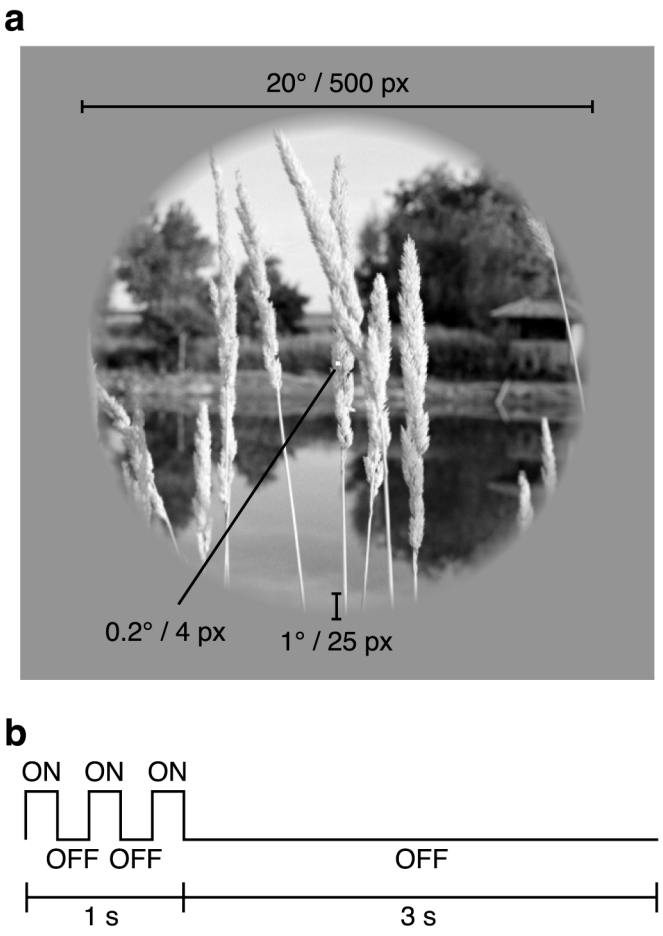
SUPPLEMENTARY INFORMATION

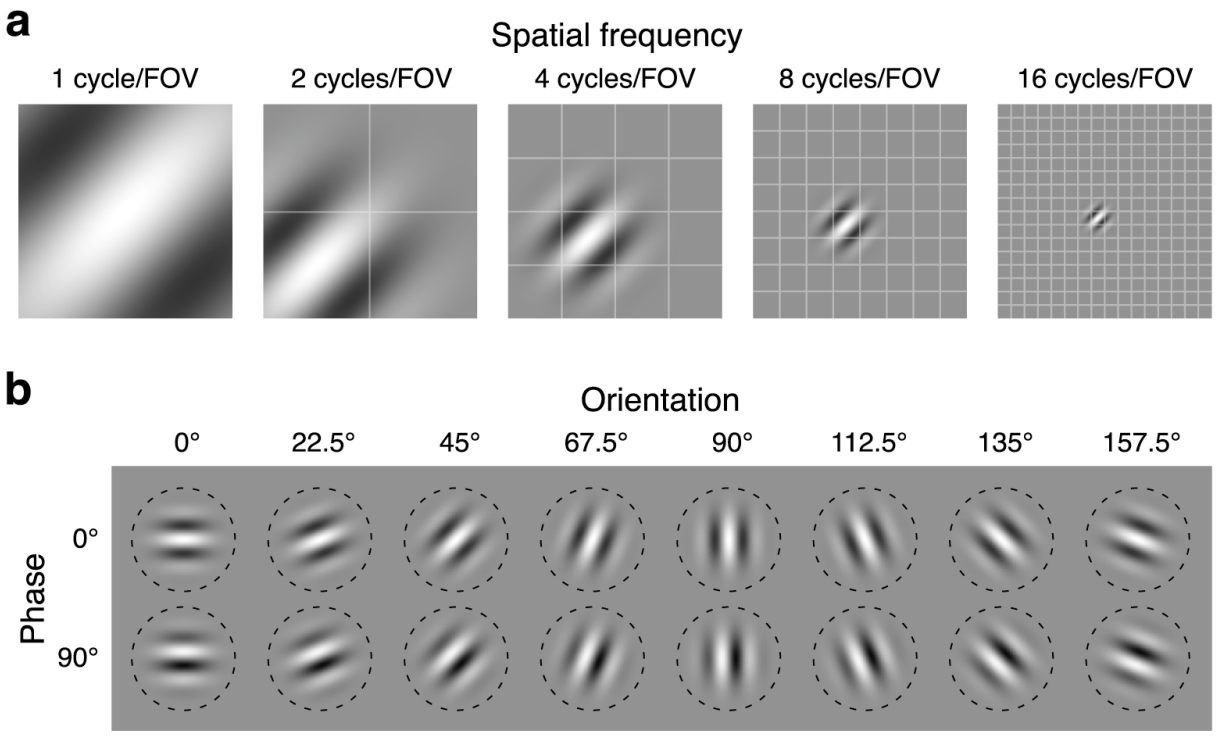
Identifying natural images from human brain activity

Kendrick N. Kay[1], Thomas Naselaris[2], Ryan J. Prenger[3] & Jack L. Gallant[1,2]
[1]Department of Psychology, [2]Helen Wills Neuroscience Institute, [3]Department of Physics,
University of California, Berkeley, California 94720, USA

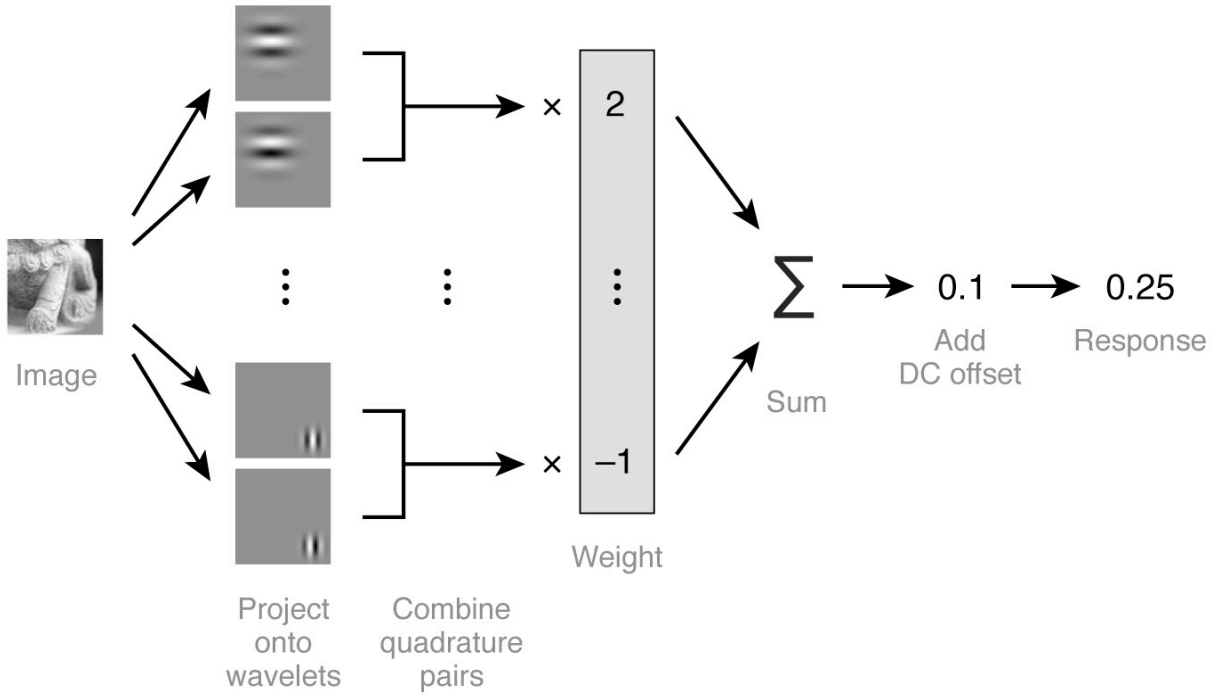
Overview



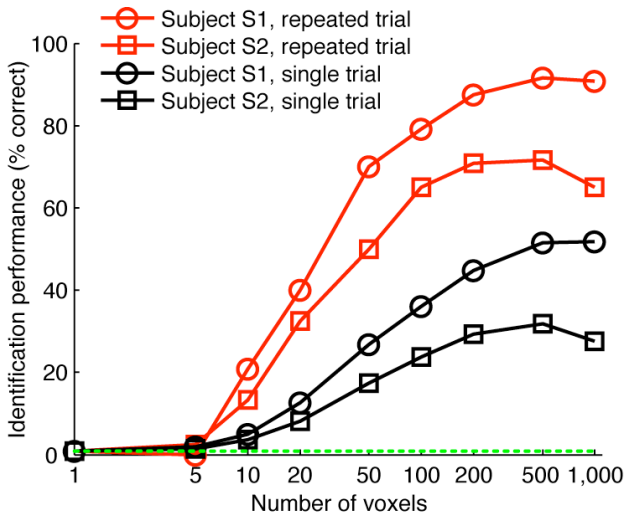
Supplementary Figure 1. Stimulus design. The stimuli consisted of sequences of grayscale natural photographs. **a**, Spatial characteristics. The photographs were masked with a circle (20° diameter) and placed on a gray background. The outer edge of each photograph (1° width) was linearly blended into the background. A central white square (0.2° side length) served as the fixation point. **b**, Temporal characteristics. The photographs were presented for 1 s with a delay of 3 s between successive photographs. Each 1-s presentation consisted of a photograph being flashed ON–OFF–ON–OFF–ON where ON corresponds to presentation of the photograph for 200 ms and OFF corresponds to presentation of the gray background for 200 ms.



