Neuron
Article

Bayesian Reconstruction of Natural Images from Human Brain Activity

Thomas Naselaris,[1] Ryan J. Prenger,[2] Kendrick N. Kay,[3] Michael Oliver,[4] and Jack L. Gallant[1,3,4,*]
[1]Helen Wills Neuroscience Institute
[2]Department of Physics
[3]Department of Psychology
[4]Vision Science Program
University of California, Berkeley, Berkeley, CA 94720, USA
*Correspondence: gallant@berkeley.edu
DOI 10.1016/j.neuron.2009.09.006

SUMMARY

Recent studies have used fMRI signals from early visual areas to reconstruct simple geometric patterns. Here, we demonstrate a new Bayesian decoder that uses fMRI signals from early and anterior visual areas to reconstruct complex natural images. Our decoder combines three elements: a structural encoding model that characterizes responses in early visual areas, a semantic encoding model that characterizes responses in anterior visual areas, and prior information about the structure and semantic content of natural images. By combining all these elements, the decoder produces reconstructions that accurately reflect both the spatial structure and semantic category of the objects contained in the observed natural image. Our results show that prior information has a substantial effect on the quality of natural image reconstructions. We also demonstrate that much of the variance in the responses of anterior visual areas to complex natural images is explained by the semantic category of the image alone.

INTRODUCTION

Functional magnetic resonance imaging (fMRI) provides a measurement of activity in the many separate brain areas that are activated by a single stimulus. This property of fMRI makes it an excellent tool for *brain reading*, in which the responses of multiple voxels are used to decode the stimulus that evoked them (Haxby et al., 2001; Carlson et al., 2002; Cox and Savoy, 2003; Haynes and Rees, 2005; Kamitani and Tong, 2005; Thirion et al., 2006; Kay et al., 2008; Miyawaki et al., 2008). The most common approach to decoding is image *classification*. In classification, a pattern of activity across multiple voxels is used to determine the discrete class from which the stimulus was drawn (Haxby et al., 2001; Carlson et al., 2002; Cox and Savoy, 2003; Haynes and Rees, 2005; Kamitani and Tong, 2005).

Two recent studies have moved beyond classification and demonstrated stimulus *reconstruction* (Thirion et al., 2006; Miyawaki et al., 2008). The goal of reconstruction is to produce a literal picture of the image that was presented. The Thirion et al. (2006) and Miyawaki et al. (2008) studies achieved reconstruction by analyzing the responses of voxels in early visual areas. To simplify the problem, both studies used geometric stimuli composed of flickering checkerboard patterns. However, a general brain-reading device should be able to reconstruct natural images (Kay and Gallant, 2009). Natural images are important targets for reconstruction because they are most relevant for daily perception and subjective processes such as imagery and dreaming. Natural images are also very challenging targets for reconstruction, because they have complex statistical structure (Field, 1987; Karklin and Lewicki, 2009; Cadieu and Olshausen, 2009) and rich semantic content (i.e., they depict meaningful objects and scenes). A method for reconstructing natural images should be able to reveal both the structure and semantic content of the images simultaneously.

In this paper, we present a Bayesian framework for brain reading that produces accurate reconstructions of the spatial structure of natural images, while simultaneously revealing their semantic content. Under the Bayesian framework used here, a reconstruction is defined as the image that has the highest posterior probability of having evoked the measured response. Two sources of information are used to calculate this probability: information about the target image that is encoded in the measured response and pre-existing, or *prior*, information about the structure and semantic content of natural images.

Information about the target image is extracted from measured responses by applying one or more *encoding models* (Nevado et al., 2004; Wu et al., 2006). An encoding model is represented mathematically by a conditional distribution, $p(\mathbf{r}|\mathbf{s})$, which gives the likelihood that the measured response **r** was evoked by the image **s** (here bold **r** denotes the collected responses of multiple voxels; italicized *r* will be used to denote the response of a single voxel). Note that functionally distinct visual areas are best characterized by different encoding models, so a reconstruction based on responses from multiple visual areas will use a distinct encoding model for each area.

Prior information about natural images is also represented as a distribution, $p(\mathbf{s})$, that assigns high probabilities to images that are most natural (Figure 1, inner bands of image samples) and low probabilities to more artificial, random, or noisy images (Figure 1, outermost band of image samples).

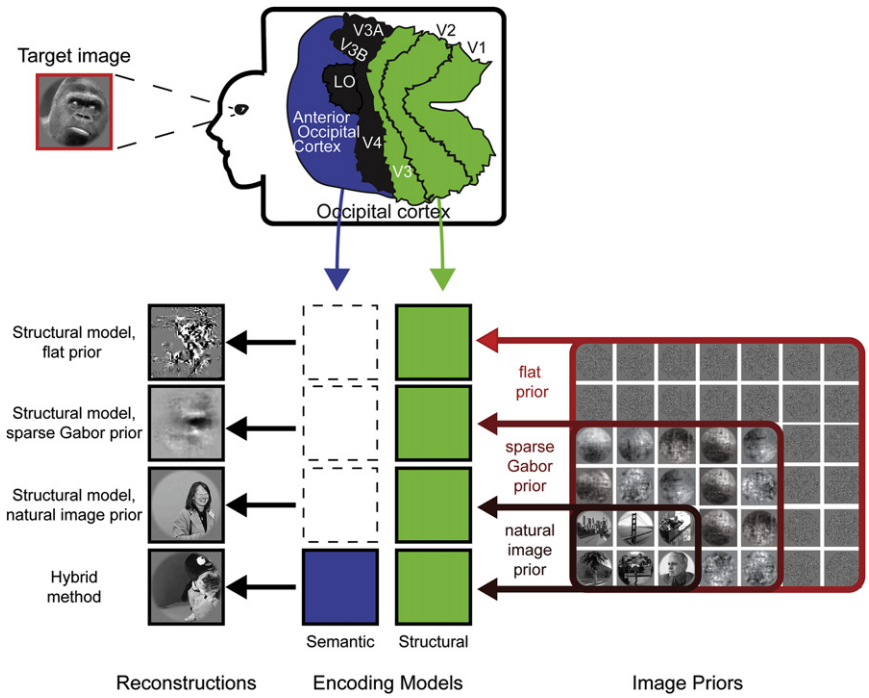


Figure 1. The Bayesian Reconstruction Framework

The goal of this experiment was to reconstruct target images from BOLD fMRI responses recorded from occipital cortex. Reconstructions were obtained by using a Bayesian framework to combine voxel responses, structural and semantic encoding models, and image priors. Target images were grayscale photographs selected at random from a large database of natural images. The fMRI slice coverage included early visual areas V1, V2, and V3; intermediate visual areas V3A, V3B, V4, and lateral occipital (labeled LO here); and a band of occipital cortex anterior to lateral occipital (here called AOC). Recorded voxel responses were used to fit two distinct encoding models: a *structural encoding model* (green) that reflects how information is encoded in early visual areas and a *semantic encoding model* (blue) that reflects how information is encoded in the AOC. Three image priors were used to bias reconstructions in favor of those with the characteristics of natural images: a flat prior that does not bias reconstructions, a sparse Gabor prior that ensures that reconstructions possess the lower-order statistical properties of natural images, and a natural image prior that ensures that reconstructions are natural images. Several different types of reconstructions were obtained by combining the encoding models and priors in different ways: the structural model and a flat prior; the structural model and a sparse Gabor prior; the structural model and a natural image prior; and the structural model, the semantic model, and a natural image prior (hybrid method). These various methods produced reconstructions with very different structural and semantic qualities, as shown in Figures 2 and 3.

The critical step in reconstruction is to calculate the probability that each possible image evoked the measured response. This is accomplished by using Bayes theorem to combine the encoding models and the image prior:

$$p(\mathbf{s}|\mathbf{r}) \propto p(\mathbf{s}) \prod_i p_i(\mathbf{r}_i|\mathbf{s}) \quad (1)$$

The *posterior distribution*, $p(\mathbf{s}|\mathbf{r})$, gives the probability that image **s** evoked response **r**. The encoding models and voxel responses from functionally distinct areas are indexed by *i*. To produce a reconstruction, $p(\mathbf{s}|\mathbf{r})$ is evaluated for a large number of images. The image with the highest $p(\mathbf{s}|\mathbf{r})$ (or *posterior probability*) is selected as the reconstruction, commonly known as the *maximum a posteriori* estimate (Zhang et al., 1998).

In a previous study, we used the structural encoding model without invoking the Bayesian framework in order to solve *image identification* (Kay et al., 2008). The goal of image identification is to determine which specific image was seen on a certain trial, when that image was drawn from a known set of images. Image identification provides an important foundation for image reconstruction, but it is a much simpler problem because the set of target images is known beforehand. Furthermore, success at image identification does not guarantee success at reconstruction, because a target image may be identified on the basis of a small number of image features that are not sufficient to produce an accurate reconstruction.

In this paper, we investigate two key factors that determine the quality of reconstructions of natural images from fMRI data: encoding models and image priors. We find that fMRI data and a structural encoding model are insufficient to support high-quality reconstructions of natural images. Combining these with an appropriate natural image prior produces reconstructions that, while structurally accurate, fail to reveal the semantic content of the target images. However, by applying an additional semantic encoding model that extracts the information present in anterior visual areas, we produce reconstructions that accurately reflect semantic content of the target images as well. A comparison of the two encoding models shows that they most accurately predict the responses of functionally distinct and anatomically separated voxels. The structural model best predicts responses of voxels in early visual areas (V1, V2, and so on), while the semantic model best predicts responses of voxels anterior to V4, V3A, V3B, and the posterior portion of lateral occipital. Furthermore, the accuracy of predictions of these models is comparable to the accuracy of predictions obtained for single neurons in area V1.

RESULTS

Blood-oxygen-level-dependent (BOLD) fMRI measurements of occipital visual areas were made while three subjects viewed a series of monochromatic natural images (Kay et al., 2008). Functional data were collected from early (V1, V2, V3) and intermediate (V3A, V3B, V4, lateral occipital) visual areas and from a band of occipital cortex directly anterior to lateral occipital that we refer to here as anterior occipital cortex (AOC). The

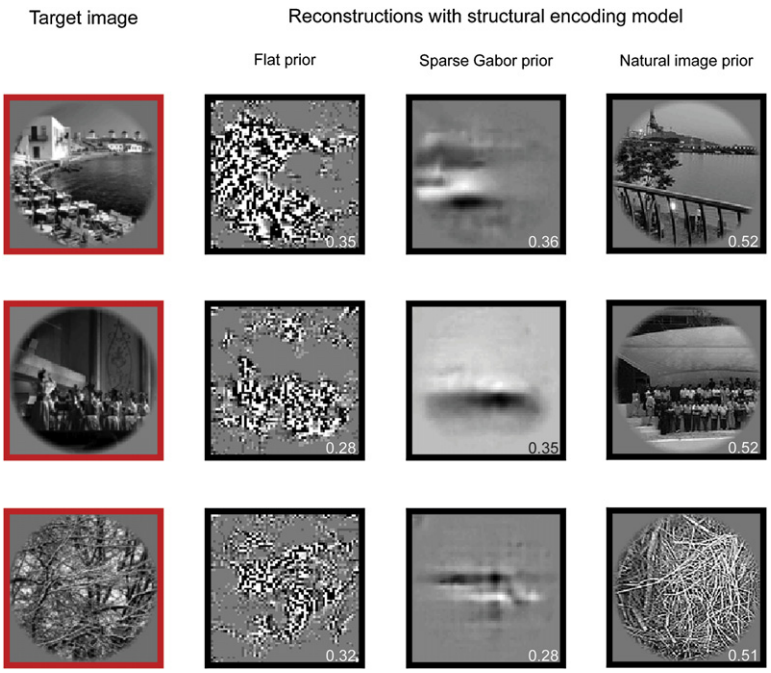


Figure 2. The Effect of Prior Information on Reconstruction with a Structural Encoding Model

Three target images are shown in the first column (red borders). The second through fourth columns show reconstructions obtained using the structural encoding model and three different types of prior information. Column two shows reconstructions obtained using a flat prior that does not bias reconstructions. Regions of the target images that have low texture contrast are depicted as smooth gray patches, and regions that have substantial texture contrast are depicted as textured patches. Thus, the flat prior reconstructions reveal the distribution of texture contrast in the target images but cannot readily be interpreted. Column three shows reconstructions obtained using a sparse Gabor prior that ensures that reconstructions possess the lower-order statistical properties of natural images. These reconstructions appear to be smoothed versions of those obtained with the flat prior, and they also cannot be readily interpreted. Column four shows reconstructions obtained using a natural image prior that ensures that reconstructions are natural images. These reconstructions accurately reflect the structure of the target images (numbers in bottom right corner of each reconstruction indicate structural accuracy, see main text for details). The example in row one is from subject TN; rows two and three are from subject SN.

experiment consisted of two stages: model estimation and image reconstruction. During model estimation, subjects viewed 1750 achromatic natural images while functional data were collected. These data were used to fit encoding models for each voxel. During image reconstruction, functional data were collected while subjects viewed 120 novel target images. These data were used to generate reconstructions.

Reconstructions that Use a Structural Encoding Model for Early Visual Areas and an Appropriate Image Prior

Our Bayesian framework requires that each voxel be fit with an appropriate encoding model. In our previous study, we showed that a *structural encoding model* based upon Gabor wavelets could be used to extract a large amount of information from individual voxels in early visual areas (Kay et al., 2008). Therefore, we began by using this model to produce reconstructions.

