Evidence for participation by object-selective visual cortex in scene category judgments

Drew Linsley

Sean P. MacEvoy

Scene recognition is a core function of the visual system, drawing both on scenes’ intrinsic global features, prominently their spatial properties, and on the identities of the objects scenes contain. Neuroimaging and neuropsychological studies have associated spatial property-based scene categorization with parahippocampal cortex, while processing of scene-relevant object information is associated with the lateral occipital complex (LOC), wherein activity patterns distinguish between categories of standalone objects and those embedded in scenes. However, despite the importance of objects to scene categorization and the role of LOC in processing them, damage or disruption to LOC that hampers object recognition has been shown to improve scene categorization. To address this paradox, we used functional magnetic resonance imaging (fMRI) to directly assess the contributions of LOC and the parahippocampal place area (PPA) to category judgments of indoor scenes that were devoid of objective identity signals. Observers were alternately cued to base judgments on scenes’ objects or spatial properties. In both LOC and PPA, multivoxel activity patterns better decoded judgments based on their typically associated features: LOC more accurately decoded object-based judgments, while PPA more accurately decoded spatial property-based judgments. The cue contingency of LOC decoding accuracy indicates that it was not an outcome of feedback from judgments and is instead consistent with dependency of judgments on the output of object processing pathways in which LOC participates.

Introduction

Identifying the local environment is critical to daily life. It allows us to navigate to a destination, know when we have arrived, and engage in situation-appropriate behavior at each point along the way. Although both navigation through and interaction with the world rely at times on the ability to identify specific places (such as our own home or office), very often we face the more general task of recognizing the kind of place we are in, such as when identifying a bathroom among the rooms of an unfamiliar house. This particular type of identification is known as scene categorization, with the term scene referring to an environmental space that is navigated through or acted within and with the term categorization referring to the assignment of such spaces to base-level taxonomic groups with labels such as kitchen, bathroom, forest, and beach (Gosselin & Schyns, 2001; Rosch, Mervis, Gray, Johnson, & Boyesbraem, 1976; Tversky & Hemenway, 1983).

Previous studies have shown that scene categorization draws on two information resources. The first and perhaps most obvious resource is formed by the kinds of objects scenes contain. Even beyond those that are essentially defined by the presence of a single class of object (e.g., bedrooms), introspection tells us that many scene categories are closely associated with particular types of objects (e.g., kitchens with stoves, refrigerators, and toasters). Indeed, scene categorization has been traditionally framed as an outcome of operations on scenes’ objects (Biederman, 1987; Biederman, Blickle, Teitelbaum, & Klatsky, 1988; Biederman, Mezzanotte, & Rabinowitz, 1982; De Graef, Christians, & d’Ydewalle, 1990; Friedman, 1979), and it comes as little surprise that the removal of strongly associated objects (e.g., placing a mask over the refrigerator in a photograph of a kitchen) hampers scene categorization (MacEvoy & Epstein, 2011), as does the deliberate insertion or accidental presence of incongruent objects (Davenport & Potter, 2004; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007).

The second resource is formed by scenes’ global features, principally their spatial properties such as
depth and openness (Fei-Fei, Iyer, Koch, & Perona, 2007; McCotter, Gosselin, Sowden, & Schyns, 2005; Oliva & Schyns, 2000; Oliva & Torralba, 2001, 2006; Schyns & Oliva, 1994; Vogel, Schwaninger, Wallraven, & Bülthoff, 2007). Observers can categorize scenes on the basis of these properties (Greene & Oliva, 2009a), and artificial scene-categorization algorithms that emphasize them tend to generate patterns of scene judgments similar to those of human observers (Greene & Oliva, 2009a; Renninger & Malik, 2004). The use of spatial properties to help categorize scenes is consistent with studies showing that scenes can be recognized essentially as quickly as the objects they contain (Joubert et al., 2007; Potter, 1975; Potter & Levy, 1969), and that scenes are more accurately recognized when viewed as coherent wholes than when cut into pieces, even pieces large enough that individual objects remain recognizable (Biederman, 1972).

Processing of these resources has been associated with distinct areas within the visual system, with object recognition most closely linked to the lateral occipital complex (LOC) (Carlson, Schrater, & He, 2003; Grill-Spector, Kushnir, Edelman, Itzchak, & Malach, 1998; James, Culham, Humphrey, Milner, & Goodale, 2003; Kourtzi & Kanwisher, 2000; Malach et al., 1995; Pitcher, Charles, Devlin, Walsh, & Duchaine, 2009) and processing of scenes’ global properties most closely associated with the parahippocampal place area (PPA) (Aguirre, Zarahn, & D’Esposito, 1998; Epstein, 2008; Epstein & Kanwisher, 1998; Harel, Kravitz, & Baker, 2012; Ishai, Ungerleider, Martin, Schouten, & Haxby, 1999; Kravitz, Peng, & Baker, 2011; Park, Brady, Greene, & Oliva, 2011). Given the importance of both objects and spatial properties to scene categorization, it stands to reason that decision processes would draw upon information encoded in each of these regions during scene categorization. Indeed, despite alternative interpretations of the affinity of PPA for scenes (E. Aminoff, Gronau, & Bar, 2007; E. M. Aminoff, Kveraga, & Bar, 2013; Mullally & Maguire, 2011), the case for its role in scene categorization is strong: Not only has multivoxel pattern analysis (MVPA) of functional magnetic resonance imaging (fMRI) data shown that PPA activity patterns differ reliably among scene categories (Epstein & Morgan, 2012; MacEvoy & Epstein, 2011; Walther, Caddigan, Fei-Fei, & Beck, 2009; Walther, Chai, Caddigan, Beck, & Fei-Fei, 2011) but damage to PPA has been shown to disrupt scene recognition (Aguirre & D’Esposito, 1999; Epstein, DeYoe, Press, Rosen, & Kanwisher, 2001; Habib & Sirigu, 1987; Hécaen, Tzortzis, & Rondot, 1980; Mendez & Cherrier, 2003). In contrast, despite the demonstrated dependence of scene recognition on scenes’ object contents, and uncontroversial evidence for an involvement of LOC in the recognition of objects, no study has yet provided direct evidence of a role for LOC in scene categorization. In fact, disruptions to LOC produced by either injury or magnetic stimulation, which predictably hamper object recognition, have been shown to actually enhance scene categorization (Mullin & Steeves, 2011; Steeves et al., 2004). These results raise the possibility that the contribution of object-based information to scene recognition may be mediated by regions other than LOC, including possibly the PPA itself, in which activity patterns have also been shown to contain information about object identity (Harel et al., 2012; MacEvoy & Epstein, 2009).