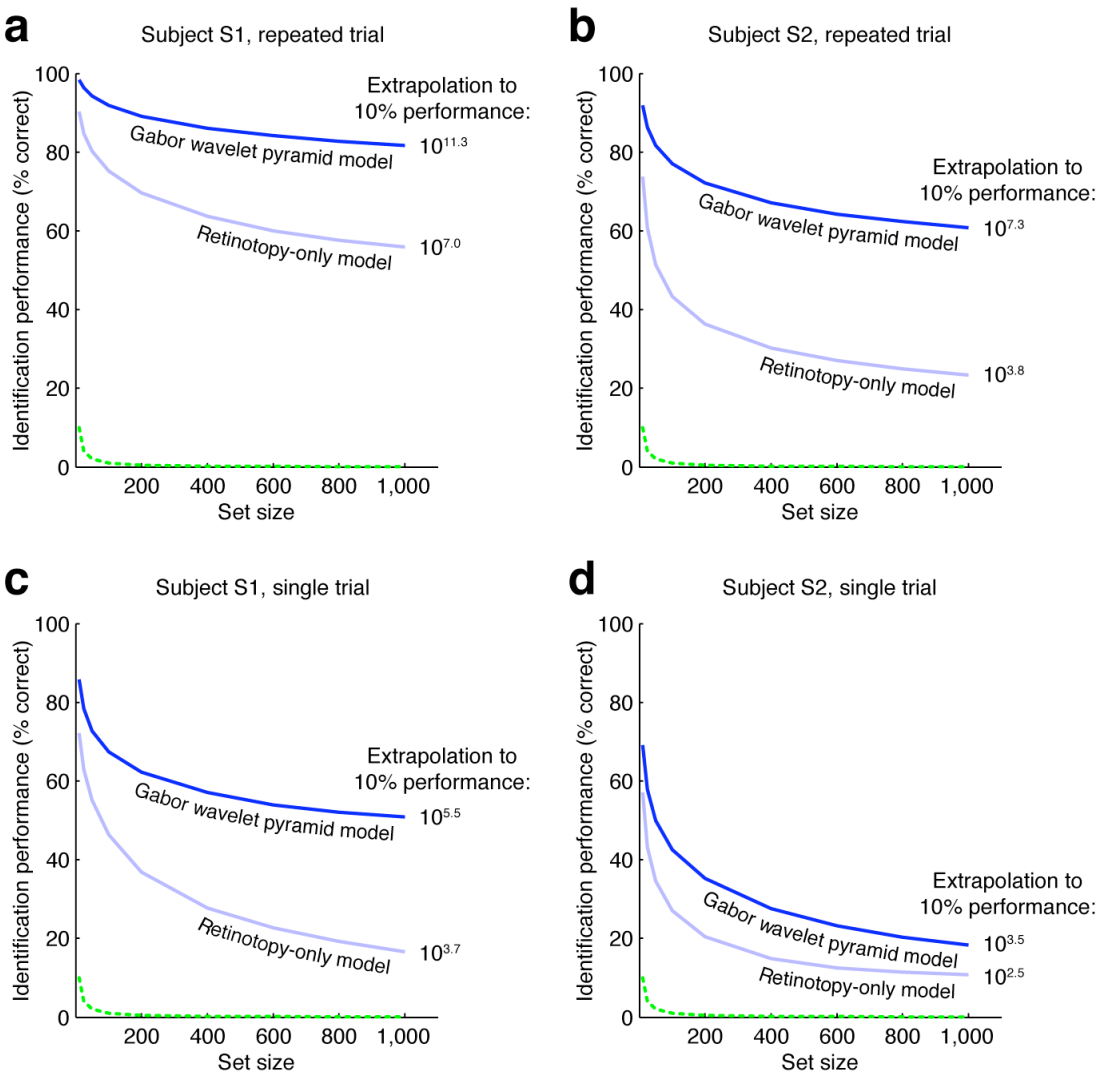
Supplementary Figure 2. Gabor wavelet pyramid design. The receptive-field model used in the present study is based on a Gabor wavelet pyramid[11–13]. **a**, Spatial frequency and position. Wavelets occur at five (or, in some cases, six) spatial frequencies. This panel depicts one wavelet at each of the first five spatial frequencies. At each spatial frequency f cycles per field-of-view (FOV), wavelets are positioned on an $f \times f$ grid, as indicated by the translucent lines. **b**, Orientation and phase. At each grid position, wavelets occur at eight orientations and two phases. This panel depicts a complete set of wavelets for a single grid position. Dashed lines indicate the bounds of the mask associated with each wavelet.



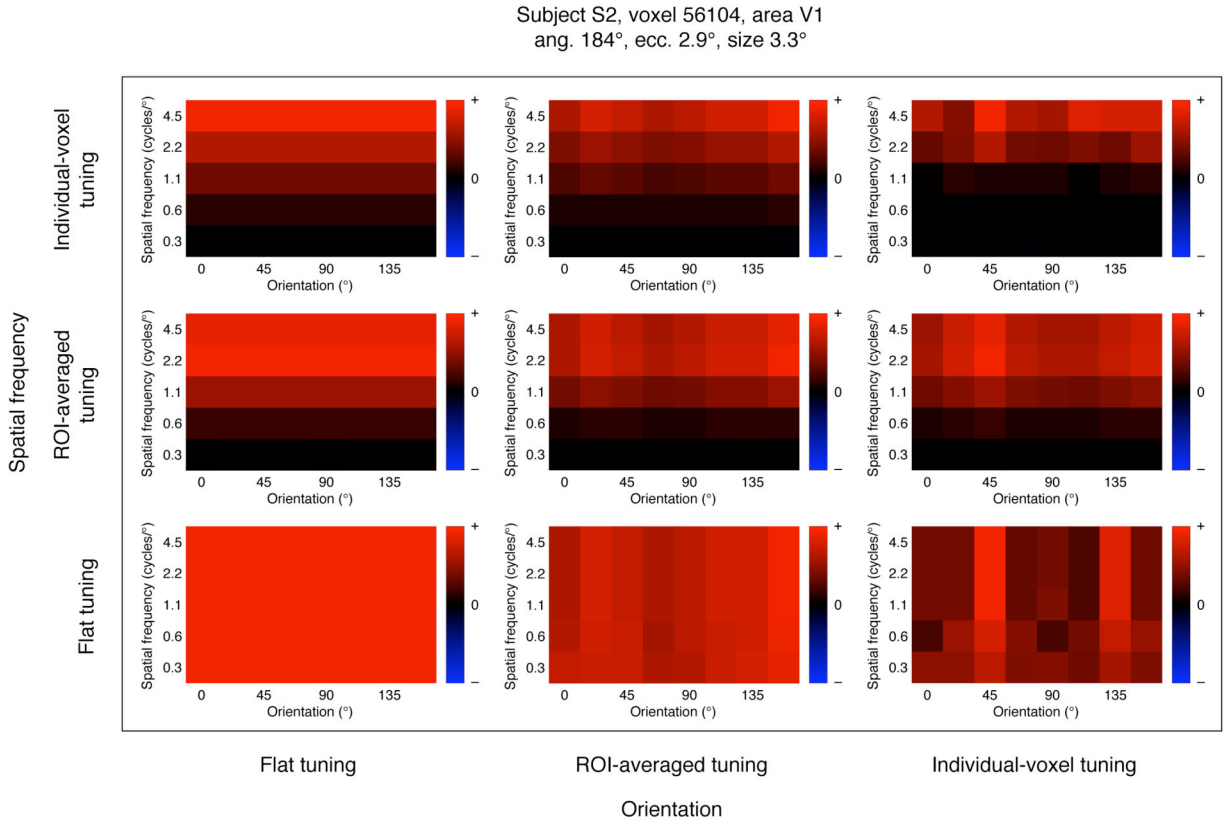
Supplementary Figure 3. Gabor wavelet pyramid model. Each image is projected onto the individual Gabor wavelets comprising the Gabor wavelet pyramid (see Supplementary Fig. 2). The projections for each quadrature pair of wavelets are squared, summed, and square-rooted, yielding a measure of contrast energy. The contrast energies for different quadrature wavelet pairs are weighted and then summed. Finally, a DC offset is added. The weights are determined by gradient descent with early stopping (see Supplementary Methods 6).



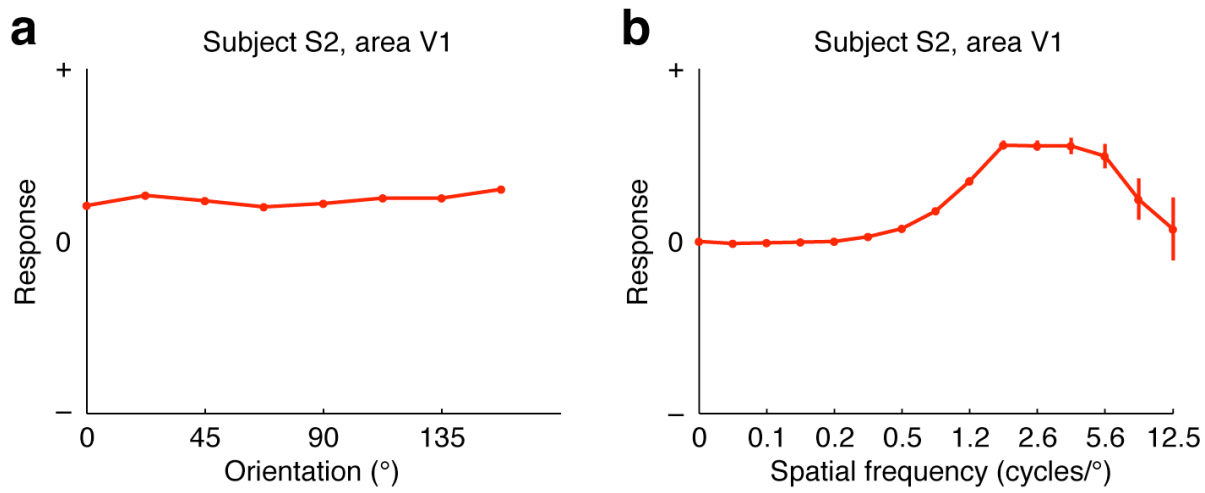
Supplementary Figure 4. Effect of number of voxels on identification performance. To optimize performance of the identification algorithm, we preferentially selected voxels whose receptive-field models had the highest predictive power (see Supplementary Methods 7). In this figure the *x* axis indicates the number of voxels selected and the *y* axis indicates identification performance. The dashed green line indicates chance performance, and results were obtained for a set size of 120 images. In all cases optimal performance was achieved using about 500 voxels. Therefore, all identification results in this study were obtained using 500 voxels.



Supplementary Figure 5. Identification performance for the retinotopy-only model. To determine whether identification is a mere consequence of the retinotopic organization of early visual areas, we evaluated an alternative retinotopy-only model that captures the location and size of each voxel's receptive field but discards orientation and spatial frequency information. **a**, Comparison of identification performance for the retinotopy-only (RO) model and the Gabor wavelet pyramid (GWP) model (results for subject S1 and repeated trials). The *x* axis indicates set size and the *y* axis indicates identification performance. The number to the right of each line gives the estimated set size at which performance declines to 10% correct, and the dashed green line indicates chance performance. Performance for the RO model was substantially lower than for the GWP model. **b**, Results for subject S2 and repeated trials. Once again the RO model performed substantially worse than the GWP model. **c–d**, Single-trial results for subjects S1 and S2. Although identification performance was poorer overall when single trials were used, the GWP model still outperformed the RO model. These results collectively indicate that spatial tuning alone does not yield optimal identification performance; identification improves substantially when orientation and spatial frequency tuning are included in the model.

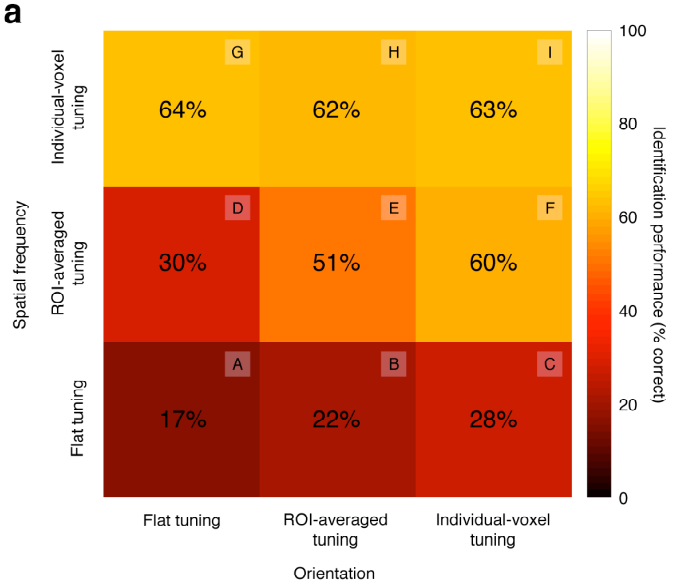


Supplementary Figure 6. Example of constraints on orientation and spatial frequency tuning. To assess the individual contributions of orientation and spatial frequency tuning to identification performance, we evaluated several constrained versions of the Gabor wavelet pyramid model. These models were constructed by fixing the spatial envelope of each voxel and then imposing different constraints on orientation and spatial frequency tuning (see Supplementary Methods 9 for details). This figure illustrates the tuning of one representative voxel under the various models. Nine plots are arranged in three columns and three rows. Each plot depicts the joint orientation and spatial frequency tuning obtained under one specific model (format is the same as in Fig. 2b). The three columns represent different constraints on orientation tuning: in the left column it is constrained to be flat; in the middle column it is constrained to match the mean orientation tuning across voxels in the corresponding region-of-interest (i.e. V1, V2, or V3); in the right column it is unconstrained (the model is allowed full flexibility in orientation tuning). The three rows represent different constraints on spatial frequency tuning: in the bottom row it is constrained to be flat; in the middle row it is constrained to match the mean spatial frequency tuning across voxels in the corresponding region-of-interest; in the top row it is unconstrained. These plots demonstrate that the models successfully incorporate the intended tuning constraints. (In the bottom-right plot orientation tuning at low spatial frequencies is not perfectly matched to the marginal orientation tuning. This is a consequence of the fact that the lowest-frequency wavelets are truncated by the field-of-view, effectively increasing their spectral bandwidth.)