Under the structural encoding model, the likelihood of a voxel's response r to an image $\mathbf{s}$ is determined by its tuning along the dimensions of space, orientation, and spatial frequency (Kay et al., 2008). The model includes a set of weights that can be adjusted to fit the specific tuning of single voxels. These weights were fit for all of the voxels in our data set using a coordinate-descent optimization procedure (see Experimental Procedures). This procedure produced a separate encoding model, $p(r|\mathbf{s})$, for each voxel. Those voxels whose responses could be predicted accurately by the model were then selected (see Experimental Procedures for specific voxel selection criteria) for use in reconstruction. The individual models for each of the selected voxels were then combined into a single multivoxel structural encoding model, $p(\mathbf{r}|\mathbf{s})$ (see Experimental Procedures for details on how individual models are combined into a multivoxel model). The majority of selected voxels were located in early visual areas (V1, V2, and V3).

The Bayesian framework also requires an appropriate prior. The reconstructions reported in Thirion et al. (2006) and Miyawaki et al. (2008) used no explicit source of prior information. To obtain comparable results to theirs, we began with a *flat prior* that assigns the same probability to all possible images. This prior makes no strong assumptions about the stimulus but instead assumes that noise patterns are just as likely as natural images (see Figure 1). Thus, when the flat prior is used, only the information encoded in the responses of the voxels is available to support reconstruction. [Formally, using the flat prior amounts to setting the prior, $p(\mathbf{s})$, in Equation 1 to a constant.]

To produce reconstructions, the structural encoding model, the flat prior, and the selected voxels were used to evaluate the posterior probability (see Equation 1) that an image $\mathbf{s}$ evoked the responses of the selected voxels. A greedy serial search algorithm was used to converge on an image with a high (relative to an initial image with all pixels set to zero) posterior probability. This image was selected as the reconstruction. Typical reconstructions are shown in the second column of Figure 2. In the example shown in row one, the target image (first column, red border) is a seaside cafe and harbor. The reconstruction (second column) depicts the shore as a textured high-contrast region and the sea and sky as smooth low-contrast regions. In row two, the target image is a group of performers on a stage, but the reconstruction depicts the performers as a single textured region on a smooth background. In row three, the target image is a patch of dense foliage, which the reconstruction depicts as a single textured region that covers much of the visual field.

All of the example reconstructions obtained using the structural model and the flat prior have similar qualities. Regions of the target images that have low contrast or little texture are depicted as smooth gray patches in the reconstructions, and regions that have significant local contrast or texture are

depicted as textured patches. Local texture and contrast are apparently the only information about natural images that can be recovered reliably from moderate-resolution BOLD fMRI measurements of activity in early visual areas. Unfortunately, reconstructions based entirely on texture and contrast do not provide enough information to reveal the identity of objects depicted in the target images.

To improve reconstructions, we sought to define a more informative image prior. A distinguishing feature of natural images is that they are composed of many smooth regions, disrupted by sharp edges. These characteristics are captured by two lower-level statistical properties of natural images: they tend to have a 1/*f* amplitude spectrum (Field, 1987), and they are sparse in the Gabor-wavelet domain (Field, 1994). In contrast, unnatural images such as white noise patterns generally have much different power spectra and are not sparse. We therefore designed a *sparse Gabor prior* that biases reconstructions in favor of images that exhibit these two well-known statistical properties (see Figure 1).

To produce a new set of reconstructions, the structural encoding model, the sparse Gabor prior, and the same set of voxels selected above were used to evaluate posterior probabilities (see Equation 1). The same greedy serial search algorithm mentioned above was used to converge on an image with a relatively high posterior probability. This image was selected as the reconstruction. Results are shown in the third column of Figure 2. The main effect of the sparse Gabor prior is to smooth out the textured patches apparent in the reconstruction with a flat prior. As a result, the reconstructions are more consistent with the lower-level statistical properties of natural images. However, these reconstructions do not depict any clearly identifiable objects or scenes and thus fail to reveal the semantic content of the target images.

Because the sparse Gabor prior did not produce reconstructions that reveal the semantic content of the target images, we sought to introduce a more sophisticated image prior. Natural images have complex statistical properties that reflect the distribution of shapes, textures, objects, and their projections onto the retina, but thus far theorists have not captured these properties in a simple mathematical formalism. We therefore employed a strategy first developed in the computer vision community to approximate these complex statistical properties (Hays and Efros, 2007; Torralba et al., 2008). We constructed an *implicit natural image prior* by compiling a database of six million natural images selected at random from the internet (see Experimental Procedures). The implicit natural image prior can be viewed as a distribution that assigns the same probability to all images in the database and zero probability to all other images.

To produce reconstructions using the natural image prior, the posterior probability was evaluated for each of the six million images in the database (note that in this case the posterior probability is proportional to the likelihood given by the encoding model); the image with the highest probability was selected as the reconstruction. Examples are shown in the fourth column of Figure 2. In row one, both the target image and the reconstruction depict a shoreline (compare row one, column one to row one, column four). In row two, both the target image and the reconstruction depict a group of performers on a stage. In row three, both the target image and the reconstruction depict a patch of foliage. In all three examples, the spatial structure and the semantic content of the reconstructions accurately reflect both the spatial structure and semantic content of the target images (also see Figure S3A). Thus, these particular reconstructions are both structurally and semantically accurate.

The examples shown in Figure 2 were selected to demonstrate the best reconstruction performance obtained with the structural encoding model and the natural image prior. However, most of the reconstructions obtained this way are not semantically accurate. Several examples of semantically inaccurate reconstructions are shown in the second column of Figure 3. In row one, the target image is a group of buildings, but the reconstruction depicts a dog. In row two, the target image is a bunch of grapes, but the reconstruction depicts a hand against a checkerboard background. In row three, the target image is a crowd of people in a corridor, but the reconstruction depicts a building. In row four, the target image is a snake, but the reconstruction depicts several buildings.

Close inspection of the reconstructions in the second column of Figure 3 suggests that they are structurally accurate. For example, the target image depicting grapes in row two has high spatial frequency, while the reconstruction in row two contains a checkerboard pattern with high spatial frequency as well. However, the reconstruction does not depict objects that are semantically similar to grapes, so it does not appear similar to the target image. This example reveals that structural similarity alone can be a poor indicator of how similar two images will appear to a human observer. Because human judgments of similarity will inevitably take semantic content into account, reconstructions should reflect both the structural and semantic aspects of the target image. Therefore, we sought to incorporate activity from brain areas known to encode information about the semantic content of images.

A Semantic Encoding Model

There is evidence that brain areas in anterior visual cortex encode information that is related to the semantic content of images (Kanwisher et al., 1997; Epstein and Kanwisher, 1998; Grill-Spector et al., 1998; Haxby et al., 2001; Grill-Spector and Malach, 2004; Downing et al., 2006; Kriegeskorte et al., 2008). In order to add accurate semantic content to our reconstructions, we designed a *semantic encoding model* that describes how voxels in these areas encode information about natural scenes. The model automatically learns—from responses evoked by a randomly chosen set of natural images—the semantic categories that are represented by the responses of a single voxel.

To fit the semantic encoding model, all 1750 natural images used to acquire the model estimation data set were first labeled by human observers with one of 23 semantic category names (see Figure S1). These categories were chosen to be mutually exclusive yet broadly defined, so that the human observers were able to assign each natural image a single category that best described it (observers were instructed to label each image with the single category they deemed most appropriate

Figure 3. The Effect of Semantic Information on Reconstructions

Four target images are shown in the first column (red borders). The second column shows reconstructions obtained using the structural encoding model and the natural image prior. These reconstructions are structurally accurate (numbers in bottom right corner indicate structural accuracy, see main text for details). However, the objects depicted in the reconstructions are not from the same semantic categories as those shown in the target images. Thus, although these reconstructions are structurally accurate they are not semantically accurate. The third column shows reconstructions obtained using the structural encoding model, the semantic encoding model, and the natural image prior (the *hybrid method*). These reconstructions are both structurally and semantically accurate. The examples in rows from one through three are from subject TN; row four is from subject SN.

and reasonable; see Experimental Procedures for details). Importantly, the images had not been chosen beforehand to fall into predefined categories; rather, the categories were designed post hoc to provide reasonable categorical descriptions of randomly selected natural images.

After the images in the model estimation set were labeled, an expectation maximization optimization algorithm (EM) was used to fit the semantic model to each voxel (see Experimental Procedures and Appendix 1 in the Supplemental Data for details). The EM algorithm learned the probability that each of the 23 categories would evoke a response either above, below, or near the average of each voxel. The resulting semantic model reflects the probability that a voxel "likes," "doesn't like," or "doesn't care about" each semantic category. This information is then used to calculate $p(r|\mathbf{s})$—the likelihood of the observed response, given a sampled image (see Figure S2 and Experimental Procedures for more details). We fit the semantic model to all of the voxels in the data set and then inspected those voxels whose responses could be predicted accurately by the model (see Experimental Procedures for specific voxel selection criteria).

Examples of the semantic encoding model fit to three voxels (one from each of the three subjects in this study) are shown in Figure 4. Gray curves show the overall distribution of responses to all images in the model estimation set. The colored curves define responses that are above (blue curve), below (red curve), or near (green curve) the average response. The bottom boxes give the probability that an image from a specific semantic category (category names at left; names are abbreviated, see Figure S1 for full names) will evoke a response above (blue boxes), below (red boxes), or near (green boxes) the average response. For each of these voxels, most categories that pertain to nonliving things—such as textures, landscapes, and buildings—are likely to evoke responses below the average. In contrast, most categories that pertain to living things—such as people, faces, and animals—are likely to evoke responses above the average. Average responses tend to be evoked by a fairly uniform distribution of categories. Thus, at a coarse level, activity in each of these voxels tends to distinguish between animate and inanimate things.

To determine how the representations of structural and semantic information are related to one another, we compared the prediction accuracy of the structural model with that of the semantic model (Figure 5, left panels). We quantified prediction accuracy as the correlation (*cc*) between the response observed in each voxel and the response predicted by each encoding model for all 120 images in the image reconstruction set. The points show the prediction accuracy of the structural encoding model (x axis) and semantic encoding model (y axis) for each voxel in our slice coverage. The distribution of points has two wings. One wing extends along the y axis, and the other extends along the x axis. This indicates that there are very few voxels whose responses are accurately predicted by both models. Most voxels whose responses are accurately predicted by the structural model (cc > 0.353; blue voxels; see Experimental Procedures for criteria used to set this threshold) are not accurately predicted by the semantic model. Most voxels whose responses are accurately predicted by the semantic model (cc > 0.353; magenta voxels) are not accurately predicted by the structural model. The wings have similar extents, indicating that the semantic model provides predictions that are as accurate as those provided by the structural model. Remarkably, the predictions for both the structural and semantic voxels can be as accurate as those obtained for single neurons in area V1 (David and Gallant, 2005; Carandini et al., 2005). Note that there is a large central mass of voxels (gray); these voxels either have poor signal quality or

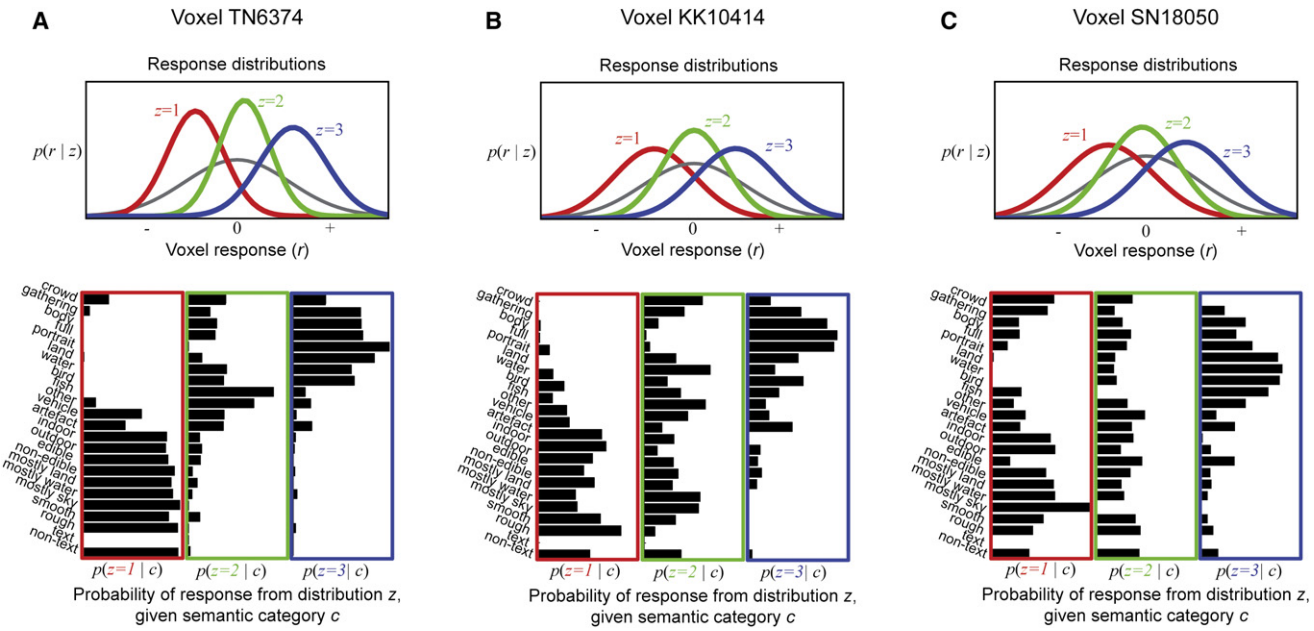


Figure 4. The Semantic Encoding Model Fit to Single Voxels from Three Subjects

(A) The top panel shows response distributions of one voxel for which the semantic encoding model produced the most accurate predictions (subject TN). The gray curve gives the distribution of z-scored responses (x axis) evoked by all images used in the model estimation data set. This distribution was modeled in terms of three underlying Gaussian distributions (colored curves labeled by the indicator variable *z*). Responses below average are shown in red ($z = 1$), responses near average in green ($z = 2$), and above average in blue ($z = 3$). The black bars in the bottom panels give the probability that each semantic category, *c*, (abbreviated labels at left) will evoke a response below the average (red box), near the average (green box), or above the average (blue box). (Note that there are no probabilities for the text category because there were no text images in the model estimation data set.) Images depicting living things tend to evoke a large response from this voxel, while those depicting nonliving things evoke a small response. Thus, this voxel discriminates between animate and inanimate semantic categories.

