Current Biology, Volume *21*

Supplemental Information

Reconstructing Visual Experiences from

Brain Activity Evoked by Natural Movies

Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant

Supplemental Inventory

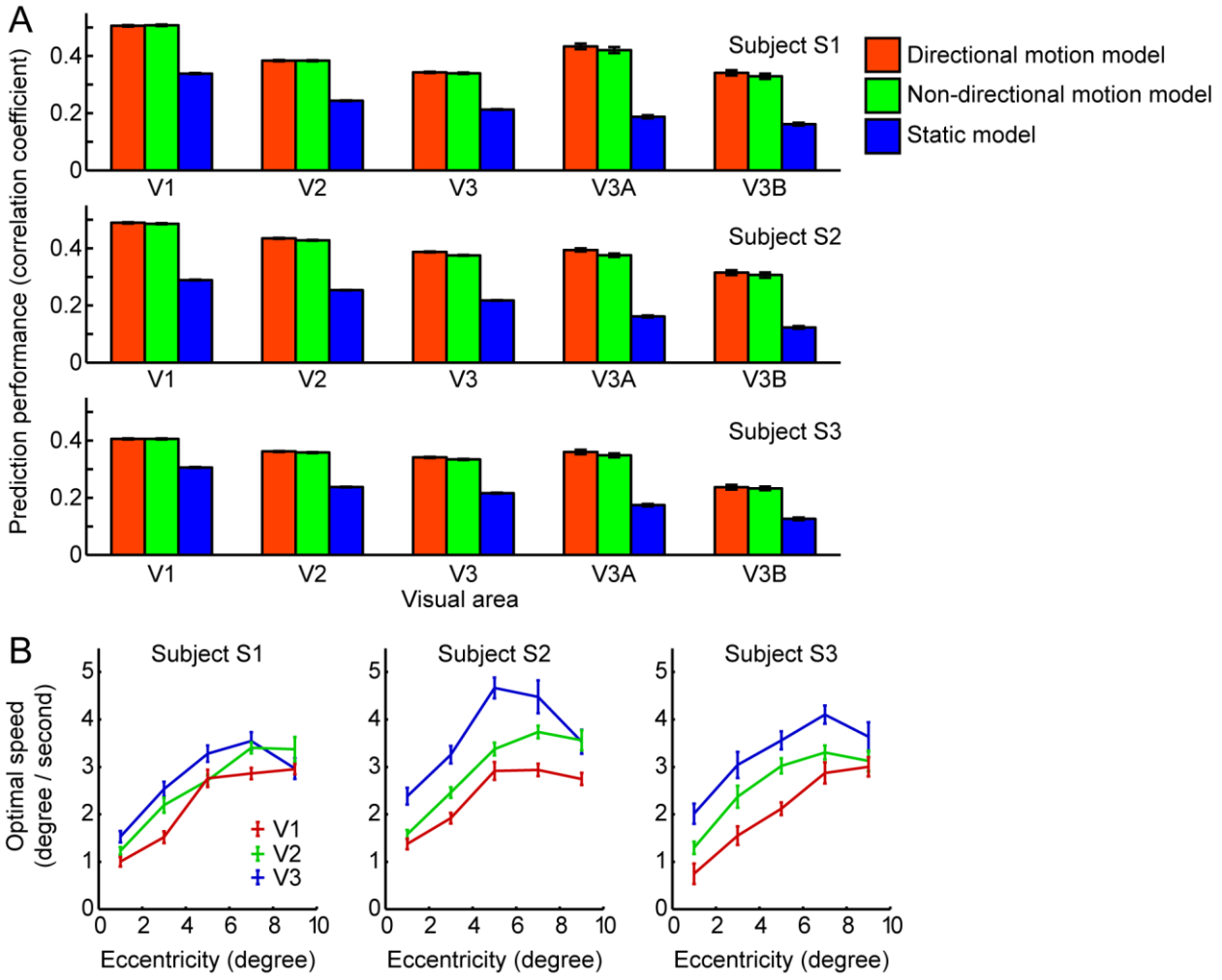

Figure S1. Encoding Model Details across Visual Areas and Subjects

(A) Prediction accuracy across visual areas and subjects. In the main text we showed that the directional motion model provides the most accurate predictions of BOLD signals to novel natural movies, and the static model provides the worst predictions, but that the difference in performance of the directional and non-directional models is minimal. We repeated the analysis to determine whether this pattern holds for individual visual areas and individual subjects. These bar graphs show prediction performance for the three models across five visual areas, for each of the three subjects. Error bars indicate ± 1 SEM across voxels (bootstrap procedure [48]). The directional and non-directional motion models perform better than the static model in every case. The best overall predictions for all three subjects ($p < 0.0001$, Wilcoxon rank-sum test) are obtained in area V1. This likely reflects the fact that the core component of the motion-energy model is a V1 complex cell model [10, 11].

(B) Speed selectivity depends on eccentricity. In the main body of the paper we showed how speed selectivity is distributed across the cortical flat map (see Figure 2J). Those data indicated that optimal speed depends on eccentricity. Here the same data are shown as the average optimal speed across voxels in visual areas (V1, V2 and V3 shown in different colors), binned in increments of two degrees of eccentricity for each of the three subjects examined in this

experiment. Optimal speed is expressed as the optimal temporal frequency divided by the optimal spatial frequency. (Voxels for which prediction accuracy of the directional motion-energy model was $p > 0.01$ or where the optimal spatial frequency was 0 cycles/degree have been omitted.) Error bars indicate ± 1 SEM across voxels for each bin (bootstrap procedure [48]). In all three subjects and all three visual areas there is a significant positive correlation between eccentricity and optimal speed ($p < 0.0001$, t test for correlation coefficient). Because high temporal frequency signals in natural movies have low energy, we estimated temporal frequency selectivity only up to 4Hz. This could bias estimates of the optimal speed toward lower values, especially for voxels in the visual periphery that are high-pass for temporal frequency (e.g., Figure 2G).