Supplementary Figure 7. Example of ROI-averaged tuning curves. Several of the constrained versions of the Gabor wavelet pyramid model involve fixing the orientation or spatial frequency tuning curve of a voxel to match the mean tuning curve across voxels in the corresponding region-of-interest (i.e. V1, V2, or V3). **a**, Example ROI-averaged orientation tuning curve for area V1. The *x* axis indicates orientation and the *y* axis indicates predicted response. Error bars indicate ± 1 s.e.m. across voxels (bootstrap procedure). The orientation tuning curve is nearly flat. **b**, Example ROI-averaged spatial frequency tuning curve for area V1. The format is the same as panel a, except that the *x* axis indicates spatial frequency. The spatial frequency tuning curve is band-pass.

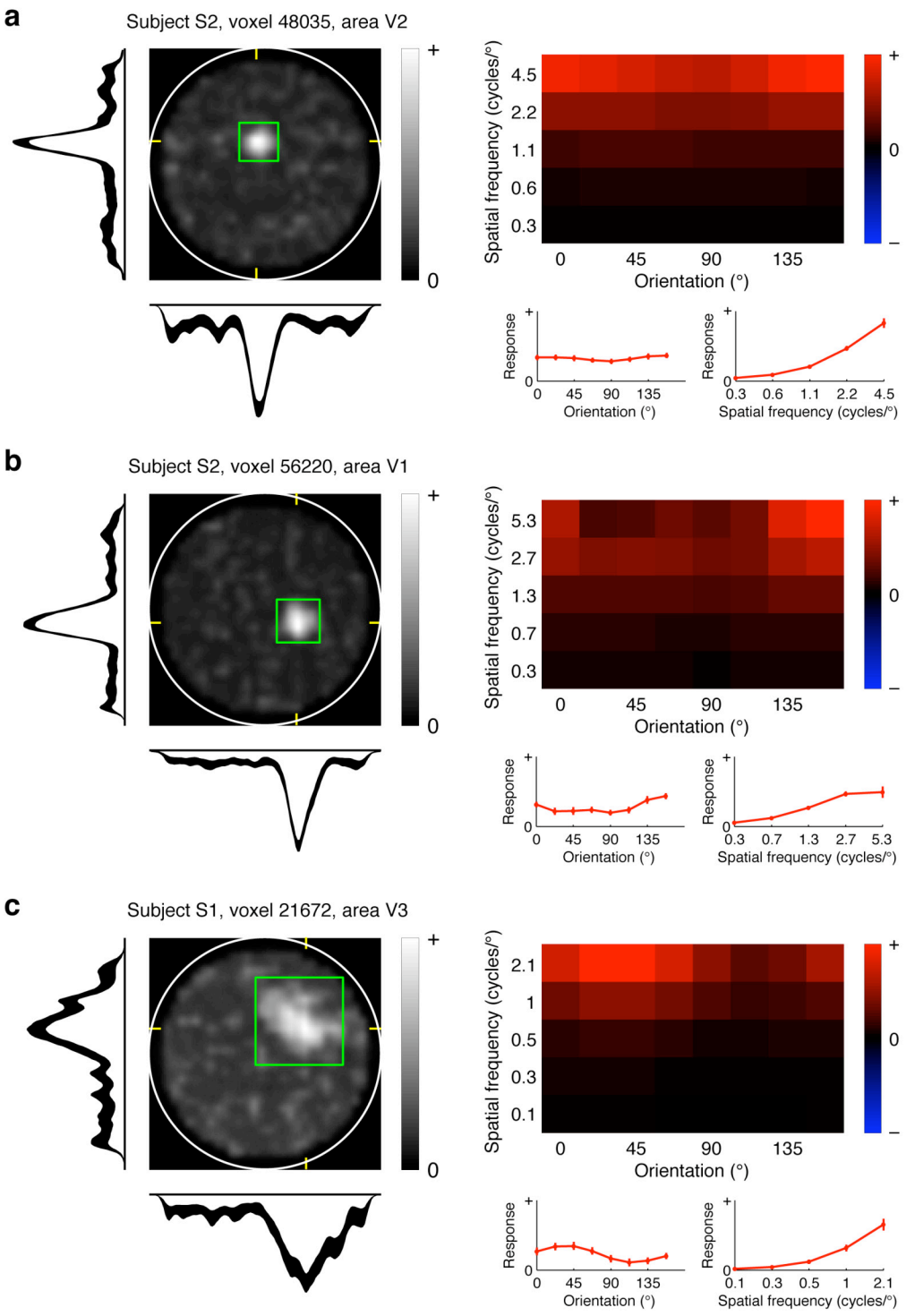
a



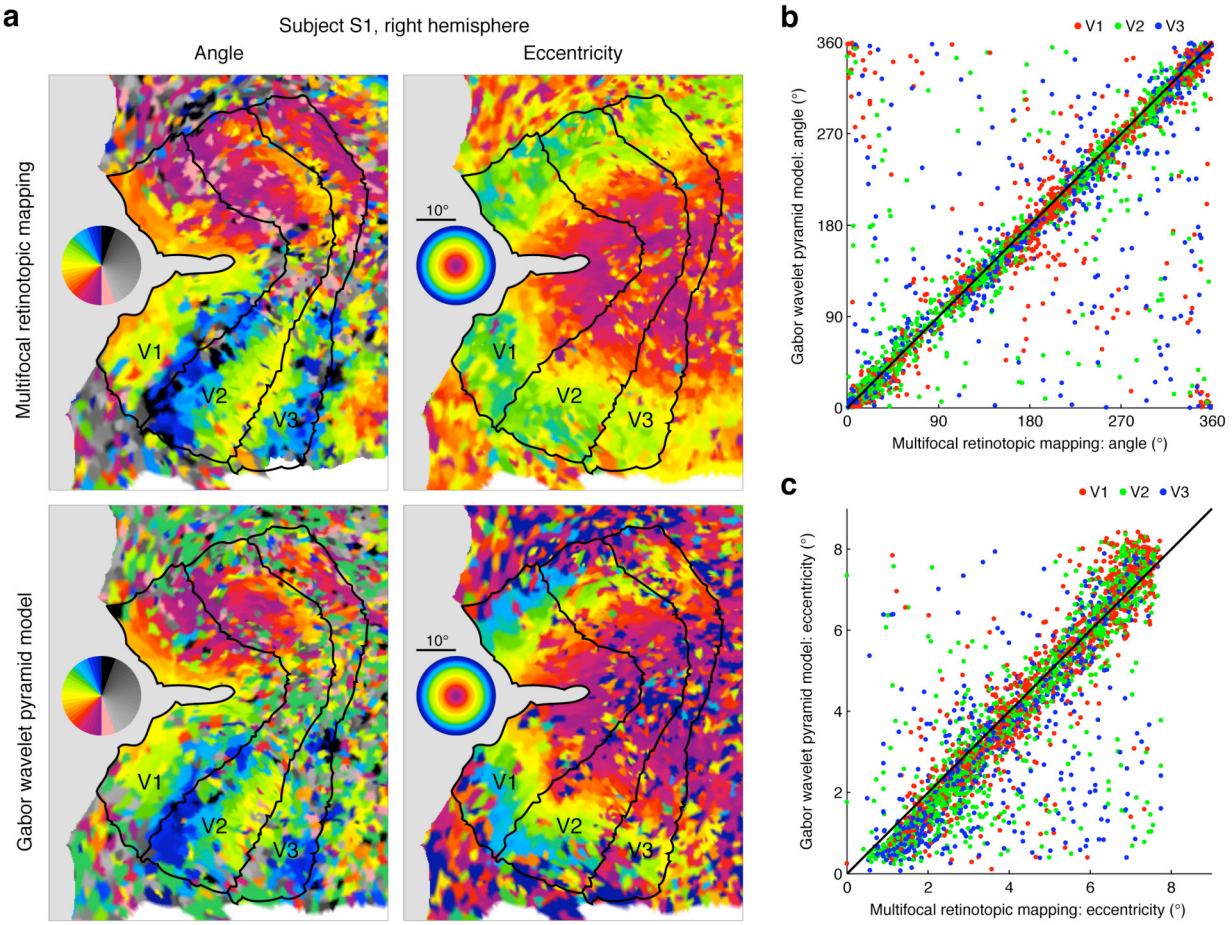
b

	A	B	C	D	E	F	G	H	I
A	–	0.03	8e-6	2e-4	7e-16	6e-21	2e-22	1e-21	2e-21
B	–	–	3e-3	0.03	4e-14	2e-18	5e-20	7e-19	3e-18
C	–	–	–	0.4	4e-10	9e-17	2e-16	2e-15	3e-15
D	–	–	–	–	1e-9	2e-16	3e-16	7e-18	3e-19
E	–	–	–	–	–	4e-4	5e-6	4e-5	8e-5
F	–	–	–	–	–	–	0.08	0.3	0.2
G	–	–	–	–	–	–	–	–	–
H	–	–	–	–	–	–	0.3	–	0.5
I	–	–	–	–	–	–	0.4	–	–

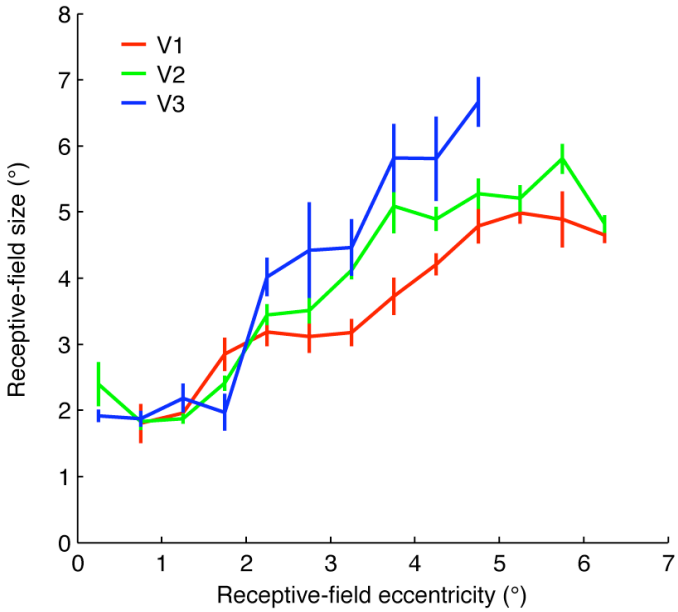
Supplementary Figure 8. Contribution of orientation and spatial frequency tuning to identification performance. Constrained versions of the Gabor wavelet pyramid model were used to investigate the individual contributions of orientation and spatial frequency tuning to identification performance (see Supplementary Fig. 6). **a**, Summary of identification performance under each model. The nine models are labeled by capital letters, and are arranged in three columns and three rows. Different columns represent different constraints on orientation tuning, and different rows represent different constraints on spatial frequency tuning (as in Supplementary Fig. 6). Colors and percentages denote identification performance achieved under each model (repeated trial, 1,000 images, performance averaged across subjects). Both orientation and spatial frequency tuning contribute to identification performance (C > A and G > A), but spatial frequency tuning is relatively more important (G > C). Voxel-to-voxel variation in orientation and spatial frequency tuning also contributes to identification performance (F > E and H > E). **b**, Statistical comparisons of identification performance. This table provides *p*-values for all pairwise model comparisons (one-tailed paired sign test, *p*-values rounded up). A red *p*-value indicates that the model in the corresponding column performed significantly better than the model in the corresponding row ($p < 0.05$), while a black *p*-value indicates that the improvement was not statistically significant ($p \geq 0.05$). The symbol '–' indicates that performance for the column model was less than or equal to that for the row model. The differences in identification performance noted in panel a are all statistically significant.



