Disentangling Scene Content from Spatial Boundary: Complementary Roles for the Parahippocampal Place Area and Lateral Occipital Complex in Representing Real-World Scenes

Behavioral/Systems/Cognitive

Disentangling Scene Content from Spatial Boundary: Complementary Roles for the Parahippocampal Place Area and Lateral Occipital Complex in Representing Real-World Scenes

Soojin Park, Timothy F. Brady, Michelle R. Greene, and Aude Oliva
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Behavioral and computational studies suggest that visual scene analysis rapidly produces a rich description of both the objects and the spatial layout of surfaces in a scene. However, there is still a large gap in our understanding of how the human brain accomplishes these diverse functions of scene understanding. Here we probe the nature of real-world scene representations using multivoxel functional magnetic resonance imaging pattern analysis. We show that natural scenes are analyzed in a distributed and complementary manner by the parahippocampal place area (PPA) and the lateral occipital complex (LOC) in particular, as well as other regions in the ventral stream. Specifically, we study the classification performance of different scene-selective regions using images that vary in spatial boundary and naturalness content. We discover that, whereas both the PPA and LOC can accurately classify scenes, they make different errors: the PPA more often confuses scenes that have the same spatial boundaries, whereas the LOC more often confuses scenes that have the same content. By demonstrating that visual scene analysis recruits distinct and complementary high-level representations, our results testify to distinct neural pathways for representing the spatial boundaries and content of a visual scene.

Introduction

Behavioral studies have shown that, in a brief glance at a scene, a rich representation is built comprising spatial layout, the degree of human manufacture, and a few prominent objects (Oliva and Torralba, 2001; Renninger and Malik, 2004; Fei-Fei et al., 2007; Greene and Oliva, 2009a,b). In parallel, neuroimaging investigations have identified specific brain regions involved in scene perception. Among these regions is the parahippocampal place area (PPA), which responds preferentially to pictures of scenes and landmarks and shows selectivity to the geometric layout of the scene but not the quantity of objects (Aguirre et al., 1998; Epstein and Kanwisher, 1998; Epstein et al., 1999; Janzen and van Turennout, 2004) and the retrosplenial complex (RSC), which also responds to scenes and navigationally relevant tasks (Epstein, 2008; Park and Chun, 2009) in addition to processing context (Bar and Aminoff, 2003). In contrast, the lateral occipital complex (LOC) has been found to represent object shapes and categories (Malach et al., 1995; Grill-Spector et al., 1998; Kourtzi and Kanwisher, 2000; Eger et al., 2008; Vinberg and Grill-Spector, 2008). Recent studies have found that activity in early visual areas PPA and LOC is discriminative enough to allow scene classification into a handful of semantic categories (Naselaris et al., 2009; Walther et al., 2009). However, we do not yet know how the brain accomplishes the diverse functions involved in scene understanding.

Here, we examined scene representation using functional magnetic resonance imaging (fMRI) pattern analysis, demonstrating the existence of high-level neural representations of visual environments that uncouple processing of the spatial boundaries of a scene from its content. Just as external shape and internal features are separable dimensions of face encoding, an environmental space can be represented by two separable and complementary descriptors (Oliva and Torralba, 2001): its spatial boundary (i.e., the shape and size of the scene's space) and its content (textures, surfaces, materials and objects). As illustrated in Figure 1, the shape of a scene may be expansive and open to the horizon, as in a field or highway, or closed and bounded by frontal and lateral surfaces, as in forests or streets. For a given spatial boundary, a scene may comprise natural or urban (manufactured) objects. Analyzing the types of errors produced by the PPA and LOC in this two-dimensional space allows us distinguish whether the PPA and LOC represent a scene in an overlapping manner (e.g., both produce similar errors when classifying scenes) or in a complementary manner (e.g., specialization in representing boundaries and content).

We show that, although both the PPA and the LOC classify scenes with the same level of accuracy, these regions show opposite patterns of classification errors. Therefore, our work provides

Received July 26, 2010; revised Oct. 26, 2010; accepted Oct. 30, 2010.

This work was funded by National Science Foundation Graduate Research Fellowships (T.F.B., M.R.G.) and National Science Foundation CAREER Award IIS 0546262 (A.O.). We thank the Athinoula A. Martinos Center at the McGovern Institute for Brain Research, Massachusetts Institute of Technology for data acquisition, and Talia Konkle and Barbara Hidalgo-Sotelo for helpful conversation and comments on this manuscript.

Correspondence should be addressed to either Soojin Park or Aude Oliva, Department of Brain and Cognitive Sciences, Room 46-4065, 77 Massachusetts Avenue, Massachusetts Institute of Technology, Cambridge, MA 02139. E-mail: sjpark31@mit.edu, oliva@mit.edu.