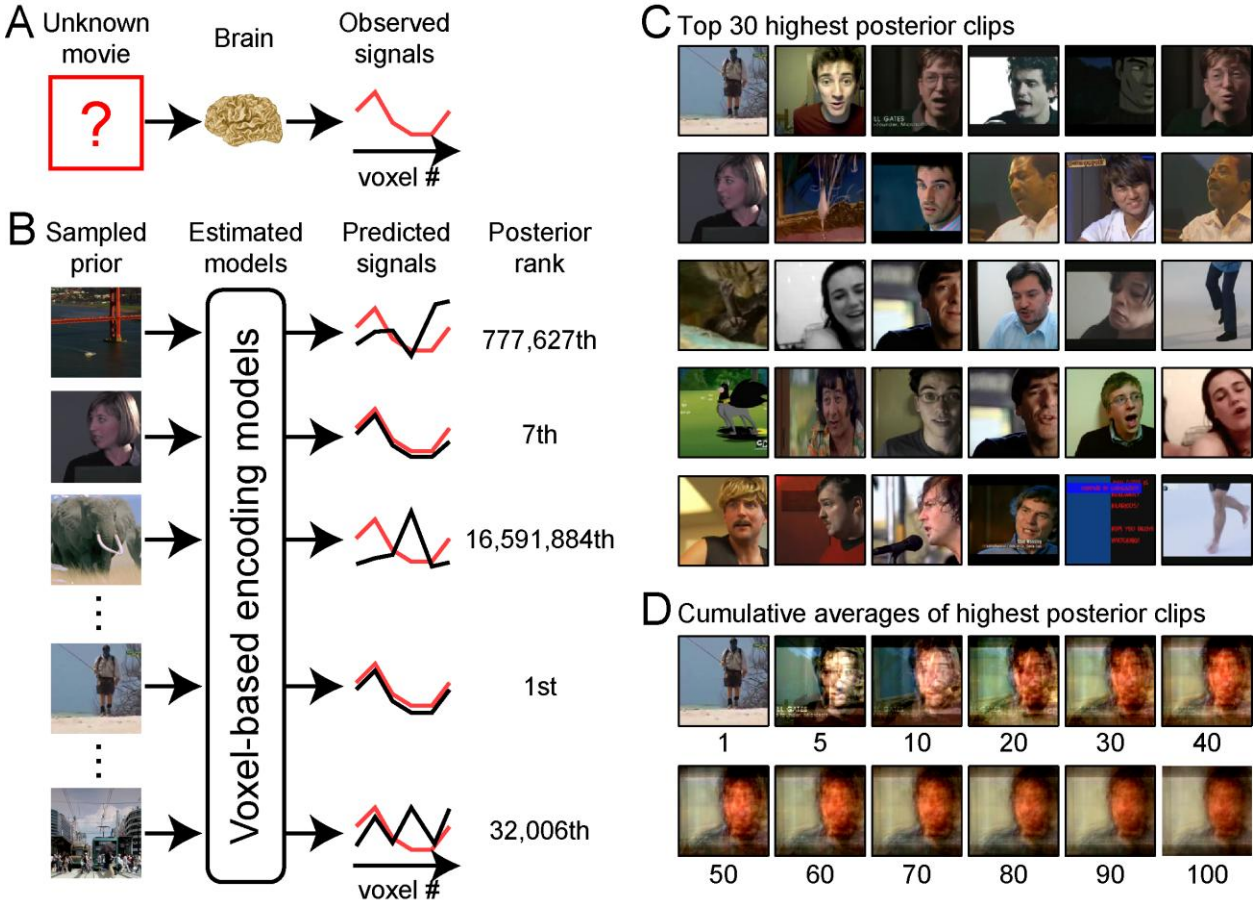

Figure S2. Schematic Diagram of Decoding Algorithm

(A) Reconstruction is a form of decoding in which the BOLD signals measured from a set of voxels are used to recreate a picture of the unknown stimulus. Here the stimulus was an unknown movie clip, and BOLD signals were recorded from a set of voxels in visual cortex.

(B) The reconstruction algorithm exploits the fact that a posterior probability is proportional to a likelihood times a prior probability [8]. We use a sampled natural movie prior, consisting of a database of ~18 million one-second movie clips drawn at random from YouTube (http://www.youtube.com; left column). To obtain the posterior, each clip in the sampled prior is first processed using the motion-energy encoding models fit to each voxel (middle column), and the predicted signals are compared to the measured signals evoked by the unknown stimulus (right column). The posterior rank of each of the clips in the sampled prior is simply the likelihood of the observed response given the clip (see Likelihood Estimation in Supplemental Information).

(C) Thirty clips from the sampled prior that had the highest posterior, given a pattern of responses evoked by the unknown clip. The clips are sorted in descending order from the highest posterior probability (top left) to the 30th (bottom right). The single clip with the highest posterior probability is the maximum a posteriori (MAP) reconstruction [8].

(D) Because the empirical prior is a sparse and relatively small sample of all possible natural movies, the MAP reconstruction may be poor. One way to simulate a denser sampling of the