Supplementary Figure 9. Additional examples of receptive-field models. a–c, Receptive-field models for three representative voxels. The format of each panel is the same as that of Fig. 2. Receptive-field (RF) location, size, orientation tuning, and spatial frequency tuning all vary substantially across voxels. The RFs also vary in reliability; for example, the RF shown in panel c exhibits less reliable spatial tuning than the RFs shown in panels a–b.



Supplementary Figure 10. Validation of retinotopic information derived from receptive-field models. Since retinotopy is a well-established property of voxels in early visual areas[14,16,18], one way to validate the Gabor wavelet pyramid (GWP) model is to confirm that it produces reasonable estimates of voxel receptive-field location. In this figure we compare angle and eccentricity estimates obtained from the GWP model with those obtained from the multifocal (MF) retinotopic mapping technique[17,31] (see Supplementary Methods 11). Note that the data used for the MF technique were completely independent of the data used for the GWP model. **a**, Comparison of retinotopic maps for a representative hemisphere. Voxel data were assigned to surface vertices using nearest neighbor interpolation, and the maps were not smoothed or thresholded. Black lines indicate the boundaries of visual areas V1, V2, and V3. (The same boundaries are replicated on each map.) Overall, the GWP maps are similar to the MF maps and exhibit the typical retinotopic organization[32,33]. The GWP maps are somewhat noisier than the MF maps, which is expected given that the MF technique is specifically optimized to provide retinotopic information. **b**, Quantitative comparison of angle estimates. Dots represent individual voxels taken across subjects (voxels for which the predictive power of the GWP model was not statistically significant at $p < 0.01$ are omitted). Notice that the MF and GWP angle estimates are well matched. **c**, Quantitative comparison of eccentricity estimates (format same as panel b). The MF and GWP eccentricity estimates are generally well matched, but there appear to be systematic discrepancies at the lowest and highest eccentricities. The likely cause of the discrepancies is the spatial granularity of the stimuli used for MF mapping[32].



Supplementary Figure 11. Relationship between receptive-field size and eccentricity. In the course of fitting the Gabor wavelet pyramid model, estimates of the location and size of each voxel's receptive field (RF) were obtained. We examined the relationship between RF size and eccentricity to see if the expected pattern of results could in fact be demonstrated. In this figure the *x* axis indicates RF eccentricity and the *y* axis indicates RF size. (RF size is defined as ± 2 s.d. of a fitted two-dimensional Gaussian; see Supplementary Methods 5.) Voxels were pooled across subjects and then binned by eccentricity. (To ensure robust results, voxels for which RF predictive power was not statistically significant at $p < 0.01$ or for which estimated RF location was not completely within the stimulus bounds were omitted before pooling.) For each bin with at least 10 voxels, the median RF size is plotted, with error bars indicating ± 1 s.e. (bootstrap procedure). RF size increases with eccentricity and across visual areas, consistent with previous fMRI studies[15,19,34–36]. The fact that our model estimation approach uncovers differences in RF size across areas suggests that it could potentially reveal other area differences.

Subject	*Visual area*	*Total number of voxels*	*High SNR (% of total)*	*High predictive power (% of total)*	*High SNR and high predictive power (% of high SNR)*
S1	V1	1331	431 (32%)	533 (40%)	406 (94%)
	V2	2208	659 (30%)	677 (31%)	558 (85%)
	V3	1973	425 (22%)	343 (17%)	260 (61%)
S2	V1	1513	275 (18%)	382 (25%)	256 (93%)
	V2	1982	369 (19%)	426 (21%)	291 (79%)
	V3	1780	223 (13%)	224 (13%)	138 (62%)

Supplementary Table 1. Signal-to-noise ratio of voxel responses and predictive power of receptive-field models. The column *High SNR (% of total)* gives the number of voxels with a signal-to-noise ratio (SNR) greater than 1.5; *High predictive power (% of total)* gives the number of voxels for which the predictive power of the best initial model was statistically significant ($p < 0.01$, bootstrap procedure); and *High SNR and high predictive power (% of high SNR)* gives the number of voxels that satisfied both criteria. (See Supplementary Methods 3 and 6 for details concerning SNR and predictive power, respectively.) Although SNR varied greatly across subjects, SNR was fairly consistent for areas V1, V2, and V3 within each subject. Predictive power generally decreased from V1 to V2 to V3, likely reflecting the fact that the Gabor wavelet pyramid model is not optimal for visual areas beyond V1.

Supplementary Discussion 1. Classification-based decoding methods cannot be used to identify novel images

Previous classification-based studies did not identify novel images

Several fMRI studies of visual cortex[4,5,37,38] have shown that classification-based decoding methods can be used to determine the category of an image seen by an observer, even if the image is a novel instance of the category. In addition, one neurophysiological study of inferotemporal neurons[7] showed that classification methods can be used to determine which object was seen by an observer, even if the object was presented at novel positions or scales. At a superficial level these results may seem to contradict our claim that classification methods cannot be used to identify novel images. However, there are two key differences between these previous studies and the present study. First, the previous studies achieved decoding for only specific kinds of novel images (e.g. novel images drawn from fixed categories). In contrast the present study achieves decoding for arbitrary novel natural images.

Second, the previous studies demonstrated classification, not identification. The goal of classification is to discriminate images belonging to a given category from those belonging to other categories. Classification thus aggregates over the individual images belonging to a given category. In contrast, the goal of identification is to discriminate an individual image from a number of other images. Identification thus treats each image as a distinct entity. To illustrate these ideas, consider a hypothetical experiment that measures brain activity evoked by an image of a dog. The goal of classification is to assign the image to one of several pre-defined categories such as *dog* or *cat*; the goal of identification is to discriminate the specific dog image from a number of other images (regardless of category membership).

Limitations of classification-based decoding methods

Classification-based decoding methods are inherently limited by the fixed set of categories that are used in training. For example, suppose a classifier is trained to discriminate brain activity evoked by dogs from that evoked by cats; without additional training the classifier would be unable to discriminate brain activity evoked by birds from that evoked by dogs or cats. This limited generality entails that classification methods cannot be used to identify novel images. To illustrate: suppose we adapt the classification framework to the problem of identification by treating each individual image as if it defines a unique category[3,7,8]. If previous measurements of brain activity evoked by each image are available for training purposes, standard classification procedures can achieve identification. However, in the case of novel images (i.e. no previous measurements of brain activity evoked by the images are available), we are faced with a critical problem: how do we perform classification for categories we have not trained for? (For additional discussion of the limitations of classification methods, see ref. 3.)

An extension of classification-based decoding methods yields poor identification performance

Is it possible to extend classification-based decoding methods to achieve identification of novel images? To address this question we developed a straightforward extension of classification methods. In this analysis we treated each image used in the model estimation stage of the

experiment as if it defined a unique category (similar to refs. 3, 7, 8). Thus, the 1,750 voxel activity patterns measured in the model estimation stage of the experiment were taken to represent 1,750 unique categories. We call these the *category activity patterns*.

For each of the 120 voxel activity patterns measured in the image identification stage of the experiment, we attempted to identify which specific image had been seen. This was accomplished by taking a given voxel activity pattern *m* and finding the category activity pattern most similar to *m* (similarity was quantified by Pearson's *r*). We call the image associated with the found category activity pattern the *matched image*. (Intuitively, the matched image is the image from the model estimation stage of the experiment that is "brain-wise" most similar to the image seen by the subject.) The matched image was then compared with each of the 120 images used in the image identification stage of the experiment, and the image most similar to the matched image was selected. Two metrics for image similarity were tested: correlation of pixel luminance and correlation of local contrast. (To calculate the local contrast of a given image, the image was divided into $n° \times n°$ blocks and root-mean-square contrast was calculated for each block. The results reported below were obtained using the value of *n* that yielded the best performance, $n = 0.6$.)

Identification performance using the pixel luminance metric was 1.7% (2/120) and 0.8% (1/120) for subjects S1 and S2, respectively (repeated trial). These values were not significantly above chance ($p \geq 0.05$, one-tailed binomial test). Identification performance using the local contrast metric was 5% (6/120) and 5.8% (7/120) for subjects S1 and S2, respectively (repeated trial). These values were above chance ($p < 0.0001$, one-tailed binomial test) but far below the performance levels achieved by the identification algorithm described in the main text (92% and 72% for subjects S1 and S2, respectively). These results suggest that classification methods cannot be easily extended to achieve accurate identification of novel images.

Supplementary Discussion 2. Comparison of classification, identification, and reconstruction

The problems of classification, identification, and reconstruction can be defined formally. Let x_1, x_2, x_3, ... represent different images. (There may be an infinite number of images.) Let l represent a function that maps images to a certain set of labels. For example, $l(x_i)$ is the label assigned to image x_i. Let p_i represent an activity pattern evoked by image x_i on a given trial. We define the following problems:

- *Classification*: given activity pattern p_i, determine $l(x_i)$.
- *Identification*: given activity pattern p_i and a finite set of images (e.g. $\{x_2\ x_7\ x_3\}$) such that x_i is a member of the set, determine x_i.
- *Reconstruction*: given activity pattern p_i, determine x_i.

At the most general level the three problems are similar: in each case the goal is to infer certain information based on a given activity pattern. In theory, identification can be considered a special case of classification where the label assigned to an image is simply the index of that image in the given set of images. However, classification normally refers to the case where a single label is assigned to multiple images, so in practice identification is distinct from classification. Furthermore, although the goal of both identification and reconstruction is to determine the specific image that had evoked a given activity pattern, in identification a set of potential images is provided whereas in reconstruction no such set is provided.

Note that these definitions do not specify what information is available to train a decoder, though this is an important issue in the present study. Unlike classification-based methods, our decoding method can achieve accurate identification of an image even when that image is novel, i.e. even when brain activity evoked by the image is not available for training.

Supplementary Discussion 3. Previous research on voxel tuning properties

The receptive-field model used in the present study is based on a Gabor wavelet pyramid (GWP). The GWP has long been regarded as the standard model of how primary visual cortex (V1) represents shape[11–13]. Under the assumption that fMRI activity reflects local pooled neural activity[1,2,39–41], it is reasonable to suppose that the GWP model is appropriate for describing voxels in early visual areas. Indeed, previous results suggest that fMRI activity in V1 reflects the average activation of a population of Gabor filters[28]. The GWP model used in the present study describes tuning along the dimensions of space, orientation, and spatial frequency. Each of these dimensions has been previously investigated in fMRI.

Spatial tuning has received considerable attention from many laboratories. The phase-encoded retinotopic mapping technique was introduced in the early days of fMRI[14,16,18] and continues to be widely used. This method provides an estimate of the location of each voxel's receptive field. Recent studies have demonstrated that estimates of voxel receptive-field size can be extracted from phase-encoded data through the use of a spatial tuning model[3,15,19,35] such as a two-dimensional Gaussian. An alternative method for estimating spatial tuning is the multifocal retinotopic mapping technique where the stimulus consists of spatial elements (e.g. wedges, rings, sectors) flashed pseudorandomly across the visual field[17,31]. This method provides a more direct estimate of the spatial envelope of a voxel receptive field, but is limited by the granularity of the stimuli and by the assumption of linear spatial summation[17].

Orientation tuning has typically been investigated in fMRI by using adaptation-based techniques[42–47] or by pooling signals across many voxels[20,48]. However, recent classification-based studies have shown that individual voxels have a slight orientation bias[1,2]. These studies are also noteworthy since they demonstrate that multivariate analysis techniques can increase the amount of information extracted from fMRI data compared to conventional univariate analysis techniques.

Spatial frequency is the final dimension represented in the GWP model. Of the various dimensions, spatial frequency has been the least studied in fMRI. A few studies have shown that fMRI signals pooled across entire visual areas exhibit some spatial frequency tuning[21,22,49]. However, these studies did not investigate potential voxel-to-voxel variation in tuning.

