Behavioral/Systems/Cognitive

Natural Scene Categories Revealed in Distributed Patterns of Activity in the Human Brain

Dirk B. Walther,[1] Eamon Caddigan,[1,2] Li Fei-Fei,[3]* and Diane M. Beck[1,2]*

[1]Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign, Urbana, Illinois, 61801-2325, [2]Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, Illinois 61820-6232, and [3]Computer Science Department, Stanford University, Stanford, California 94305-9025

Human subjects are extremely efficient at categorizing natural scenes, despite the fact that different classes of natural scenes often share similar image statistics. Thus far, however, it is unknown where and how complex natural scene categories are encoded and discriminated in the brain. We used functional magnetic resonance imaging (fMRI) and distributed pattern analysis to ask what regions of the brain can differentiate natural scene categories (such as forests vs mountains vs beaches). Using completely different exemplars of six natural scene categories for training and testing ensured that the classification algorithm was learning patterns associated with the category in general and not specific exemplars. We found that area V1, the parahippocampal place area (PPA), retrosplenial cortex (RSC), and lateral occipital complex (LOC) all contain information that distinguishes among natural scene categories. More importantly, correlations with human behavioral experiments suggest that the information present in the PPA, RSC, and LOC is likely to contribute to natural scene categorization by humans. Specifically, error patterns of predictions based on fMRI signals in these areas were significantly correlated with the behavioral errors of the subjects. Furthermore, both behavioral categorization performance and predictions from PPA exhibited a significant decrease in accuracy when scenes were presented up-down inverted. Together these results suggest that a network of regions, including the PPA, RSC, and LOC, contribute to the human ability to categorize natural scenes.

Introduction

Consider for a moment the range of images one might categorize as a picture of a "beach." Add to that the large number of categories we must master to function in familiar and new environments and perform critical visual tasks (Tversky and Hemenway, 1983). Although natural scene categorization is a difficult problem requiring subtle distinctions between heterogeneous sets of images, humans are remarkably proficient at it. They can recognize natural scenes with exposures as brief as 100 ms (Potter and Levy, 1969), with processing times as short as 150 ms (Thorpe et al., 1996; VanRullen and Thorpe, 2001), in the near-absence of attention (Li et al., 2002; Fei-Fei et al., 2005), and with little time to prepare for the categorization tasks (Walther and Fei-Fei, 2007). They can also access many details of natural scenes in a single glance (Fei-Fei et al., 2007).

What is the neural basis of this astonishing feat of visual processing? The difficulty in answering this question is due in part to limitations of conventional univariate functional magnetic resonance imaging (fMRI) analysis, in which each voxel is treated as an independent unit, and activity is typically averaged across voxels. Because of the complexity of the images, natural scene categories are likely to be encoded in patterns of activation that can only be detected using multivariate statistical techniques. In fact, multivoxel pattern analysis (MVPA) of fMRI activity has been used to decode representations that are distributed both within and across brain regions (Haxby et al., 2001; Carlson et al., 2003; Cox and Savoy, 2003), including representations that might exist at a smaller spatial scale than the size of functional voxels (Haynes and Rees, 2005; Kamitani and Tong, 2005). Such a technique should therefore be well suited for identifying regions that may contribute to natural scene categorization.

Although MVPA can potentially make predictions of natural scene categories from fMRI activity, it does not necessarily provide evidence that the brain uses this same information in performing the same categorization task. It is therefore central to our approach that we compare the predictions from the fMRI activity with behavioral data to identify those areas that do not simply contain scene category-specific information but are also the most likely to contribute to the categorization decisions made by humans.

Several areas along the visual processing pathway could potentially participate in the task of natural scene categorization. Conventional analysis of fMRI data has revealed the parahippocampal place area (PPA) and the retrosplenial cortex (RSC) to be more active for pictures of places (such as landscapes, houses, or rooms) than for faces or isolated objects (Aguirre et al., 1996; Epstein and Kanwisher, 1998; O'Craven and Kanwisher, 2000). It is entirely unknown, however, whether the information represented in the PPA and RSC is sufficient to distinguish among the

Received Jan. 30, 2009; revised May 21, 2009; accepted July 8, 2009.

This work is funded by National Institutes of Health Grant 1 R01 EY019429 (to L.F.-F., D.M.B., D.B.W.), a Beckman Postdoctoral Fellowship (to D.B.W.), a Microsoft Research New Faculty Fellowship (to L.F.-F.), and the Frank Moss Gift Fund (to L.F-F.). We thank Daniel Simons and Sabine Kastner for feedback on earlier versions of this manuscript.

*L.F.-F. and D.M.B. contributed equally to this work.

Correspondence should be addressed to Dr. Dirk B. Walther, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, 405 North Mathews Avenue, Urbana, IL 61801-2325. E-mail: walther@illinois.edu.

DOI:10.1523/JNEUROSCI.0559-09.2009

Copyright © 2009 Society for Neuroscience 0270-6474/09/2910573-09$15.00/0

different scene categories. We use MVPA to ask whether these as well as other brain regions contain information that not only discriminates among scene categories but also correlate with human behavioral performance on a natural scene categorization task.

Materials and Methods

Subjects. Five subjects (two females; age: 21–38 years; including three of the authors) participated in the study, which was approved by the Institutional Review Board of the University of Illinois. All participants were in good health with no past history of psychiatric or neurological diseases and gave their written informed consent. Subjects had normal or corrected-to-normal vision.

Visual stimuli and experimental design. Test stimuli consisted of 120 color images from each of six categories: beaches, buildings, forests, highways, industry, and mountains (see Fig. 1 for examples) downloaded from the Internet. Photographs were chosen to capture the high variability within each scene category. The dataset is available for download at vision.stanford.edu/fmriscenes/resources.html.

In the behavioral experiment, 360 of those images (60 from each category) were presented at a resolution of 800 by 600 pixels (23° × 18° of visual angle), centered over a 50% gray background, on a CRT monitor running at a resolution of 1024 by 768 pixels at 89 Hz. Stimulus presentation and response recording were controlled using the open source Vision Egg package. On each trial, a fixation cross was displayed for 500 ms, followed by the brief (11–45 ms; determined separately for each subject) presentation of an image, which was replaced by a perceptual mask (see Fig. 1, bottom row) for 500 ms, and finally a blank screen appeared for 2000 ms. Subjects performed six-alternative forced-choice categorization of the images by pressing one of six buttons on the keyboard. The mapping of categories to buttons was counterbalanced across subjects, and subjects received training on that mapping before the staircasing part of the experiment.