(B) The same analysis shown in (A) applied to the single voxel from subject KK for which the semantic encoding model produced the most accurate predictions. Semantic tuning for this voxel is similar to the one shown in (A).

(C) The same analysis shown in (A) and (B) applied to the single voxel from subject SN for which the semantic encoding model produced the most accurate predictions. Semantic tuning for this voxel is similar to those shown in (A) and (B).

represent information not captured by either the structural or semantic models.

In order to determine the anatomical locations of the voxels in the two separate wings, we projected voxels whose responses are accurately predicted by the structural (blue) and semantic (magenta) models onto flat maps of the right and left occipital cortex (Figure 5, right panels). Most of the voxels whose responses are accurately predicted by the structural model are located in early visual areas V1, V2, and V3. In contrast, most of the voxels whose responses are accurately predicted by the semantic model are located in the AOC, at the anterior edge of our slice coverage.

Our results show that the semantic encoding model accurately characterizes a set of voxels in anterior visual cortex that are functionally distinct and anatomically separated from the structural voxels located in early visual cortex. The structural voxels in early visual areas encode information about local contrast and texture, while the semantic voxels in anterior portions of lateral occipital and in the AOC encode information related to the semantic content of natural images. Therefore, a reconstruction method that uses the structural and semantic encoding models to extract information from both sets of voxels should produce reconstructions that reveal both the structure and semantic content of the target images.

Reconstructions Using Structural and Semantic Models and a Natural Image Prior

To incorporate the semantic encoding model into the reconstruction algorithm, we first selected all of the voxels for which the semantic encoding model provided accurate predictions. Most of these voxels were located in the anterior portion of lateral occipital and in the AOC (see Experimental Procedures for details on voxel selection). The individual models for each selected voxel were then combined to into a single, multivoxel semantic encoding model, $p(\mathbf{r}|\mathbf{s})$ (see Experimental Procedures for details).

To produce reconstructions, the semantic and structural encoding models (with their corresponding selected voxels) were used to evaluate the posterior probability (see Equation 1) of each of the six million images in the natural image prior. For convenience, we refer to the use of the structural model, semantic model and natural image prior as the *hybrid method*.

Reconstructions obtained using the hybrid method are shown in the third column of Figure 3. In contrast to the reconstructions produced using the structural encoding model and natural image prior, the hybrid method produces reconstructions that are both structurally and semantically accurate. In the example shown in row one, both the target image and the reconstruction depict buildings. In row two, the target image is a bunch of grapes, and the reconstruction depicts a bunch of berries. In row three,

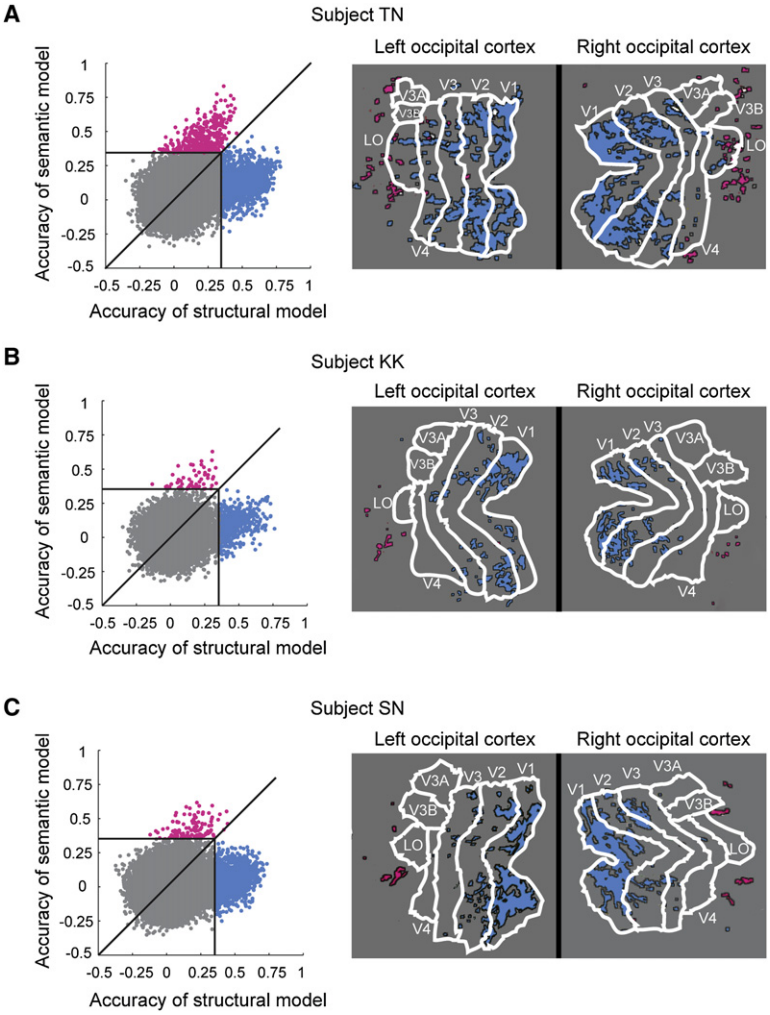


Figure 5. Structural versus Semantic Encoding Models

(A) The left panel compares the accuracy of the structural encoding model (x axis) versus the semantic encoding model (y axis) for every voxel within the slice coverage (subject TN). Here accuracy is defined as the correlation (*cc*) between the response observed in each voxel and the response predicted by each encoding model for all 120 images in the image reconstruction set. The distribution of points has two wings. One wing extends along the y axis, and another extends along the x axis, indicating that very few voxels are accurately predicted by both models. The voxels whose responses are accurately predicted by the structural model but not the semantic model are shown in blue ($cc > 0.353$, $p < 3.9*10^{-5}$; see Experimental Procedures for criteria used to set this threshold). The voxels whose responses are accurately predicted by the semantic model but not the structural model are shown in magenta (same statistical threshold as above). Most voxels are poorly predicted by both models (gray), either because neither model is appropriate or because of poor signal quality. The right panel shows flat maps of the left and right hemispheres of this subject. Visual areas identified using a retinotopic mapping procedure (see Experimental Procedures) are outlined in white. Voxels whose responses are accurately predicted by the structural (blue) or semantic (magenta) models are plotted on the flat maps (the few voxels for which both models are accurate are shown in white). Most structural voxels are located in early visual areas V1, V2, and V3. Most semantic voxels are located in the anterior portion of lateral occipital (labeled LO) and in the anterior occipital cortex.

(B) Data for subject KK, format same as in (A). Most structural voxels are located in early visual areas V1, V2, and V3. Semantic voxels are located in the anterior occipital cortex.

(C) Data for subject SN, format same as in (A) and (B). Structural voxels are located in early visual areas V1, V2, and V3. Semantic voxels are located in the anterior occipital cortex.

the target image depicts a crowd of people in a corridor, and the reconstruction depicts a crowd of people on a narrow street. In row four, the target image depicts a snake crossing the visual field at an angle, while the reconstruction depicts a caterpillar crossing the visual field at a similar angle. (In Figure S3, we also present the second and third most probable images in the natural image prior. The spatial structure and semantic content of these alternative reconstructions is consistent with the best reconstruction.)

Objective Assessment of Reconstruction Accuracy

To quantify the spatial similarity of the reconstructions and the target images, we used a standard image similarity metric proposed previously (Brooks and Pappas, 2006). This metric reflects the complex wavelet-domain correlation between the reconstruction and the target image. We applied this metric to the four types of reconstruction presented in Figures 2 and 3. As shown by the plots on the left side of Figure 6, the structural accuracy of all the reconstruction methods that use a non-flat prior is significantly greater than chance for all three subjects ($p < 0.01$, t test; comparison is for each individual subject). Reconstruction with the structural model and the natural image prior is significantly more accurate than reconstruction with a sparse Gabor prior, ($p < 0.01$, t test; comparison is for each individual subject). These results indicate that prior information is important for obtaining accurate image reconstructions. The structural accuracy of the structural model with natural image prior and the hybrid method are not significantly different ($p > 0.3$, t test; comparison is for each individual subject), so structural accuracy is not affected by the addition of the semantic model.

To quantify the semantic similarity of the reconstructions and the target images, we formulated a semantic accuracy metric. In this case, we estimated the probability that a reconstruction obtained using some specific reconstruction method would belong to the same semantic category as the target image. (Because calculating semantic accuracy requires time-consuming labeling of many images, we calculated semantic accuracy for only the first 30 images in the image reconstruction set; see Experimental Procedures for details.) We considered semantic categories at four different levels of specificity, from two broadly defined categories ("mostly animate" versus

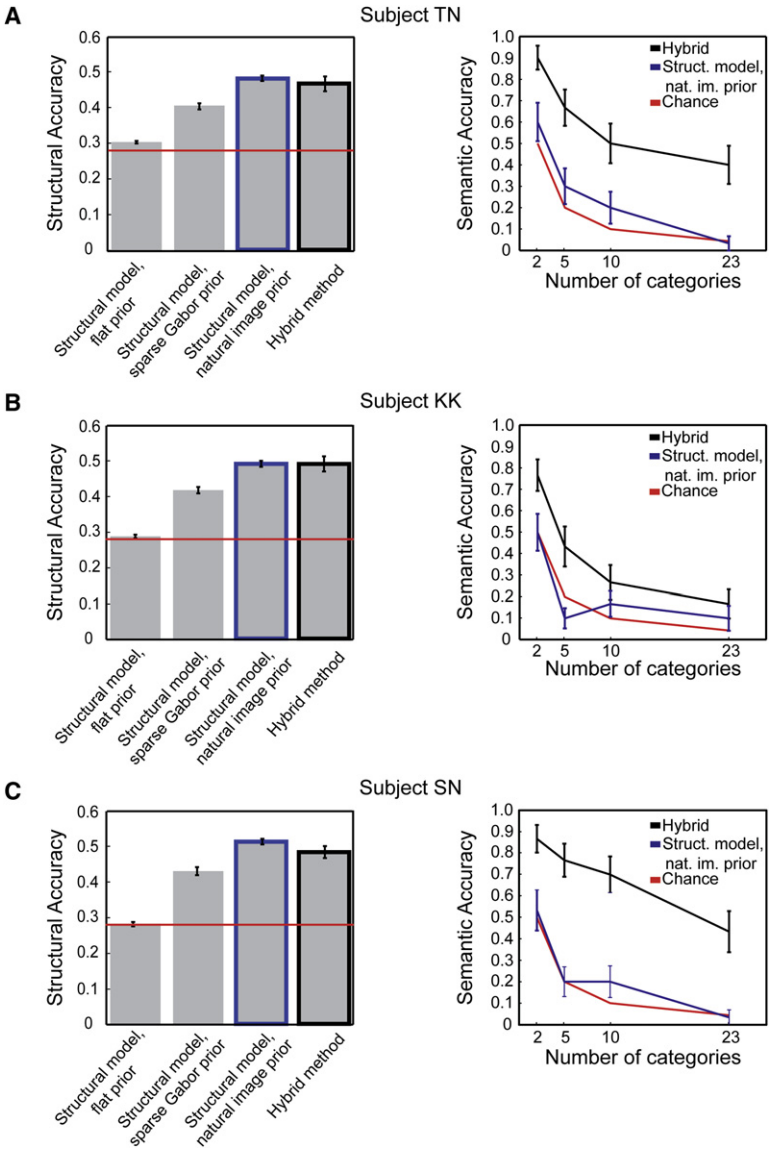


Figure 6. Structural and Semantic Accuracy of Reconstructions

(A) The left panel shows the structural accuracy of reconstructions using several different methods (subject TN). In each case, structural reconstruction accuracy (y axis) is quantified using a similarity metric that ranges from 0.0 to 1.0. From left to right, the bars give the structural similarity between the target image and reconstruction (mean ± SEM, image reconstruction data set) for the structural model with a flat prior; the structural model with a sparse Gabor prior; the structural model with a natural image prior; and the *hybrid method* consisting of the structural model, the semantic model, and the natural image prior. The red line indicates chance performance. Reconstructions produced using the sparse Gabor or natural image prior are significantly more accurate than chance ($p < 0.01$, t test; for this subject only, the reconstructions produced using a flat prior are also significant at this level). Reconstruction with the structural model and the natural image prior is significantly more accurate than reconstruction with a sparse Gabor prior ($p < 0.01$, t test). These results indicate that prior information is important for obtaining structurally accurate image reconstructions. The structural accuracy of the structural model with natural image prior and the hybrid method are not significantly different ($p > 0.3$, t test), so structural accuracy is not affected by the addition of the semantic model. The right panel shows semantic accuracy of reconstructions obtained using the structural model with natural image prior (blue) and the hybrid method (black). In each case, semantic reconstruction accuracy (y axis) is quantified in terms of the probability that a reconstruction will belong to the same semantic category as the target image (error bars indicate bootstrapped estimate of SD). The number of semantic categories varies from two broadly defined categories to the 23 specific categories shown in Figure 4 (x axis). The red curve indicates chance performance. The semantic accuracy of the reconstructions obtained using the structural model and natural image prior are rarely significantly greater than chance ($p > 0.3$, binomial test). However, the semantic accuracy of the hybrid method is significantly greater than chance regardless of the number of semantic categories ($p < 10^{-5}$, binomial test).