Most fMRI experiments measure tuning along one dimension at a time. This approach assumes that stimulus dimensions are separable and that they can be measured independently of one another. In addition, fMRI experiments usually measure tuning using artificial stimuli such as gratings and checkerboard patterns (but see exceptions[21,28,50]). In the present study the GWP model is fit to voxel responses evoked by natural images. This approach measures tuning along multiple dimensions simultaneously, and produces a unified description of how images are mapped onto fMRI activity.

Supplementary Methods 1. Design of model estimation and image identification runs

The experiment consisted of two distinct stages, model estimation and image identification. Model estimation runs and image identification runs were conducted in the same fMRI scan sessions. Each estimation run used 70 distinct images presented 2 times each. Each run consisted of 168 trials, and had a duration of 168 trials × 4 s = 11.2 min. The first four and last four trials were null trials (no images presented). For the remaining 160 trials, every 8th trial was also a null trial. The presentation order of the images was determined by randomly generating a large number of sequences under the constraint that same image could not be presented on consecutive trials, and then choosing the sequence that yielded the greatest estimation efficiency[51].

Each identification run used 12 distinct images presented 13 times each. The presentation order of the images was determined by an m-sequence[52] of level 13, order 2, and length $13^2 - 1 = 168$. The m-sequence included 12 null trials (no images presented). Code for m-sequence generation was provided by T. Liu (http://fmriserver.ucsd.edu/ttliu/mttfmri_toolbox.html). During stimulus presentation the first 6 trials were repeated at the end of the 168-trial sequence. In the repeated-trial analysis, data collected during the initial 6 trials were ignored[53,54]. In the single-trial analysis, all data were used. Each run had a duration of 174 trials × 4 s = 11.6 min.

Supplementary Methods 2. Reconstruction and co-registration of brain volumes

Functional and anatomical brain volumes were reconstructed using the ReconTools software package (https://cirl.berkeley.edu/view/BIC/ReconTools). For functional volumes, a phase correction was applied to reduce Nyquist ghosting and image distortion, and differences in slice acquisition times were corrected by sinc interpolation.

All functional volumes acquired for a given subject were registered to a single spatial reference frame. Automated motion correction procedures (SPM99, http://www.fil.ion.ucl.ac.uk/spm/) were used to correct differences in head positioning within scan sessions by rigid-body transformations. Manual co-registration procedures (in-house software) were used to correct differences in head positioning across scan sessions by affine transformations. Each functional volume was resampled only once (by sinc interpolation); this minimized interpolation errors that could accumulate over multiple resamplings. No additional spatial filtering was applied to the functional volumes.

Supplementary Methods 3. Time-series pre-processing

The time-series data for each voxel were pre-processed prior to the model estimation and image identification stages of the experiment. The primary purpose of the pre-processing was to estimate and deconvolve voxel-specific response timecourses from the time-series data. This decreased the computational requirements of subsequent analyses by reducing the effective number of data points. Pre-processing was based on the basis-restricted separable (BRS) model (see Supplementary Methods 4). In brief, the BRS model uses a set of basis functions to characterize the shape of the response timecourse and a set of parameters to characterize the amplitudes of responses to different images.

During pre-processing the time-series data were analyzed both as repeated trials and as single trials. The repeated-trial analysis produced, for each voxel, an estimate of the amplitude of the response (a single value) evoked by each distinct image used in the model estimation and image identification runs. In this case each estimate reflects data from multiple image presentations. The single-trial analysis produced, for each voxel, an estimate of the amplitude of the response (a single value) evoked by each trial of the model estimation and image identification runs. In this case each estimate reflects data from a single image presentation.

Repeated-trial analysis

The following procedure was performed for each voxel in each scan session. First, the BRS model was fit to the time-series data from the model estimation runs. A set of Fourier basis functions was used to characterize the shape of the response timecourse, and a separate parameter was used to characterize the amplitude of the response to each distinct image. Fitting the BRS model produced an estimated timecourse and a set of estimated response amplitudes. If necessary, the estimated timecourse and estimated response amplitudes were multiplied by –1 so that the estimated timecourse had a positive value at a time lag of 5 s. (This prevented ambiguity with respect to the sign of the response amplitudes.) We refer to the estimated timecourse as the *hemodynamic response function* (HRF), and the estimated response amplitudes as the *model estimation responses*.

Second, the BRS model was fit to the time-series data from the image identification runs. One basis function was used to characterize the shape of the response timecourse; this basis function was simply the HRF calculated in step 1. A separate parameter was used to characterize the amplitude of the response to each distinct image. Fitting the BRS model produced a set of estimated response amplitudes. We refer to the estimated response amplitudes as the *image identification responses*.

Third, the model estimation responses were standardized, and the same transformation (i.e. the same mean and standard deviation) was applied to the image identification responses. Standardization improved the consistency of responses across scan sessions (data not shown).

After this procedure was performed for each voxel in each scan session, model estimation responses and image identification responses were aggregated across scan sessions. For each model estimation response, the ratio between the absolute value of the response and its standard

error was calculated. For a given voxel the median ratio across model estimation responses was taken as the signal-to-noise ratio (SNR) of that voxel.

Single-trial analysis

The following procedure was performed for each voxel in each scan session. First, the BRS model was fit to the time-series data from the model estimation and image identification runs. One basis function was used to characterize the shape of the response timecourse; this basis function was simply the HRF calculated in the repeated-trial analysis. A separate parameter was used to characterize the amplitude of the response to each trial. Fitting the BRS model produced a set of estimated response amplitudes. We refer to the estimated response amplitudes for the model estimation and image identification runs as the *single-trial model estimation responses* and *single-trial image identification responses*, respectively. Next, the single-trial model estimation responses were standardized, and the same transformation (i.e. the same mean and standard deviation) was applied to the single-trial image identification responses. After this procedure was performed for each voxel in each scan session, single-trial model estimation responses and single-trial image identification responses were aggregated across scan sessions.

Analysis for additional scan sessions

In addition to the scan sessions for the main experiment, three additional scan sessions were conducted (see Methods in the main text). Each of these scan sessions consisted solely of image identification runs. To analyze the time-series data from these scan sessions, the following procedure was performed for each voxel in each scan session. First, the BRS model was fit to the time-series data from the image identification runs using the procedure described in step 1 of the repeated-trial analysis. This produced a set of image identification responses. Second, the image identification responses were standardized. Third, the BRS model was fit to the time-series data from the image identification runs using the procedure described in step 1 of the single-trial analysis. This produced a set of single-trial image identification responses. Fourth, the single-trial image identification responses were standardized. After this procedure was performed for each voxel in each scan session, image identification responses and single-trial identification responses were aggregated across scan sessions.

Construction of voxel activity patterns

The results of the repeated-trial and single-trial analyses were used to construct the voxel activity patterns used in the image identification stage of the experiment. Each voxel activity pattern represents the ensemble voxel response to an image. *Repeated-trial activity patterns* reflect data from multiple image presentations, and were constructed by concatenating individual voxels' estimated response amplitudes for an image. *Single-trial activity patterns* reflect data from single image presentations, and were constructed by concatenating individual voxels' estimated response amplitudes for a single trial.

Supplementary Methods 4. Basis-restricted separable model

The basis-restricted separable (BRS) model was used to pre-process the time-series data for each voxel (see Supplementary Methods 3). The BRS model assumes that each distinct image evokes a fixed response and that responses to different images sum over time. In addition, the model assumes that the response timecourses elicited by different images differ by only a scale factor[53]. To account for stimulus-related effects, the BRS model uses a set of basis functions to characterize the shape of the response timecourse[51] and a set of parameters to characterize the amplitudes of responses to different images. To account for noise-related effects, the model uses a set of polynomials[53] of degrees 0 through 3 and a first-order autoregressive noise model[55].

Let t be the number of time-series data points, e be the number of distinct images or trials, l be the number of points in the response timecourse, m be the number of timecourse basis functions, and p be the number of polynomial regressors. The time-series data for a given voxel are modeled as

$$\mathbf{y} = (\mathbf{X} * (\mathbf{Lc}))\mathbf{h} + \mathbf{Sb} + \mathbf{n}$$

where $\mathbf{y}$ is the data ($t \times 1$), $\mathbf{X}$ is the stimulus matrix ($t \times e$), $\mathbf{L}$ is the set of timecourse basis functions ($l \times m$), $\mathbf{c}$ is a set of parameters ($m \times 1$), $*$ denotes convolution, $\mathbf{h}$ is a set of response amplitudes ($e \times 1$), $\mathbf{S}$ is the set of polynomial regressors ($t \times p$), $\mathbf{b}$ is a set of parameters ($p \times 1$), and $\mathbf{n}$ is a noise term ($t \times 1$).

For the analysis of the time-series data as repeated trials, $\mathbf{X}$ consisted of one column per distinct image, where each column was a binary sequence with ones indicating the onsets of an image. For the analysis of the time-series data as single trials, $\mathbf{X}$ consisted of one column per trial, where each column was a binary sequence with a one indicating the onset of a trial. In cases where the shape of the timecourse was unknown, $\mathbf{L}$ was a set of Fourier basis functions consisting of a constant function and sine and cosine functions with 1, 2, and 3 cycles. These basis functions extended from 1 to 16 s after image onset. In cases where an estimate of the shape of the timecourse was available, $\mathbf{L}$ was simply taken to be that estimate.

Model parameters were estimated using a variant of the Cochrane-Orcutt procedure[55]. After initializing $\mathbf{h}$ to all ones, iterations alternated between ordinary least-squares estimation of $\mathbf{c}$ and $\mathbf{b}$ while holding $\mathbf{h}$ fixed and ordinary least-squares estimation of $\mathbf{h}$ and $\mathbf{b}$ while holding $\mathbf{c}$ fixed. After each iteration autoregressive noise parameters were estimated from the residuals of the model fit. These autoregressive noise parameter estimates were used to transform the data and design matrix prior to the next iteration. Fitting proceeded until convergence of parameter estimates.

Supplementary Methods 5. Model estimation

In the model estimation stage of the experiment, a receptive field was estimated for each voxel using the Gabor wavelet pyramid (GWP) model. The model estimation procedure is complicated because it involves multiple uses of the GWP model. For this reason we provide a high-level description of the procedure in this section, and present specific details of the GWP model in Supplementary Methods 6.

Rough localization of the receptive field

The first step of the model estimation procedure was to obtain a rough localization of the receptive field (RF). This was accomplished by fitting several initial models to the data. Each of the initial models covered a specific region of the stimulus (called the *field-of-view*), and was an instantiation of the GWP model at a resolution of 128 px × 128 px. One model covered the full 20° × 20° extent of the stimulus. In this case performance was limited by the fact that the maximum wavelet spatial frequency was 1.6 cycles/°. To better characterize voxels tuned to higher spatial frequencies, two additional models were used. One covered the central 10.1° × 10.1° of the stimulus, and the other covered the central 5.2° × 5.2° of the stimulus. In these cases the maximum wavelet spatial frequencies were 3.2 cycles/° and 6.2 cycles/°, respectively. (Voxels tuned to higher spatial frequencies tended to be found in more central regions of the visual field; data not shown.)

For each of the initial models, the RF was constrained to be orientation invariant. This was accomplished by summing over groups of input channels that differ in orientation but share the same spatial frequency and position, prior to fitting the model. The orientation invariance constraint reduced the number of free parameters and improved predictive power (data not shown). (Note that the final model was not constrained to be orientation invariant; see below.) There were a total of 1,367 free parameters for each initial model.