To understand the contribution of LOC to scene category judgments, we used fMRI to record patterns of brain activity while participants performed a two-alternative forced-choice categorization task on computer-generated scenes. Adapting the design of a previous study targeting brain regions involved in motion judgments (Serences & Boynton, 2007), a large majority of scenes were configured to unambiguously be either kitchens or bathrooms, and the activity patterns evoked by these were used to train a pattern classifier to decode participants’ judgments of rarely shown scenes that were used to train a pattern classifier to decode participants’ judgments of rarely ambiguous category identity. Critically, participants were periodically instructed to base their category judgments on either scenes’ object contents or their spatial properties. Varying the cued resource in this way allowed us to differentiate between cortical regions whose activity patterns merely followed category judgments and those that contributed either object- or spatial property-based information to judgments. Specifically, we predicted that a general role for LOC in contributing information to scene category judgments would be evident in significantly greater behavioral decoding accuracy when participants were cued to base judgments on scenes’ object contents versus their spatial properties.

Materials and methods

Participants

Eleven participants (eight female, aged 19–29 years) gave written informed consent in compliance with procedures approved by the Boston College Institutional Review Board. In addition to satisfying typical selection criteria including normal or corrected-to-normal visual acuity, right handedness, and no history of neurological disease, these participants also passed a web-based screening procedure, described in the next section, to ensure that they distinguished between kitchens and bathrooms on the basis of size. Participants were paid $60.
Visual stimuli

The general strategy of our experiment was to record functional brain volumes while participants judged the category identities of rooms whose object contents and spatial properties each either contained a signal for scene category or did not and while we varied which of those information resources participants were cued to base their judgments on. To reduce this task to a two-alternative format, we first needed to identify two scene categories that were distinguishable from each other by both their spatial properties and object contents. Crowdsourced ratings of the perceived sizes of common indoor rooms, collected as part of another study (Linsley & MacEvoy, 2014), showed that kitchens and bathrooms have different, albeit overlapping, distributions of real-world sizes (Supplementary Figure 1A), making room size a potentially useful cue when distinguishing between exemplars of these categories. Because these two categories are also associated with very different object sets (Linsley & MacEvoy, 2014), they formed an appropriate pair for our experiment. However, because it is possible that not all people distinguish between these categories on the basis of spatial properties, potential participants were directed to a website where they were asked to judge whether the size of each of 100 computer-generated rooms was more similar to that of a kitchen or a bathroom. These scenes were selected randomly from a library of 300 empty rooms ranging in simulated floor area from 4.5 square meters to 15 square meters in 0.035 square meter increments. Rooms were rendered using Trimble Sketchup (www.sketchup.com), IRender nXt 4.0 (www.renderplus.com), and custom Ruby scripts. Each room was populated with a wavelet-scrambled (Honey, Kirchner, & VanRullen, 2008) version of a couch, lamp, and dresser in order to enhance its realism by providing some object-like forms that nonetheless carried no objective information relevant to the kitchen/bathroom task. Each potential participant’s proportion of kitchen judgments as a function of room floor area was fit with a four-parameter sigmoidal curve. Potential participants were invited to participate in the fMRI study if their fitted curves increased monotonically with floor area. Average data for the 11 participants who met this requirement are shown in Supplementary Figure 1B.

Because the room size perceived as intermediate between that of the average kitchen and of the average bathroom might differ among individuals, each invited participant’s online screening results were next used to customize the spatial properties of the scenes s/he would see in the scanner. For each participant, the room size at which his/her fitted function was halfway between its minimum and maximum values was taken as his/her spatial neutral point between bathrooms and kitchens. (As explained in following paragraphs, however, the design of our experiment did not ultimately require that determinations of neutral points be precise.) This room size was used to divide the original set of 300 rendered rooms into three groups: a spatially neutral group consisting of at least 50 consecutively sized rooms in a range centered on the neutral point and two spatially biased groups consisting of the rooms with sizes above and below this range, and therefore closer to each participant’s judgment of the size of the average kitchen and bathroom, respectively. These size groups were used to generate the five types of rooms that were shown during fMRI scans (Figure 1A): (a) rooms from the neutral group filled with kitchen-associated objects (spatially neutral unambiguous kitchens); (b) rooms from the neutral group filled with bathroom-associated objects (spatially neutral unambiguous bathrooms); (c) rooms from the spatially biased towards kitchens (i.e., larger than neutral) and filled with kitchen-associated objects (spatially biased unambiguous kitchens); (d) rooms from the group spatially biased towards bathrooms and filled with bathroom-associated objects (spatially biased unambiguous bathrooms); and (e) rooms from the neutral group filled with objects that were associated with neither kitchens nor bathrooms and obscured by wavelet scrambling (ambiguous rooms). Kitchen-associated objects were a refrigerator and combination stove/sink/cabinet unit and bathroom-associated objects were a toilet, sink, and shower stall. Objects associated with neither category were a couch, lamp, and chair.