In the main experiment, trials were grouped into 18 blocks of 20 images, and alternating blocks consisted of upright or inverted (reflected about the horizontal axis) images. Subjects viewed each image once in an upright and once in an inverted block. The duration of image presentation was adjusted for each subject with a staircasing procedure using the Quest algorithm (King-Smith et al., 1994) on a separate set of 120 images, which were all displayed upright. Presentation duration was staircased for each subject individually to a classification accuracy of 65%. Staircasing was terminated when the SD of the display times over a block was less than the refresh period of the monitor (1/89 s). An 800 Hz, 100 ms tone alerted participants to incorrect responses during staircasing. No feedback was given during the main experiment.

In the fMRI experiment, images (625 × 469 pixels; subtending 23 × 18° of visual angle) were presented in the center of the display, using MR-compatible liquid crystal display goggles (Resonance Technologies) operating at a resolution of 800 × 600 pixels at 60 Hz. Stimuli were arranged into blocks of 10 images from the same natural scene category. A fixation cross was presented throughout each block, and subjects were instructed to maintain fixation. Each image was displayed for 1.6 s. A run was composed of 6 blocks, one for each natural scene category, interleaved with 12 s fixation periods to allow for the hemodynamic response to return to baseline levels. A session contained 12 such runs, and the order of categories was randomized across blocks. Images were presented upright or inverted in alternating runs, with each inverted run preserving the image and category order used in the preceding upright run. Each subject performed two sessions (12 runs each), on separate days, of passive viewing with separate sets of images. In the fMRI experiment, subjects saw each image once upright and once inverted. The behavioral experiments were performed at least 6 weeks after the fMRI experiment to minimize effects of image repetition and familiarity between the two experiments.

MRI acquisition and preprocessing. Imaging data were acquired with a 3 tesla Siemens Allegra Scanner equipped for echo planar imaging. A gradient echo, echo-planar sequence was used to obtain functional images [volume repetition time (TR), 2 s; echo time (TE), 30 ms; flip angle, 90°; matrix, 64 × 64 voxels; FOV, 22 cm; 34 axial 3 mm slices with 1 mm gap; in-plane resolution, 3.44 × 3.44 mm). We collected a high-resolution (1.25 × 1.25 × 1.25 mm voxels) structural scan (MPRAGE; TR, 2 s; TE, 2.22 ms, flip angle, 8°) in each scanning session to assist in registering our echo planar imaging images across sessions.

Pattern analysis. Functional data were motion corrected and normalized to the temporal mean of each run using the AFNI software suite (Cox, 1996). No other smoothing or normalization steps were performed. The 1152 brain volumes acquired during the viewing of the scene images (2 sessions × 12 runs × 6 blocks × 10 images × 1.6 s presentation time/2 s TR) were extracted from the time series with a time lag of 4 s to approximate the lag in the hemodynamic response. A support vector machine (SVM) classifier (linear kernel, using LIBSVM, Chang and Lin, 2001; http://www.csie.ntu.edu.tw/~cjlin/libsvm/) was trained to assign the correct scene category labels to the voxel activation patterns of the individual brain volumes in one of five regions of interest (ROIs) (see below) from 11 of the 12 upright runs. The data from the left-out upright run were presented to the trained classifier, which generated predictions of the class labels for each brain acquisition. To resolve disagreements in the labeling of the eight fMRI volumes coming from the same block, we applied a standard majority voting scheme. The label predicted most frequently among the eight volumes was assumed as the label for the entire block. Ties were broken by adopting the label with the highest decision value in the SVM classifier. In 12 repetitions of this procedure, each of the 12 upright runs was left out once (leave-one-run-out cross validation, LORO). Accuracy of the decoding was computed as the fraction of labels that were correctly predicted over all 12 repetitions. One-tailed *t* tests were used to determine whether decoding accuracy in the five subjects was significantly above chance level of 1/6.

The pattern analysis procedure was slightly modified for determining the effects of scene inversion. In addition to the procedure outlined above, the classifier trained on 11 upright runs was also tested on the inverted run corresponding to the left-out upright run, and the difference in decoding accuracy was determined. Furthermore, to control for the effect of having different image geometries in training and testing, a separate classifier was trained on 11 of the 12 inverted runs and tested on the left-out inverted run. Again, the difference in decoding accuracy with the original classifier was computed. Significance of differences in accuracy was established using paired, one-tailed *t* tests over the five subjects.

ROIs. We identified five regions of interest: V1, PPA, RSC, the fusiform face area (FFA), and the lateral occipital complex (LOC). Area V1 was identified as a region spanning the calcarine fissure delineated by a representation of the vertical meridian, which forms the border of V1 and V2. The vertical meridian was determined using a rotating hemifield procedure (Schneider et al., 2004) in a separate scanning session. The four other ROIs were identified from linear contrasts in separate localizer scans as sets of contiguous voxels that responded preferentially to faces, scenes, and objects. Briefly, localizer scans consisted of blocks of face, object, scrambled object, landscape, and cityscape images. Each block consisted of 20 images presented for 450 ms each with a 330 ms interstimulus interval. Each of the five types of stimuli was presented four times during a run, with 12 s fixation periods after two or three blocks. Participants completed two runs, performing a one-back task during the localizer by pressing a button every time an image was repeated. The PPA and the RSC were identified in both hemispheres by a (cityscapes and landscapes) > (objects and faces) contrast. The FFA was identified by a faces > (objects, cityscapes, and landscapes) contrast, and the LOC by an objects > scrambled objects contrast. For all localizer contrasts, a maximum threshold of $p < 2 \times 10^{-3}$ (uncorrected) was applied. Stricter thresholds were used when necessary to break clusters. There was no overlap between any of the ROIs, and all ROI voxels were used for the pattern analysis without any further voxel selection.