Precise localization of the receptive field

The second step of the model estimation procedure was to obtain a more precise estimate of the RF location. This was accomplished by fitting an isotropic two-dimensional Gaussian function to the spatial envelope associated with each initial model. The RF location was estimated as the region bounded by ± 2 s.d. of the fitted Gaussian. (The RF size was taken to be the size of this region.) For the 10.1° × 10.1° and 5.2° × 5.2° models, the estimated RF location was considered valid only if the 2-s.d. region was completely within the field-of-view of the model. This criterion excluded models artificially truncated by the field-of-view.

Of all the initial models that yielded a valid estimate of RF location, the model that achieved the least squared error on a separate stopping set was chosen (see Supplementary Methods 6). We refer to this model as the *best initial model*, and it was taken as providing the best estimate of the RF location. To reduce computational demands, subsequent analyses included only those voxels for which the predictive power of the best initial model was statistically significant (see Supplementary Methods 6).

Final estimate of the receptive field

The last step of the model estimation procedure was to obtain a final estimate of the RF. This was accomplished by fitting a GWP model that was specifically tailored to the estimated RF location. This model had a resolution of 64 px × 64 px, and was not constrained to be orientation invariant. There were a total of 2,730 free parameters in this final model.

Supplementary Methods 6. Gabor wavelet pyramid model

In Supplementary Methods 5 we described how the Gabor wavelet pyramid (GWP) model was used to estimate the receptive field of each voxel. In this section we provide specific details of the GWP model, such as how model parameters are determined.

Basic framework

The GWP model is applied to a specific region of the stimulus, called the *field-of-view* (FOV). The resolution is typically 64 px × 64 px, though in some cases, a resolution of 128 px × 128 px is used. The model describes how the portion of the stimulus within the FOV (henceforth simply referred to as the *image*) is transformed into a predicted response. Note that the GWP model does not include a temporal component because voxel-specific response timecourses are removed from the time-series data in pre-processing.

Stimulus pre-processing

To accommodate a variety of FOVs and resolutions, the stimuli used in the experiment were pre-processed at multiple resolutions. The dimensions of the pre-processed stimuli were given by $\min(500,\text{round}(2^{9-x/8}))$ where x ranges from 0 to 24. For example, for $x = 0$ the stimuli were left at the original resolution of 500 px × 500 px, and for $x = 10$ the stimuli were downsampled to a resolution of 215 px × 215 px. Stimuli were converted to luminance values using the measured luminance response of the goggles (see below). The mean luminance across all stimuli was then subtracted.

The luminance response of the goggles was measured with a Minolta LS-110 photometer (Konica Minolta Photo Imaging, Mahwah, NJ). The luminance response of the left-eye display was slightly different from that of the right-eye display; for analysis, the average of the two luminance responses was assumed. The minimum, maximum, and mean luminance was 0.8 cd/m^2, 11.1 cd/m^2, and 6.3 cd/m^2, respectively.

Design of the wavelet pyramid

The Gabor wavelet pyramid is illustrated in Supplementary Fig. 2. For the 64 px × 64 px model resolution, wavelets occur at five spatial frequencies: 1, 2, 4, 8, and 16 cycles per FOV. (For the 128 px × 128 px model resolution, wavelets occur at six spatial frequencies: 1, 2, 4, 8, 16, and 32 cycles per FOV.) At each spatial frequency f cycles per FOV, wavelets are positioned on an $f \times f$ grid. At each grid position wavelets occur at eight orientations, 0, 22.5°, 45°, ..., and 157.5°, and two quadrature phases, 0° and 90°. An isotropic Gaussian mask is used for each wavelet, and its size relative to spatial frequency is such that all wavelets have a spatial frequency bandwidth of 1 octave and an orientation bandwidth of 41°. A luminance-only wavelet that covers the entire image is also included.

Wavelets are truncated to lie within the bounds of the image, and are restricted in spatial extent by setting to zero the portions of the masks whose values are less than 0.01 of the peak value

(Supplementary Fig. 2b). Each wavelet is made zero-mean and unit-length within the bounds of its associated mask.

Transformation from image to predicted response

The following steps transform a given image into the predicted response from the GWP model (Supplementary Fig. 3). First, the image is projected onto the set of Gabor wavelets. The projections for each quadrature pair of wavelets are then squared, summed, and square-rooted, yielding a set of *input channels*. These input channels reflect the contrast energy contained in quadrature wavelet pairs. (For the luminance-only wavelet, the projection is squared, multiplied by 2, and square-rooted.) Next, the input channels are weighted by a set of values called the *kernel* and then summed. Finally, a DC offset is added to the result.

Wavelets positioned near the edge of the circular stimulus mask (see Supplementary Fig. 1) yield artifactually small projections. To avoid instability in parameter estimation, the projection for a given wavelet is set to zero if more than half of its associated mask lies beyond 90% of the stimulus radius.

The quantification of contrast energy is a nonlinear operation that transforms the stimulus into a space where the relationship between stimulus and response is more linear; for this reason, the GWP model is termed a *linearized model*[10]. A purely linear model that characterizes the voxel response as a weighted sum of the raw wavelet projections yields very poor fits (data not shown).

Estimation of model parameters

Responses to the images used in the model estimation runs of the experiment are used to fit the GWP model. Formally, let p be the number of images, and q be the number of input channels. The voxel responses were modeled as

$$\mathbf{y} = \mathbf{X}\mathbf{h} + c\mathbf{1} + \mathbf{n}$$

where $\mathbf{y}$ is the set of responses ($p \times 1$), $\mathbf{X}$ is the set of input channels ($p \times q$), $\mathbf{h}$ is the kernel ($q \times 1$), c is the DC offset (1×1), $\mathbf{1}$ is a vector of ones ($p \times 1$), and $\mathbf{n}$ is a noise term ($p \times 1$). Model parameters were estimated using gradient descent with early stopping[56]. Gradient descent is an iterative fitting technique where the difference between the model fit and the data is gradually reduced. Early stopping is a form of regularization[10] where the magnitude of model parameter estimates are shrunk in order to prevent overfitting.

The specific procedure was as follows. A randomly selected 20% of the responses were removed and kept as a stopping set. The mean of the remaining responses $\mathbf{y}_\mu$ (1×1) was subtracted, yielding responses $\tilde{\mathbf{y}}$ ($p \times 1$). The mean of each input channel $\mathbf{X}_\mu$ ($1 \times q$) was subtracted and the standard deviation of each input channel $\mathbf{X}_\sigma$ ($1 \times q$) was divided out, yielding input channels $\tilde{\mathbf{X}}$ ($p \times q$). The kernel was initialized to all zeros ($\mathbf{h}_1 = \mathbf{0}$) and then iteratively updated using gradient descent:

$$\mathbf{h}_{i+1} = \mathbf{h}_i - \varepsilon \mathbf{g}_i$$

where $\mathbf{h}_i$ is the kernel at iteration i, $\mathbf{g}_i$ is the normalized error gradient at iteration i, and $\varepsilon = 0.001$ is the step size. The normalized error gradient is given by

$$\mathbf{g}_i = \left[\left[\tilde{\mathbf{X}}^{\mathrm{T}}\left(\tilde{\mathbf{X}}\mathbf{h}_i - \tilde{\mathbf{y}}\right)\right] + \alpha\mathbf{g}_{i-1}\right]$$

where $[\mathbf{x}] = \mathbf{x} / \|\mathbf{x}\|$ represents vector length normalization, $\alpha = 0.9$ is a momentum parameter[57], and $\mathbf{g}_0 = \mathbf{0}$. Iterations proceeded until the squared error on the stopping set no longer decreased, or until the squared error on the responses no longer decreased. The final estimate of the kernel was calculated as

$$\hat{\mathbf{h}} = \mathbf{h}_{final} ./ \mathbf{X}_{\sigma}^{\mathrm{T}}$$

where $\mathbf{h}_{final}$ is the kernel at the last iteration and $./$ denotes element-by-element division. The final estimate of the DC offset was calculated as

$$\hat{c} = \tilde{\mathbf{y}} - \mathbf{X}_{\mu}\hat{\mathbf{h}}$$

where the symbols are as defined earlier.

Estimation of variance

One hundred bootstrap samples were drawn from the original set of responses, and parameter estimates were obtained for each bootstrap sample. (The size of each bootstrap sample was equal to the number of responses, and the stopping set was selected after each bootstrap sample was drawn.) Standard errors on parameter estimates were calculated as the standard deviation across bootstraps. Final parameter estimates were calculated as the mean across bootstraps.

To prevent artificially high variance of parameter estimates, the number of fitting iterations was held constant across bootstrap samples. This was accomplished as follows. Prior to bootstrapping, parameter estimates were obtained using gradient descent with early stopping on the original set of responses. The number of fitting iterations n was recorded. Then, for each bootstrap sample, parameter estimates were obtained using gradient descent for n iterations.

Quantification of predictive power

An objective measure of the quality of a receptive-field model is how well the model predicts responses to images not used for model estimation[10]. Here, the predictive power of a receptive-field model was calculated as the correlation (Pearson's r) between measured and predicted responses for the images used in the image identification runs of the experiment. (There were 120 images used in the image identification runs, and these were distinct from the 1,750 images used in the model estimation runs; see Methods in the main text.) A bootstrap procedure was used to estimate statistical significance of predictive power ($r > 0$, one-tailed p-values).

Calculation of tuning curves

Tuning curves for space, orientation, and spatial frequency were calculated for each receptive-field (RF) model. To calculate the spatial tuning curve, i.e. spatial envelope, of an RF, the wavelet mask associated with each input channel was normalized to sum to 1, and was then scaled by the absolute value of the kernel weight associated with that input channel. The spatial

envelope was obtained by summing all wavelet masks. To calculate the orientation and spatial frequency tuning curves of an RF, a set of sinusoidal gratings were constructed at the same orientations and spatial frequencies used in the GWP model. At each combination of orientation and spatial frequency, gratings were constructed at multiple phases. The response of the RF to each grating was calculated, and tuning curves were obtained by averaging responses over one or more of the dimensions of orientation, spatial frequency, and phase.

Supplementary Methods 7. Image identification

Voxel selection

In our experiment approximately 5000 voxels were located in the stimulated portions of visual areas V1, V2, and V3 (see Supplementary Fig. 10 and Supplementary Table 1). There was substantial variation in the predictive power of the receptive-field models obtained for different voxels. Therefore, to optimize performance of the identification algorithm, we preferentially selected voxels whose receptive-field models had the highest predictive power. (Predictive power was quantified as how well a given model predicts responses to images not used to estimate the model; see Supplementary Methods 6). Note that the image to be identified was not included in the calculation of predictive power; this prevented voxel selection bias.

All identification results in this study were obtained using 500 voxels, as that number yields optimal performance (Supplementary Fig. 4). Most of these voxels were located in area V1 where predictive power was highest (Supplementary Table 1).

For measurement of identification performance under the Gabor wavelet pyramid and retinotopy-only models, voxels were selected based on the predictive power of the specific model under consideration. This ensured that each model had the best possible chance at performing well. For measurement of identification performance under the various constrained versions of the Gabor wavelet pyramid model, a single, fixed set of voxels was used. (The voxels in this set were selected based on the predictive power of the model that imposed no constraints on orientation and spatial frequency tuning.) Fixing the set of voxels used ensured that differences in identification performance directly reflect the different constraints imposed by the models.

Identification performance for different set sizes

To measure identification performance for set sizes up to 1,000 images, the following procedure was used. First, a library of 999 images was constructed. These images were randomly selected and were different from the images used in the model estimation and image identification stages of the experiment. Then, for set size s and measured voxel activity pattern m, identification performance was calculated as the probability that the predicted voxel activity pattern for the correct image is more correlated with m than the predicted voxel activity patterns for $s-1$ images drawn randomly from the library:

$$f(m,s) = \prod_{i=1}^{s-1} \frac{1{,}000 - g(m) - i}{1{,}000 - i}$$

where $f(m,s)$ is identification performance and $g(m)$ is the number of library images whose predicted voxel activity patterns were more correlated with m than with the correct image. Finally, identification performance was averaged over all measured voxel activity patterns m.