It was critical to the design of our experiment that the scrambled objects in each ambiguous scene provided no visual information that might consistently bias judgments of the scene’s identity toward one category or the other. For example, at the extreme it was theoretically possible that wavelet scrambling could have made the chair in an ambiguous scene closely resemble a stove, which would mean that the scene was not actually visually ambiguous at all. Perhaps more likely, it was also possible by chance that the low-level statistical properties (e.g., spatial frequency spectra) of scrambled objects in some ambiguous scenes made those scenes more kitchen-like in terms of those properties, while making others more bathroom-like; if observers differentiated between kitchens and bathrooms on the basis of these properties, these scenes too could not appropriately be considered truly ambiguous. To validate ambiguous scenes’ category neutrality, we asked paid observers recruited through Amazon Mechanical Turk (n = 25) to categorize them as either bathrooms or kitchens based solely on their object contents. Our prediction was that any scene whose masked objects carried an objective signal along some visual dimension separating kitchens...
and bathrooms should elicit a significantly biased distribution of category judgments. Among the set of 81 ambiguous scenes that comprised all those used across all fMRI participants, none showed any such bias (binomial z-test, all uncorrected $p$ values > 0.0719).

Note that this procedure only confirmed that there was no category-relevant information contained in ambiguous’ scenes objects. Because ambiguous scenes were drawn from a range of room sizes bracketing each participant’s kitchen/bathroom spatial neutral point (as described previously), they were technically not ambiguous with respect to spatial properties. That is, the nonzero value of each room’s size deviation from a participant’s neutral point necessarily amounted to some spatial signal for scene identity. This was by design: Variation in spatial size was included in order to provide participants with some objective spatial signal that participants could use to make judgments when cued to attend to spatial properties (as described in the next section) in order to ensure that they did not revert to object-based judgments out of frustration. This emphasis on ensuring that attention was drawn away from objects, even at the expense of introducing a spatial signal to ambiguous scenes, stemmed from our main focus on measuring the contribution of LOC to scene category judgments. In practice, however, participants’ scene judgments were not generally sensitive

Figure 1. Visual stimuli and procedure. (A) Images of scene exemplars used in the fMRI experiment. During fMRI scans, participants made two-alternative forced categorical judgments on scenes populated by kitchen-associated objects (unambiguous kitchens), bathroom-associated objects (unambiguous bathrooms), or by scrambled versions of objects associated with neither category (ambiguous scenes). Unambiguous kitchens were sized to match the size (in terms of simulated floor area) considered by each participant to be neutral between kitchens and bathrooms, or to a larger size more similar to the average kitchen; unambiguous bathrooms were sized to the neutral point or to a smaller size similar to the average bathroom. Ambiguous scenes were always spatially neutral. (B) Experimental procedure. Stimulus events were organized into 66-s blocks, each of which began with text cueing subjects to base judgments of scene category either on rooms’ object contents (object blocks) or on their spaciousness (size blocks). Individual trials began with a subliminal reminder of the cued feature, followed by the scene to be judged, a 1/8 noise mask, and finally a notification to register a scene category decision by button press.
to variations in room size within the neutral range, as explained in the Results. The use of a range of sizes also insulated us against difficulties that might have arisen from inevitable inaccuracies in our measurements of participants’ neutral points. Whether due to measurement error or shortcomings of our function fitting, our determination of each participant’s spatial neutral point doubtlessly fell somewhere on one side of his or her true neutral point. As such, if we had always matched the sizes of ambiguous scenes to a single erroneous neutral value that was, for example, on the bathroom side of a participant’s neutral point, it might have relieved him or her from actively evaluating each scene in those groups.

Procedure

Subjects participated in eight fMRI scan runs. The five types of room stimuli described above, along with three-second null events, were ordered according to third-order counterbalanced de Bruijn sequences (Aguirre, Mattar, & Magis-Weinberg, 2011; MacEvoy & Yang, 2012), which provided pseudorandom sequences of the minimum length required to achieve three-back counterbalancing of the five stimulus and one null event types. Scenes were shown for 150 ms, followed by an 80 ms 1/f noise mask, and a 1238 ms period during which participants responded as quickly as possible by button press with their judgments of each scene’s category (Figure 1B; including the 32 ms instruction reminder discussed below, stimulus events lasted 1500 ms). The next stimulus or null event immediately followed this response period. Participants were told that some scenes would be more difficult to judge than others, but that all had a correct answer. No feedback was given. Each scan run contained 36 repetitions of each scene type and lasted 7 min 6 s, including 30 s of fixation at the end. Unique stimulus sequences were constructed for all eight scan runs for each subject.

In the critical manipulation, each scan run was divided into six 66-s blocks differing in the scene features participants were instructed to use when making scene category decisions. At the beginning of each block, participants were shown text lasting 3 s that instructed them to “Base decisions on rooms’ objects” or “Base decisions on rooms’ spaciousness.” An audio tone was delivered at the onset of instruction screens to alert participants to the instruction. To unobtrusively help participants keep track of which scene feature to focus on, each stimulus event within blocks began with display of one of the words “OBJECTS” or “SIZE” for 32 ms. Task block order was randomized within each run.