Whole-brain analysis. To explore brain regions outside of our predefined ROIs, we performed a searchlight analysis (Kriegeskorte et al., 2006) of the whole brain. For this purpose, we defined a spherical template (diameter of 5 voxels), which contained 81 voxels, making it similar in size to the ROIs obtained from localizer scans. We centered the template on each voxel in turn and performed the same LORO cross-validation procedure as above on the 81 voxels in the template. Voxels that fell outside the brain were omitted from the analysis. The decoding accuracy for each

Figure 1. Example images of the six natural scene categories used in this study (from top to bottom): beaches, buildings, forests, highways, industry, and mountains, as well as four randomly selected examples of the perceptual masks used in the behavioral experiments (bottom row).

template location was stored at the center voxel. By repeating this process for every voxel, we obtained a brain mask of decoding accuracies for each subject. For group analysis, we registered the decoding accuracy maps into Montreal Neurological Institute (MNI) space using FLIRT (Jenkinson et al., 2002) and smoothed them with a Gaussian kernel (full-width at half-maximal = 8 mm). We tested whether decoding accuracy was above chance (1/6) with a voxelwise t test, thresholded at $p < 0.01$ (uncorrected), and then corrected for multiple comparisons at the cluster-level ($p < 0.05$). The minimum cluster size of 19 voxels, estimated with AlphaSim from the AFNI toolbox (Cox, 1996), accounted for voxel dependencies present in the data due to the nature of the hemodynamic signal or introduced by the smoothing process. The resulting regions were transformed back into individual subject space, in which their overlap with the individual ROIs was computed as percentage of ROI voxels that are part of the searchlight regions.

Image analysis. To assess the physical similarity of the images used in our experiments, we subsampled the images to a resolution of 320 × 240 pixels and computed the pixel-wise correlations of each pair of RGB images. We arrived at a correlation matrix for image categories by averaging over the correlation coefficients for all pairs of images representing any given pair of categories. We then computed the correlation between the off-diagonal elements of this correlation matrix and the off-diagonal elements of the confusion matrices derived from the decoding analysis in the ROIs.

Results

Subjects were scanned while passively viewing real-world photographs of six natural scene categories (beaches, buildings, forests, highways, industry, and mountains) (Fig. 1). To determine whether it was possible to make predictions of the natural scene category from a particular ROI, we used fMRI voxels from functionally defined ROIs as the input to a "decoder" for natural scene categories. The decoder was constructed by training a SVM classification algorithm to assign the correct category labels to the fMRI data in an LORO cross-validation procedure (see Materials and Methods for details).

We used separate localizer and retinotopic mapping scans (see Materials and Methods) to identify five ROIs: the PPA (90 ± 23 voxels), the RSC (60 ± 19 voxels), the LOC (100 ± 74 voxels), the FFA (67 ± 64 voxels), and the primary visual cortex (V1; 447 ± 121 voxels). However, unlike traditional univariate methods of analysis, which ask whether the average activity level in a ROI differs as a function of category, we were interested in category-dependent differences of patterns of responses within these areas.

The LOC, which has been shown to be sensitive to a variety of objects (Malach et al., 1995), was included in the analysis, because natural scenes can be thought of as compositions of objects. Although originally identified as an area specifically selective for faces (Kanwisher et al., 1997), the FFA has also been reported to be activated for other kinds of objects (Grill-Spector et al., 1999; Gauthier et al., 2000; Tarr and Gauthier, 2000; Haxby et al., 2001). We therefore include it as yet another ROI associated with processing complex visual information and hence potentially involved in natural scene classification.

Scene categories may also differ in the global distribution of spatial frequencies in the image (Oliva and Torralba, 2001) or distributed local feature information such as oriented textures (Fei-Fei and Perona, 2005; Bosch et al., 2006). Thus, with its selectivity for specific orientations and spatial frequencies (Hubel and Wiesel, 1962; De Valois and De Valois, 1980), we selected V1 as another region that might participate in the representation of scene categories.

Decoding accuracy

Applying the LORO cross-validation procedure to the PPA voxels in the upright runs resulted in a decoding accuracy of 31% (i.e., rate of correctly predicting the scene categories shown to subjects from the pattern of voxel activity), which is significantly above the chance level of 1/6 ($t_5 = 4.17$; $p = 0.0070$). Decoding from RSC, although less accurate (27%) than decoding from PPA, was still significantly better than chance ($t_5 = 3.24$, $p = 0.016$). In other words, we show that information about the category being viewed is present in the pattern of activity across voxels in both PPA and RSC (Table 1, first column).

We were also able to decode scene category from activity in the object-sensitive area LOC with 24% accuracy ($t_5 = 2.27$; $p = 0.043$). One possible explanation is that LOC may contribute to the categorization of natural scenes by detecting objects that are consistent with a particular type of scene. Such objects have been shown to improve fast recognition of scenes (Davenport and Potter, 2004), and LOC was also shown to be involved in the processing of visual context (Bar and Aminoff, 2003).

Although PPA, RSC, and, to some extent, LOC were included as likely candidates for contributing to natural scene categorization, we also asked whether some information regarding scene categories might be contained within FFA. However, decoding accuracy in the FFA was not significantly above chance (22%; $t_5 = 1.73$), consistent with the claim that it primarily encodes faces.

Decoding accuracy from area V1 was significantly above chance at 26% ($t_5 = 2.64$; $p = 0.029$), leaving open the possibility that scene categories could be distinguished on the basis of low-level features computed in V1. This finding is also consistent with recent data showing that the identity of individual natural images can be successfully decoded from V1 voxel activity (Kay et al., 2008). We note, however, that decoding the category of an image is a very different process from decoding the identity of an image (as a specific image). Images belonging to the same scene category (e.g., a beach) can display distinctively different color, texture, illumination, spatial layout, individual object components, and so forth. Such geometric and photometric variability requires a more abstract representation of scene categories; that is, the decoder must extract the information that distinguishes a beach from a forest across multiple and variable instances of beaches and forests. In short, although individual instances of a category may be easily distinguishable in V1, it is more remarkable that patterns of activity in V1 can reliably predict the general category.

To exclude the possibility that differences in decoding accuracy are due to different numbers of voxels in the ROIs, we repeated the analysis with 20 randomly drawn voxels from each ROI. The mean decoding accuracy over 20 independent random draws was significantly above chance in all ROIs, except for the FFA, confirming the pattern of results obtained from the full ROIs.

Table 1. Summary of main results

ROI	Decoding accuracy	Error correlation	Image similarity correlation	Inversion effect
V1	26%*	0.21	0.46**	0%
FFA	22%	0.10	0.03	2%
LOC	24%*	0.42*	−0.22	3%
RSC	27%*	0.34†	−0.24	2%
PPA	31%**	0.57**	−0.07	7%*

Decoding accuracy is measured in percentage of blocks predicted correctly, and significance is assessed relative to chance (17%). Error correlation establishes a correlation between misclassifications (off-diagonal entries in the confusion matrices) (Figs. 2, 3) between decoding from ROIs and human behavior. Image similarity correlation correlates the image similarities matrix with the confusion matrix from fMRI decoding. The inversion effect is defined as the difference in accuracy of a decoder trained and tested with upright versus trained and tested with inverted scene presentations. PPA shows significant effects in all analyses except for the image similarity correlation. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{\dagger}p = 0.069$.