DOI:10.1523/JNEUROSCI.3885-10.2011

 0270-6474/11/311333-08$15.00/0

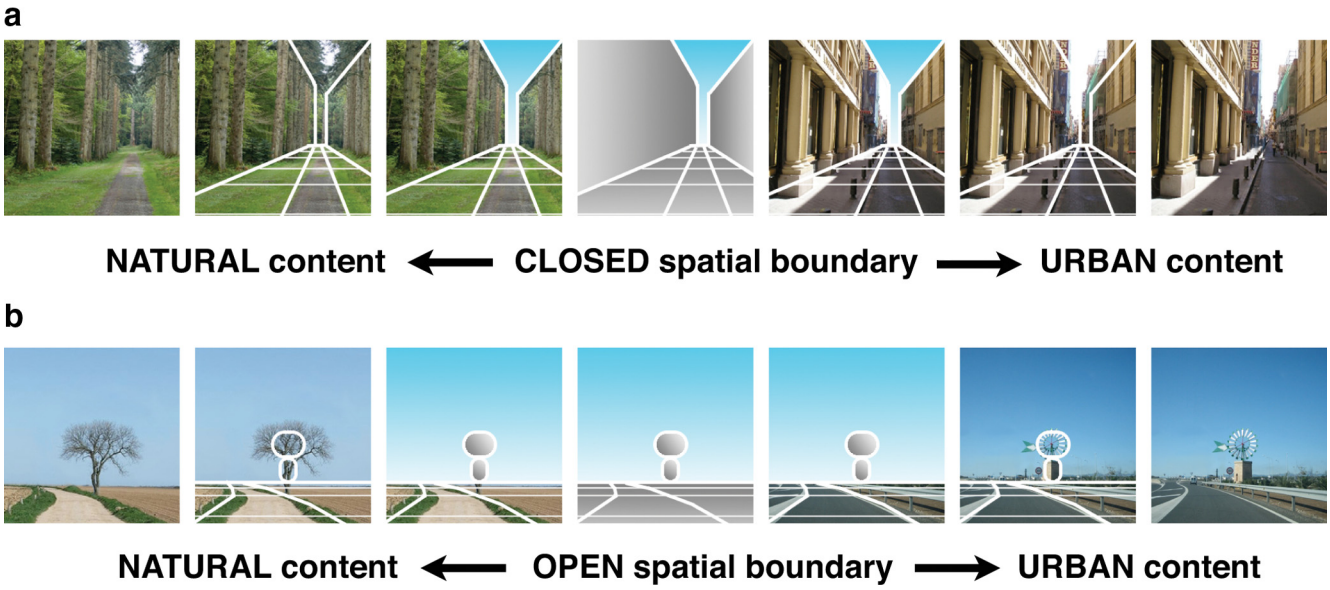

Figure 1. A schematic illustration of how pictures of real-world scenes can be uniquely defined by their spatial boundary information and content. Note that the configuration, size, and locations of components were corresponding between natural and urban environments. ***a***, Keeping the enclosed spatial boundary, if we strip off the natural content of a forest and fill the space with urban contents, then the scene becomes an urban street scene. ***b***, Keeping the open spatial boundary, if we strip off the natural content of a field and fill the space with urban contents, then the scene becomes an urban parking lot.

the first evidence that multiple brain regions perform distinct and complementary analysis of a visual scene, similar in spirit to that proposed by computational models of scene understanding (Oliva and Torralba, 2001; Vogel and Schiele, 2007; Greene and Oliva, 2009a) and visual search (Torralba, 2003; Torralba et al., 2006).

Materials and Methods

Subjects. Eight participants (two females; one left-handed; ages, 19–28 years) for the main experiment, six participants (three females; ages, 18–29 years) for the first control experiment (using phase-scrambled images), and seven participants (three females; ages, 20–35 years) for the second control experiment (using added vertical and horizontal bars) were recruited from the Massachusetts Institute of Technology community for financial compensation. All had normal or corrected-to-normal vision. Informed consent was obtained, and the study protocol was approved by the Institutional Review Board of the Massachusetts Institute of Technology. One participant for the main experiment was excluded from the analyses because of excessive head movement (over 8 mm across runs).

Visual stimuli. Scenes were carefully chosen to represent each of the following four scene groups: "open natural" images, "closed natural" images, "open urban" images, and "closed urban" images (Oliva and Torralba, 2001; Greene and Oliva, 2009a). Images were visually matched for spatial boundary and content across these groups (see examples in Figs. 1, 3*a*). Importantly, each scene group included multiple basic-level scene categories. For example, open and closed natural images included different views of fields, oceans, forests, creeks, mountains, and deserts, whereas open and closed urban images included views of highways, parking lots, streets, city canals, buildings, and airports. There were 140 test images per scene group. In the main experiment, photographs were 256 × 256 pixels resolution (4.5° × 4.5° of visual angle) and were presented in grayscale with a mean luminance averaging 127 (on a 0–255 luminance scale). In the first control experiment, the same images were phase scrambled, so as to keep second-order image statistics but remove high-level scene information. The grayscale images from experiment 1 were first Fourier transformed to decompose them into their amplitude spectrum and phase. Next, the phase at each frequency was replaced with a random phase. Finally, the image was reconstructed from this modified Fourier space and then rescaled so the luminance of each pixel ranged from 0 to 255, and the overall image had mean luminance of 127. In the second control experiment, horizontal or vertical lines were superposed on top of the images and were presented 500 × 500 pixels resolution to maximize the visibility of lines (for control experiment stimuli, see Fig. 6). The same orientation lines were added on top of either all of the natural scenes or all of the urban scenes to increase the within-content low-level image similarity across these two sets of conditions. Images were presented in the scanner using a Hitachi (CP-X1200 series) rear-projection screen.

Experimental design. Twenty images from an image group were presented in blocks of 20 s each. The order of block conditions was randomized within each run. Each block was followed by a 10 s fixation period. Within a block, each scene was displayed for 800 ms, followed by 200 ms blank. The entire image set (560 images) was presented across two runs with a break between: the first run was composed of 16 blocks with four blocks per condition, acquiring 245 image volumes; the second run was composed of 12 blocks with three blocks per condition, acquiring 185 image volumes. This set of two runs was repeated four times within a session, totaling eight runs, to increase the number of samples and power. Accordingly, participants saw the same image four times across runs at different time points per each run. Twenty-eight blocks per each condition were acquired and used as training samples throughout the experiment. Participants performed a one-back repetition detection task to maintain attention.

The experimental design for the first and second control experiments were identical to the main experiment except that participants in the first control experiment performed a red-frame detection task rather than a one-back repetition task to maintain attention on the phase scrambled images.

MRI acquisition and preprocessing. Imaging data were acquired with a 3 T Siemens fMRI scanner with 32-channel phased-array head coil (Siemens) at the Martinos Center at the McGovern Institute for Brain Research at Massachusetts Institute of Technology. Anatomical images were acquired using a high-resolution (1 × 1 × 1 mm voxel) magnetization-prepared rapid-acquisition gradient echo structural scan. Functional images were acquired with a gradient echo-planar T2* sequence [repetition time (TR), 2 s; echo time, 30 ms; field of view, 200 mm; 64 × 64 matrix; flip angle, 90°; in-plane resolution, 3.1 × 3.1 × 3.1 mm; 33 axial 3.1 mm slices with no gap; acquired parallel to the anterior commissure–posterior commissure line].