(B) Data for subject KK, format same as in (A). Prior information is important for obtaining structurally accurate image reconstructions (p values of structural accuracy comparisons same as in A). The semantic accuracy of the hybrid method is significantly greater than chance ($p < .002$, binomial test).

(C) Data for subject SN, format same as in (A). Prior information is important for obtaining structurally accurate image reconstructions (p values of structural accuracy comparisons same as in A). The semantic accuracy of the hybrid method is significantly greater than chance ($p < 10^{-5}$, binomial test).

"mostly inanimate") to 23 narrowly defined categories (see Figure S1 for complete list). Semantic accuracies for the structural model with natural image prior and the hybrid method are shown by the plots on the right side of Figure 6 (note that semantic accuracy cannot be determined for methods that did not use the natural image prior). The semantic accuracy of the hybrid method is significantly greater than chance for all three subjects, and at all levels of specificity ($p < 10^{-5}$, binomial test, for subjects TN and SN; $p < 0.002$, binomial test, for subject KK). The semantic accuracy of the reconstructions obtained using the structural model and natural image prior are rarely significantly greater than chance for all three subjects ($p > 0.3$, binomial test). The hybrid method is quite semantically accurate. When two categories are considered, accuracy is 90% (for subject TN), and when the full 23 categories are considered, accuracy is still 40%. In other words, reconstructions produced using the hybrid method will correctly depict a scene whose animacy is consistent with the target image 90% of the time and will correctly depict the specific semantic category of the target image 40% of the time.

DISCUSSION

We have presented reconstructions of natural images from BOLD fMRI measurements of human brain activity. These reconstructions were produced by a Bayesian reconstruction framework that uses two different encoding models to integrate information from functionally distinct visual areas: a structural model

that describes how information is represented in early visual areas and a semantic encoding model that describes how information is represented in anterior visual areas. The framework also incorporates image priors that reflect the structural and semantic statistics of natural images. The resulting reconstructions accurately reflect the spatial structure and semantic content of the target images.

Relationship to Previous Reconstruction Studies

Two previous fMRI decoding papers presented algorithms for reconstructing the spatial layout of simple geometrical patterns composed of high-contrast flicker patches (Thirion et al., 2006; Miyawaki et al., 2008). Both these studies used some form of structural model that reflected the retinotopic organization of early visual cortex, but neither explored the role of semantic content or prior information. Our previous study on image identification from brain activity (Kay et al., 2008) used a more sophisticated voxel-based structural encoding model that reflects the way that spatial frequency and orientation information are encoded in brain activity measured in early visual areas. However, the image identification task does not require the use of semantic information.

The study reported here presents a solution to a more general problem: reconstructing arbitrary natural images from fMRI signals. It is much more difficult to reconstruct natural images than flickering geometrical patterns because natural images have a complex statistical structure and evoke signals with relatively low signal to noise. Our study employed a structural encoding model similar to that used in our earlier image identification study (Kay et al., 2008), but we found that this model is insufficient for reconstructing natural images, given the fMRI signals collected in our study. Successful reconstruction requires two additional components: a natural image prior and a semantic model. The natural image prior ensures that potential reconstructions will satisfy all of the lower- and higher-order statistical properties of natural images. The semantic encoding model reflects the way that information about semantic categories is represented in brain responses measured in AOC. Our study is the first to integrate structural and semantic models with a natural image prior to produce reconstructions of natural images.

Under the Bayesian framework, each of the separate sources of information used for reconstruction are represented by a separate encoding model or image prior. This property of the framework makes it an efficient method for integrating information from disparate sources in order to optimize reconstruction. For example, adding the semantic model to the reconstruction process merely required adding an additional term to Equation 1. However, this property also has value even beyond its use in optimizing reconstructions. Because the sources of structural and semantic information are represented by separate models, the Bayesian framework makes it possible to disentangle the contributions of functionally distinct visual areas and prior information to reconstructing the structural and semantic content of natural images (see Figure 6).

But Is This Really Reconstruction?

Reconstruction using the natural image prior is accomplished by sampling from a large database of natural images. One obvious difference between this sampling approach and the methods used in previous studies (Thirion et al., 2006; Miyawaki et al., 2008) is that reconstructions will always correspond to an image that is already in the database. If the target image is not contained within the natural image prior then an exact reconstruction of the target image cannot be achieved. The database used in our study contains only six million images, and with a set this small it is extremely unlikely that any target image (chosen from an independent image set) can be reconstructed exactly. However, as the size of the database (i.e., the natural image prior) grows, it becomes more likely that any target image will be structurally and/or semantically indistinguishable from one of the images in the database. For example, if the database contained many images of one person's personal environment, it would be possible to reconstruct a specific picture of her mother using a similar picture of her mother. In this case, the fact that the reconstruction was not an exact replica of the target image would be irrelevant.

It is important to emphasize that in practice exact reconstructions are impossible to achieve by *any* reconstruction algorithm on the basis of brain activity signals acquired by fMRI. This is because all reconstructions will inevitably be limited by inaccuracies in the encoding models and noise in the measured signals. Our results demonstrate that the natural image prior is a powerful (if unconventional) tool for mitigating the effects of these fundamental limitations. A natural image prior with only six million images is sufficient to produce reconstructions that are structurally and semantically similar to a target image. There are many other potential natural image priors that could be used for this process, and some of these may be able to produce reconstructions even better than those demonstrated in this study. Exploration of alternative priors for image reconstruction and other brain decoding problems will be an important direction for future research.

New Insights from the Semantic Encoding Model

Many previous fMRI studies have investigated representations in the anterior regions of visual cortex, beginning in the region we have defined as AOC and extending beyond the slice coverage used here to more anterior areas such as the fusiform face area and the parahippocampal place area (Kanwisher et al., 1997; Epstein and Kanwisher, 1998). These anterior regions are more activated by whole images than by scrambled images (Malach et al., 1995; Grill-Spector et al., 1998), and some specialized regions appear to be most activated by specific object or scene categories (Kanwisher et al., 1997; Epstein and Kanwisher, 1998; Downing et al., 2001; Downing et al., 2006). A recent study using sophisticated multivariate techniques revealed a rough taxonomy of object representations within inferior temporal cortex (Kriegeskorte et al., 2008). Together, these studies indicate that portions of anterior visual cortex represent information related to meaningful objects and scenes—what we have referred to here as "semantic content."

This result forms the inspiration for our semantic encoding model, which assigns a unique semantic category to each natural image in order to predict voxel responses. This aspect of the model permits us to address one very basic and important

question that has not been addressed by previous studies: what proportion of the variance in the responses evoked by a natural image within a single voxel can be explained solely by the semantic category of the image? Our results show that for voxels in the region we have defined as AOC, semantic category alone can explain as much as 55% of the response variance (see Figure 5). An important direction for future research will be to apply the semantic encoding model to voxels in cortical regions that are anterior to our slice coverage. Recent work on a competing model that is conceptually similar to our semantic encoding model (Mitchell et al., 2008) suggests that the semantic encoding model will be useful for predicting brain activity in these more anterior areas.

The results in Figure 5 show that the particular structural features used to build the structural encoding model are very weakly correlated with semantic categories. However, it is important to bear in mind that all semantic categories are correlated with some set of underlying structural features. Although structural features underlying some categories of natural landscape (Greene and Oliva, 2009) have been discovered, the structural features underlying most semantic categories are still unknown (Griffin et al., 2007). Thus, it is convenient at this point to treat semantic categories as a form of representation that is qualitatively different from the structural features used for the structural encoding model.

One notable gap in our current results is that neither the structural nor semantic models can adequately explain voxel responses in intermediate visual areas such as area V4 (see Figure 5). These intermediate areas are thought to represent higher-order statistical features of natural images (Gallant et al., 1993). Because the structural model used here only captures the lower-order statistical structure of natural images (Field, 1987; Field, 1994) it does not provide accurate predictions of responses in these intermediate visual areas. Development of a new encoding model that accurately predicts the responses of individual voxels in intermediate visual areas would provide an important new tool for vision research and would likely further improve reconstruction accuracy.

Future Directions

Much of the excitement surrounding the recent work on visual reconstruction is motivated by the ultimate goal of directly picturing subjective mental phenomena such as visual imagery (Thirion et al., 2006) or dreams. Although the prospect of reconstructing dreams still remains distant, the capability of reconstructing natural images is an essential step toward this ultimate goal. Future advances in brain signal measurement, the development of more sophisticated encoding models, and a better understanding of the structure of natural images will eventually make this goal a reality. Such brain-reading technologies would have many important practical uses for brain-augmented communication, direct brain control of machines and computers, and for monitoring and diagnosis of disease states. However, such technology also has the potential for abuse. Therefore, we believe that researchers in this field should begin to develop ethical guidelines for the application of brain-reading technology.

EXPERIMENTAL PROCEDURES

Data Collection

The MRI parameters, stimuli, experimental design, and data preprocessing are identical to those presented in a previous publication from our laboratory (Kay et al., 2008). Here, we briefly describe the most pertinent details.

MRI Parameters

All MRI data were collected at the Brain Imaging Center at UC-Berkeley, using a 4 T INOVA MR (Varian, Inc., Palo Alto, CA) scanner and a quadrature transmit/receive surface coil (Midwest RF, LLC, Hartland, WI). Data were acquired in 18 coronal slices that covered occipital cortex (slice thickness 2.25 mm, slice gap 0.25 mm, field of view 128 × 128 mm^2). A gradient-echo EPI pulse sequence was used for functional data (matrix size 64 × 64, TR 1 s, TE 28 ms, flip angle 20°, spatial resolution 2 × 2 × 2.5 mm^3).

Stimuli

All stimuli were grayscale natural images selected randomly from several photographic collections. The size of the images was 20° × 20° (500 px × 500 px). A central white square served as the fixation point(0.2° × 0.2°; 4 px × 4 px). Images were presented in successive 4 s trials. In each trial, a photo was flashed at 200 ms intervals (200 ON, 200 OFF) for 1 s, followed by 3 s of gray background.

Experimental Design

Data for the model estimation and image reconstruction stages of the experiment were collected in the same scan sessions. Three subjects were used: TN, KK, and SN. For each subject, five scan sessions of data were collected. Scan sessions consisted of five model estimation runs and two image reconstruction runs. Runs used for model estimation were 11 min each and consisted of 70 distinct images presented two times each. Runs used for image reconstruction were 12 min each and consisted of 12 distinct images presented 13 times each. Images were randomly selected for each run and were not repeated across runs. The total number of distinct images used in the model estimation runs was 1750. For image reconstruction runs, the total was 120.

Data Preprocessing

Functional brain volumes were reconstructed and then coregistered across scan sessions. The time series data was used to estimate a voxel-specific response time course; deconvolving this time course from the data produced, for each voxel, an estimate of the amplitude of the response (a single value) to each image used in the model estimation and image reconstruction runs. Retinotopic mapping data collected in separate scan sessions was used to assign voxels to their respective visual areas based on criteria presented in Hansen et al. (2007).

Notation

All sections below use the same notational conventions. The response of a single voxel is denoted r. Bold notation is used to denote the collected responses of N separate voxels in an $Nx1$ voxel response vector: $\mathbf{r} = (r1,\ldots,r_N)^T$. Subscripts i applied to voxel response vectors, $\mathbf{r}_i$, are used to distinguish between functionally distinct brain areas. In practice, images are treated as vectors of pixel values, denoted $\mathbf{s}$. These vectors are formed by columnwise concatenation of the original 2D image; unless otherwise noted, $\mathbf{s}$ is a $128^2 \times 1$ column vector.

Encoding Models

Our reconstruction algorithm requires an encoding model for each voxel. Each encoding model describes the voxel's dependence upon a particular set of image features. Formally, this dependence is given by a distribution over the possible responses of the voxel to an image: $p(r|\mathbf{s})$. We presented two different types of encoding models: a structural encoding model and a semantic encoding model. Each model is defined by a transformation of the original image $\mathbf{s}$ into a set of one or more features. For the structural encoding model, these features are spatially localized orientations and spatial frequencies (which can be described using Gabor wavelets). For the semantic model, these features are semantic categories assigned to the images by human labelers. The conditional distributions $p(r|\mathbf{s})$ for both models are defined by one or more Gaussian distributions. For the structural models, $p(r|\mathbf{s})$ is a Gaussian distribution whose mean is a function of the image $\mathbf{s}$; for the semantic model, $p(r|\mathbf{s})$ is a *weighted sum* of Gaussian distributions, each of whose means is

a function of the image. Each model can be used to predict the specific response of a voxel to an image by taking the expected value of r with respect to $p(r|\mathbf{s})$. Note that if a voxel has a very weak dependence on the features assumed by the model, the expected value of r with respect to $p(r|\mathbf{s})$ will poorly predict the actual response of the voxel. Note also that both the structural and semantic models have a number of free parameters that must be estimated using a suitable fitting procedure.

