecting to area MT in macaque monkeys. J Neurosci 16:7733–7741.
Movshon JA, Adelson EH, Gizzi MS, Newsome WT (1985) The analysis of moving visual patterns. In: Pattern recognition mechanisms (Chagas C, Gattass R, Gross C, eds), p 117. New York: Springer.
Nishimoto S, Ishida T, Ohzawa I (2006) Receptive field properties of neurons in the early visual cortex revealed by local spectral reverse correlation. J Neurosci 26:3269–3280.
Okamoto H, Kawakami S, Saito H, Hida E, Odajima K, Tamanoi D, Ohno H (1999) MT neurons in the macaque exhibited two types of bimodal direction tuning as predicted by a model for visual motion detection. Vision Res 39:3465–3479.
Olmos A, Kingdom FAA (2004) A biologically inspired algorithm for the recovery of shading and reflectance images. Perception 33:1463–1473.
Pack CC, Born RT (2001) Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. Nature 409:1040–1042.
Perrone JA, Thiele A (2001) Speed skills: measuring the visual speed analyzing properties of primate MT neurons. Nat Neurosci 4:526–532.
Priebe NJ, Cassanello CR, Lisberger SG (2003) The neural representation of speed in macaque area MT/V5. J Neurosci 23:5650–5661.
Rodman HR, Albright TD (1987) Coding of visual stimulus velocity in area MT of the macaque. Vision Res 27:2035–2048.
Rust NC, Movshon JA (2005) In praise of artifice. Nat Neurosci 8: 1647–1650.
Rust NC, Mante V, Simoncelli EP, Movshon JA (2006) How MT cells analyze the motion of visual patterns. Nat Neurosci 9:1421–1431.
Schrater PR, Simoncelli EP (1998) Local velocity representation: evidence from motion adaptation. Vision Res 38:3899–3912.
Schrater PR, Knill DC, Simoncelli EP (2000) Mechanisms of visual motion detection. Nat Neurosci 3:64–68.
Sharpee TO, Sugihara H, Kurgansky AV, Rebrik SP, Stryker MP, Miller KD (2006) Adaptive filtering enhances information transmission in visual cortex. Nature 439:936–942.
Simoncelli EP, Heeger DJ (1998) A model of neuronal responses in visual area MT. Vision Res 38:743–761.
Smith MA, Majaj NJ, Movshon JA (2005) Dynamics of motion signaling by neurons in macaque area MT. Nat Neurosci 8:220–228.
Snowden RJ, Treue S, Erickson RG, Andersen RA (1991) The response of area MT and V1 neurons to transparent motion. J Neurosci 11:2768–2785.
Stanley GB (2008) Au naturel. Neuron 58:467–469.
Theunissen FE, Sen K, Doupe AJ (2000) Spectral-temporal receptive fields of non-linear auditory neurons obtained using natural sounds. J Neurosci 20:2315–2331.
Watson AB, Ahumada AJ Jr (1985) Model of human visual-motion sensing. J Opt Soc Am A 2:322–341.
Willmore BD, Prenger RJ, Gallant JL (2010) Neural representation of natural images in visual area V2. J Neurosci 30:2102–2114.
Wu MC, David SV, Gallant JL (2006) Complete functional characterization of sensory neurons by system identification. Annu Rev Neurosci 29:477–505.