Scan sessions also included two functional localizer scans lasting 7 min 48 s each, during which subjects viewed blocks of color photographs of scenes, faces, common objects, and scrambled objects presented at a rate of 1.33 pictures per second (Epstein & Higgins, 2006). Localizer stimuli occupied the central 15° of visual space. Stimulus presentation and behavioral data collection was managed by custom MATLAB code using the Psychophysics Toolbox (Brainard, 1997).

MRI acquisition

All scan sessions were conducted at the Brown University MRI Research Facility using a 3T Siemens Trio scanner with a 32-channel head coil. Structural T1* weighted images for anatomical localization were acquired using a 3-D MPRAGE pulse sequences (TR = 1620 ms, TE = 3 ms, TI = 950 ms, voxel size = 0.9766 × 0.9766 × 1 mm, matrix size = 192 × 256 × 160). T2* weighted scans sensitive to blood oxygenation level-dependent (BOLD) contrasts were acquired using a gradient-echo echo-planar pulse sequence (TR = 3000 ms, TE = 30 ms, voxel size = 3 × 3 × 3 mm, matrix size = 64 × 64 × 45). Visual stimuli were rear projected onto a screen at the head end of the scanner bore and viewed through a mirror affixed to the head coil. The entire projected field subtended 24° × 18° at 1024 × 768 pixel resolution. Scene stimuli in the main experiment occupied the central 9.3° of visual space.

fMRI data analysis

Functional volumes were subjected to standard preprocessing routines in SPM8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/), including resampling slices in time to match the first slice of each volume, spatial realignment to the first volume of each scan, and spatial normalization to the Montreal Neurological Institute (MNI) template. Voxel time series were upsampling via cubic interpolation to 500 ms resolution and back-shifted by 4.5 s to compensate for hemodynamic delay. Single-trial voxel responses were then stored as the average signal at 0, 500, and 1000 ms post-stimulus. Voxelswise responses across trials were standardized to zero mean and unit standard deviation independently within each scan run.

Our general analysis strategy was to use MVPA to decode participants’ judgments of ambiguous scenes from activity patterns in several regions of interest (ROIs; see below for definitions). To maximize our potential ability to do so, we first used a feature selection algorithm to identify those voxels in each ROI that best discriminated between the actual categories of unambiguous scenes. For a set of spherical 3 mm
searchlight masks centered iteratively on each voxel (Kriegeskorte, Goebel, & Bandettini, 2006) in each ROI, we calculated the Mahalanobis distance between the activity pattern evoked by a single unambiguous scene in one of the eight experimental scans and each of two pattern reference clusters, defined as the sets of patterns accumulated across all presentations of unambiguous kitchens and all unambiguous bathroom from the remaining seven scans. For each searchlight position, average values of [between category distance – within category distance] were computed separately for unambiguous kitchens and unambiguous bathrooms in each held-out scan. These values were then averaged across held-out scans. Each searchlight mask in an ROI was thus associated with two values: the average [between category – within category] distance contrast for unambiguous kitchens stimuli, and the corresponding value for unambiguous bathrooms. The resulting set of two-dimensional vectors was subjected to k-means clustering to generate four clusters, theoretically corresponding to (a) high values along both category dimensions (i.e., high sensitivity to both bathrooms and kitchens), (b) high values along the bathroom dimension and low values along the kitchen dimension, (c) high values along the kitchen dimension and low values along the bathroom dimension, and (d) low values along both dimensions. In practice, the clusters that emerged did not necessarily fit these descriptions. All voxels in the cluster showing the highest average sensitivity to bathrooms and the cluster showing the highest average sensitivity to kitchens formed the set of voxels passed to further analysis, with the restriction that the passed set consist of at least seven voxels. If it did not, the cluster showing the second highest sensitivity to kitchens or to bathrooms was randomly chosen and its contents added to the set passed to further analysis. (This procedure was the outcome of our attempt to find a selection procedure less arbitrary than one based on a fixed target ROI size, e.g., 100 most discriminative voxels, or voxel statistical threshold, e.g., all voxels discriminating between categories with a $p$ value of less than 0.05). The identification of four clusters was based conceptually on a high/low split in sensitivity along each of the kitchen and bathroom sensitivity dimensions, while the seven voxel minimum was based on the number of voxels in each 3 mm searchlight mask. Note that while this feature selection procedure identified the voxels in each ROI that differentiated between unambiguous bathrooms or kitchens, it did not draw on any patterns evoked by the ambiguous scenes that were the targets of our analyses, and therefore cannot have introduced any bias into the measured relationship between those patterns and participants’ perceptual judgments.

To measure the ability of an ROI to decode each participant’s categorical judgments of ambiguous scenes, patterns consisting of selected voxels were passed to a leave-one-out classification procedure in which patterns evoked by correctly-identified unambiguous bathrooms and kitchens in seven of the eight scans formed reference sets to which patterns evoked by ambiguous scenes in a single held-out test scan were compared. The classifier was considered to have correctly decoded the participant’s response to an ambiguous scene if the reference set with the shorter Mahalanobis distance matched the participant’s decision.

**Statistical procedures**

Behavioral decoding accuracy was computed for patterns pooled across object and size task blocks, as well as within each block type. The statistical significance of overall classification accuracy for each ROI was assessed by accumulating accuracy scores over 10,000 permutations of participants’ judgments of ambiguous scenes. An ROI’s subject-averaged accuracy was considered significant if it exceeded the 95th percentile of the distribution of subject-averaged accuracies across label permutations; this single-tailed test was appropriate because accuracy rates lower than chance (50%) logically have no inferential significance. As elaborated in the Results, the critical measure of the contribution of an ROI to behavioral decisions was a difference between accuracies in object and size blocks that was in the direction of the ROI’s preferred feature (i.e., greater accuracy in object blocks for object-encoding ROIs, greater in size blocks for ROIs associated with scenes’ spatial properties). Because we thus had a clear hypothesis about the sign of such differences for almost all ROIs, values were considered significant if they fell within the $p < 0.05$ region of one tail of the distribution of differences accumulated across label permutations. A two-tailed test was used only for early visual cortex, for which we had no hypothesis about the sign of the block accuracy difference.