To measure identification performance for larger set sizes, an extrapolation method was used. First, the correlation between the measured voxel activity pattern m and the predicted voxel activity pattern for each library image was calculated. This produced a distribution of 999 correlation values. Next, the distribution was smoothed using a Gaussian kernel. Kernel width

was chosen by pseudo-likelihood cross-validation[58] using code provided by A. Ihler (http://ttic.uchicago.edu/~ihler/code/). (Smoothing in this way produces a better estimate of the true underlying distribution, and is reasonable given that the library images were randomly selected.) Identification performance was then calculated as

$$f(m,s) = \left(1 - h(m)\right)^{s-1}$$

where $f(m,s)$ is identification performance and $h(m)$ is the fraction of the smoothed distribution larger than the correlation between m and the predicted voxel activity pattern for the correct image. This equation quantifies the probability that the predicted voxel activity pattern for the correct image is more correlated with m than with the predicted voxel activity patterns for $s - 1$ images drawn randomly from all possible images. Finally, identification performance was averaged over all measured voxel activity patterns m. To validate the described extrapolation method, we calculated empirical performance levels for a set size of six million images, and confirmed that these values are accurately estimated by extrapolation.

Estimation of the noise ceiling

The noise ceiling on identification performance was estimated in order to determine whether differences in identification performance across subjects could be attributed to differences in signal-to-noise ratio (see Fig. 4a). The noise ceiling is the theoretical maximum performance that could ever be achieved, given the level of noise in the data. To estimate the noise ceiling, 25 bootstrap-like simulations were conducted for each of the 120 images used in the image identification stage of the experiment. In each simulation, the first step was to generate a measured voxel activity pattern for the correct image. This was accomplished by taking the mean of a random sample drawn from the single-trial activity patterns evoked by the correct image. The next step was to generate a predicted voxel activity pattern for each potential image the subject could have seen. This was accomplished by taking the mean of a random sample drawn from the single-trial activity patterns evoked by each potential image. (The intuition here is that the quality of the predicted voxel activity patterns is limited only by intrinsic measurement variability, not by the predictive power of receptive-field models.) Finally, the image whose predicted voxel activity pattern was most correlated with the measured voxel activity pattern was selected. The noise ceiling was calculated as the percentage of simulations where identification was successful.

Supplementary Methods 8. Retinotopy-only model

The retinotopy-only (RO) model characterizes the response of each voxel as a function of the luminance and contrast of a specific region of the stimulus (this region is henceforth simply referred to as the *image*). There are two input channels. The luminance channel represents absolute deviation from mean luminance. (It has been shown that changes in uniform illumination evoke fMRI activity in early visual areas[27].) The contrast channel represents the total energy contained in the image excluding overall luminance. Note that the RO model is invariant to the particular orientations and spatial frequencies present in the image.

The RO model provides a plausible functional description of a voxel in early visual areas, and it is similar to recently proposed models of phase-encoded retinotopic mapping data[3,15,19,35]. Since the RO model captures only spatial tuning, it serves as a way of testing whether the additional orientation and spatial frequency tuning captured by the Gabor wavelet pyramid (GWP) model have a significant impact on identification performance. If orientation and spatial frequency tuning are irrelevant for identification or if they cannot be estimated reliably from voxel responses, then performance for the RO model should be at least as good as performance for the GWP model.

To ensure that the RO and GWP models are compared fairly, the RO model was applied to the same estimated receptive-field location as used for the GWP model (see Supplementary Methods 5), and both models were fit using the same gradient descent method (see Supplementary Methods 6).

Spatially-weighted metrics for luminance and contrast

To implement the RO model we must choose metrics for luminance and contrast. The standard metrics for luminance and contrast are the mean and standard deviation of the pixel luminance values, respectively. These metrics are spatially homogenous in the sense that all portions of the image contribute equally. However, it is reasonable to presume that the receptive field of a voxel exhibits spatial gradation such that portions of the image near the center of the receptive field contribute more strongly to the response than portions of the image near the periphery of the receptive field. Indeed, previous studies[3,15,35] have proposed a two-dimensional Gaussian model of the spatial envelope of a voxel receptive field. Moreover, the receptive fields obtained in the present study under the GWP model do appear to be spatially graded (see Fig. 2 and Supplementary Fig. 9).

To accommodate spatial gradation the RO model uses the following metrics. The *spatially-weighted luminance* of an image is given by

$$L = \frac{\sum_i w_i x_i}{\sum_i w_i}$$

where L is the spatially-weighted luminance, w_i is the weight on pixel i, and x_i is the luminance of pixel i. The *spatially-weighted contrast* of an image is given by

$$C = \sqrt{\frac{\sum_i w_i \left(x_i - L\right)^2}{\sum_i w_i}}$$

where C is the spatially-weighted contrast and the other symbols are as defined earlier. These metrics are calculated at the original stimulus resolution (downsampling is not necessary). Note that in the case where all weights are equal to one, the spatially-weighted metrics for luminance and contrast reduce to the standard metrics for luminance and contrast.

Transformation from image to predicted response

Let G represent the two-dimensional Gaussian fit to the spatial envelope associated with the best initial model (as described in Supplementary Methods 5). The following steps transform a given image into the predicted response from the RO model. First, the spatially-weighted luminance of the image is calculated using the weights provided by G. The absolute value of the result constitutes the first input channel. (This full-wave rectification parallels how luminance is treated in the GWP model.) Next, the spatially-weighted contrast of the image is calculated using the weights provided by G. The result constitutes the second input channel. The two input channels are then weighted by a set of values and summed. Finally, a DC offset is added to the result.

Validation of the spatially-weighted metrics

To verify that the spatially-weighted metrics yield reasonable results, we compared identification performance achieved using the spatially-weighted metrics with that achieved using the standard metrics. (The standard metrics were calculated using the region of the stimulus bounded by ± 2 s.d. of the Gaussian function G.) Identification performance was 55% and 42% for the spatially-weighted metrics and standard metrics, respectively (repeated trial, 120 images, performance averaged across subjects). This validates the spatially-weighted metrics and indicates that we have cast the RO model in the best possible light.

Supplementary Methods 9. Constrained versions of the Gabor wavelet pyramid model

Several constrained versions of the Gabor wavelet pyramid (GWP) model were constructed in order to assess the individual contributions of orientation and spatial frequency tuning to identification performance. These models are based on the GWP model instantiated at a resolution of 64 px × 64 px, and impose various constraints on orientation and spatial frequency tuning. The models were applied to the same estimated receptive-field location as used for the GWP model (see Supplementary Methods 5) and were fit using the same gradient descent method (see Supplementary Methods 6).

Model simplification

To facilitate manipulation of orientation and spatial frequency tuning, the spatial envelope of the GWP model was first fixed. This was accomplished by weighting the image with the two-dimensional Gaussian associated with the estimated receptive-field location (see Supplementary Methods 5), and summing over input channels that differ in position but share the same orientation and spatial frequency. (Note that different voxels had different spatial envelopes.) To facilitate imposition of tuning constraints, input channels were linearly transformed such that weights on input channels directly reflect how the model responds to sinusoidal gratings (details of the transformation are provided in a later section).

Constraints on orientation and spatial frequency tuning

Systematic constraints were imposed on orientation and spatial frequency tuning. Three different constraints were used for each dimension, yielding a total of 3 × 3 = 9 different models. Under the constraint of *flat tuning*, the tuning curve of each voxel is constrained to be entirely flat. Under *ROI-averaged tuning*, the tuning curve of each voxel is constrained to match the mean tuning curve across voxels in the corresponding region-of-interest (i.e. V1, V2, or V3), and any voxel-to-voxel variation in tuning is ignored. Under *individual-voxel tuning*, each voxel is allowed full flexibility in tuning, so voxel-to-voxel variation in tuning is captured. (For an illustrative example, see Supplementary Fig. 6.)

Tuning constraints were achieved by applying marginalization operations to input channels. To achieve flat tuning, input channels were summed across the relevant dimension; to achieve ROI-averaged tuning, input channels were multiplied by the appropriate ROI-averaged tuning curve (see Supplementary Fig. 7) and then summed across the relevant dimension; and to achieve individual-voxel tuning, input channels were left as-is. To illustrate, consider the model that imposes flat orientation tuning and ROI-averaged spatial frequency tuning. This model was constructed by summing over input channels that differ in orientation but share the same spatial frequency, multiplying the resulting channels by the ROI-averaged spatial frequency tuning curve, and summing the results.

Comparison with the retinotopy-only model

Like the retinotopy-only (RO) model, the constrained versions of the GWP model help assess the contribution of orientation and spatial frequency tuning to identification performance. The RO

model serves as a simple, plausible alternative to the GWP model, and assesses the overall importance of orientation and spatial frequency. In contrast, the constrained versions of the GWP model examine the individual contributions made by orientation and spatial frequency, and constitute a more direct investigation of the orientation and spatial frequency information conveyed by the GWP model.

The model that imposes flat orientation and spatial frequency tuning is similar to the RO model in that both models capture spatial tuning but discard orientation and spatial frequency information. However, the models are not equivalent: they are constructed from different image bases (Gabor wavelet basis vs. pixel basis) and incorporate different kinds of nonlinearities. The models also differ in how they handle overall luminance, so their predicted responses can diverge substantially at very low spatial frequencies.

Technical detail on the transformation of input channels

Before transformation, weights on input channels do not reflect how the model responds to sinusoidal gratings. This is due to the fact that the Gabor wavelets have overlapping spectra and the fact that there are different numbers of wavelets at different spatial frequency levels of the Gabor wavelet pyramid.

The crux of the transformation lies in simulating the response of the model to gratings. Gratings are constructed at 16 equally spaced phases at each combination of orientation and spatial frequency used in the GWP model. This yields a total of 8 orientations × 5 spatial frequencies × 16 phases = 640 gratings. The gratings are then used to construct a set of input channels $\mathbf{G}$ ($n \times q$) where $n = 640$ is the number of gratings and $q = 41$ is the number of input channels.

Suppose that responses evoked by the gratings were actually measured. These responses could be modeled as

$$\mathbf{r} = \mathbf{G}\mathbf{k} + \mathbf{n}$$

where $\mathbf{r}$ ($n \times 1$) is the set of grating responses, $\mathbf{k}$ is the kernel ($q \times 1$), and $\mathbf{n}$ is a noise term ($n \times 1$). Then, ordinary least-squares estimation could be used to determine the kernel that minimizes the squared error between the model prediction and the measured responses:

$$\hat{\mathbf{k}} = \left(\mathbf{G}^{\mathrm{T}}\mathbf{G}\right)^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{r}$$

where the symbols are as defined earlier. Intuitively, $\hat{\mathbf{k}}$ can be viewed as the kernel that best achieves the measured grating responses $\mathbf{r}$ under the least-squares (LS) criterion. (In practice, $\hat{\mathbf{k}}$ achieves very good approximations of $\mathbf{r}$—see Supplementary Fig. 6.)

Now, observe that the images shown in the actual experiment can be used to construct a set of input channels $\mathbf{X}$ ($p \times q$) where p is the number of images. Under the assumption that the kernel is equal to $\hat{\mathbf{k}}$, the predicted responses to the images are given by $\mathbf{X}\hat{\mathbf{k}} + c\mathbf{1}$ where c is a DC offset (1×1) and $\mathbf{1}$ is a vector of ones ($p \times 1$). This expression can be rewritten as $\tilde{\mathbf{X}}\mathbf{r} + c\mathbf{1}$ where $\tilde{\mathbf{X}} = \mathbf{X}\left(\mathbf{G}^{\mathrm{T}}\mathbf{G}\right)^{-1}\mathbf{G}^{\mathrm{T}}$ is a set of transformed input channels ($p \times n$). Thus, implicit here is the following model of the responses to the images:

$$\mathbf{y} = \tilde{\mathbf{X}}\mathbf{r} + c\mathbf{1} + \mathbf{n}$$

where $\mathbf{y}$ is the set of image responses ($p \times 1$) and $\mathbf{n}$ is a noise term ($p \times 1$).