posterior is to simply average over the clips near the peak of the posterior. Here averages over 1-100 clips are shown. Note that to equalize the contributions from each clip we prenormalized the pixel values of each clip to have a unit standard deviation before averaging. After averaging we post-normalized the averaged clip so that its mean and standard deviation were equal to those of the average of the top 100 clips. We found in practice that averaging over 100 clips near the peak of the posterior yields robust and stable reconstructions. We call this the averaged high posterior (AHP) reconstruction (Figure 4).

Supplemental Experimental Procedures

List of Movies Used as Stimuli

To minimize potential biases in the stimulus set, the movies used for the experiment were drawn from a wide variety of different sources. The bulk of the movies were taken from trailers for the following movies: “Australia”, “Bolt”, “Bride Wars”, “Changeling”, “Duplicity”, “Fuel”, “Hotel for Dogs”, “Ink Heart”, “King Lines”, “Mall Cop”, “Madagascar 2”, “Pink Panther 2”, “Proud American”, “Role Models”, “Shark Water”, “Star Trek”, “The Tale of Despereaux”, “Warren Miller Higher Ground” and “Yes Man”. Additional movies were taken from following libraries: “Artbeats HD”, “BBC Motion Gallery”, “Mammoth HD” and “The Macaulay Library”. These movies were supplemented with high-definition movies drawn from YouTube: “IGN Game of the Year 2008”, “JAL Boeing 747 landing Kai Tak”, “The American Recovery and Reinvestment Plan” and “Where the hell is Matt?”.