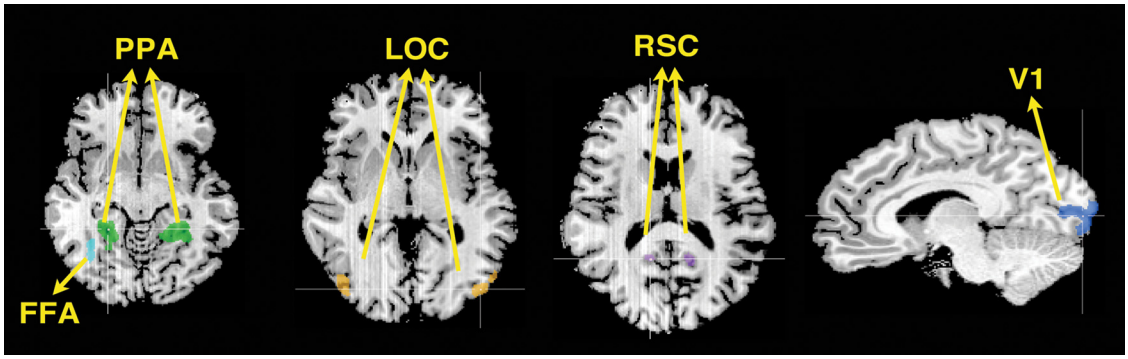

Figure 2. Five regions of interests are shown on a representative participant's brain. Talairach coordinates for peak voxels of each ROI shown are as follows: LPPA, −21, −43, −5; RPPA, 21, −43, −5; LLOC, −45, −79, 1; RLOC, 39, −73, 7; RFFA, 36, −52, −5; LRSC, −15, −52, 13; RRSC, 12, −52, 13; LV1, −14, −94, 1; RV1, 14, −92, 4. Note that our SVM analysis was conducted on spatially nontransformed individual brain. However, for comparison purposes, we are reporting Talairach coordinates of a representative subject's ROIs.

Functional data were preprocessed using Brain Voyager QX software (Brain Innovation). Preprocessing included slice scan-time correction, linear trend removal, and three-dimensional motion correction. No additional spatial or temporal smoothing was performed. For retinotopic analysis, we mapped data from a functional retinotopic localizer (see below) onto a surface-rendered brain of each individual.

Pattern analysis. For multivariate pattern analysis, we extracted MRI signal intensity from each voxel within the region of interest (ROI) across all time points and transformed the MRI signal intensity within each voxel into *z*-scores so that the mean activity is set to 0 and the SD is set to 1. This helps mitigate overall differences in fMRI signal across different regions of interests and across runs and sessions (Kamitani and Tong, 2005). Activity level for each block of each individual voxel was labeled with condition, which spanned 20 s (10 TR) in all experiments. In labeling the condition, we added 4 s (2 TR) to each time points, to account for the hemodynamic delay of the blood oxygenation level-dependent (BOLD) response (4–6 s) so that we are targeting the correct period corresponding to each block of samples. We did not do a voxel selection, and the entire cluster that passed the threshold of an ROI localizer was used. This eliminates problems with "double-dipping" as a result of nonindependence in feature selection (Kriegeskorte et al., 2009; Vul et al., 2009). The average number of voxels for each of the ROIs in the main experiment were as follows (after mapping into the structural $1 \times 1 \times 1$ voxel space; for voxel counts in the original functional space, divide by 3.1^3, or 29.8): left (L) PPA, 2702 voxels; right (R) PPA, 2455 voxels; LLOC, 3010 voxels; RLOC, 2291 voxels; fusiform face area (FFA), 1292 voxels; LRSC, 699 voxels; RRSC, 1242 voxels; L primary visual area (V1), 4357 voxels; and RV1, 3827 voxels. The average number of voxels for the PPA and V1 in control experiments 1 and 2 were as follows: LPPA, 2719 voxels; RPPA, 2656 voxels; LV1, 5102 voxels; RV1, 4011 voxels for experiment 1; LPPA, 2024 voxels; RPPA, 2207 voxels; LV1, 5150 voxels; and RV1, 4420 voxels for experiment 2. Although the bigger stimuli size in the second control experiment led to more V1 voxels, we found that the support vector machine (SVM) performance was not influenced by the number of voxels included in the analysis beyond a plateau point of ~500 voxels.

For each region in each participant, we used a separate linear SVM classifier based on LIBSVM (http://sourceforge.net/projects/svm/). For each block, we computed the average pattern of activity across the voxels. An SVM classifier was trained to classify the four conditions using all blocks but one and then was tested on the remaining block. Percentage correct classification for each subject and each ROI was calculated as the average performance over all leave-one-block-out classifications.

Regions of interest. Regions of interest were defined for each participant using independent localizers (Fig. 2). Scene localizer runs presented blocks of scene pictures (representing outdoor natural and urban scenes) and faces (half female and half male), whereas object localizer runs presented blocks of real-world objects and scrambled versions of these pictures. Scrambled images were created by dividing intact object images into a 16×16 square grid and by scrambling the positions of squares based on eccentricity (Kourtzi and Kanwisher, 2000). Scene localizer runs presented six blocks of scenes and six blocks of faces that alternated with each other with 10 s blank periods between. Object localizer runs presented four blocks of objects and four blocks of scrambled objects that alternated with each other with 10 s blank periods between. Each block presented 12 images, and participants performed a repetition detection task on consecutive repetitions (main and control experiment 2) or performed a red-frame detection task that appeared around the stimuli (control experiment 1). A retinotopic localizer presented vertical and horizontal visual field meridians to delineate borders of retinotopic areas (Sereno et al., 1995; Spiridon and Kanwisher, 2002). Triangular wedges of black and white checkerboards were presented either vertically (upper or lower vertical meridians) or horizontally (left or right horizontal meridians) in 16 s blocks, alternated with 16 s blanks. Participants were instructed to fixate on a small central fixation dot.

The left and right PPA were defined by contrasting brain activity of scene blocks > face blocks and finding clusters between the posterior parahippocampal gyrus and anterior lingual gyrus. This contrast also defined the RSC near the posterior cingulated cortex, which was present in five of seven participants. The left and right LOC were defined by contrasting brain activity of objects > scrambled objects blocks in the lateral occipital lobe. The FFA was defined by contrasting brain activity of face blocks > scene blocks and finding clusters in the right fusiform gyrus of the occipitotemporal cortex. Only the right FFA was included in the analyses because it was consistently found across all participants (Kanwisher et al., 1997). The retinotopic borders of left and right V1 were defined with a contrast between vertical and horizontal meridians. For each region in each participant, a separate classifier was used. Classification performance for bilateral regions did not differ from each other, so the classification performance was averaged across left and right in the PPA, LOC, V1, and RSC for all analyses presented here.

Results

We extracted multivoxel fMRI activity from our ROIs (PPA, LOC, V1, RSC, and FFA). For each region in each participant, we used a separate linear SVM to classify each block of scenes as closed natural images, closed urban images, open natural images, or open urban images (Fig. 3*a*). As shown in Figure 3*b*, classification accuracy in all of the ROIs was significantly above chance (25%): PPA, LOC, V1, RSC, and FFA had respective classification accuracies of 51% ($t_{(6)} = 7.1$, $p < 0.001$), 45% ($t_{(6)} = 4.0$, $p < 0.01$), 50% ($t_{(6)} = 6.1$, $p < 0.001$), 37% ($t_{(4)} = 4.4$, $p < 0.05$), and 43% ($t_{(6)} = 4.1$, $p < 0.01$). These results show that these regions can distinguish between scenes varying in spatial and content properties. Furthermore, these results suggest that the analysis of scene information is not processed by a single scene-specific region but is organized in a distributed manner across multiple visual areas.