Viewed image category	Subjects' response: beaches	buildings	forests	highways	industry	mountains	Decoder prediction (PPA): beaches	buildings	forests	highways	industry	mountains
beaches	**0.70**	0.02	0.03	0.10	0.03	0.04	**0.28**	0.08	0.10	0.20	0.10	0.23
buildings	0.02	**0.88**	0.02	0.00	0.03	0.01	0.12	**0.27**	0.07	0.27	0.15	0.13
forests	0.06	0.04	**0.78**	0.01	0.02	0.05	0.15	0.12	**0.33**	0.13	0.15	0.12
highways	0.03	0.03	0.01	**0.82**	0.02	0.03	0.18	0.17	0.10	**0.40**	0.10	0.05
industry	0.02	0.12	0.04	0.02	**0.71**	0.02	0.08	0.30	0.10	0.12	**0.25**	0.15
mountains	0.08	0.02	0.06	0.02	0.02	**0.74**	0.18	0.08	0.23	0.07	0.08	**0.35**

Figure 2. Confusion matrices for behavioral responses (left) and decoder predictions of fMRI activity in PPA (right). The rows of this matrix indicate the scene categories presented to the subjects (ground truth), and the columns the subjects' behavioral response (left) and the predictions by the decoder (right). An ideal confusion matrix would have 1 everywhere on the diagonal (correct classifications) and 0 in the off-diagonal entries (errors). Frequent confusions are highlighted in yellow.

Correlation of error patterns with behavior

Having found that information relevant to scene categories is contained within PPA, RSC, LOC, and V1 voxels, we can ask whether human subjects use that same information for scene categorization. To address this question, we compared decoder performance with that of human subjects categorizing the same scenes in a separate experiment. Subjects were asked to indicate the category of briefly presented (11–45 ms, followed by a perceptual mask) scenes by pressing one of six buttons. Subjects identified the correct scene categories in this fast six alternative forced-choice behavioral task with 77% accuracy, which is significantly above the chance level of 1/6 ($t_5 = 20.32$; $p = 0.00002$). Moreover, categorization accuracy was significantly above chance ($t_5 \geq 8.0$; $p < 0.001$) for each of the six categories.

Although presentation times in the behavioral experiment were considerably shorter than in the fMRI experiment, if both experiments rely on the same category-specific signal in the brain, we should see a correspondence between the errors that the humans make and the errors made by the decoder. The weaker or less distinct this signal is, the more errors the humans should make and the harder it will be to read it out with fMRI, leading to more decoding mistakes.

Not surprisingly, categorization accuracy by human subjects is much higher than from fMRI decoding, presumably because of the limited spatial resolution of the fMRI signal and the limited number of voxels used. In contrast, the behavioral decision is likely to be affected by the firing patterns of neurons below the resolution of our decoder, and these neural signals are presumably not restricted to local ROIs. But humans were not perfect in their behavior, and their pattern of errors can be captured in a confusion matrix (Fig. 2). The rows of this matrix indicate the image category presented in the experiment. The cells in each row contain the proportion of trials in which subjects responded with the category indicated by the column. The diagonal entries in this matrix are correct responses, and the off-diagonal entries are erroneous responses. The errors in the fMRI analysis can be summarized by a similar confusion matrix with the scene categories presented to the subjects in the rows and the categories predicted by the fMRI decoder in the columns.

The patterns of behavioral as well as decoding errors offer us an opportunity to correlate the MVPA results with human behavior. A comparison of the confusion matrices in Figure 2 shows some interesting similarities. Specifically, the confusion matrix for analyzing PPA activity (Fig. 2, right) indicates a high number of misclassifications of industry and buildings, buildings and

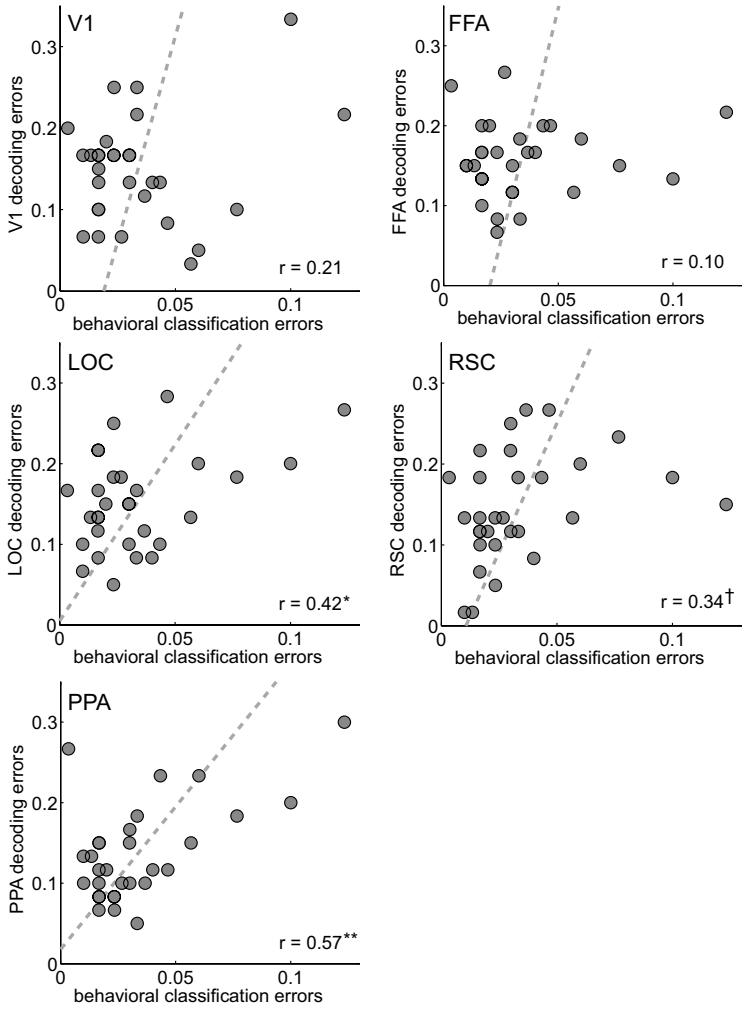

Figure 3. Correlations of error patterns. The 30 off-diagonal entries (errors) in the fMRI-decoding confusion matrices (Fig. 2) are plotted over the errors in the behavioral experiment. The dashed lines show least-squares fits of linear relationships. Agreement of the error patterns was assessed with the Pearson product–moment correlation coefficient. High correlation was found for PPA, RSC, and LOC. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{\dagger}p = 0.069$.

highways, mountains and forests, beaches and mountains, and beaches and highways. Some of the confusions, such as between buildings and highways or between beaches and highways, may reflect similarities in the image structure and low-level image statistics (large, horizontally oriented areas, sky-quality textures). Other confusions, for example, between mountains and forests or between industry and buildings are particularly understandable given that these pairs of categories not only share low-level image statistics but also substantial semantic overlap in the form of wooded mountains and industrial buildings, respectively. Importantly, we observe some of the same confusions in the behavioral results (Fig. 2, left), albeit with lower values because of higher overall decoding accuracy.