**Regions of interest**

ROIs were defined with an algorithmic approach to data from localizer scans (Julian, Fedorenko, Webster, & Kanwisher, 2012). Briefly, a group-level whole-brain map was generated for each contrast of interest (e.g., objects > [scenes & scrambled objects & faces]) in which each voxel was tagged with the proportion of individual participant’s $t$ values at that voxel that exceeded the 95th percentile of that participant’s whole-brain distribution of $t$ values. This volume was smoothed with a 3 mm FWHM Gaussian kernel then
parcellated with a watershed algorithm implemented in MATLAB. Resulting group-ROI volumes contained parcels corresponding to locations of commonly shared activations across subjects. To reduce individual subject influence on the group-ROI volumes, parcels generated from the activations of less than 50% of subjects were removed. For contrast volumes expected to contain more than one ROI, a k-means algorithm was applied to voxel distances to determine ROI borders, with k equaling the number of ROIs expected. Resulting ROI identities were determined through visual inspection by the authors.

Individual subject ROIs associated with a given contrast were defined from the intersection between the shared activation volume and each subject’s contrast map thresholded to keep only the top 5% of values. This procedure was applied to the contrasts of objects > [scenes & scrambled objects & faces] to define the lateral occipital (LO) and posterior fusiform (pF) subregions of LOC; to the contrast of scenes > [objects & scrambled objects & faces] to identify the PPA, retrosplenial complex (RSC), and transverse occipital sulcus (TOS); and scrambled objects > [scenes & objects & faces] to identify early visual cortex (EVC). After applying the feature selection algorithm described in the previous section, average ROI sizes in voxels were: right LO, 148; left LO, 109; right pF, 47; left pF, 14; right PPA, 42; left PPA, 39; right RSC, 116; left RSC, 118; right TOS, 42; left TOS, 40; right EVC, 75; left EVC, 96.

Results

fmRI

In a rapid event-related design (Figure 1B), participants viewed images of computer-generated rooms configured as unambiguous kitchens or bathrooms (80% of stimulus events), or as category-ambiguous scenes (20% of events); participants were asked to judge whether each was a kitchen or bathroom, periodically receiving instructions to base their judgments on scenes’ object contents (object blocks) or spatial scales (size blocks). Multivoxel activity patterns evoked by correctly judged unambiguous scenes in each ROI were used to train a pattern classifier to decode participants’ category judgments of ambiguous scenes from the patterns they evoked. An ROI was considered to contain information about participants’ decision states if the classifier output matched participants’ categorical judgments of ambiguous scenes at a rate significantly greater than chance. For data pooled across object and size blocks, activity patterns in a majority of ROIs satisfied this criterion (Figure 2). Among object-selective ROIs, these included the LO subdivision of LOC in both hemispheres, and the pF subdivision in the left hemisphere.

There are three possible reasons why patterns from an object-selective ROI would be able to decode participants’ judgments of ambiguous scenes, each of which makes a different prediction for relative decoding accuracy in object and size blocks. First, signals arising directly from participants’ decision output, at a cognitive or motor stage, might have fed back onto patterns in the ROI. If this were the case, the ROI should show equivalent decoding accuracies for object and size blocks, since the decisions themselves and corresponding motor outputs were identical in both. Second, activity patterns and behavioral decisions may both have been yoked to the spatial variability of ambiguous scenes, but independent of each other. As detailed in the Methods, ambiguous scene size varied randomly around each participant’s kitchen/bathroom spatial neutral point. Thus if decoding were driven by
scene spatial variability, we expected accuracy to be greater during size blocks, when spatial properties were attended, than during object blocks. Finally, behavioral judgments may have been determined, at least in part, by fluctuations in the state of neural activity in the ROI between more kitchen-like or bathroom-like states, driven by random neural events or by some unmeasured internal variable (e.g., the expected category of the stimulus). If this were the case, we expected decoding accuracy to be greater during object blocks than during size blocks, reflecting participants’ decision processes deferring to neural activity in regions that typically encode object information.

Consistent with this final explanation, decoding accuracy in right LO was significantly higher in object blocks than in size blocks (permutation test, \( p < 0.001 \); Figure 3C). It is critical to emphasize that this difference could not have resulted simply from a common dependence of activity patterns and scene judgments on some signal for scene category embedded in scenes’ objects. Objects in ambiguous scenes carried no information about scene category since they were not associated with either bathrooms or kitchens and furthermore had their identities obscured with a mask. In the absence of any such object-based category signal in the stimulus itself, the ability of right LO patterns to decode behavioral judgments in object blocks, but not size blocks, suggests that those judgments were at least in part determined by the state of neural activity in LO.

Figure 2. Decoding accuracies of subject category decisions for ambiguous scenes. A classifier was trained on multivoxel activity patterns elicited in each ROI by correctly identified bathrooms and kitchens across both block types and tested on patterns evoked by ambiguous scenes. Significantly above-chance decoding of subject judgments of ambiguous scenes’ identities was found in a majority of ROIs (circles). Gradient-filled bars mark central 95% of the distribution of accuracies generated across 10,000 permutations of decision labels for ambiguous scenes. Data marker diameters denote ROI sizes (inset); see Methods for average voxel counts. Error bars are \( SEM \* = p < 0.05; ** = p < 0.01; *** = p < 0.001 \).