Classification errors

Classification accuracy does not inform about the nature of the scene representation in each region. For example, the spatial boundary (open or closed) and the content (natural or urban) may have contributed equally to the classification performance across different regions. Conversely, the spatial boundary and the content may have contributed differently to the classification performance across different areas. We examined this contribution by studying the types of errors made by the classifier. If the classifier shows systematic confusions between two conditions in a particular brain region, this suggests that the region has similar representations of the scenes from these two conditions.

Figure 4*a* illustrates two possible orthogonal patterns of errors. First, if a particular brain region is sensitive to the spatial boundaries of a scene, it should confuse images with the same global structure, regardless of the content or objects in the scene. For instance, scenes representing closed urban environments (e.g., street, buildings, or city canal) might be confused with closed natural scenes that have similar spatial boundaries (e.g., forest, mountain, or canyon) but not with open urban scenes that have similar content (e.g., highways, parking lot, or airport). In Figure 4*a*, we refer to these errors as confusions within the same spatial boundary. Conversely, if a given brain region is more sensitive to the content of a scene than its structure, it should confuse images with the same content, regardless of their spatial boundaries (e.g., fields, deserts, or ocean would be confused with forest, mountain, or canyons but not with highways, parking lots, or airports that have urban contents). We refer to these errors as confusion within the same content.

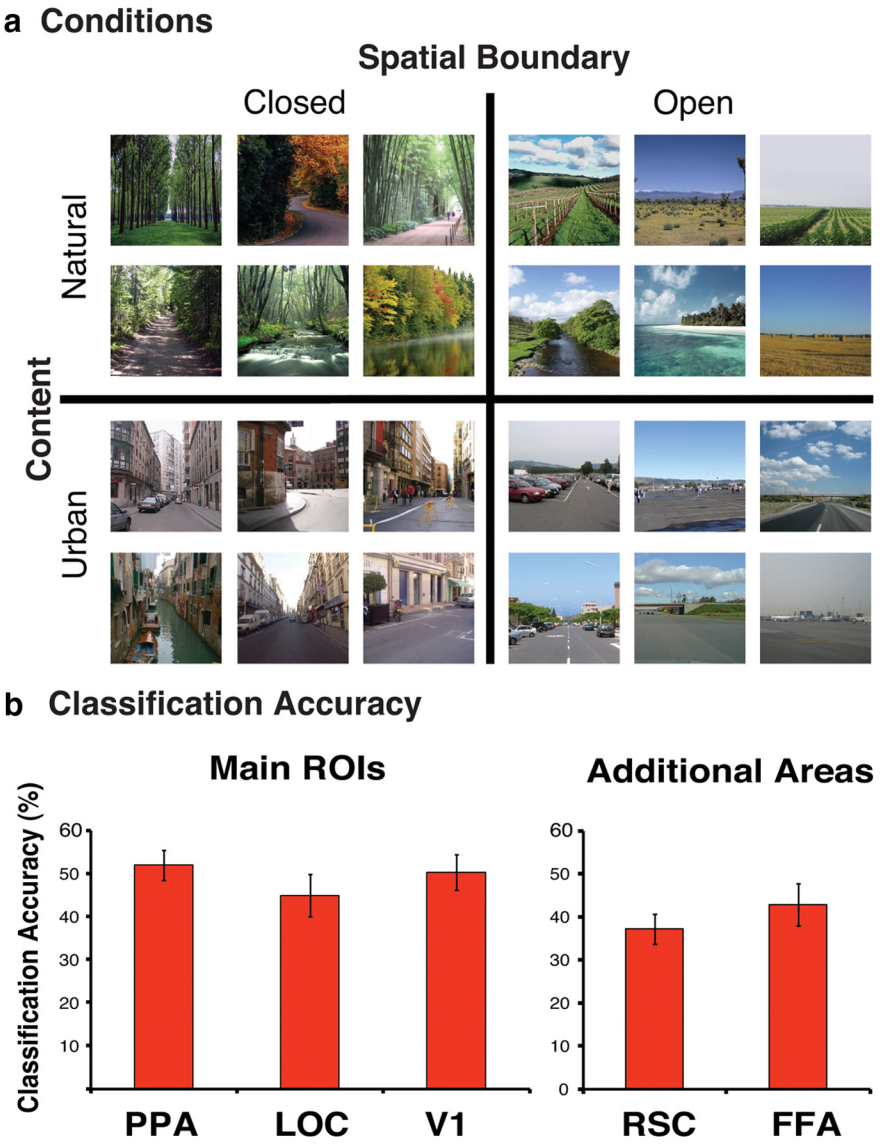

Figure 3. ***a***, Examples scenes for the four conditions. Scenes were defined by their spatial boundary and content properties: closed natural images, closed urban images, open natural images, and open urban images. ***b***, The average classification accuracy (percentage) in the PPA, LOC, V1, RSC, and FFA. All areas showed classification accuracy that was significantly above chance (25%). Error bars represent ±1 SEM.

Figure 4*b* shows the types of errors made by different regions of interest. First, as we hypothesized, we observed a striking dissociation between the main regions under investigation, the PPA and the LOC: whereas the PPA made the most errors between scenes that shared the same spatial boundaries, regardless of content (e.g., confusing open natural and open urban scenes), the LOC made the most errors between scenes that shared the same content, regardless of the spatial boundary (e.g., confusing open natural and closed natural scenes). Most importantly, there was a significant interaction across area (PPA or LOC) and types of errors (confusion within the same boundary vs confusion within the same content, $F_{(1,6)} = 69.8$, $p < 0.001$). We conducted planned *t* tests between the two types of errors (confusion within the same boundary vs confusion within the same content), which were significant in both the PPA ($t_{(6)} = 4.6$, $p < 0.005$) and in the opposite direction in the LOC ($t_{(6)} = -6.6$, $p < 0.001$). Additionally, other regions of interest also show a specific pattern of errors: the FFA made more confusion errors within the same scene content, similar to the LOC ($t_{(6)} = -3.3$, $p < 0.05$). RSC showed a numerical pattern of greater confusion within the same spatial layout, like the PPA pattern, but not as dominant as found in the PPA, and this pattern was not significant ($t_{(4)} = 1.2$, $p > 0.2$). On average, more than 72% of errors were confusions within the same spatial boundary or within the same content. Double errors, in which both content and layout were incorrect, were the least frequent (21, 27, 16, 25, and 23% in the PPA, LOC, V1, RSC, and FFA, respectively) and did not show any interaction across areas.