We can quantify the similarity of the decoding and behavioral error patterns by computing the pairwise correlation of the errors (off-diagonal elements of the confusion matrices) from the fMRI results for each ROI with the behavioral results (Fig. 3). The Pearson correlation coefficient was significant for PPA ($r = 0.57$; $p = 0.0011$) and LOC ($r = 0.42$; $p = 0.021$), marginally significant for RSC ($r = 0.34$; $p = 0.069$), but not significant for V1 ($r = 0.21$; $p = 0.21$) and FFA ($r = 0.10$; $p = 0.60$). That is to say, the error patterns of the decoder using voxels from the PPA, LOC, and, to a lesser extent, RSC were similar to the error patterns of our human subjects in the behavioral paradigm (Table 1, second column).

We would like to note that it is rather remarkable that we see a significant correlation between behavior and activity in our local regions at all. As noted before, the human is almost certainly arriving at his or her decision using neural activity patterns distributed across the entire brain, many of which will be below the resolution of the decoder. To our knowledge, this is the first evidence of a positive correlation between the complex error patterns of humans categorizing six classes of natural scenes and those of a decoder classifying fMRI activity, although similar methods have been applied to other visual tasks (Aguirre, 2007; Williams et al., 2007; Haushofer et al., 2008).

Having established the compatibility of the scene category-specific neural signal in our later ROIs (PPA, RSC, and LOC) with human behavior, we can also test to what extent both the neural signal and human behavior are accounted for by the physical similarity of the stimuli. To investigate this issue, we computed the pixel-wise correlations of all image pairs and sorted and averaged them according to their category pair (e.g., beaches vs forests). As a result, we obtain a correlation matrix for image category similarity. Interestingly, the off-diagonal elements of this image similarity matrix are not correlated with the behavioral errors ($r = -0.11$; $p = 0.55$), suggesting that behavioral image categorization is not primarily driven by physical similarity. Comparisons of image similarity with the fMRI error patterns did, however, show a highly significant correlation with decoding from V1 but not with any of the later visual areas (Table 1, third column). Together with the fact that error patterns in V1 did not correlate with behavioral confusions, a picture of natural scene categorization emerges in which V1 preserves the physical similarity relations among the images, whereas the later areas, in particular LOC, RSC, and PPA, contain information more compatible with human categorization behavior.

Scene inversion

To strengthen our findings that PPA, RSC, and LOC may contribute to natural scene categorization, we performed a second experiment to further substantiate the relationship between fMRI decoding and the behavior of the subjects: the experiment and procedures were identical to those just described except that the same images were presented inverted (i.e., mirrored across the horizontal axis). The images were presented in alternating upright and inverted blocks, in both the imaging and behavioral experiments. Our expectation was that subjects would find inverted images more difficult to categorize than upright images. We would then ask whether we see a similar decrease in decoding accuracy for inverted images in our ROIs. As predicted, subjects

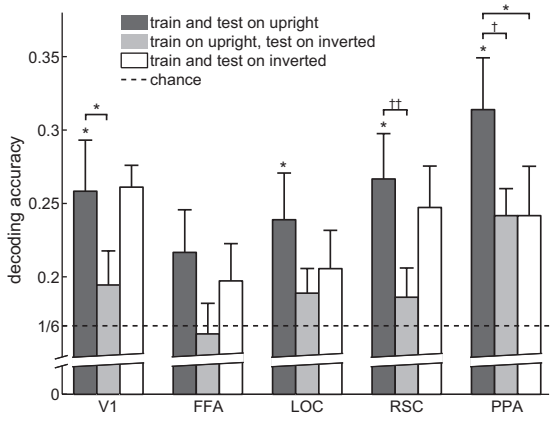

Figure 4. Effects of scene inversion. A decoder that was trained on fMRI activity from upright scenes showed significantly lower accuracy when decoding fMRI activity from inverted (light gray bars) than upright (dark gray bars) scenes in V1, PPA, and, marginally, RSC. When comparing the upright decoder tested on activity from upright scenes (dark gray bars) with a decoder that was trained and tested on fMRI activity from inverted scenes (white bars), only PPA showed a significant decrease in decoding accuracy. Error bars are SEM over five subjects. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{\dagger}p = 0.052$, and $^{\dagger\dagger}p = 0.072$.

were significantly less accurate at identifying natural scene categories ($t_5 = 6.07$; $p = 0.0019$) for inverted than for upright scenes.

We probed for an inversion effect in the fMRI analysis in two ways. The first possibility is to test the decoder, which was trained on all but one of the upright runs, on the inverted run corresponding to the left-out upright run in the LORO cross-validation procedure. In other words, we asked how well the decoder trained on upright runs would transfer to inverted runs (Fig. 4). We found decreased decoding accuracy for inverted scenes in PPA ($t_5 = 2.10$; $p = 0.052$), V1 ($t_5 = 2.23$; $p = 0.045$), and, marginally, in RSC ($t_5 = 1.81$; $p = 0.072$), but not in the FFA ($t_5 = 1.33$; $p = 0.13$) and LOC ($t_5 = 1.45$; $p = 0.11$). However, this result can be interpreted as reflecting the fact that the global image statistics between the training (upright) and testing (inverted) stimuli differ (e.g., sky in the top half versus sky in the bottom half), rather than because of inversion per se.