Data Preprocessing

BOLD signals were preprocessed as described in earlier publications [8, 19]. Briefly, motion compensation was performed using SPM '99 (http://www.fil.ion.ucl.ac.uk/spm), and supplemented by additional custom algorithms. For each 10 minute run and each individual voxel, drift in BOLD signals was first removed by fitting a third-degree polynomial, and signals were then normalized to mean 0.0 and standard deviation 1.0. Retinotopic mapping data collected from the same subjects in separate scan sessions was used to assign voxels to visual areas [47].

To compensate for hemodynamic transients caused by movie onset, we presented 10 seconds of dummy movies before each 10 minute block. The dummy movies were identical to the final 10 seconds of movies for each block. Data collected during this initial 10 seconds were excluded from data analysis.

Motion-Energy Encoding Model

Our motion-energy encoding model describes BOLD signals as a linear weighted sum of local, nonlinear motion-energy filters. The model has two main steps (see Figure 1). Movies first pass through a bank of nonlinear motion-energy filters, and these transformed signals then pass through a bank of temporal hemodynamic response filters. The nonlinear motion-energy filter bank itself consists of several stages of processing (Figure 1A). To minimize the computational burden all movie frames are first spatially down-sampled to 96x96 pixels. The RGB pixel values are then converted into Commission internationale de l'éclairage (CIE) L*A*B* color space and color information is discarded. The luminance patterns then pass through a bank of three-dimensional spatiotemporal Gabor wavelet filters, where two dimensions represent space and one represents time (see Gabor Wavelet Basis Set). The output of each quadrature pair of filters (i.e., filters of two orthogonal phases) is squared and summed to yield local motion-energy measurements [10, 11]. Motion-energy signals are then compressed by a log-transform and temporally down-sampled from the original frequency of the movie (15 Hz) to the sampling rate used to measure BOLD signals (1 Hz). Each motion-energy signal is then normalized across time by a Z-score transformation so that each has mean 0.0 and standard deviation 1.0. Any motion-energy signal outliers more than 3.0 standard deviations from the mean are truncated to 3.0 in order to improve stability in the model estimation procedure. Finally, the output of each motion-energy filter is temporally convolved with one specific hemodynamic response filter, and all

channels are summed linearly. The shape of each hemodynamic response filter is fit separately using data from the training set (see Model Fitting). To minimize computational time we restricted the temporal window of the hemodynamic response filters to a period 3-6 seconds (4 time samples) before BOLD signals. To simplify the association between each BOLD signal and each one second movie clip during reconstruction we refit the encoding model after shrinking the window so that it included only the single delay of 4 seconds (one time sample).

Note that in theory the hemodynamic convolution could be applied before down-sampling the filtered stimuli. Although this would reproduce more faithfully the underlying process that generates BOLD signals, it is computationally more efficient to perform the convolution after down-sampling.

Gabor Wavelet Basis Set

One important component of the motion-energy encoding model is a bank of three-dimensional spatiotemporal Gabor wavelet filters (Figure 1). The complete spatiotemporal Gabor wavelet basis set contains 6,555 separate three-dimensional Gabor filters. Each filter is constructed by multiplying a three-dimensional spatiotemporal (2 dimensions for space, 1 dimension for time) sinusoid by a three-dimensional spatiotemporal Gaussian envelope [49, 50]. Filters occur at six spatial frequencies (0, 2, 4, 8, 16 and 32 cycles/image), three temporal frequencies (0, 2 and 4 Hz) and eight directions (0, 45,…, 315 degrees). The zero temporal frequency filters occur at only four orientations (0, 45, 90 and 135 degrees) and the zero spatial frequency filters occur only once (no orientation). Filters are positioned on a square grid that covers the movie screen. Grid spacing is determined separately for filters at each spatial frequency so that adjacent Gabor wavelets are separated by 3.5 standard deviations of the spatial Gaussian envelope. To facilitate the motion-energy computation [10, 11] each filter occurs at two quadratic phases (0 and 90 degrees).