The overall performance was the same when we have normalized the number of voxels across the different regions of interest (Fig. 5). We randomly sampled different numbers of voxels (from 10 to 1000) from each ROI, and this process was repeated 200 times for each number of voxels within each ROI in each observer. The classification accuracy plateaued at 500 voxels and was equal to the classification performance when all the voxels

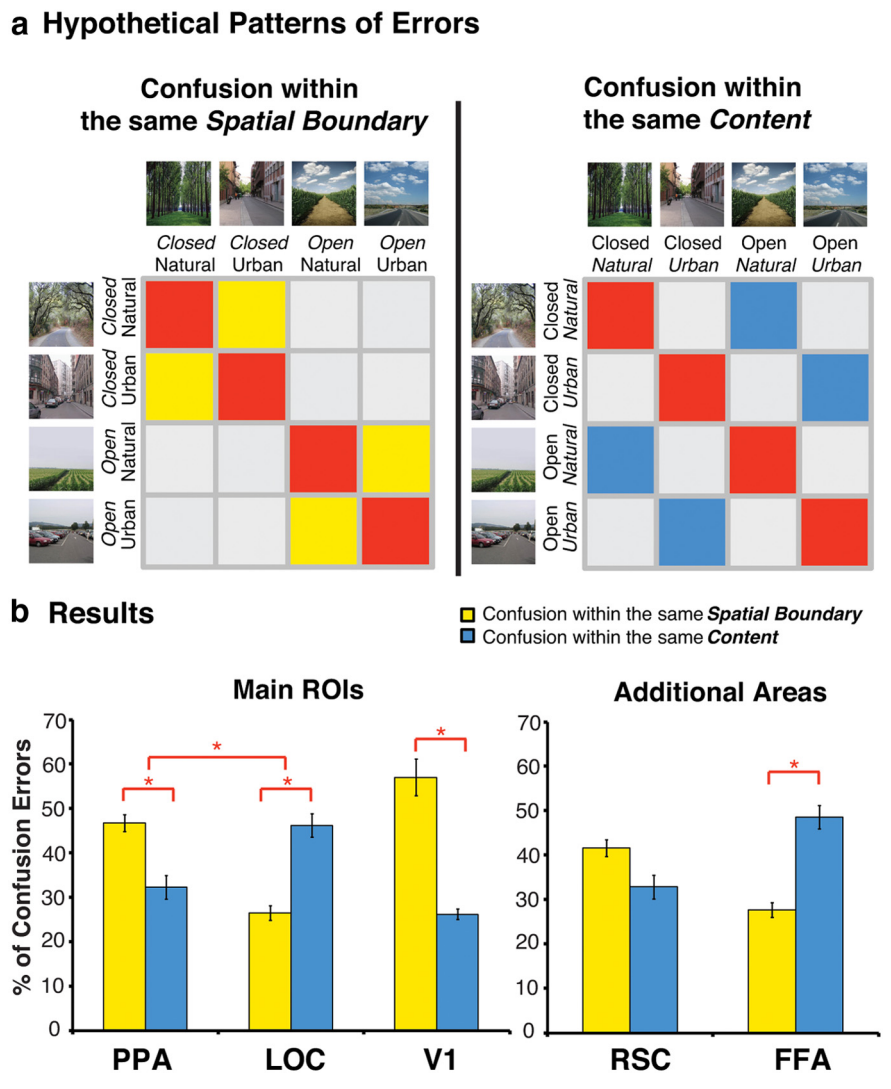

Figure 4. ***a***, Hypothetical patterns of errors that could be made by the classifier. The rows represent the scene image conditions as presented to the participants, and the columns represent the scene condition that the classifier predicted from the fMRI patterns of activity. If spatial boundary properties of scenes are represented in a particular brain area, we expect high confusions within scenes that share the same spatial boundary (marked in yellow). If content properties of scenes are important for classification, we expect high confusions within scenes that share the same content (blue cells). ***b***, Confusion errors (percentage) are shown for main ROIs (the PPA, LOC, and V1) on the left and for additional regions (RSC and FFA) on the right. Yellow bars represent the average percentage of confusions within the same spatial boundary; blue bars represent the average percentage of confusions within the same content. Error bars represent SEM.

within ROIs were accounted (PPA, LOC, V1, RSC, and FFA classification accuracies were, respectively, 50.4, 45, 49, 37, and 42%).

Moreover, we observed significantly higher classification accuracy for a combined region of interest merging PPA and LOC voxels (an average of 56%, after randomly sampling 500 voxels from the combined ROI and repeating this process 200 times) compared with when only 500 PPA voxels were included in the original analysis (50.4%, $t_{(6)} = 2.5$, $p = 0.05$) or when only 500 LOC voxels were included (45%, $t_{(6)} = 3.4$, $p < 0.05$). The fact that the addition of LOC voxels to PPA voxels increases classification accuracy suggests that the neural representation of scenes within these regions is complementary. Altogether, the results provide the first evidence that the PPA and LOC represent complementary properties of real-world scenes, with PPA holding a representation of spatial boundaries and LOC simultaneously holding a representation of scene content.

Correlation analysis

In addition to looking at the confusions made by a classifier, we can also examine the similarity in the patterns across voxels in the different conditions using a correlation-based analysis (Haxby et al., 2001). When we conducted the split-half correlation analysis, results showed similar conclusions to the classifier-based analysis. Within-condition correlation for the PPA, RSC, LOC, FFA, and V1, respectively, were 0.74, 0.64, 0.64, 0.44, and 0.83. Most importantly, the across-condition correlation patterns showed the same pattern as in the SVM confusion error analysis. There was a significant crossover interaction between the PPA and LOC, such that the correlation of images that shared the spatial boundary is higher in PPA and the correlation of images with shared content is higher in LOC ($F_{(1,6)} = 12.8$, $p < 0.05$). Thus, the two results provide similar conclusions. They differ in that the classifier confusions weigh informative voxels only, do not discard differences between the mean activity level of the two conditions, and, in addition, provide interpretable effect sizes (in terms of percentage correct) rather than differences in correlation r values (which do not indicate how separable the sets actually are).