To control for this confound, we compared the performance of a decoder trained and tested on upright runs with a decoder trained and tested on inverted runs, thus equating global image statistics across training and testing (Fig. 4). If the quality of the category representation in our ROIs suffered as a result of inversion, then the decoder should do less well when it only had access to inverted images than when it only had access to upright images. Again, we found a significant drop in decoding accuracy for inverted scenes compared with upright scenes in the PPA ($t_5 = 2.33$; $p = 0.040$), but not in V1 ($t_5 < 1$), FFA ($t_5 < 1$), LOC ($t_5 = 1.24$; $p = 0.14$), or RSC ($t_5 < 1$). It should be noted that finding a significant decrease of decoding accuracy for inverted scenes is to some extent contingent on finding high decoding accuracy for upright scenes. It is therefore conceivable that the inversion effect for LOC, for instance, might have reached significance had decoding accuracy been higher to begin with.

The different patterns of results for the two inversion analyses underscore the importance of the second analysis. In area V1, for example, we found a significant drop in accuracy when training the decoder on the fMRI activity acquired during the presentation of upright scenes and testing on inverted scenes. This result is consistent with the view of V1 as a retinotopic area whose neural representation reflects the orientation of the visual input, leading to poor transfer from training the decoder on upright to testing it on inverted images. There was no difference in accuracy, however, when training and testing on inverted scenes, indicating that the representation in V1 is indifferent to the correct orientation of the scenes, leading to good decoder accuracy as long as the decoder was trained and tested on the same image orientation.

The only region to continue to show a significant inversion effect when test and training were better matched was the PPA (Table 1, fourth column). Only the PPA showed a similar sensitivity to scene inversion as the human subjects, once again implicating the PPA in the human ability to categorize natural scenes. We note that the decrease in decoding accuracy for inverted scenes is not because of differences in the general activity level of voxels in the PPA. Unlike a previous study (Epstein et al., 2006), we did not find a significant difference in the mean voxel activation between upright and inverted scenes in any of the ROIs ($t_5 < 1$). Instead, this effect was only apparent in the pattern of activity in PPA.

It could be argued that the inversion effect might be due in part to the repetition of stimuli, since the blocks with inverted images were always shown after the blocks with upright images. Such an adaptation effect, however, should be observable in a higher mean activity level for upright than for inverted scenes within our ROIs. As mentioned above, we did not find any such difference in mean activity level.

Whole-brain analysis

To explore brain regions that might be involved in natural scene categorization beyond our predefined ROIs, we performed a whole-brain analysis. To this end, we defined a spherical "searchlight" region similar in size to the ROIs (81 voxels) that we positioned at all possible locations in the brain (Kriegeskorte et al., 2006). Within each local region, we performed the same LORO cross-validation procedure as for the ROIs and stored the decoding accuracies in a brain map. Group analysis over the decoding accuracy maps of the individual subjects yielded a map with mean decoding accuracies as well as a map of t values for testing whether decoding accuracies are above chance. After thresholding at $p < 0.01$, we performed a cluster-level correction for multiple comparisons, resulting in a minimum cluster size of 19 voxels (obtained using α probability simulation). Figure 5 shows the resulting brain maps in yellow.

Decoding accuracy was significantly above chance in a large cluster of voxels in the right ventral visual cortex, including the right parahippocampal and fusiform cortex, and extending laterally and posteriorly from there. The left parahippocampus was also active, but to a lesser extent than the right parahippocampal activity. We found other visually active areas showing high decoding accuracy in the left occipital cortex, in the right precuneus, and in the left inferior temporal gyrus. Interestingly, we also found activity in frontal regions, most prominently in both inferior frontal gyri, encompassing Brodmann areas 44 and 45. This region is also known as Broca's area, which is credited with language processing (Geschwind, 1970; Grodzinsky and Santi, 2008). Although subjects were not instructed to name the natural scene category that they were viewing, subvocalization of the category name may be automatic and unavoidable, possibly explaining activity in Broca's area.

As a quantification of the agreement of the searchlight analysis with the ROI-based analysis we computed the percentage of voxels in each ROI that overlap with the searchlight analysis. Since location and shape of ROIs tend to differ between subjects, we performed this analysis for each subject individually. Table 2 shows the summary statistics over five subjects, and Figure 5

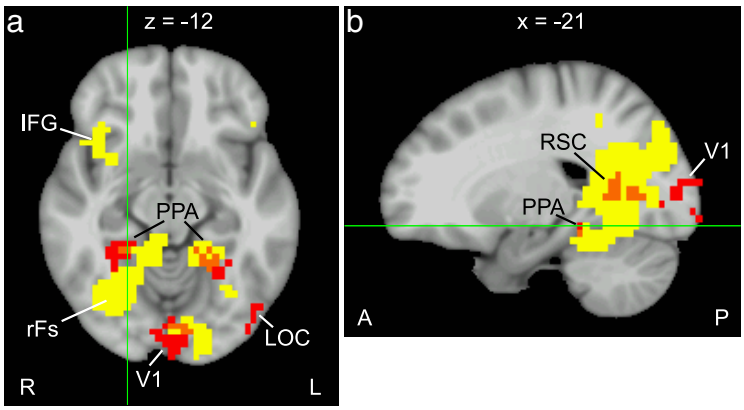

Figure 5. Axial (***a***) and sagittal (***b***) view of the whole-brain searchlight analysis. Areas in yellow show decoding accuracy significantly above chance ($p < 0.01$, corrected at the cluster level). Localizer-based ROIs from a single subject are marked in red, and overlap between the searchlight result and ROIs is shown in orange. In addition to visual areas around the PPA and the right fusiform gyrus (rFs), the inferior frontal gyrus (IFG) showed decoding accuracy significantly above chance.

Table 2. Percentage overlap of ROI voxels with searchlight activity

ROI	Mean	SD
V1	12.3%	3.7%
FFA	6.5%	14.6%
LOC	3.4%	3.3%
RSC	60.0%	13.8%
PPA	39.7%	13.1%

shows the overlap for one subject. Not surprisingly, the ROIs with the highest decoding accuracies in the ROI-based analysis (PPA and RSC) also show the largest amount of overlap with the searchlight analysis. Variability in the location and extent of the ROIs between subjects leads to large variations in the overlap with the searchlight analysis. Furthermore, a searchlight region located near the edge of an ROI includes some voxels belonging to the ROI but also many voxels from outside the ROI. Thus, we would not expect perfect agreement between the results of the searchlight group analysis and ROIs determined individually for each subject in separate localizer sessions.