Figure 3. Differential ROI behavioral decoding accuracies in object and size blocks. (A, B) Decoding accuracies for each ROI with above-chance accuracy for data collapsed across blocks (Figure 2), for object and size blocks, respectively. No hypothesis tests were executed on these data. (C) Differences between decoding accuracies in object and size blocks. Positive values denote higher accuracy in object blocks. Patterns in right LO were significantly better at decoding scene judgments that were based on object contents than on scenes’ sizes. PPA showed a significant difference for the reverse contrast. No other ROI demonstrated a significant accuracy difference between block types. Gradient bars reflect 95% range of permutation distributions for the quantity in each panel. Error bars are \( SEM \* = p < 0.05; *** = p < 0.001 \).
or at least in the object recognition network within which LO participates. (The importance of the distinction between these two possibilities is explored in the Discussion.)

It is unclear what variable determined whether LO assumed more bathroom-like or kitchen-like states in response to ambiguous scenes. As mentioned above, random neural events are one possibility. Another, however, is that LO patterns evoked by ambiguous scenes tended to adhere to the pattern evoked by the preceding unambiguous scene. While such a tendency by itself is not inconsistent with a dependence of behavioral judgments on neural activity in LO, it raises the possibility of an interpretive error. Specifically, it was possible that the ability of right LO patterns to decode behavior simply reflected adherence of LO patterns evoked by ambiguous scenes to those evoked by immediately preceding unambiguous scenes, correlated with a similar tendency of judgments of ambiguous scenes to adhere to those of preceding unambiguous ones, without any actual dependence of decisions on LO activity. This interpretation was lent plausibility by the fact that over 60% of judgments of ambiguous scenes matched those of their preceding unambiguous scenes.

To eliminate any contribution of such correlation to our results, we measured behavioral decoding accuracy in subsets of each participant’s trials that consisted of all ambiguous scenes judged different from the preceding unambiguous scene combined with an equal number of randomly drawn trials in which ambiguous scenes were judged the same as the preceding unambiguous scene. Because the resulting trial samples necessarily possessed no correlation between judgments of ambiguous scenes and preceding unambiguous scenes (i.e., no behavioral history effect), they could not possess any correlation between those judgments and LO patterns that was dependent upon a behavioral history effect. (In other words, where two variables B and C are both dependent upon variable A but otherwise unlinked, the correlation between B and C will be destroyed if the correlation between B and A is eliminated.) As such, if the accuracy differential between object and size blocks in right LO arose from an independent tendency of judgments and LO patterns to adhere to history, it should have been absent from the new samples. Instead, we found that the distribution of accuracy differentials accumulated over 1,000 samples contained no element with an object-minus-size block differential equal or less than zero. This does not allow us to infer that there was no contribution of behavioral response history to the ability of LO to decode behavioral responses. However, by analogy to the method of drawing inferences from bootstrap confidence intervals, it does allow us to infer that the probability that the true contribution of factors other than history was zero is less than 0.001.

Our results are not only consistent with a contribution to scene judgments by neural activity in LO, but indicate that the magnitude of this contribution varies with task demands. Mechanistically, this flexibility could have arisen in two ways. First, participants’ decision processes may have simply increased the weight given to neural signals from object-processing regions during object blocks. In this model, the amplitude of deviations by LO and associated regions into more kitchen-consistent or bathroom-consistent states was the same in object and size blocks, but those deviations were given greater weight by decision processes in object blocks than in size blocks. The alternative is that the weight given object codes by decision processes was identical across blocks, but that the magnitude of random activity deviations in LO was greater in object blocks, perhaps as a consequence of attentional mechanisms. To differentiate between these alternatives, for each participant we tagged the pattern evoked by each ambiguous scene with its corresponding perceptual judgment, and computed the average Euclidean distance, within each block type, across (a) all pairs of patterns with opposite judgment tags and (b) all pairs of patterns with the same judgment tag. We find that the [opposite minus same] average distance contrast for LO patterns was significantly greater during object blocks than during size blocks (two-tailed t test, \( t(10) = 2.43, p = 0.036 \)). While this does not exclude the possibility that the greater ability of right LO patterns to decode judgments made during object blocks resulted from greater weight given to LO-associated object codes by decision processes, it suggests that at least some portion of the enhancement arose from greater variability in object codes at their source during object blocks. There was no significant difference between blocks in the magnitude of overall BOLD signal in right LO, \( t(10) = 0.58, p = 0.575 \).

In contrast to right LO, both left LO and pF showed no significant differences between judgment decoding accuracies in object and size blocks. The absence of such differences does not warrant the conclusion that left LO and pF did not contribute to category judgments; it is possible that the absence of an object/size difference means that they contributed equally to judgments on both blocks types. However, it does leave open the possibility that their activity patterns simply followed participants’ behavior. Our experimental design cannot differentiate between these possibilities.