Contribution of low-level visual properties

As shown in Figures 3*b* and 4*b*, V1 had the same overall pattern of classification and confusion errors as the PPA, confusing scenes with similar boundaries more often than scenes with similar content. The high level of accuracy of V1 is not surprising given that previous studies observed high classification for natural images using patterns of activity of retinotopic areas (Kay et al., 2008; Naselaris et al., 2009) and the fact that the poles of the two spatial boundary properties contain distinctive and diagnostic low-level features (Oliva and Torralba, 2001; Torralba and Oliva, 2003). Given the dissociation in the types of confusion between the PPA and LOC (and other ROIs), it is unlikely that the patterns of activity of PPA result only from the PPA mirroring a straightforward influence of V1. Previous studies have demonstrated the high-level nature of the responses of PPA, including its sensitivity to spatial layout (Epstein, 2005) and its responses during binocular rivalry of scenes and faces, which reflects the perceived representation rather than direct physical input (Tong et al., 1998). Nevertheless, to empirically address the possibility that the greater confusion of PPA within spatial layout was a byproduct of the low-level properties of the stimuli, we conducted two control experiments. These experiments manipulated the low-level properties of the images to decipher the nature of the representations in the PPA and V1. Specifically, in the first control experiment, we expect the removal of high-level im-

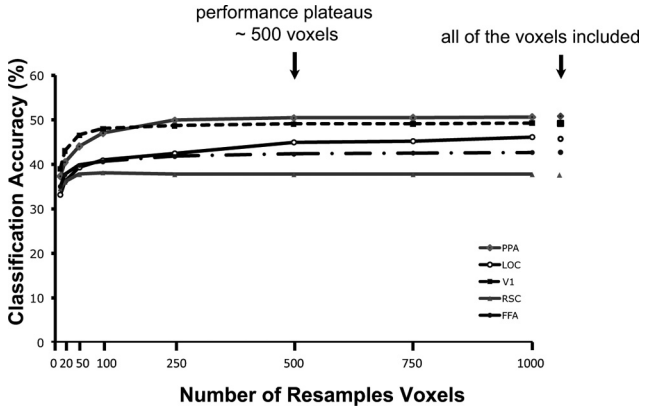

Figure 5. The classification performance as a function of the number of voxels resampled. The classification performance plateaus at ~500 voxels, equaling the classification accuracy when all of the voxels within the ROI are included.

age information to disrupt the classification within the PPA but not within V1, whereas in the second control experiment, we expect the addition of accidental low-level features to the images to improve V1 performance but not the PPA (Fig. 6).

The first control experiment was designed to remove higher-order image information while preserving a class of low-level visual features: the power spectrum, or energy, at all spatial frequencies and orientations. To do so, the phase of each image was randomized. If V1, but not the PPA, is sensitive to these low-level image statistics alone, we expect V1 accuracy to remain significantly above chance, whereas PPA classification performance should drop to chance level. Indeed, this is exactly what we found: although both the PPA and V1 dropped in performance with phase-scrambled images compared with original intact images (main effect of experiment, $F_{(1,11)} = 22.5, p < 0.05$), the drop was much greater in the PPA (significant interaction between PPA and V1 classification performance, $F_{(1,11)} = 5.7, p < 0.05$), suggesting that the higher-order information removed with phase-scrambled images were more important for the classification of PPA than V1. Although V1 performance was lower than in the main experiment, conceivably because phase scrambling removes even some low-level statistics of the scenes about which V1 cares, V1 still classified the four scene groups above chance with phase-scrambled images (accuracy: 36%, $t_{(5)} = 5.3, p < 0.01$), whereas the PPA classification accuracy was at chance (accuracy: 28%, $t_{(5)} = 1.7, p > 0.1$) and significantly lower than V1 ($t_{(5)} = 3.2, p < 0.05$). Because there was no object or scene information in the phase-scrambled images and the classification performance was at chance, no areas showed interpretable patterns of confusion errors (t test between types of errors: PPA, $t_{(5)} = 1.3$, $p = 0.28$; LOC, $t_{(5)} = 1.45$, $p = 0.2$). As expected, V1 showed higher confusion errors between the phase-scrambled versions of open and closed scenes, which was the information most available in the amplitude spectrum ($t_{(5)} = 5.1, p < 0.01$). This is consistent with the idea that PPA relies on higher-order scene information, whereas low-level statistics are sufficient for scene classification in V1.

In the second control experiment, we tested the opposite dissociation. We artificially added horizontal or vertical lines onto the images to add accidental low-level features, so that the low-level visual discriminability across the natural and urban images (that have different orientation of lines on top) will be as large as the already large low-level visual discriminability across open and closed images. If both the PPA and V1 classification relied on the low-level visual differences across conditions, then we would expect the classification performance in both areas to significantly improve compared with the original images that had less visual difference for the four conditions overall. Contrary to this prediction, we observed a significant boost of performance in V1 compared with the original experiment (64 vs 50% in the main experiment) but not in the PPA (54 vs 51% in the main experiment; interaction of the PPA and V1 across experiments was significant, $F_{(1,12)} = 5.6, p < 0.05$). V1 accuracy was also significantly higher than the PPA in control experiment 2 ($t_{(6)} = 2.7$, $p < 0.05$). Thus, the fact that adding low-level image features, such as oriented bars, improved V1 but not PPA performance strengthens the conclusion that the PPA is sensitive to higher-level scene information, not simply the low-level statistics of the images (Epstein, 2005). Moreover, the confusion errors for the PPA and LOC completely replicated the main experiment, because these areas are essentially invariant to the addition of the lines. There was a significant interaction across area (PPA or LOC) and types of errors (confusion within the same boundary vs confusion within the same content, $F_{(1,6)} = 24.7, p < 0.005$). The t tests between the two types of errors (confusion within the same boundary vs confusion within the same content) within each ROI was also significant in the PPA (48% confusion within the same boundary vs 32% confusion within the same content; $t_{(6)} = 3.2$, $p < 0.05$) and in the opposite direction in the LOC (30% confusion within the same boundary vs 49% confusion within the same content; $t_{(6)} = -2.6, p < 0.05$).

Together, these additional experiments show that, although V1 classification primarily relies on low-level image statistics independent of the scene spatial layout or content, the PPA is sensitive to the removal of higher-order information (control experiment 1) but not the addition of low-level features (control experiment 2). Therefore, the results support our hypothesis that the PPA represents higher-level scene information, such as the spatial boundaries, and does not merely reflect V1 representations.