Two simplified encoding models were also used in this study. The non-directional motion model is identical to the directional model except the outputs of anti-directional filters (e.g., 0 degrees and 180 degrees) are summed at each spatial position, spatial orientation and temporal frequency. The static model includes only the subset of filters with zero temporal frequency.

Model Fitting

The motion-energy encoding model was fit to each voxel individually (Figure 1A) by means of a set of linear temporal filters meant to model the hemodynamic response and its coupling with neural activity. The encoding model for the *i*-th voxel can be written in linear vector form:

$$\hat{r}_i = \mathbf{s} * \mathbf{w}_i$$

$$\hat{r}_i = \begin{bmatrix} s_{d1} & s_{d2} & \cdots & s_{dK} \end{bmatrix} * \begin{bmatrix} h_{i,d1} \\ h_{i,d2} \\ \vdots \\ h_{i,dK} \end{bmatrix}$$

where $\hat{r}_i$ is the predicted BOLD signal, $\mathbf{s}$ is a motion-energy filtered stimuli and $\mathbf{w}_i$ is a linear weight vector that represents the motion-energy specific hemodynamic response filters. In this schematic each rectangle represents a vector (or scalar). Brackets indicate that matrices are concatenated. To capture temporal delays of the BOLD signals in the model, the vector $\mathbf{s}$ is constructed by concatenating motion-energy filtered stimulus vectors at various temporal delays. Here, $\mathbf{s}_{dx}$ is a [1 x F] vector (F is # of filters) representing the motion-energy filtered stimuli shifted by d_x seconds, while $\mathbf{s}$ is a concatenated vector $[\mathbf{s}_{d1} \ldots \mathbf{s}_{dK}]$ where d_x (x=1…K) are the temporal delays of interest. The resulting vector $\mathbf{s}$ is of size [1 x M], where M is # of parameters that is given by F x K. The weight vector $\mathbf{w}_i$ consists of multiple linear weight vectors $\mathbf{h}_{i,dx}$, where each $\mathbf{h}_{i,dx}$ is a weight vector for each motion-energy at the specific delay d_x.

In this study L1-regularized least squares regression procedure was used to obtain the linear weights $\mathbf{w}_i$ [15, 16]. Note that the matrix multiplication between the temporally shifted stimulus vector ($\mathbf{s}$) and the weight vector ($\mathbf{w}_i$) is functionally equivalent to linear temporal convolution.

The training data consisted of 12 separate blocks of 10 minutes each. The first 6 seconds of each 10 minute block were discarded. (The assignment scheme described above assumes implicitly that these signals are not causally related to the stimuli, so they can be discarded safely.) The total number of samples in the training data was therefore (600-6) x 12 = 7128. The test data consisted of 9 separate blocks of 1 minute each. The first 6 seconds of each test block were also discarded. The total number of samples in the test data was therefore (60-6) x 9 = 486.

Selectivity Estimation

Once the motion-energy encoding model was estimated for each voxel a visualization procedure was used to recover the estimated spatial receptive field (Figures 2F and 2G left), spatial and temporal frequency tuning (Figures 2F and 2G right) for each voxel. Visualization of the receptive field is complicated by the fact that the motion-energy encoding model consists of many Gabor wavelets at multiple positions and scales, along with hemodynamic delays that are unique to each motion-energy filter and each voxel.

To estimate spatial selectivity we used a simulated system identification procedure in which each voxel was stimulated with a two-dimensional dynamic Gaussian white noise pattern, presented at various positions across the virtual display. The noise tiled the screen in a 17x17 grid. The motion-energy encoding model estimated for each voxel was used to obtain predicted responses. Predictions to uniform gray stimuli were obtained to determine the response baseline. These predicted responses describe the sensitivity of each voxel to each spatial position, and spatial responses for each voxel were aggregated together into a two-dimensional spatial selectivity map for visualization (Figures 2F and 2G left). A two-dimensional Gaussian was fit to the spatial receptive field estimated for each voxel and the center of the fitted Gaussian gave the angle and eccentricity for each voxel. These values were aggregated across voxels to form angle and eccentricity maps (Figures 2H and 2I). Voxel data were assigned to surface vertices using nearest neighbor interpolation and the maps were not smoothed. Voxels whose prediction accuracy was $p > 0.01$ are shown as gray in the Figures 2H-2J.