Structural Encoding Model

The structural encoding model used in this work is similar to the Gabor Wavelet Pyramid model described in our previous publication (Kay et al., 2008). The model describes the spatial frequency and orientation tuning of each voxel. These attributes can be efficiently described by Gabor wavelets. A Gabor wavelet is a spatially localized filter with a specific orientation and spatial frequency. To construct the structural encoding model, all images are first filtered by a set of Gabor wavelets that cover many spatial locations, orientation, frequencies, and scales. The filtered signals are then passed through a fixed nonlinearity. This nonlinear transformation of the image defines the *feature set* for the structural encoding model. Formally, the features are defined as $\boldsymbol{f}(\mathbf{s}) = \log(|W^T\mathbf{s}| + 1)$, where $\boldsymbol{f}$ is an $F \times 1$ vector containing the features (F = 10,921, the number of wavelets used for the model), and W denotes a matrix of complex Gabor wavelets. W has as many rows as there are pixels in $\mathbf{s}$, and each column contains a different Gabor wavelet; thus, its dimension is $128^2 \times 10921$. The features are the log of the magnitudes obtained after filtering the image by each wavelet. The log is applied because we have found that a compressive nonlinearity improves prediction accuracy.

The wavelets in W occur at six spatial frequencies: 1, 2, 4, 8, 16, and 32 cycles per field of view (FOV = 20°; images were presented at a resolution of 500 × 500 pixels but were downsampled to 128 × 128 pixels for this analysis). At each spatial frequency of n cycles per FOV, wavelets are positioned on an $n \times n$ grid that tiles the full FOV. At each grid position, wavelets occur at eight orientations, 0°, 22.5°, 45°, …, and 157.5°. 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.

As mentioned above, the conditional distribution for the structural encoding model is a Gaussian distribution whose mean is defined by a weighted sum of the features $\boldsymbol{f}$:

$$p(r|\mathbf{s}) \propto \exp\left(-\frac{(r - \mathbf{h}^T\boldsymbol{f}(\mathbf{s}))^2}{2\sigma^2}\right)$$

where $\mathbf{h}$ ($F \times 1$) is the set of weighting parameters, and σ (a scalar) is the standard deviation of voxel responses.

The encoding model's predicted response to an image $\mathbf{s}$ is defined as the mean response with respect to $p(r|\mathbf{s})$. This mean is just the feature transform (which is the same for all voxels) multiplied by the weighting parameters $\mathbf{h}$ (which are fit independently for all voxels): $\hat{\mu}(\mathbf{s}) = \mathbf{h}^T\boldsymbol{f}(\mathbf{s})$.

Coordinate descent with early stopping was used to find the parameters $\mathbf{h}$ that minimized the sum-of-square-error between the actual and predicted responses. For each voxel, this minimization was performed on three sets of M-I training samples (M = 1750, $I = M$*0.1), selected randomly without replacement. Each set produced a separate estimate $\mathbf{h}_j$ (j = [1, 2, 3]). $\mathbf{h}$ was set equal to the arithmetic mean of ($\mathbf{h}_1$, $\mathbf{h}_2$, $\mathbf{h}_3$).

Semantic Encoding Model

The features used for the semantic encoding model are quite different from those used for the structural encoding model. Instead of features that are defined by a wavelet transformation, the features for the semantic model are semantic categories assigned to each image by human labelers.

The semantic categories used for the model were drawn from a *semantic basis*. The semantic basis is a set of categories designed to satisfy two key properties. First, the categories are broad enough that any natural image can be assigned to *at least* one of them. Second, categories in the semantic basis are nonoverlapping, so that a human observer can confidently assign any arbitrary image to *only* one of them. We developed the semantic basis using a tree of categories, shown in Figure S1. In the first layer of the tree, all possible images are divided into mutually exclusive categories: "mostly animate" and "mostly inanimate." In subsequent layers, each category is again divided into two or three exclusive categories. At the bottom of the tree (rightmost layer in Figure S1) is a set of 23 categories; this is the semantic basis. The inputs to the model are natural images that have been labeled with one of these 23 categories by two human observers. The observers did not know whether the images they labeled were target images, training images, or potential reconstructions sampled from the natural image prior. Observers were instructed to label each image by working their way down the semantic tree: first they assigned the correct label from the first level, then the second level, and so on until reaching the bottom of the tree. In a few cases the labels assigned by different labelers were inconsistent, and in these cases they discussed the images between themselves in order to arrive at a consistent conclusion.

The form of the conditional distribution for the semantic encoding model is slightly more complex than for the structural model. In order to make the model easily interpretable, we designed it so that it would clearly delineate the semantic categories that a voxel likes (categories that are likely to evoke above average responses), doesn't like (categories that are likely to evoke below average responses), or doesn't care about (categories that are likely to evoke near average responses). This clear delineation is achieved by decomposing the overall voxel response distribution (the gray curves in the top panels of Figure 4) into a mixture of subdistributions that span the above, below, and near average response ranges:

$$p(r|c(\mathbf{s})) = \sum_z p(r|z)p(z|c(\mathbf{s}))$$

where $z \in$ [1,2,3] is an indicator variable used to delineate the ranges, and $c(\mathbf{s})$ denotes the semantic category assigned to the image $\mathbf{s}$ (This notation is used throughout to make the model's dependence on semantic categories explicit. However, the more general notation used for the structural model, $p(r|\mathbf{s})$, is applicable here as well). Each of the subdistributions, $p(r|z)$ is a Gaussian (colored curves in top panels of Figure 4) with its own mean and variance, μ_z and σ_z:

$$p(r|z) \propto \exp\left(-\frac{(r - \mu_z)^2}{2\sigma_z^2}\right)$$

Each of these Gaussian subdistributions are weighted by a multinomial mixing distribution, $p(z|c(\mathbf{s}))$, that gives the probability that the voxel's response will be driven into response range z when presented with an image from category c (bar charts in bottom panels of Figure 4).

The predicted response of the semantic model to an image $\mathbf{s}$ is the mean of $p(r|c(\mathbf{s}))$. This mean is a weighted sum of the means of each of subdistribution:

$$\hat{\mu}(c(\mathbf{s})) = \sum_z \mu_z p(z|c(\mathbf{s}))$$

The free parameters of the semantic encoding models are the mean and variance of $p(r|z)$ for each value of z, and the parameters of the multinomial mixture distribution $p(z|c(\mathbf{s}))$. We estimated these parameters for each voxel using an expectation maximization algorithm that we present in Appendix 1 (see Supplemental Data).

Voxel Selection and Multivoxel Encoding Models

To perform reconstruction using the responses of many voxels, it is necessary to first select a set of voxels for use in reconstruction and then combine the individual encoding model for each selected voxel into a single multivoxel encoding model for the entire set.

Voxels were selected for reconstruction on the basis of the predictive accuracy of their encoding models. If the prediction accuracy of the structural encoding model was above threshold (see below), it was considered a *structural voxel,* and it was used for all reconstructions that involved the structural encoding model. If the prediction accuracy of the semantic encoding model for a voxel was above threshold (see below) it was considered a *semantic voxel,* and it was used for all reconstructions that involved the semantic encoding model. In the rare cases where both the structural *and* the semantic encoding models were above the selection thresholds for both models, the voxel was used for both structural and semantic reconstruction.

For the structural model, the threshold was a correlation coefficient of >0.353. This correlation coefficient corresponds to a p value < 3.9×10^{-5}, which is roughly the inverse of the number of voxels in our data set. For the semantic model, the threshold was a correlation coefficient of >0.26. This correlation coefficient was chosen because it optimized semantic accuracy on an additional set of 12 experimental trials obtained for subject TN (none of these trials were part of the model estimation or image reconstruction sets used here).

In order to control for a possible selection bias, the correlation values for both the structural and semantic encoding models were calculated separately for each of the image reconstruction trials. To reconstruct the j^{th} image, the correlation coefficients were calculated using the remaining 119 image reconstruction trials. Thus, a slightly different set of voxels was selected for each reconstruction trial. The average number of voxels selected by the structural model was 788 (average taken across all three subjects and all reconstruction trials). The average number of voxels selected by the semantic model was 579. The average number of voxels selected by both was 73.

Once voxels were selected for each reconstruction trial, multivoxel versions of the structural and semantic encoding models were constructed using the univariate model for each of the selected voxels. The multivoxel versions of the structural and semantic encoding models are given by the following distribution:

$$p(\mathbf{r}|\mathbf{s}) \propto \exp\left[-\frac{1}{2}(\mathbf{r}' - \hat{\mathbf{r}}(\mathbf{s}))^T \Lambda^{-1} (\mathbf{r}' - \hat{\mathbf{r}}(\mathbf{s}))\right]$$

where Λ is a covariance matrix. Let $\hat{\mu}_i(\mathbf{s}) := \langle r_i|\mathbf{s}\rangle$ be the predicted response for the i^{th} voxel, given an image $\mathbf{s}$ (the predicted mean response for the structural and semantic encoding models are defined above). Let $\hat{\boldsymbol{\mu}}(\mathbf{s}) = (\hat{\mu}_1(\mathbf{s}), \ldots, \hat{\mu}_N(\mathbf{s}))^T$ be the collection of predicted mean responses for N voxels. We define $\hat{\mathbf{r}}$ as the *normalized* predicted mean response vector:

$$\hat{\mathbf{r}}(\mathbf{s}) = \frac{P^T \hat{\boldsymbol{\mu}}(\mathbf{s})}{\|P^T \hat{\boldsymbol{\mu}}(\mathbf{s})\|}$$

where the sidebars denote vector normalization and the columns of the matrix P contain the first p principal components of the distribution over $\hat{\boldsymbol{\mu}}$ (p = 45 for the structural model; p = 21 for the semantic model. For all subjects and both models, these values of p occur at or near the inflection point of the plot of rank-ordered eigenvalues). The prime notation denotes the same linear transformation and scaling of measured response vectors:

$$\mathbf{r}' = \frac{P^T \mathbf{r}}{\|P^T \mathbf{r}\|}$$

To estimate P for the structural encoding model, we generated predicted mean response vectors to a gallery of 12,000 natural images and applied standard principal components analysis to this sample. For the semantic encoding model, we used a smaller gallery of 3000 images labeled according to the scene categories shown in Figure S1 (rightmost layer of the tree). The reduction of dimensionality achieved by projection onto the first p principal components of the predicted responses, and the normalization after projection, act to stabilize the inverse of the covariance matrix, Λ. The elements of Λ give the covariance of the residuals (i.e., the difference between the responses and predictions, $r' - \hat{r}(\mathbf{s})$). We used the 1750 trials in the model estimation set to estimate Λ.

General Reconstruction Algorithm

All of the reconstructions presented in the paper are special cases of a general Bayesian algorithm, summarized by the following equation:

$$p(\mathbf{s}|\mathbf{r}) \propto p(\mathbf{s}) \prod_i p_i(\mathbf{r}_i|\mathbf{s})$$

On the left-hand side is the posterior distribution, $p(\mathbf{s}|\mathbf{r})$. The posterior gives the probability that an image $\mathbf{s}$ evoked the measured response $\mathbf{r}$. The goal of reconstruction is to find the image with the highest posterior probability, given the responses (this is often referred to as *maximum a posteriori* decoding). The formula on the right-hand side shows how the posterior probability is calculated. The first term, $p(\mathbf{s})$, is the image prior. It reflects preexisting, general knowledge about natural images and is independent of the responses. We consider three separate priors in this study: the flat prior, the sparse Gabor prior, and natural image prior. The image prior is followed by a product of encoding models, p_i, each of which is applied to the responses, $\mathbf{r}_i$, of voxels in a functionally distinct brain area. To produce reconstructions, we used either one (structural) or two (structural and semantic) encoding models.

The four different reconstruction methods presented in the main text differ only by the particular choice of priors and encoding models used to calculate the posterior probability. For reconstructions that use the structural model and a flat prior, the posterior is $p(\mathbf{s}|\mathbf{r}_1) \propto p_1(\mathbf{r}_1|\mathbf{s})$, where p_1 is the structural encoding model, and $\mathbf{r}_1$ are the structural voxels (selected according to the voxel selection procedure defined above: see "Voxel Selection and Multivoxel Encoding Models"). Note that the image prior, $p(\mathbf{s})$, does not appear here because the flat prior is simply a constant that is independent of both images and responses.

For reconstruction with the structural encoding model and sparse Gabor prior, the posterior is $p(\mathbf{s} \mid \mathbf{r}_1) \propto p_1(\mathbf{r}_1 \mid \mathbf{s})\, p_{SG}(\mathbf{s})$, where $p_{SG}(\mathbf{s})$ is the sparse Gabor prior described in detail below.

For reconstructions with the structural model and natural image prior, the posterior is $p(\mathbf{s} \mid \mathbf{r}_1) \propto p_1(\mathbf{r}_1 \mid \mathbf{s})\, p_{NIP}(\mathbf{s})$, where $p_{NIP}(\mathbf{s})$ is the natural image prior.

Finally, for the hybrid reconstructions, the posterior incorporates two encoding models: $p(\mathbf{s} \mid \mathbf{r}) \propto p_1(\mathbf{r}_1 \mid \mathbf{s})\, p_2(\mathbf{r}_2 \mid \mathbf{s}) p_{NIP}(\mathbf{s})$, where p_2 is the semantic encoding model, and $\mathbf{r}_2$ are the semantic voxels (selected according to the procedure defined above: see "Voxel Selection and Multivoxel Encoding Models").