Among scene-selective ROIs, patterns in bilateral PPA and bilateral retrosplenial complex (RSC) decoded judgments of scene category at rates greater than
chance for data collapsed across object and size blocks (Figure 2). Using the reverse of the logic we applied to object-selective ROIs, we reasoned that if the ability of patterns in scene-selective ROIs to decode behavior reflected a contribution of those regions to decision processes, then they should have exhibited greater decoding accuracy during size blocks than during object blocks. PPA satisfied this condition in both hemispheres (Right PPA, \(p = 0.031\); Left PPA, \(p = 0.049\); Figure 3). However, the fact that ambiguous scenes varied in spatial properties makes it possible that the higher decoding accuracy in size blocks might reflect common enhanced dependence of both PPA patterns and behavioral judgments on spatial properties during size blocks, without any dependence of one on the other. For this explanation to be plausible, however, categorical judgments of ambiguous scenes in size blocks should show some consistent relationship to scenes’ sizes. The software that controlled stimulus presentation during the fMRI experiment did not store the sizes of rooms on each ambiguous trial, making a direct assessment of this relationship impossible. However, a retrospective analysis of decisions made during each participant’s preliminary screening stage showed no significant correlation between category judgments and scene sizes within the range of sizes ultimately selected to form each participant’s ambiguous scene set in the fMRI experiment, average Fisher-transformed \(R = 0.06, t(10) = 0.62, p = 0.551\). The minimal contribution of scene size variation over this range to behavioral judgments suggests that the differential between PPA decoding accuracy during object and size blocks reflects a dependence of scene decisions on PPA activity rather than a common dependency of PPA patterns and judgments on a stimulus-based signal.

The impact of history effects on PPA decoding accuracy of behavioral decisions was measured using the same analysis detailed above for LO. This resampling procedure allowed us to measure the range of PPA decoding task differentials after destroying task history correlations. As expected, none of the resampled object/size accuracy differentials were greater than zero, demonstrating that the increased impact of PPA patterns on decisions during size blocks was not a result of autocorrelation.

Unlike in LO, the enhanced influence of PPA activity patterns during size blocks did not coincide with increased variability in PPA patterns, i.e., the average same-decision pairwise pattern distance was not different from the average different-decision pairwise pattern distance, \(t(10) = 0.68, p = 0.253\). This suggests that the increased influence of PPA on scene decisions may depend more on a reweighting of PPA inputs to downstream decision processes than an increase in the variability of random PPA activity states.

In addition to object- and scene-selective ROIs, we also examined the ability of patterns in early visual cortex (EVC) to decode behavior. For data combined across blocks, patterns from bilateral EVC did decode participants’ judgments of ambiguous scenes at rates significantly above chance (Figure 2). However, we observed no significant difference in decoding accuracy in either hemisphere between object and size blocks (Figure 3).

Finally, we used a searchlight analysis to identify any brain areas outside our predefined ROIs that may have showed significant differences between behavioral decoding accuracy in object and size blocks. Within occipitotemporal cortex, the only cluster showing a significant difference was within the boundaries of the consensus right LO (Figure 4). Consistent with our ROI analysis, this cluster showed higher accuracy in object blocks than size blocks. The absence of a PPA cluster surviving multiple comparison adjustment is consistent with the weaker difference between blocks types shown by PPA compared to LO in our ROI analysis. We would like to emphasize again, however, that while these results are consistent with a causal relationship between LO activity patterns and category decisions, the absence of significant clusters in other brain areas is in no way prejudicial to a hypothesized role for them in scene categorization.
Behavioral

It is possible that participants may not have categorized scenes at all during object blocks, but simply adopted the strategy of categorizing one or more objects (or, in ambiguous scenes, scrambled objects) and then selecting the appropriate associated scene response. Although this possibility does not challenge the validity of our conclusion that participants' behaviors in object blocks were contingent on activity states in right LO and/or its associated object-processing networks, it does raise the concern that our results do not apply to scene recognition under normal circumstances, i.e., in the absence of explicit cues to attend to particular features. We provide a more detailed treatment of this issue in the Discussion. However, we observed no significant difference in reaction times for judgments of unambiguous scenes between object blocks and size blocks (681 ms vs. 661 ms, respectively; t(10) = 1.62, p = 0.135), both of which were similar to latencies reported previously for a similar categorization task (MacEvoy & Epstein, 2011). This suggests that scene judgments in object blocks did not suffer from unnatural detours through object-based decision pathways.

Discussion

A role for LOC in scene categorization is intuitive from the dependence of scene categorization on scenes' object contents (Biederman, 1972, 1987; Biederman et al., 1982; Davenport, 2007; Davenport & Potter, 2004; De Graef et al., 1990; Friedman, 1979; Joubert et al., 2007) coupled to the role of LOC in object processing (Diana, Yonelinas, & Ranganath, 2008; Grill-Spector, 2003; Grill-Spector, Kourtzi, & Kanwisher, 2001; Haxby et al., 2001; MacEvoy & Epstein, 2009, 2011; Malach et al., 1995; Park et al., 2011; Peelen, Fei-Fei, & Kastner, 2009; Sayres & Grill-Spector, 2008). Previous research had moreover shown that activity patterns evoked in LOC by scenes can be decoded from patterns evoked by isolated objects associated with those scenes (MacEvoy & Epstein, 2011; Peelen et al., 2009), indicating that neural activity in LOC was a rich information resource for object-based scene categorization. Whether this resource was actually used during scene categorization, however, was unclear. The ability we observed of right LO patterns to decode object-based decisions in the absence of any objective object-based signal for scene identity indicates that it is.

It is critical that we very clearly explain exactly how we arrive at this inference. The scene judgments participants made in our study were, as all behaviors of course are, products of particular brain states. The proximal cortical cause of any visually-guided behavior is obviously activity in motor cortex, while the ultimate cause is the pattern of neural activity on the retina(s), with the two connected by a long chain of neural events that is also subject to influence by internal variables (i.e., those not related to the retinal image). Any particular task will cause this chain to be configured such that a retinal image containing an objective signal favoring a particular behavioral response will perturb this chain into a state that produces the appropriate activity profile in motor cortex. As such, activity in an area that acts as a link in this chain (or set of chains, in the likely scenario that behavior depends on visually-evoked activity in multiple diverging/converging pathways) will be able to decode visually-guided behavior. It would be a mistake, however, to make the reverse inference that the ability of activity in a brain area to decode behavior marks that area as a link in the chain that converts visual information into behavior. This is due to the common input problem: Both behavior and activity in an area may be yoked to the retinal signal and therefore predict each other, without any causal relationship between them. An extreme manifestation of this is that, in an experiment similar to ours, activity patterns in the brain of one observer will undeniably decode another observer’s judgments of unambiguous scenes, so long as the two observers viewed the same sequence of scenes.