A similar procedure was used to estimate spatial and temporal frequency selectivity for each voxel (Figure 2F and 2G right). In this case the probe stimuli consisted of a set of full-field drifting gratings with the same set of directions, spatial and temporal frequencies as the Gabor wavelet basis set used in the motion-energy encoding model. Predicted responses were then estimated for each of the gratings. The spatiotemporal frequency selectivity map was obtained by

averaging predicted responses across all directions. Predictions of responses to a uniform gray field were used to determine the response baseline.

Likelihood Estimation

For identification and recostruction analysis, we calculate likelihood of stimuli given observed BOLD signals and estimated voxel-wise models. Let $\mathbf{r}$ denote the collection of observed BOLD signals ($\mathbf{r}=[r_i, \ldots, r_N]$, N is the number of voxels) and $\mathbf{s}$ denote motion-energy filtered stimuli (see Model Fitting). Assuming that BOLD signals are affected by Gaussian additive noise, the likelihood of the response $\mathbf{r}$ given the (motion-energy filtered) stimulus $\mathbf{s}$, or $p(\mathbf{r}\,|\,\mathbf{s})$, can be expressed by a multivariate Gaussian distribution [1]:

$$p(\mathbf{r}\,|\,\mathbf{s}) \propto \exp\{(\mathbf{r}-\hat{\mathbf{r}}(\mathbf{s}))\Sigma^{-1}(\mathbf{r}-\hat{\mathbf{r}}(\mathbf{s}))'\},$$

where $\hat{\mathbf{r}}(\mathbf{s})$ is the collection of predicted BOLD signals for each of the N voxels ($\hat{\mathbf{r}}(\mathbf{s})=[\hat{r}_i(\mathbf{s}), \ldots, \hat{r}_N(\mathbf{s})]$, see *model fitting* in Supplemental Information) given the stimulus $\mathbf{s}$ and the noise covariance matrix Σ for the training samples:

$$\Sigma = \left\langle (\mathbf{r}-\hat{\mathbf{r}}(\mathbf{s}))'(\mathbf{r}-\hat{\mathbf{r}}(\mathbf{s}))\right\rangle.$$

In most cases the matrix Σ is singular or close to singular. In these cases it is not possible to calculate the inverse of Σ in a stable manner. To overcome this problem we used Tikhonov regularization (equivalently ridge regularization) to estimate the inverse [51].

Voxel Selection

The scanning protocol produced data from about 15,000 voxels located in occipital cortex. Of these, we restricted our analysis to about 4,500 voxels located in the stimulated portions of visual areas V1, V2, V3, V3A and V3B. However, there was substantial variation in the predictive power of the motion-energy models obtained for these voxels. Therefore, to obtain optimal reconstructions for each subject we used only the 2,000 voxels that produced the most accurate predictions. The same voxel selection procedure was applied for identification analysis.

We used the following procedure to estimate prediction accuracy for each voxel. First, 90% of the samples in the training data set were used to fit a motion-energy encoding model for each voxel, and the remaining 10% of the training data were used to evaluate predictions of the fit model. The held out 10% of the data were chosen by first dividing the training data set into 50 second blocks and then choosing blocks at random until 10% of the samples were chosen. This procedure ensured that prediction accuracy was estimated using movies that were independent of those used for reconstruction.

Additional Notes on Reconstruction

The goal of the decoding analysis is to identify or reconstruct the stimulus that was most likely to have evoked measured BOLD signals. The motion-energy encoding model provides a mapping between stimuli and evoked BOLD signals. We can use the encoding model as a likelihood function to invert the mapping and recover the most likely stimuli from the BOLD signals, under some prior beliefs or constraints on the nature of stimuli observed. Because the motion-energy encoding model involves a non-linear convolution, decoding corresponds to a Bayesian deconvolution of BOLD signals (similar in concept to the approach used in dynamic causal

modeling [52-54]). Full Bayesian deconvolution would involve a mapping between a sequence of movie clips and a sequence of BOLD signals, which would cause a combinatorial explosion that would make decoding much more difficult. Therefore, to simplify numerical calculation we assume that the convolution is simply a delay in the hemodynamic response. This assumption allows us to convert the Bayesian deconvolution problem into a simpler problem, in which the causes of the current BOLD signal can be expressed in terms of stimuli presented at a fixed temporal delay (here four seconds). Furthermore, this assumption allows us to decode BOLD signals on a second by second basis, and to assess the decoding accuracy in terms of short (one second) movie sequences.