Once a posterior distribution is defined, a reconstruction is produced by finding an image that has a high posterior probability. In general, it is not possible to determine the image that maximizes the posterior distribution analytically. Thus, a search algorithm must be applied to search the space of possible images for potential reconstructions that have high posterior probability.

Reconstructions Using the Structural Encoding Model and a Flat Prior

For the reconstructions presented in the second column of Figure 2, a set of voxels located primarily in the early visual areas V1, V2, and V3 (see above for explanation of how these voxel were selected) and a multivoxel structural encoding model were used.

The prior used for this type of reconstruction was the trivial or "flat" prior: p(**s**) = constant.

This prior assigns the same value to all possible images, including those with randomly selected pixel values. In this case, the posterior probability $p(\mathbf{s}|\mathbf{r})$, and the likelihood, $p(\mathbf{r}|\mathbf{s})$, are proportional.

To produce reconstructions, a greedy serial search algorithm was used to maximize the posterior distribution. At each iteration of the algorithm, a small group of pixel values in the reconstruction was updated. If the newly updated pixel values increased the posterior probability, they were retained as part of the reconstructed image. Otherwise, they were rejected. The procedure halted when the change in posterior probability remained below a small threshold (4 bits) for a number of iterations. A formal description of the procedure is given in Appendix 2 (see Supplemental Data).

Reconstructions Using the Structural Encoding Model and a Sparse Gabor Prior

For the reconstructions presented in the third column of Figure 2, we used the same selected voxels and structural encoding model as above (see "Reconstructions Using the Structural Encoding Model and a Flat Prior"). Instead of a flat prior, we used a sparse Gabor prior. The sparse Gabor prior places high probability on images that have the 1/*f* amplitude spectrum characteristic of natural images. In other words, this prior prefers images in which nearby pixels are somewhat correlated. The distribution also assigns high probability to images that are sparse in the Gabor wavelet domain. Suppose that $\mathbf{a} = G\mathbf{s}$ is the transformation of an image $\mathbf{s}$ into the Gabor domain, where G is a matrix of real-valued Gabor wavelets, and $\mathbf{a} = (a_1, \ldots, a_g)$, is a vector whose elements, a_i,

denote the "activation" of the i^{th} Gabor wavelet in G. To say that images are "sparse" in the Gabor domain means that their Gabor activations obey a distribution with a sharp peak and a steep falloff (in other words, a distribution with high kurtosis). This aspect of natural images was captured using a Laplace distribution:

$$p(a_i) = \frac{1}{2\beta_i}\exp\left(-\left|\frac{a_i - u_i}{\beta_i}\right|\right)$$

where u_i, and β_i determine the mean and variance of the distribution.

To generate an image from the sparse Gabor prior, activations for all of the wavelets in the Gabor basis G are sampled independently from the above Laplace. The activations are then linearly transformed back into the pixel domain to obtain an image. Finally, this image is transformed again by an "unwhitening" matrix, U, and offset by μ_s, the mean of all natural images: $\mathbf{s} = UG^{-1}\mathbf{a} + \mu_s$.

Effectively, the application of U smooths the image so that it possesses the $1/f$ structure characteristic of natural images.

The Laplace distribution, along with the transformation from Gabor activations into the pixel domain, together define an explicit formula for the sparse Gabor prior:

$$p_{SG}(\mathbf{s}) = \int_{\mathbf{a}} p(\mathbf{s}|\mathbf{a})p(\mathbf{a})d\mathbf{a}$$

where $p(\mathbf{s} \mid \mathbf{a}) = 1$ whenever $\mathbf{s} = UG^{-1}\mathbf{a} + \mu_s$ and is set to zero otherwise. The Gabor activations are assumed to be independent of each other, so

$$p(\mathbf{a}) = \prod_i p(a_i).$$

This equation is just a formal way of stating that under the sparse Gabor prior, the probability of sampling an image $\mathbf{s}$ is proportional to the probability of sampling its underlying Gabor activations $\mathbf{a}$. (Note that this model has a number of free parameters that must be chosen or estimated empirically. Explicit formulas for estimating these parameters are given in Appendix 3 of the Supplemental Data.)

Reconstructions with the structural model and sparse Gabor prior were generated using a search algorithm identical to the one used for reconstruction with a flat prior, except that in this case, reconstructions were updated at each iteration of the algorithm by incrementing the activation for a single Gabor by ±0.1. The updated reconstruction was transformed into pixel space, and its posterior probability was evaluated using the sparse Gabor prior.

Reconstructions Using the Structural Encoding Model and a Natural Image Prior

To produce the reconstructions shown in the fourth column of Figure 2, the second column of Figure 3, and Figure S3A, we used the structural encoding model and corresponding structural voxels. Instead of a sparse Gabor prior, we used an implicit natural image prior, $p_{NIP}(\mathbf{s})$. Informally, the natural image prior is simply a large (6 million samples) database of natural images. Formally, it is a distribution that assigns a fixed value to all the images in the database, and a zero value to all images that are not:

$$p_{NIP}(\mathbf{s}) = \frac{1}{C}\sum_{i=1}^{C}\delta_{\mathbf{s}^{(i)}}(\mathbf{s})$$

where C is the total number of images in the database, and $\delta_{\mathbf{s}^{(i)}}$ is the delta function that returns 1 whenever $\mathbf{s} = \mathbf{s}^{(i)}$ (the i^{th} image in the database) and a 0 otherwise.

Reconstruction was performed by simply evaluating the posterior probability for each of the images in the database [note that for images in the natural image prior, the posterior is proportional to the likelihood $p(\mathbf{r} \mid \mathbf{s})$] and choosing the one that resulted in the highest posterior probability. Evaluating the posterior is computationally intensive. As a time-saving approximation, we first evaluated each image in the database using the voxel-wise correlation between the measured responses and the responses predicted by the encoding model. This metric was used in Kay et al. (2008) for image identification. For each target image, we retained the 100 images with the highest correlation. We then evaluated each of these 100 images under $p(\mathbf{s} \mid \mathbf{r})$. The image with the highest $p(\mathbf{s} \mid \mathbf{r})$ was retained as the reconstruction.

Reconstructions Using the Structural Encoding Model, the Semantic Encoding Model, and the Natural Image Prior (Hybrid Method)

To produce reconstructions shown in the third column of Figure 3 (and in Figure S3B) the posterior probabilities were evaluated for each of the images in the natural image prior using both the structural encoding model and the semantic encoding model. Reconstruction using this *hybrid method* was performed for 30 of the image reconstruction trials (all of these were from the first scan session).

To evaluate the semantic encoding model for a given image $\mathbf{s}$, the image must be assigned a semantic category from the semantic basis set (Figure S1). Because it is not feasible to label all 6 million images in the natural image prior, we labeled only those images with relatively high likelihoods under the structural encoding model. For a single reconstruction trial this set of images was defined as $S_r = \{\mathbf{s} : \mathbf{s} \in S, \textit{and}\ p_1(\mathbf{r}_1 \mid \mathbf{s}) > \beta_r\}$, where S is the database of images. β_r was chosen so that S_r contained 100 images.

Reconstruction Accuracy

Structural Accuracy

Structural accuracy of the reconstructions was assessed using the weighted complex wavelet structural similarity metric (Brooks and Pappas, 2006). The metric uses the coefficients of a complex wavelet decomposition of two images in order to compute a single number describing the degree of structural similarity between the two images. To produce the structural accuracy metrics in Figure 6 (left panels), the similarity between each target image and its reconstruction was averaged across all 120 reconstruction trials. (For the case of the hybrid method, averages were taken over the smaller set of 30 reconstructions.) Note that this metric is not the same as the posterior probability of an image. Thus, a rank-ordering of images according to this metric may not perfectly correspond to a rank-ordering of images according to their posterior probabilities (as in Figure S3).

Semantic Accuracy

To assess the semantic accuracy of the reconstructions, the probability of a semantic category, α, given a response, $\mathbf{r}$, was calculated for each of the 30 reconstruction trials used for the hybrid method. If the most probable category was also the category of the target image, the trial was considered to be semantically accurate. Semantic accuracy for each type of reconstruction (Figure 6, right panels) is the fraction of semantically accurate reconstruction trials.

The probability of a semantic category given a response, $p(\alpha \mid \mathbf{r})$, is calculated via a linear operation on the encoding models:

$$p(\alpha|\mathbf{r}) \propto p(\alpha)\sum_{s\in S_\mathbf{r}} p(\mathbf{r}|\mathbf{s})p(\mathbf{s}|\alpha)$$

where $p(\mathbf{r}|\mathbf{s})$ can refer to either the structural encoding model, the semantic encoding model, or the product of the two (as in Equation 1). The distribution $p(\mathbf{s} \mid \alpha)$ is the probability of an image, given a category. This probability was set to 0 if the image was not a member of the category, and a constant value otherwise. The prior on categories, $p(\alpha)$, was assumed to be flat. Thus, the above equation has a very intuitive explanation: the probability of a semantic category given a response is proportional to the average likelihood of all the images from that category within a subset ($S_\mathbf{r}$; see above for definition) of the database.

SUPPLEMENTAL DATA

Supplemental Data include four figures and three mathematical appendices and can be found with this article online at http://www.cell.com/neuron/supplemental/S0896-6273(09)00685-0.

ACKNOWLEDGMENTS

This work was supported by an NRSA postdoctoral fellowship (T.N.), the National Institutes of Health, and University of California, Berkeley, intramural

funds. We thank B. Inglis for assistance with MRI, K. Hansen for assistance with retinotopic mapping, D. Woods and X. Kang for acquisition of whole-brain anatomical data, and A. Rokem for assistance with scanner operation. We thank A. Berg for assistance with the natural image database, B. Yu and T. Griffiths for consultation on the mathematical analyses, and S. Nishimoto, and D. Stansbury for their help in various aspects of this research.

Accepted: September 9, 2009
Published: September 23, 2009

REFERENCES

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.

Cadieu, C., and Olshausen, B. (2009). Learning transformational invariants from time-varying natural movies. Proc. Adv. Neural Inform. Process. Syst. *21*, 209–216.

Carandini, M., Demb, J.B., Mante, V., Tolhurst, D.J., Dan, Y., Olshausen, B.A., Gallant, J.L., and Rust, N.C. (2005). Do we know what the early visual system does? J. Neurosci. *25*, 10577–10597.

Carlson, T.A., Schrater, P., and He, S. (2002). Patterns of activity in the categorical representations of objects. J. Cogn. Neurosci. *15*, 704–717.

Cox, D.D., and Savoy, R.L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage *19*, 261–270.

David, S.V., and Gallant, J.L. (2005). Predicting neuronal responses during natural vision. Network *16*, 239–260.

Downing, P.E., Jiang, Y., Shuman, M., and Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. Science *293*, 2470–2473.

Downing, P.E., Chan, A.W., Peelen, M.V., Dodds, C.M., and Kanwisher, N. (2006). Domain specificity in visual cortex. Cereb. Cortex *16*, 1453–1461.

Epstein, R., and Kanwisher, N. (1998). A cortical representation of the local visual environment. Nature *392*, 598–601.

Field, D.J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. J. Opt. Soc. Am. Optic. Image. Sci. Vis. *4*, 2379–2394.

Field, D.J. (1994). What is the goal of sensory coding? Neural Comput. *6*, 559–601.

Gallant, J.L., Braun, J., and Van Essen, D.C. (1993). Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. Science *259*, 100–103.

Griffin, G., Holub, A.D., and Perona, P. (2007). The Caltech-256. Caltech Technical Report 2007.

Grill-Spector, K., and Malach, R. (2004). The human visual cortex. Annu. Rev. Neurosci. *27*, 649–677.

Grill-Spector, K., Kushnir, T., Hendler, T., Edelman, S., Itzchak, Y., and Malach, R. (1998). A sequence of object-processing stages revealed by fMRI in the human occipital lobe. Hum. Brain Mapp. *6*, 316–328.

Greene, R., and Oliva, A. (2009). Recognition of natural scenes from global properties: seeing the forest without representing the trees. Cog. Psych. *58*, 137–176.

Hansen, K.A., Kay, K.N., and Gallant, J.L. (2007). Topographic organization in and near human visual area V4. J. Neurosci. *27*, 11896–11911.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science *293*, 2425–2430.

Haynes, J.D., and Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat. Neurosci. *8*, 686–691.

Hays, J., and Efros, A.A. (2007). Scene completion using millions of photographs. ACM. Trans. Graph (SIGGRAPH 2007) *26*, 1–5.

Kamitani, Y., and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. Nat. Neurosci. *8*, 679–685.

Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. *17*, 4302–4311.

Karklin, Y., and Lewicki, M.S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. Nature *457*, 83–86.

Kay, K.N., and Gallant, J.L. (2009). I can see what you see. Nat. Neurosci. *12*, 245.

Kay, K.N., Naselaris, T., Prenger, R.J., and Gallant, J.L. (2008). Identifying natural images from human brain activity. Nature *452*, 352–355.

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P.A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron *60*, 1126–1141.

Malach, R., Reppas, J.B., Benson, R.R., Kwong, K.K., Jiang, H., Kennedy, W.A., Ledden, P.J., Brady, T.J., Rosens, B.R., and Tootell, R.B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proc. Natl. Acad. Sci. USA *92*, 8135–8139.