Discussion

Using multivoxel pattern analysis to examine the neural basis of natural scene representation, we found a differential coding of scene features throughout the ventral visual cortex, including V1, PPA, LOC, RSC, and FFA. Interestingly, all these regions accurately classified the spatial boundary and content of a scene, across different semantic categories, suggesting that these brain regions are sensitive to both of these high-level scene properties. Importantly, the degree to which a specific property was used to classify scenes was different across the regions of interest. We found that spatial boundary information was primarily represented in the PPA and RSC and scene content primarily represented in the LOC and FFA. This suggests a complementary role for scene-selective areas and object-related areas in scene representation.

The possibility of a complementary scene representation with objects and scene content processed in a separate pathway from the spatial boundary is similar to that proposed by recent models of scene recognition and visual search (Torralba, 2003; Torralba et al., 2006; Vogel and Schiele, 2007; Greene and Oliva, 2009a). In such models, both spatial boundary and object information is extracted from a scene, often in parallel, and then integrated to arrive at a decision about the identity of the scene or where to search for a particular object. This convergence between fMRI data and scene recognition models suggests a possible computa-

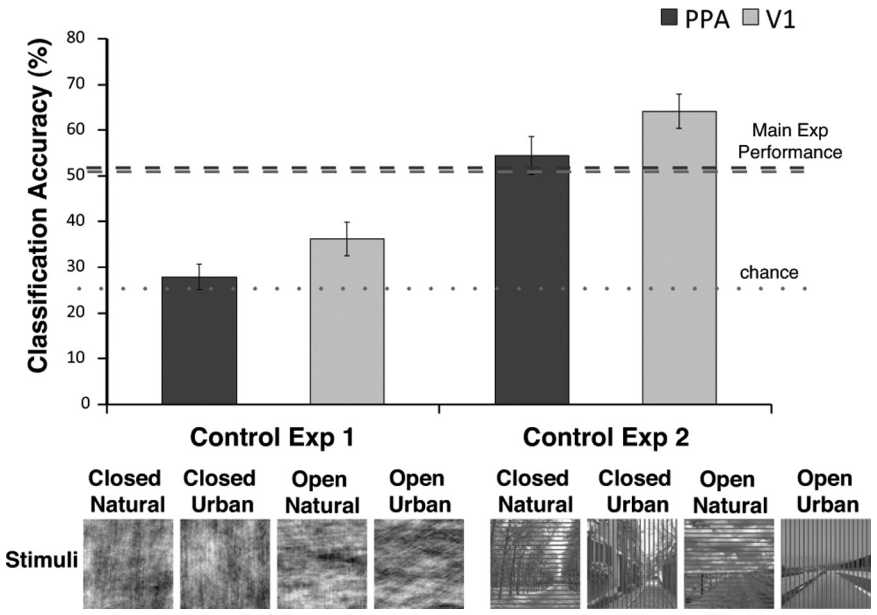

Figure 6. Control experiment 1: classification for the phase-scrambled images significantly decreased. Whereas the PPA performance dropped to chance, V1 still performed above chance classification. Control experiment 2: when lines were added to create low-level differences across the images, V1 classification performance significantly increased, whereas PPA performance did not change, suggesting that the PPA classification uses information beyond the low-level visual similarity. Examples of stimuli for each condition are shown.

tional role for the separate representations observed in many studies of the LOC and PPA.

Multivoxel pattern analysis and confusion errors

Previous studies using multivoxel pattern classification found that activity patterns in fMRI data can accurately classify between oriented lines (Kamitani and Tong, 2005; Kamitani and Tong, 2006), object categories (Haxby et al., 2001), and even object exemplars within a category (Eger et al., 2008). Recent fMRI studies with natural images have shown that high-level scene processing areas can classify across basic-level scene categories, such as mountains, beaches, or buildings (Walther et al., 2009). In addition, using generative models of the fMRI response in early visual cortex has allowed the prediction of a viewed image from brain data (Kay et al., 2008) and even some reconstruction of images from these data (Naselaris et al., 2009).

In the present study, rather than attempt to classify images into semantic categories or to recognize their identity exactly, we focus on global properties of scenes (Oliva and Torralba, 2001; Greene and Oliva 2009a). We thus provide the first neural evidence of a coding scheme of visual scenes based on structural and content properties. In other words, we show that scene and object areas in the brain not only represent the semantic category of a scene (forest, mountain) but also its spatial boundaries (open, closed) and degree of naturalness (urban, natural).

Analyzing the errors made by a classifier demonstrates that it is possible to probe the nature of neural representations in various ventral visual areas beyond the level of classification accuracy. By examining the different pattern of errors for our conditions, we can dissociate multiple levels of image representation coexisting within a single image and test the extent to which a specific property is coded in the brain. Reliable patterns of errors between scenes with a shared visual property (such as similar boundaries or content) provide evidence for that property being coded.

Naturalness and the role of objects

Throughout the paper, we have focused on naturalness as expressing the content of the scene: tree, bush, rock, grass, etc., make the scene content natural, whereas manufactured surfaces and man-made objects create urban content. However, we are not suggesting that the naturalness dimension reflects solely object processing. Naturalness is correlated with visual features at all levels of processing, from different distributions of orientations (Torralba and Oliva, 2003; Kaping et al., 2007) to different textural and surface components (Torralba and Oliva, 2002) in addition to different objects.

Recent behavioral work suggests a prominent role for global properties in the rapid analysis of visual scenes (Renninger and Malik, 2004; Greene and Oliva, 2009a,b). Indeed, global properties of scene shape boundary, such as openness, perspective, or mean depth, and global properties of scene content, such as the naturalness of an environment can influence rapid basic-level scene categorization. Human observers are more likely to confuse scenes that share a similar global scene property (e.g., a field and a coast are both open environments) than scenes that share same regions (e.g., trees can be shared by field and forest). In addition, human observers can classify the naturalness of a scene after a briefer exposure than is required for individual objects to be recognized (Joubert et al., 2007; Greene and Oliva, 2009b), and the naturalness of a scene is a property that is selectively adaptable (Greene and Oliva, 2010), suggesting that, at a high level of visual processing, naturalness is a unified construct.

The fact that LOC is both responsive to scenes and can accurately classify scene categories based on either content or spatial boundary provides some suggestive evidence that this region might play a more integrative role in scene understanding than simply representing object form information.