The sampled prior used in this study consisted of many dynamic movies. However, in some cases the movie was relatively static and did not change for many seconds (e.g., long-lasting static scenes). In preliminary studies we found that the average high posterior (AHP) reconstruction sometimes picked up many seconds of these static clips in a row, which visually biased the reconstruction. To avoid choosing similar clips too many times in succession, once we chose a clip from a single movie we discarded the subsequent five seconds of that movie from the selection process.

In preliminary studies we also explored reconstructions in which the 100 clips with the highest posterior probability were weighted according to their likelihood before averaging. However, we found that the weighted average tended to be dominated by one or two clips and the resulting reconstruction was worse than the MAP reconstruction. (This likely occurs because the sampled movie prior is relatively sparse.) Therefore, in the current study we simply averaged across all 100 of the clips with the highest posterior probability.

Evaluating the Accuracy of Reconstructions

To evaluate reconstruction accuracy we quantified the structural similarity between the natural movies used as stimuli in the experiment and their reconstructions. Structural similarity was quantified by calculating the correlation between the original movie stimuli and reconstructions within the motion-energy feature space. Although this study is the first to assess similarity in the motion-energy space, other studies have assessed similarity in a static complex wavelet feature space [8, 55].

To estimate structural similarity between the test movies and the MAP reconstructions, we first processed the test movies with the motion-energy filter bank (up to the temporal down-sampling stage, absent the hemodynamic coupling component used in the encoding models (Figure 1)). This produced a vector of motion-energy weights for each one second segment of the movie. The MAP reconstructions were treated the same way, giving a vector of motion-energy weights for each one second reconstruction. The similarity of the original movie and the MAP reconstruction was then taken as the motion-energy domain correlation between these two vectors, at a resolution of one second. The same procedure was applied to AHP reconstruction to obtain structural similarity between the test movies and the AHP reconstruction. In both cases a correlation of 1.0 indicates that the reconstruction captures all of the motion energy in the original stimulus, while a correlation of 0.0 indicates that the reconstruction is unrelated to the original stimulus.

To test significance of the reconstructions we compared the measured correlations to the distribution of motion-energy filter-domain correlations between the original movies and a set of clips drawn at random from the natural movie prior. A Wilcoxon rank-sum test was used to examine statistical significance between the correlation values from the actual reconstructions

and those from random clips. The Wilcoxon signed-rank test was also used to determine whether there was any significant difference in quality between MAP and AHP reconstructions. The chance performance was shown as the 99th percentile of the null distribution (Figure 4D, dashed line).

Supplemental References

48. Efron, B., and Tibshirani, R. (1993). An Introduction to the Bootstrap (New York: Chapman & Hall).
49. Jones, J.P., and Palmer, L.A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J Neurophysiol *58*, 1233-1258.
50. DeAngelis, G.C., Ohzawa, I., and Freeman, R.D. (1993). Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development. J Neurophysiol *69*, 1091-1117.
51. Marquardt, D.W. (1970). Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. Technometrics *12*, 591-612.
52. Friston, K.J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. NeuroImage *19*, 1273-1302.
53. Penny, W., Ghahramani, Z., and Friston, K. (2005). Bilinear dynamical systems. Philosophical transactions of the Royal Society of London. Series B, Biological sciences *360*, 983-993.
54. Makni, S., Beckmann, C., Smith, S., and Woolrich, M. (2008). Bayesian deconvolution of fMRI data using bilinear dynamical systems. NeuroImage *42*, 1381-1396.
55. Brooks, A.C., and Pappas, T.N. (2006). Structural similarity quality metrics in a coding context: exploring the space of realistic distortions. Proc. SPIE *6057*, 299-310.