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K., Malave, V.L., Mason, R.A., and Just, M.A. (2008). Predicting human brain activity associated with the meanings of nouns. Science *320*, 1191–1195.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M., Morito, Y., Tanabe, H.C., Sadato, N., and Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron *60*, 915–929.

Nevado, A., Young, M.P., and Panzeri, S. (2004). Functional imaging and neural information coding. Neuroimage *21*, 1083–1095.

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.B., Lebihan, D., and Dehaene, S. (2006). Inverse retinotopy: inferring the visual content of images from brain activation patterns. Neuroimage *33*, 1104–1116.

Torralba, A., Fergus, R., and Freeman, W.T. (2008). 80 million tiny images: a large data set for nonparametric object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. *30*, 1958–1970.

Wu, M.C.K., David, S.V., and Gallant, J.L. (2006). Complete functional characterization of sensory neurons by system identification. Annu. Rev. Neurosci. *29*, 477–505.

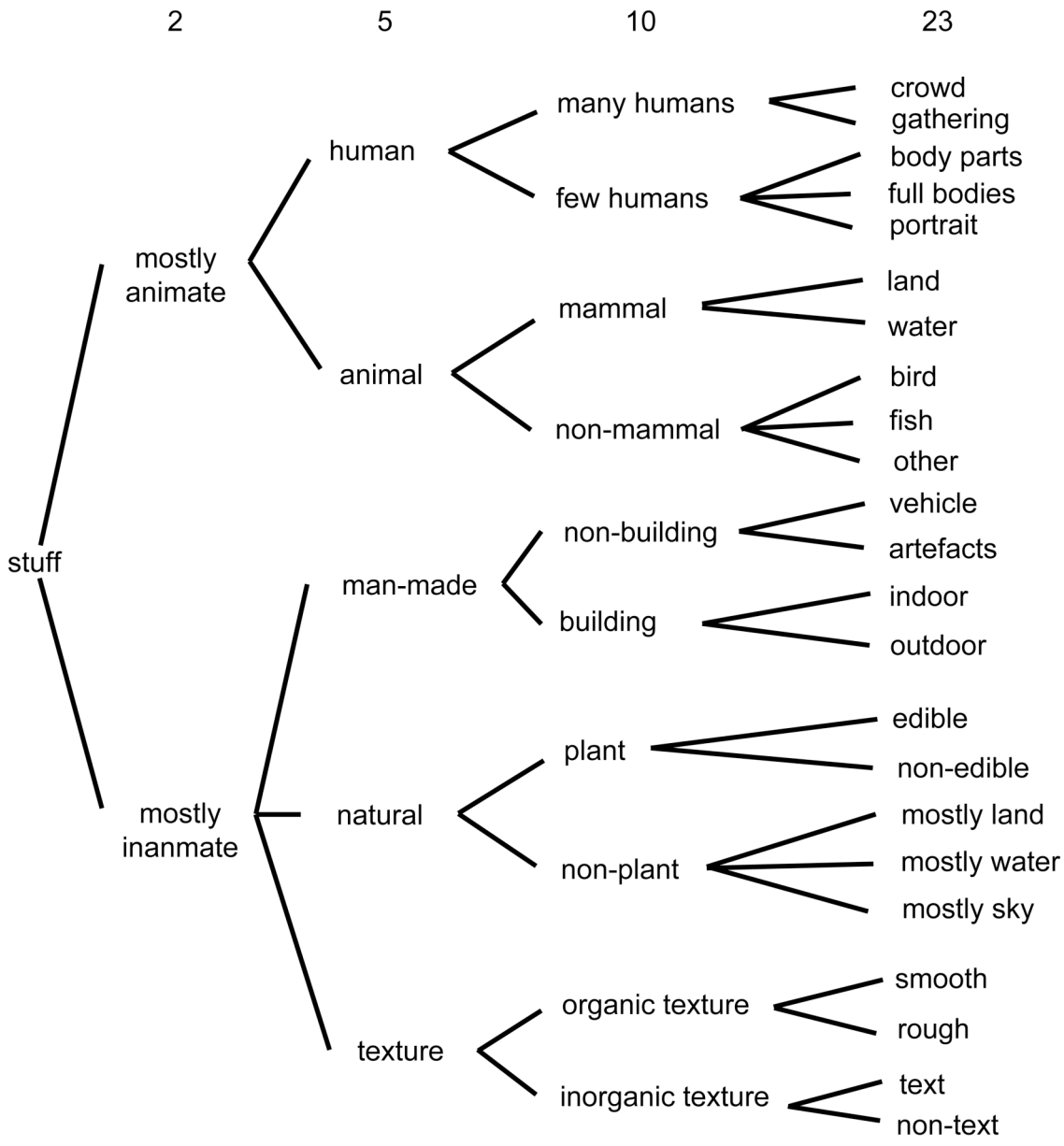
Zhang, K., Ginzburg, I., McNaughton, B.L., and Sejnowski, T.J. (1998). Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. J. Neurophysiol. *79*, 1017–1044.

Supplemental Data for:

Bayesian reconstruction of natural images from human brain activity

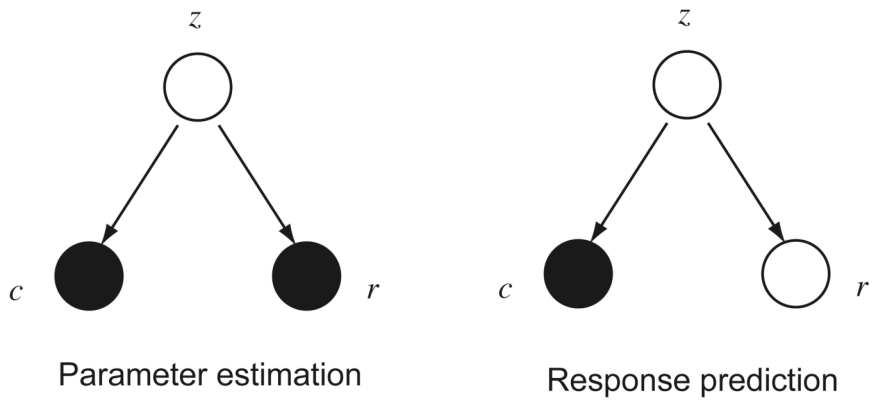
Thomas Naselaris, Ryan Prenger, Kendrick Kay, Michael Oliver, Jack Gallant

Neuron, Volume 63

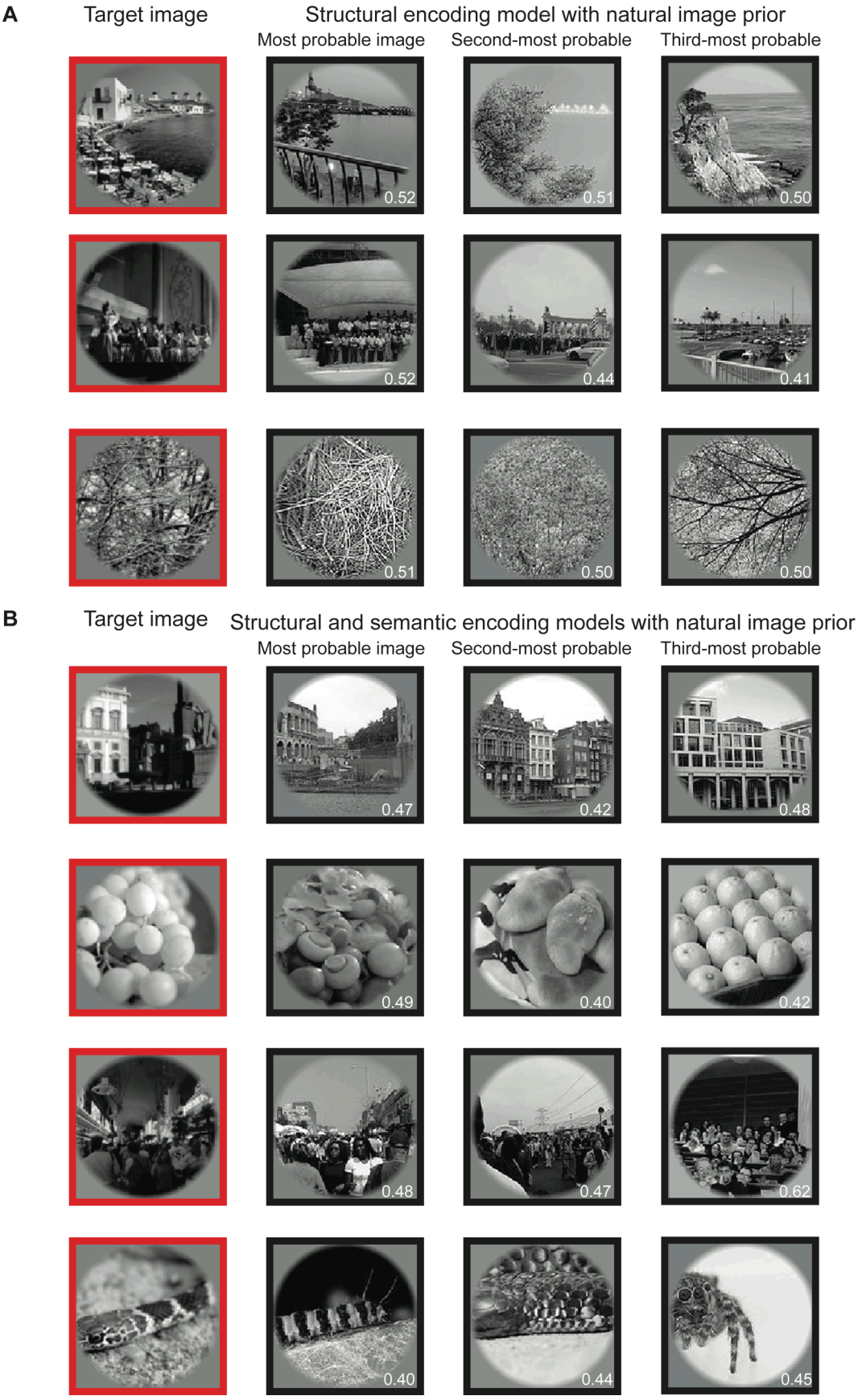


Supplemental Figure 1. The semantic basis used for the semantic encoding model.
The basis used for the semantic model consists of 23 mutually exclusive categories, represented here as the branches of a semantic tree. The total number of categories in each layer of the tree is indicated at top (2, 5, 10, 23). By following the tree from the root to the branches, each image can be assigned one unique semantic label. The semantic tree was constructed in an attempt to ensure that categories at the same level of the tree would

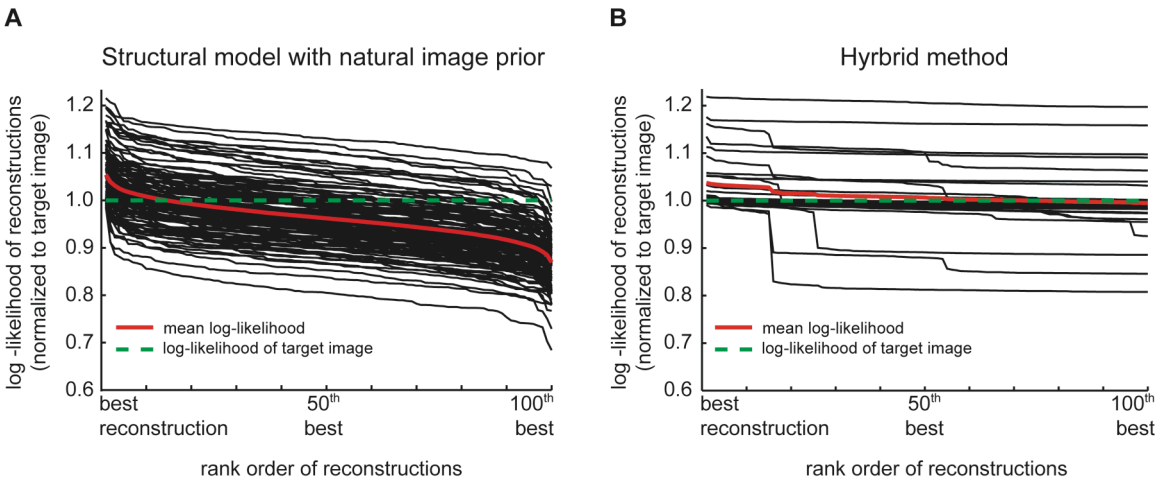
have a roughly equal number of exemplars in a database of randomly selected natural images. Images were labeled by observers who were naïve to the details of the model and experiment. The semantic categories shown here were also used to measure semantic accuracy.



Supplemental Figure 2. Formal graphical depiction of the semantic encoding model. The semantic encoding model is derived from a joint distribution over semantic categories *c*, voxel responses *r* and a latent variable *z*. The latent variable *z* links semantic categories to the responses. For a given value of *z*, the responses *r* are assumed to be Gaussian distributed, while categories *c* are assumed to follow a multinomial distribution. To estimate the parameters of this joint distribution for a single voxel (left panel), we observe the semantic categories and responses (filled nodes indicate that the variables are directly observed) for all of the trials in the model estimation set, and apply an Expectation Maximization algorithm. To predict voxel responses (right panel), we observe only *c* and integrate over the hidden states to obtain a distribution on the possible responses, $p(r|c)$. This conditional distribution is the semantic encoding model. The mean of this distribution is taken as the predicted response of the voxel.