The role of other regions

It is interesting to note that all the visual regions we examined, including those that were not part of our original hypothesis, showed above-chance classification for scene properties. Of particular interest is the RSC, which has been found to be involved in processing navigationally relevant properties of scenes (Epstein, 2008; Park and Chun, 2009; Vann et al., 2009) and context (Bar and Aminoff, 2003). Indeed, the patterns of confusion errors in RSC were similar to the pattern observed in the PPA. However, the overall classification performance of RSC was significantly lower than the PPA ($t_{(4)} = 6$, $p < 0.01$), and the difference in confusion errors between spatial boundary confusions and content confusions were not significant. This suggests that, although the RSC may be involved in representing spatial boundary or correlated dimensions, it is less consistent across observers than the PPA.

The FFA also performed above-chance level for scene classification, with more confusion errors among scene content. This fits with similar results finding above-chance classification performance for object categories in FFA that have been reported by

Haxby et al. (2001) and Reddy and Kanwisher (2007). However, we should remain cautious about potential involvement of the FFA in scene processing. For example, the current result could potentially be triggered by the presence of pedestrians in urban scenes (indeed, people are a consistent object for outdoor urban environments). In addition, there is strong evidence in the literature for the face selectivity of the FFA (for review, Kanwisher, 2010), and it is important to note that the overall BOLD response in FFA for scenes is well below that for faces (in both our own data and that of many previous studies, including Kanwisher et al., 1997). In contrast, regions such as LOC respond nearly as strongly to scenes as to objects, even in data specifically selected to show a large response to objects (data from localizer vs main experiment).

Conclusion

In summary, the current study demonstrates a distributed and complementary neural coding of scene information, mediated via global properties of spatial boundary and content. By studying the patterns of errors made by a classifier, we conclude that the spatial boundary information was primarily used in the PPA for scene classification, whereas the content was primarily used in the LOC. Such complementary functions of these regions support the idea that observers concurrently process the global spatial boundary of a scene, which is important for navigation, as well as the content properties of a scene, which may be important for object identification and directed action.

References

Aguirre GK, Zarahn E, D'Esposito M (1998) An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. Neuron 21:373–383.

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

Eger E, Ashburner J, Haynes JD, Dolan RJ, Rees G (2008) fMRI activity patterns in human LOC carry information about object exemplars within category. J Cogn Neurosci 20:356–370.

Epstein RA (2005) The cortical basis of visual scene processing. Vis Cogn 12:954–978.

Epstein RA (2008) Parahippocampal and retrosplenial contributions to human spatial navigation. Trends Cogn Sci 12:388–396.

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

Epstein R, Harris A, Stanley D, Kanwisher N (1999) The parahippocampal place area: Recognition, navigation, or encoding? Neuron 23:115–125.

Fei-Fei L, Iyer A, Koch C, Perona P (2007) What do we perceive in a glance of a real-world scene? J Vis 7:1–29.

Greene MR, Oliva A (2009a) Recognition of natural scenes from global properties: seeing the forest without representing the trees. Cogn Psychol 58:137–176.

Greene MR, Oliva A (2009b) The briefest of glances: the time course of natural scene understanding. Psychol Sci 20:464–472.

Greene MR, Oliva A (2010) High-level aftereffects to global scene properties. J Exp Psychol Hum Percept Perform 36:1430–1442.

Grill-Spector K, Kushnir T, Edelman S, Itzchak Y, Malach R (1998) Cue-invariant activation in object-related areas of the human occipital lobe. Neuron 21:191–202.

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.

Janzen G, van Turennout M (2004) Selective neural representation of objects relevant for navigation. Nat Neurosci 7:673–677.

Joubert OR, Rousselet GA, Fize D, Fabre-Thorpe M (2007) Processing scene context: fast categorization and object interference. Vis Res 47:3286–3297.

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

Kamitani Y, Tong F (2006) Decoding seen and attended motion directions from activity in the human visual cortex. Curr Biol 16:1096–1102.

Kanwisher N (2010) Functional specificity in the human brain: a window into the functional architecture of the mind. Proc Natl Acad Sci U S A 107:11163–11170.

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.

Kaping D, Tzvetanov T, Treue S (2007) Adaptation to statistical properties of visual scenes biases rapid categorization. Vis Cogn 15:12–19.

Kay KN, Naselaris T, Prenger RJ, Gallant JL (2008) Identifying natural images from human brain activity. Nature 452:352–355.

Kourtzi Z, Kanwisher N (2000) Cortical regions involved in processing object shape. J Neurosci 20:3310–3318.

Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci 12:535–540.

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.

Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL (2009) Bayesian reconstruction of natural images from human brain activity. Neuron 63:902–915.

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, Chun MM (2009) Different roles of the parahippocampal place area (PPA) and retrosplenial cortex (RSC) in scene perception. Neuroimage 47:1747–1756.

Reddy L, Kanwisher N (2007) Category selectivity in the ventral visual pathway confers robustness to clutter and diverted attention. Curr Biol 17:2067–2072.

Renninger LW, Malik J (2004) When is scene identification just texture recognition? Vis Res 44:2301–2311.

Sereno MI, Dale AM, Reppas JB, Kwong KK, Belliveau JW, Brady TJ, Rosen BR, Tootell RB (1995) Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. Science 268:889–893.

Spiridon M, Kanwisher N (2002) How distributed is visual category information in human occipital-temporal cortex? An fMRI study. Neuron 35:1157–1165.

Tong F, Nakayama K, Vaughan JT, Kanwisher N (1998) Binocular rivalry and visual awareness in human extrastriate cortex. Neuron 21:753–759.

Torralba A (2003) Contextual priming for object detection. Int J Comput Vis 53:169–191.

Torralba A, Oliva A (2002) Depth estimation from image structure. IEEE Trans Pattern Anal Mach Intell 24:1226–1238.

Torralba A, Oliva A (2003) Statistics of natural images categories. Network 14:391–412.

Torralba A, Oliva A, Castelhano MS, Henderson JM (2006) Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol Rev 113:766–786.

Vann SD, Aggleton JP, Maguire EA (2009) What does the retrosplenial cortex do? Nat Rev Neurosci 10:792–802.

Vinberg J, Grill-Spector K (2008) Representation of shapes, edges, and surfaces across multiple cues in the human visual cortex. J Neurophysiol 99:1380–1393.

Vogel J, Schiele B (2007) Semantic modeling of natural scenes for content-based image retrieval. Int J Comput Vis 72:133–157.

Vul E, Harris C, Winkielman P, Pashler H (2009) Puzzlingly high correlations in fMRI studies of emotion, personality and social cognition. Perspect Psychol Sci 4:274–290.

Walther DB, Caddigan E, Fei-Fei L, Beck DM (2009) Natural scene categories revealed in distributed patterns of activity in the human brain. J Neurosci 29:10573–10581.