Discussion

We used multivoxel pattern recognition to move beyond the question of whether a brain region is sensitive to images of natural scenes and ask whether these regions contain information that can discriminate between different categories of natural scenes. We found that activity in V1, PPA, RSC, and LOC allowed us to predict the categories of previously unseen images significantly above chance (Table 1). This above-chance performance is even more remarkable given the great variability in the appearance of exemplars from natural scene categories (e.g., Fig. 1). Indeed, despite four decades of research into the features that may distinguish natural scene categories (Potter and Levy, 1969; Biederman, 1972; Delorme et al., 2000; Oliva and Torralba, 2001), we still know very little about the critical features that the human uses to make such categorizations, presumably because of the high dimensionality and complex nature of the feature space.

One of the concerns with pattern recognition techniques is that there is no guarantee that the algorithm is using the same information that is used by the humans. Knowing that the information is present in the brain, however, gets us one step closer to this goal. We still do not know, of course, if a particular pattern is ultimately contributing to the human subject's ability to categorize scenes. We take a further step to guard against this concern by correlating decoding performance with human behavior. The response patterns of PPA, and to a slightly lesser extent RSC and LOC, agreed with behavioral scene categorization by human subjects in two ways (Table 1). First, the erroneous categorizations made by the decoder in these regions were correlated with the errors made by subjects when categorizing briefly presented scenes. Second, in parallel with the behavioral results, the accuracy of decoding scenes from PPA was lower for inverted than for upright scenes. RSC and LOC, however, failed to show such an inversion effect.

In contrast, physical similarity of the images was strongly correlated with the error pattern in V1 but not with other areas or with behavior. This implies that V1 preserves the similarity relations between images, but that representation of scenes in higher visual areas, namely PPA, RSC, and LOC, more closely tracks human behavior rather than physical similarity.

These results are not only compatible with the preference of PPA and RSC for natural scenes as opposed to faces or other objects (Aguirre et al., 1996; Epstein and Kanwisher, 1998; O'Craven and Kanwisher, 2000; Yi et al., 2006), but they further suggest that PPA and RSC play a role in categorizing those scenes. At first glance, this may seem at odds with the data of Epstein and Higgins (2007), which showed less activity in PPA when subjects categorized natural scenes (e.g., this is a parking lot) than when they identified them (e.g., this is the Franklin Building). However, as we have argued earlier, information regarding the category of a natural scene may reside in the pattern of activity in the PPA rather than the mean level of activity. Furthermore, as subjects were identifying known landmarks in the identification task in the Epstein and Higgins (2007) study, the greater activity in PPA may also have reflected greater familiarity with the stimuli used in the identification task.

The lower decoding accuracy from the PPA that we observed for inverted than for upright scenes is consistent with the PPA's role in scene layout (Aguirre et al., 1996; Epstein et al., 2007; Park et al., 2007), which is considerably disrupted by scene inversion. In contrast, V1 showed high decoding accuracy without regard to the image orientation as long as training and test data had the same orientation, which is consistent with V1's role in representing the spatial distribution of local features rather than global scene layout. Interestingly, RSC, which is thought to be involved in navigation (Maguire, 2001) and placing objects within a visual context (Bar and Aminoff, 2003), does not show this same scene inversion effect.

The pattern of results in the LOC is also noteworthy. It showed both significant decoding accuracy and a clear correlation with behavior. The role typically attributed to the LOC is the representation of objects (Malach et al., 1995; Grill-Spector et al., 1999). How might the LOC relate to scene categorization then? In many instances, objects can indicate particular scene categories (Hollingworth and Henderson, 2002; Davenport and Potter, 2004). For instance, the presence of a car could indicate a highway and the presence of trees a forest. In fact, activity in LOC was previously reported in a comparison of strong versus weak associations between objects and scenes (Bar and Aminoff, 2003).

However, the gist of a scene is also thought to provide context for object detection (Biederman, 1972; Bar, 2004), suggesting that a top-down signal originating in scene-sensitive regions such as the PPA and RSC might modulate neural activity in object-selective regions such as the LOC in ways consistent with the scene category. It is therefore possible that the sensitivity of LOC voxels for natural scene categories is because of this modulation signal rather than the computations originating in LOC itself. Further experiments will be necessary to determine the relative contributions and the nature of the information flow between these areas.

It is important to note that not all ROIs tested were able to discriminate natural scenes successfully. We did not decode natural scene categories significantly above chance from the face-sensitive FFA, suggesting that it does not play an important role in scene categorization. This negative result for FFA further highlights the importance of PPA, RSC, and LOC in scene categorization.

We were able to decode scene categories from the fMRI activity in V1 significantly above chance. Given V1's retinotopic organization and sensitivity to local orientations and spatial frequencies (Hubel and Wiesel, 1962; De Valois and De Valois, 1980), this finding could be interpreted as evidence for a representation of scene categories based on the global distribution of spatial frequencies in the image (Oliva and Torralba, 2001) or distributed local feature information such as oriented textures (Fei-Fei and Perona, 2005; Bosch et al., 2006). Indeed, such an early representation could explain the speed with which scene categorization is accomplished (Thorpe et al., 1996; VanRullen and Thorpe, 2001). However, our results show that scene categories as experienced by human observers go beyond low-level, V1-like features. The pattern of decoding errors in V1 did not correlate with behavioral errors by human subjects. We conclude that fMRI activity patterns in V1, while enabling the decoding of scene category significantly above chance, play a less direct role in humans' ability to categorize natural scenes than the activity patterns in PPA, RSC, and LOC, at least in the case of the relatively long presentation durations (1.6 s) used in the fMRI design.

A whole-brain searchlight analysis of decoding accuracy largely confirmed the involvement of PPA and RSC in natural scene categorization and showed regions in visual cortex extending beyond these narrowly defined ROIs. Above-chance decoding accuracy in Broca's area could be because of subjects subvocalizing the names of natural scene categories during the passive-viewing fMRI experiment.

In summary, our findings suggest that scene categorization depends on a hierarchy of multiple regions within visual cortex. Scene categories differ as early as V1, because of differences in the distribution of spatial frequencies and orientations in the images. However, the scene category information present in the PPA, RSC, and LOC are more closely related to the ultimate behavior of the human. On the basis of the previously suggested functions of the PPA and RSC, we propose that they are responsible for extracting differences in spatial layout among different categories of scenes. LOC, however, may play a role in extracting scene specific objects, which can then bias scene categorization in other regions. Ultimately, the information in the PPA, RSC, and in the LOC is presumably passed on to higher-level areas involved in decision making. More work will be needed to verify these roles, as well as to illuminate the specific flow of information between regions. But at least now we have a rudimentary knowledge of a network of regions that participate in natural scene categorization.