A
Target image
Structural encoding model with natural image prior
Most probable image
Second-most probable
Third-most probable
0.52
0.51
0.50
0.52
0.44
0.41
0.51
0.50
0.50
B
Target image
Structural and semantic encoding models with natural image prior
Most probable image
Second-most probable
Third-most probable
0.47
0.42
0.48
0.49
0.40
0.42
0.48
0.47
0.62
0.40
0.44
0.45

Supplemental Figure 3. Gallery with nearly optimal reconstructions. **A,** The three target images from Figure 2 are shown in the first column (red borders). The second through fourth columns show reconstructions obtained using the structural encoding model and the natural image prior. The second column is the reconstruction with the highest posterior probability (the *maximum a posteriori* estimate); the third column is the second most probable reconstruction; and the fourth column is the third most probable reconstruction. For all three target images the three most probable reconstructions are structurally accurate (numbers in bottom right corner indicate structural accuracy, see Experimental Procedures for details). However, some of the reconstructions are not semantically accurate (see second and third most probable reconstructions in row 2). **B,** The four target images from Figure 3 are shown in the first column (red borders). The second through fourth columns show reconstructions obtained using the structural encoding model, the semantic encoding model and the natural image prior (i.e., the hybrid method). The second through fourth columns show the three reconstructions nearest to the peak of the posterior distribution, as in panel A. For all four target images the three most probable reconstructions are both structurally and semantically accurate (numbers in bottom right corner indicate structural accuracy, see main text for details).



Supplemental Figure 4: Analysis of sub-optimal reconstructions. The natural image prior consisted of 6 million natural images. The decoding algorithm assigned a posterior probability to each one of these 6 million images, and the maximum of the posterior distribution was taken as the best reconstruction. **A,** Normalized log-likelihoods under the structural encoding model for the 100 most probable reconstructions from a single image reconstruction trial (120 trials total; all data from subject TN). (Note that log-likelihood and posterior-probabilities will produce the same ranking; here log-likelihoods for the reconstruction were divided by the log-likelihood of the target image.) On each trial the "best reconstruction" is the image with the highest log-likelihood (referred to as the "reconstruction" in the main text) and the "100th best reconstruction" is the image with the 100th highest log-likelihood. When the structural encoding model is used alone the log-likelihoods decay smoothly away from the best reconstruction. The best reconstruction usually has a higher log-likelihood than the target image (compare red line to dashed green), but log-likelihoods decay quickly below the log-likelihood of the target image (~10th best image). Thus, it appears that the size of the database used in our study was large enough to ensure that the best reconstruction had a high log-likelihood relative to the target image. **B**, Normalized log-likelihoods under the structural and semantic encoding models for the 100 most probable reconstructions from a single image reconstruction trial (30 trials total; all data from subject TN). The best reconstruction usually has a higher log-likelihood than the target image, indicating that the size of the database was large enough to ensure that the best reconstruction had a high log-likelihood

relative to the target image. Here the log-likelihoods decay in discrete steps, reflecting the discrete semantic category boundaries.

Mathematical Appendices

APPENDIX 1: Derivation of Expectation Maximization algorithm for the semantic encoding model

The semantic encoding model describes a relationship between three variables: *c*, the semantic category; *r*, the voxel response; and *z*, a discrete latent variable that decomposes the overall distribution of responses into a set of sub-distributions. The relationship between these variables is depicted by the graphical model in Supplemental Figure 2. The model specifies the joint distribution over *c*, *r*, and *z* as:

$$p(r,c,z) = p(z)p(c \mid z)p(r \mid z) \qquad \text{(A1)}$$

To obtain an encoding model, $p(r|c)$, we integrate out the *z* variable and normalize:

$$p(r \mid c) = \frac{1}{p(c)} \sum_{z=1}^{K} p(z)p(r \mid z)p(c \mid z)$$

Notice that, after applying Bayes theorem to $p(z|c)$, this formulation of the encoding model is equivalent to the formulation presented in the section *Semantic Encoding Model* in **Experimental Procedures**.

Equation A1 shows that the semantic encoding model depends upon three underlying distributions. First is a multinomial prior on the variable *z*:

$$p(z) = \prod_i^K \pi_i^{z_i}, \quad \pi_i \geq 0, \quad and \quad \sum_i \pi_i = 1$$

Here, the meaning of the *z* notation is slightly modified, so that $z = (z_1, \ldots, z_K)$, $z_i \in [0,1]$, is a binary string used to encode the *K* possible values of *z* (in our work, *K* = 3). The π_i's are parameters that determine the prior probability of each of the $i \in [1, K]$ states.

Second is a multinomial distribution that relates *c* and *z*:

$$p(c \mid z) = \prod_i^K \prod_j^L \gamma_{ij}^{c_j z_i}, \quad \gamma_{ij} \geq 0, \quad and \quad \sum_j \gamma_{ij} = 1$$

where the γ_{ij}'s are parameters that determine the probability of each of the $j \in [1, L]$ semantic categories (in our work, *L* = 23), given z_i=1. Here, the meaning of the *c* notation has also been modified, so that $c = (c_1, \ldots, c_L)$, $c_j \in [0,1]$,

Third, is a Gaussian voxel response distribution:

$$p(r \mid z) = \prod_i^K \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(r-\mu_i)^2}{2\sigma_i^2} \right]^{z_i} \qquad \text{(A2)}$$

where the μ_i's and σ_i's are the mean and variance of the voxel responses, given z_i=1.

We derived an Expectation Maximization (EM) procedure to estimate the parameters for these distributions. Let Θ denote the set of free parameters for the three distributions described above:

$$\Theta = (\pi, \gamma, \mu, \sigma).$$

The EM procedure maximizes the expected log-likelihood of the parameters:

$$l_{EM}(\Theta) = E[\log p(\mathbf{r}, \mathbf{c}, \mathbf{z} \mid \Theta)]$$

where E denotes expectation taken with respect to z and it's posterior distribution $p(z \mid r, c)$. In this context, the bold notation $\mathbf{r} = (r^1, \ldots, r^M)$ refers to a set of M independent samples of responses for a single voxel (notice the abuse of notation: **r** is not a collection single responses from many voxels in this case, but a collection of many responses from a single voxel.) The same convention applies to **c** and **z**.

To write out $l_{EM}(\Theta)$ explicitly we first apply the logarithm to the joint distribution, thus converting all of the products to sums:

$$\log p(\mathbf{r},\mathbf{c},\mathbf{z}\mid\Theta)=\sum_{m}^{M}\sum_{i}^{K}\log\pi_i^{z_i^m}+\log\mathcal{N}(r^m\mid\mu_i,\sigma_i^2)^{z_i^m}+\sum_{j}^{L}\log\gamma_{ij}^{c_j^m z_i^m}$$

where $\mathcal{N}(r^m\mid\mu_i,\sigma_i^2)$ is used as shorthand notation for the voxel response distribution in Equation A2 (note that the superscript *m* is used to denote individual samples, and does not denote exponentiation).

Taking the log of the individual parameters moves the variables z_i^m down from the exponent. Upon taking the expectation with respect to *z*, the formula becomes:

$$l_{EM}(\Theta)=\sum_{m}^{M}\sum_{i}^{K}\left\langle z_i^m\right\rangle\left(\log\pi_i+\log\mathcal{N}(r^m\mid\mu_i,\sigma_i^2)+\sum_{j}^{L}\log\gamma_{ij}^{c_j^m}\right)\qquad\text{(A3)}$$

Because z_i^m is an indicator variable, its expected value is given by simply evaluating its posterior distribution at z_i^m. Thus, we introduce the following notation:

$$\tau_i^m\coloneqq\left\langle z_i^m\right\rangle=p(z_i^m\mid r^m,c^m,\Theta)=\frac{\pi_i\gamma_{ij(m)}\mathcal{N}(r^m\mid\mu_i,\sigma_i^2)}{\sum_{k}^{K}\pi_k\gamma_{kj(m)}\mathcal{N}(r^m\mid\mu_k,\sigma_k^2)}$$

where $j(m)\in[1,L]$ picks out the semantic category of the m^{th} sample image. Substituting with this notation in Equation A3 gives the final simplified expression for the expected

log-likelihood:

$$l_{EM}(\Theta) = \sum_{m}^{M} \sum_{i}^{K} \tau_i^m \left(\log \pi_i + \log \mathcal{N}(r^m \mid \mu_i, \sigma_i^2) + \sum_{j}^{L} \log \gamma_{ij}^{c_j^m} \right)$$

To fit the semantic encoding model, the expected log-likelihood function must be maximized with respect to the free parameters Θ. The EM procedure prescribes the following steps:

1) initialize the parameters Θ
2) calculate the τ_i^m's
3) maximize $l_{EM}(\Theta)$ w.r.t. Θ (this is an elementary optimization problem; general solutions are not provided here)
4) evaluate $l_{EM}(\Theta)$
5) If stopping criterion is met, halt. Otherwise, goto 2.

Parameters of the two multinomial distributions were initialized to give uniform probabilities to each category *c* and each state of the latent variable *z*. The neural response distributions are initialized with mean 0 and standard deviation 1. The procedure halted when the absolute difference between all of the τ_i^m's on two successive iterations fell below a small threshold value.

APPENDIX 2: Greedy search algorithm for producing reconstructions using the flat prior or sparse Gabor prior.

Let $\mathbf{s}_t$ be the current best reconstruction at the t^{th} iteration of the algorithm. The steps for the algorithm are as follows:

1) Initialize a seed reconstruction by setting all pixel-values to 0: $\mathbf{s}_0 = \mathbf{0}$.
2) Chose 256 pixels from the current reconstruction, $\mathbf{s}_t$, spaced at regular intervals Δ on a 16x16 grid. Let $\mathbf{s}_t^+ = (s_i^+, s_{i+\Delta}^+, \ldots, s_{i+256\Delta}^+)_t$ denote the chosen pixels. Let $\mathbf{s}_t^-$ denote all remaining pixels.
3) Clamping all other pixels at their current value (*e.g.*, the value of the pixels in $\mathbf{s}_t$,), evaluate $p(\mathbf{s}|\mathbf{r})$ on the current image with s_i^+ set to each of its 10 permissible values (evaluation requires up-sampling $\mathbf{s}_t$ to 128x128 pixels). Retain $x_i^{\max}$, the value of s_i^+ resulting in the highest $p(\mathbf{s}|\mathbf{r})$ (referred to as the "best" value for pixel s_i^+). Repeat step 3 for all other pixels in $\mathbf{s}_t^+$.
4) Update current image by setting all pixels in $\mathbf{s}_t^+$ to their best value:

$$\mathbf{s}_{t+1} = \mathbf{s}_t^- \bigcup (s_i^+ = x_i^{\max})_t$$

5) Evaluate $p(\mathbf{s}_{t+1}|\mathbf{r})$
6) If $p(\mathbf{s}_{t+1}|\mathbf{r}) > p(\mathbf{s}_t|\mathbf{r}) + b$, ($b$ = 4 bits), translate 16x16 sampling grid by 1 pixel, and goto 2. Otherwise, halt.

This algorithm is computationally intensive. Therefore, we worked with down-sampled 64x64 images (note that the dimensions of the inputs to the structural encoding model are 128x128), and partitioned the range of values (determined by the maximum and minimum of values over a database of 10,000 images) permissible for each pixel into 10 uniform bins.

APPENDIX 3: Estimation of parameters of the sparse Gabor prior

All notation in this appendix is introduced in the main text (see subsection of **Experimental Procedures** titled **Reconstructions using the structural encoding model and a sparse Gabor prior**). The Gabor-wavelet basis, *G*, used to construct the sparse Gabor prior was distinct from the basis, *W*, used for the structural encoding model. *G* had 5 different scales, 6 different orientations, 2 cycles per standard deviation, and both even and odd Gabors. The centers of the largest Gabors were aligned with the center of the image, while the smaller ones were repeated at various positions throughout the image.

Once the Gabor basis is fixed, the free parameters of the distributions $p(\mathbf{s}|\mathbf{a})$ and $p(\mathbf{a})$ must be estimated. We estimated each of these parameters using $M = 11962$ randomly selected natural images. The mean of the images was calculated as:

$$\boldsymbol{\mu}_s = \sum_i^M \frac{\mathbf{s}_i}{M}$$

The unwhitening matrix U is calculated by taking a Cholesky decomposition of the image covariance matrix:

$$U = Chol\left(\frac{\sum_i^M (\mathbf{s}_i - \boldsymbol{\mu}_s)(\mathbf{s}_i - \boldsymbol{\mu}_s)^T}{M} \right)$$

Finally, the mean and variance of the distribution over Gabor activations, $p(\mathbf{a})$, are calculated by transforming the images into the space of Gabor coefficients, and then taking the maximum likelihood values of the parameters. For the mean, this procedure gives:

$$\mathbf{u}_a = \left[u_1, \ldots u_g\right] = \sum_i^M \frac{GU^{-1}\left(\mathbf{s}_i - \boldsymbol{\mu}_s\right)}{M}$$

For the variance of parameters, the procedure gives;

$$\left[\beta_1, \ldots \beta_g\right] = \sum_i^M \frac{\left|GU^{-1}\left(\mathbf{s}_i - \boldsymbol{\mu}_s\right) - \mathbf{u}_a\right|}{M}$$

where g is the total number of wavelets in G.

Note that the Gabor transform matrix G is not square, and the unwhitening matrix U is ill-conditioned. All matrix inversions in the above formulas denote pseudo-inverses. This is the equivalent of assuming all dimensions of the images that are not defined by our Gabor basis are set to 0.