The above considerations demonstrate that under the LS criterion, responses to the images shown in the actual experiment can be modeled using the transformed input channels $\tilde{\mathbf{X}}$ and the set of weights $\mathbf{r}$. Thus, the transformation of $\mathbf{X}$ into $\tilde{\mathbf{X}}$ achieves the desired condition that weights on input channels directly reflect the response of the model to gratings. The final step is to sum over the input channels of $\tilde{\mathbf{X}}$ that represent different grating phases but the same grating orientation and spatial frequency. This is reasonable since phase information is discarded when quadrature pairs of wavelets are combined.

Supplementary Methods 10. Visual area localization

Construction of cortical surface representation

High-resolution anatomical data were acquired on a 1.5 T Philips Eclipse MR scanner (Philips Medical Systems, N.A., Bothell, WA). A T1-weighted MPRAGE pulse sequence was used: TR 15 ms, TE 4.47 ms, flip angle 35°, field-of-view 240 mm × 240 mm × 275.6 mm, matrix size 256 × 256 × 212, resolution 0.9375 mm × 0.9375 mm × 1.3 mm. Two anatomical volumes were acquired for each subject. The volumes were resampled to isotropic 1 mm × 1 mm × 1 mm voxels, manually co-registered using a rigid-body transformation, and averaged together to increase the contrast-to-noise ratio. The SureFit BETA v4.45 software package[59] was used to construct a triangulated mesh at the boundary between white and gray matter. The Caret v5.1 software package[59] was used to flatten this surface representation using a cut along the calcarine sulcus. (See http://brainmap.wustl.edu/caret/ for more information on SureFit and Caret.)

Registration of functional volumes

In the main experiment, an in-plane anatomical volume was acquired in the spatial reference frame to which all functional volumes for a given subject were registered. This in-plane anatomical volume was manually registered to the high-resolution anatomical volume (described above) using a rigid-body transformation. The parameters for this transformation were then used as an initial guess for the registration of the functional volumes to the high-resolution anatomical volume. This registration was subsequently improved by manually adjusting scaling and translation along the in-plane image dimensions. This resulted in an affine transformation that described the registration of the functional volumes to the cortical surface representation.

Localization of visual areas

In separate scan sessions fMRI data were collected using the multifocal retinotopic mapping technique[17,31] (see Supplementary Methods 11). These data were used to generate flattened maps of receptive-field angle and eccentricity. Visual areas V1, V2, and V3 were selected on these surface maps, and were represented as mutually disjoint sets of vertices. For assignment of voxels to visual areas, only voxels within 4 mm of surface vertices were considered. Each voxel was assigned to the visual area associated with the vertex closest to the voxel. Voxels outside of areas V1, V2, and V3 were discarded and not used in this study.

Supplementary Methods 11. Multifocal retinotopic mapping

The multifocal retinotopic mapping technique[17,31] was used to localize visual areas and to validate retinotopic information derived from the Gabor wavelet pyramid model. Estimates of retinotopic tuning provided by the multifocal technique have been shown to be similar to those provided by the more conventional phase-encoded technique[31,60,61].

Stimulus

The stimulus size was 20° × 20° (500 px × 500 px). A central white square served as the fixation point, and its size was 0.2° × 0.2° (4 px × 4 px). The stimulus was composed of 33 spatial components: a central circle and surrounding sectors defined by the intersections of 8 wedges and 4 rings. The boundaries of the wedges were positioned at angles of 0°, 45°, 90°, ..., and 315°, and the boundaries of the rings were positioned at eccentricities of 0.5°, 1.3°, 2.8°, 5.4°, and 10°. Each spatial component had one of two states. In the ON state, the spatial component was filled with a grayscale texture composed of non-Cartesian gratings[32]. The texture switched to different random configurations at a rate of 4 Hz. In the OFF state, the spatial component was filled with the gray background. The luminance of the gray background was set to the mean luminance of the texture.

The ON/OFF patterns for the spatial components were determined by an m-sequence[52] of level 5, order 4, and length $5^4 - 1 = 624$. Code for m-sequence generation was provided by T. Liu (http://fmriserver.ucsd.edu/ttliu/mttfmri_toolbox.html). One level of the m-sequence was associated with the ON state, and the other levels were associated with the OFF state. The m-sequence was repeatedly cyclically shifted by four elements to produce the ON/OFF pattern for each spatial component. Each element was assigned a duration of 4 s, and the total stimulus duration was 624 elements × 4 s = 41.6 min. For the purposes of data collection, the stimulus was divided into three consecutive segments (13.9 min each).

Data collection

Retinotopic mapping data were collected in one scan session from each subject. The same stimulus presentation setup and MRI parameters were used as in the main experiment. Each scan session consisted of three runs (13.9 min each), corresponding to the three segments of the stimulus.

Data analysis

Functional brain volumes were reconstructed and co-registered as in the main experiment. The time-series data for each voxel were then analyzed using the basis-restricted separable model (see Supplementary Methods 4). A set of basis functions was used to characterize the shape of the response timecourse, and a free parameter was used to characterize the amplitude of the response to each spatial component. Note that this model assumes linear spatial summation[17] in the sense that the response of a voxel to a combination of spatial components is assumed to equal the sum of the responses of the voxel to each individual spatial component.

For each voxel the estimated response amplitudes to each spatial component were used to calculate estimates of the angle and eccentricity of the voxel's receptive field. For angle, a vector summation procedure[32] was used:

$$A = \arg\left(\sum_i |a_i|^+ e^{j\theta_i} \right)$$

where A is the estimated receptive-field angle, i ranges over each spatial component except the central circle, a_i is the estimated response amplitude to spatial component i, $|\ |^+$ represents positive half-wave rectification, and θ_i is the mean angle of spatial component i. For eccentricity, a center-of-mass weighting procedure[32] was used:

$$E = \frac{\sum_i |a_i|^+ k_i}{\sum_i |a_i|^+}$$

where E is the estimated receptive-field eccentricity, i ranges over each spatial component, k_i is the mean eccentricity of spatial component i, and other symbols are as defined earlier.

Supplementary Note 1. Additional references

31. Vanni, S., Henriksson, L. & James, A. C. Multifocal fMRI mapping of visual cortical areas. *Neuroimage* **27**, 95–105 (2005).
32. Hansen, K. A., Kay, K. N. & Gallant, J. L. Topographic organization in and near human visual area V4. *J. Neurosci*. **27**, 11896–11911 (2007).
33. Wandell, B. A., Dumoulin, S. O. & Brewer, A. A. Visual field maps in human cortex. *Neuron* **56**, 366–383 (2007).
34. Kraft, A. *et al*. fMRI localizer technique: efficient acquisition and functional properties of single retinotopic positions in the human visual cortex. *Neuroimage* **28**, 453–463 (2005).
35. Larsson, J. & Heeger, D. J. Two retinotopic visual areas in human lateral occipital cortex. *J. Neurosci*. **26**, 13128–13142 (2006).
36. Tootell, R. B. *et al*. Functional analysis of V3A and related areas in human visual cortex. *J. Neurosci*. **17**, 7060–7078 (1997).
37. O'Craven, K. M. & Kanwisher, N. Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *J. Cogn. Neurosci*. **12**, 1013–1023 (2000).
38. O'Toole, A. J., Jiang, F., Abdi, H. & Haxby, J. V. Partially distributed representations of objects and faces in ventral temporal cortex. *J. Cogn. Neurosci*. **17**, 580–590 (2005).
39. Boynton, G. M., Engel, S. A., Glover, G. H. & Heeger, D. J. Linear systems analysis of functional magnetic resonance imaging in human V1. *J. Neurosci*. **16**, 4207–4221 (1996).
40. Heeger, D. J., Huk, A. C., Geisler, W. S. & Albrecht, D. G. Spikes versus BOLD: what does neuroimaging tell us about neuronal activity? *Nature Neurosci*. **3**, 631–633 (2000).
41. Rees, G., Friston, K. & Koch, C. A direct quantitative relationship between the functional properties of human and macaque V5. *Nature Neurosci*. **3**, 716–723 (2000).
42. Boynton, G. M. & Finney, E. M. Orientation-specific adaptation in human visual cortex. *J. Neurosci*. **23**, 8781–8787 (2003).
43. Engel, S. A. Adaptation of oriented and unoriented color-selective neurons in human visual areas. *Neuron* **45**, 613–623 (2005).
44. Fang, F., Murray, S. O., Kersten, D. & He, S. Orientation-tuned fMRI adaptation in human visual cortex. *J. Neurophysiol*. **94**, 4188–4195 (2005).
45. Larsson, J., Landy, M. S. & Heeger, D. J. Orientation-selective adaptation to first- and second-order patterns in human visual cortex. *J. Neurophysiol*. **95**, 862–881 (2006).
46. Murray, S. O., Olman, C. A. & Kersten, D. Spatially specific fMRI repetition effects in human visual cortex. *J. Neurophysiol*. **95**, 2439–2445 (2006).
47. Tootell, R. B. *et al*. Functional analysis of primary visual cortex (V1) in humans. *Proc. Natl Acad. Sci. USA* **95**, 811–817 (1998).
48. Furmanski, C. S. & Engel, S. A. An oblique effect in human primary visual cortex. *Nature Neurosci*. **3**, 535–536 (2000).
49. Goodyear, B. G., Nicolle, D. A., Humphrey, G. K. & Menon, R. S. BOLD fMRI response of early visual areas to perceived contrast in human amblyopia. *J. Neurophysiol*. **84**, 1907–1913 (2000).
50. Sereno, M. I. & Huang, R. S. A human parietal face area contains aligned head-centered visual and tactile maps. *Nature Neurosci*. **9**, 1337–1343 (2006).
51. Dale, A. M. Optimal experimental design for event-related fMRI. *Hum. Brain Mapp*. **8**, 109–114 (1999).
52. Buracas, G. T. & Boynton, G. M. Efficient design of event-related fMRI experiments

using m-sequences. *Neuroimage* **16**, 801–813 (2002).
53. Kay, K. N., David, S. V., Prenger, R. J., Hansen, K. A. & Gallant, J. L. Modeling low-frequency fluctuation and hemodynamic response timecourse in event-related fMRI. *Hum. Brain Mapp.* **29**, 142–156 (2008).
54. Kellman, P., van Gelderen, P., de Zwart, J. A. & Duyn, J. H. Method for functional MRI mapping of nonlinear response. *Neuroimage* **19**, 190–199 (2003).
55. Bullmore, E. *et al*. Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains. *Hum. Brain Mapp.* **12**, 61–78 (2001).
56. Skouras, K., Goutis, C. & Bramson, M. J. Estimation in linear models using gradient descent with early stopping. *Stat. Comput.* **4**, 271–278 (1994).
57. Qian, N. On the momentum term in gradient descent learning algorithms. *Neural Networks* **12**, 145–151 (1999).
58. Cao, R., Cuevas, A. & Manteiga, W. G. A comparative study of several smoothing methods in density estimation. *Comput. Stat. Data An.* **17**, 153–176 (1994).
59. Van Essen, D. C. *et al*. An integrated software suite for surface-based analyses of cerebral cortex. *J. Am. Med. Inform. Assn.* **8**, 443–459 (2001).
60. Fukunaga, M., van Gelderen, P., de Zwart, J. A., Jansma, J. M. & Duyn, J. H. Retinotopic fMRI mapping with pseudo-random stimulus presentation using the m-sequence paradigm. *Soc. Neurosci. Abstr.* 693.2 (2004).
61. Kay, K. N., Hansen, K. A., David, S. V. & Gallant, J. L. Artifacts in phase-encoded fMRI retinotopic mapping. *Soc. Neurosci. Abstr.* 508.12 (2005).