Finally, the methods presented here have implications beyond the question of natural scene categorization. We expect our multivoxel pattern recognition methods to be useful in identifying neural representations in a variety of other contexts previously thought to be beyond the resolution of fMRI. Going beyond reports of generally high activity in particular brain regions, our procedure of correlating differences in patterns of activity with human behavior allows us to not only determine brain regions that contain information relevant to a complex visual behavior, but also whether the information contained within those areas are likely to contribute to the observed behavior.

References

Aguirre GK (2007) Continuous carry-over designs for fMRI. Neuroimage 35:1480–1494.

Aguirre GK, Detre JA, Alsop DC, D'Esposito M (1996) The parahippocampus subserves topographical learning in man. Cereb Cortex 6:823–829.

Bar M (2004) Visual objects in context. Nat Rev Neurosci 5:617–629.

Bar M, Aminoff E (2003) Cortical analysis of visual context. Neuron 38:347–358.

Biederman I (1972) Perceiving real-world scenes. Science 177:77–80.

Bosch A, Zisserman A, Munoz X (2006) Scene classification via pLSA. Paper presented at European Conference of Computer Vision, Graz, Austria, May.

Carlson TA, Schrater P, He S (2003) Patterns of activity in the categorical representations of objects. J Cogn Neurosci 15:704–717.

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

Cox RW (1996) AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162–173.

Davenport JL, Potter MC (2004) Scene consistency in object and background perception. Psychol Sci 15:559–564.

Delorme A, Richard G, Fabre-Thorpe M (2000) Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans. Vision Res 40:2187–2200.

De Valois RL, De Valois KK (1980) Spatial vision. Annu Rev Psychol 31:309–341.

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

Epstein RA, Higgins JS (2007) Differential parahippocampal and retrosplenial involvement in three types of visual scene recognition. Cereb Cortex 17:1680–1693.

Epstein RA, Higgins JS, Parker W, Aguirre GK, Cooperman S (2006) Cortical correlates of face and scene inversion: a comparison. Neuropsychologia 44:1145–1158.

Epstein RA, Parker WE, Feiler AM (2007) Where am I now? Distinct roles for parahippocampal and retrosplenial cortices in place recognition. J Neurosci 27:6141–6149.

Fei-Fei L, Perona P (2005) A Bayesian hierarchical model for learning natural scene categories. Paper presented at IEEE International Conference on Computer Vision and Pattern Recognition, San Diego, June.

Fei-Fei L, VanRullen R, Koch C, Perona P (2005) Why does natural scene categorization require little attention? Exploring attentional requirements for natural and synthetic stimuli. Vis Cogn 12:893–924.

Fei-Fei L, Iyer A, Koch C, Perona P (2007) What do we perceive in a glance of a real-world scene? J Vis 7: 10:11–29.

Gauthier I, Skudlarski P, Gore JC, Anderson AW (2000) Expertise for cars and birds recruits brain areas involved in face recognition. Nat Neurosci 3:191–197.

Geschwind N (1970) The organization of language and the brain. Science 170:940–944.

Grill-Spector K, Kushnir T, Edelman S, Avidan G, Itzchak Y, Malach R (1999) Differential processing of objects under various viewing conditions in the human lateral occipital complex. Neuron 24:187–203.

Grodzinsky Y, Santi A (2008) The battle for Broca's region. Trends Cogn Sci 12:474–480.

Haushofer J, Livingstone MS, Kanwisher N (2008) Multivariate patterns in object-selective cortex dissociate perceptual and physical shape similarity. PLoS Biol 6:e187.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001)

Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425–2430.
Haynes JD, Rees G (2005) Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat Neurosci 8:686–691.
Hollingworth A, Henderson JM (2002) Accurate visual memory for previously attended objects in natural scenes. J Exp Psychol Hum Percept Perform 28:113–136.
Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160:106–154.
Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825–841.
Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. Nat Neurosci 8:679–685.
Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 17:4302–4311.
Kay KN, Naselaris T, Prenger RJ, Gallant JL (2008) Identifying natural images from human brain activity. Nature 452:352–355.
King-Smith PE, Grigsby SS, Vingrys AJ, Benes SC, Supowit A (1994) Efficient and unbiased modifications of the QUEST threshold method: theory, simulations, experimental evaluation and practical implementation. Vision Res 34:885–912.
Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. Proc Natl Acad Sci U S A 103:3863–3868.
Li FF, VanRullen R, Koch C, Perona P (2002) Rapid natural scene categorization in the near absence of attention. Proc Natl Acad Sci U S A 99:9596–9601.
Maguire EA (2001) The retrosplenial contribution to human navigation: a review of lesion and neuroimaging findings. Scand J Psychol 42:225–238.
Malach R, Reppas JB, Benson RR, Kwong KK, Jiang H, Kennedy WA, Ledden PJ, Brady TJ, Rosen BR, Tootell RB (1995) Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proc Natl Acad Sci U S A 92:8135–8139.
O'Craven KM, Kanwisher N (2000) Mental imagery of faces and places activates corresponding stimulus-specific brain regions. J Cogn Neurosci 12:1013–1023.
Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 42:145–175.
Park S, Intraub H, Yi DJ, Widders D, Chun MM (2007) Beyond the edges of a view: boundary extension in human scene-selective visual cortex. Neuron 54:335–342.
Potter MC, Levy EI (1969) Recognition memory for a rapid sequence of pictures. J Exp Psychol 81:10–15.
Schneider KA, Richter MC, Kastner S (2004) Retinotopic organization and functional subdivisions of the human lateral geniculate nucleus: a high-resolution functional magnetic resonance imaging study. J Neurosci 24:8975–8985.
Tarr MJ, Gauthier I (2000) FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. Nat Neurosci 3:764–769.
Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system. Nature 381:520–522.
Tversky B, Hemenway K (1983) Categories of environmental scenes. Cogn Psychol 15:121–149.
VanRullen R, Thorpe SJ (2001) Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. Perception 30:655–668.
Walther DB, Fei-Fei L (2007) Task-set switching with natural scenes: measuring the cost of deploying top-down attention. J Vis 7: 9:1–12.
Williams MA, Dang S, Kanwisher NG (2007) Only some spatial patterns of fMRI response are read out in task performance. Nat Neurosci 10:685–686.
Yi DJ, Kelley TA, Marois R, Chun MM (2006) Attentional modulation of repetition attenuation is anatomically dissociable for scenes and faces. Brain Res 1080:53–62.