In our study, we avoided the common input problem by focusing upon the ability of neural activity patterns to predict judgments of ambiguous scenes that were verified to contain no useful objective signal for scene category. In the absence of any such signal, participants’ judgments necessarily resulted from perturbations of the behavior-generating networks by nonvisual variables; these may have been the product of random neuro-chemical events or cognitive factors such as stimulus expectation. (This is not to say that observers did not genuinely perceive some ambiguous scenes as kitchens and others as bathrooms, but that such perceptions were themselves necessarily the outcomes of internal variables. That is, while a participant may have felt that a tall whitish object in a scene was likely a refrigerator and the scene containing it therefore a kitchen, it must have been some internal variable that led her to feel that the object was more refrigerator-like than, say, shower-like in the first place.) We designed our study specifically to ensure, as much as possible, the networks generating decisions of ambiguous scenes were the same as those generating decisions of unambiguous ones: Ambiguous scenes were relatively rare and unpredictably timed, and participants were led to believe that ambiguous scenes contained real, albeit degraded, information about scene identity. The only difference between the two conditions is the origin of the dominant variable that perturbed the networks formed by these areas into kitchen-like or
bathroom-like states: For unambiguous scenes it was some component of the retinal image, while for ambiguous scenes it was some internal variable, or combination of such variables.

Even after ensuring that decisions are based on internal variables, however, it would be a mistake to infer that activity in a brain area determined judgments of ambiguous scenes simply because its patterns predicted them. This is due to the possibility that activity patterns in that area might simply have followed behavior. Our experiment was designed specifically to allow us to avoid this confound. Because the behavioral output was the same in both object and size blocks, simple feedback from scene decisions would have led to equivalent prediction accuracies for both block types. (This is based on the assumption that while behavioral outputs are selectively coupled to neural activity in those visual networks that typically contain the most information about the cued visual variable [scenes’ objects or spatial properties], feedback from behavior is not similarly gated by task.) This is why the critical diagnostic feature identifying LO as a link in the chain between visual input and behavior, instead of simply as a follower of behavior, was significantly greater decoding accuracy in object blocks over size blocks.

It is critical to emphasize that this conclusion does not involve nor derive from any inference that LO was the source or point of entry of the internal variable that ultimately determined participants’ responses to ambiguous scenes; indeed, any such inference would be unsupported. It remains possible and perhaps even likely that the ability of LO to predict scene judgments reflected feedback from downstream regions in which the perturbing effect of those variables was initially brought to bear. These may have been brain areas encoding higher level object features than encoded in LO, or even regions encoding semantic labels of the perceived identities of objects in ambiguous scenes. Regardless of the exact origin of the perturbations, however, we can safely conclude that it must have been within the object-specific processing chain in which LO sits. This is the only explanation for why LO patterns predicted judgments only in object blocks, and necessarily leads to the conclusion that object-based judgments of ambiguous scenes were at least in part dependent on the output of this chain. Inasmuch as the role of LO as a conduit for visual information into this object-recognition chain is uncontroversial, we may therefore infer, by transitivity, that LO codes participate in object-based scene categorization when scenes’ retinal images contain object-based signals for their identity.

Although this conclusion may seem obvious in light of previous studies linking LOC to object recognition, including judgments of ambiguous or imagined objects (Cichy, Heinzle, & Haynes, 2012; Reddy, Tsuchiya, & Serre, 2010; Stokes, Thompson, Cusack, & Duncan, 2009; Williams, Dang, & Kanwisher, 2007), our experiment was necessary in light of a pair of studies that called into question the contribution of LOC by demonstrating that scene categorization accuracy improved following damage (Steeves et al., 2004) or disruptive transcranial magnetic stimulation (Mullin & Steeves, 2011) to LOC that simultaneously degraded object recognition. One interpretation of these prior results is that the contribution of objects to scene categorization is simply not mediated by LO. This interpretation is lent plausibility by the presence of information about objects’ identities in PPA activity patterns (Harel et al., 2012; MacEvoy & Epstein, 2009), which could therefore potentially support object-based scene categorization without LOC involvement. Our results do not support this interpretation, and instead suggest that the benefit to scene recognition accompanying LOC disruption likely reflected the particular scene categorization task used in the studies that reported it, in which participants were solely asked to judge whether scenes were man made or natural. This superordinate categorization task is likely to have drawn chiefly on scenes’ global properties rather than their object contents (Greene & Oliva, 2009a, 2009b; Torralba & Oliva, 2003). That TMS to LO during such a task should not have impaired scene recognition is consistent with our finding of no ability of LO to predict scene judgments when participants were cued to base judgments on scenes’ spatial properties.

As mentioned in the Results, it could be argued that our results do not allow us to infer a contribution to real-world scene categorization since we provided an explicit cue for observers to base judgments on scenes’ objects (or spatial properties) that is absent during normal behavior. This argument does not deny that objects provide cues to real-world scene categorization, nor even that scene categorization may flexibly reweight the relative influence of objects and spatial properties, but simply contends that the explicit instructions we provided to our participants linked their behavioral responses to object recognition pathways in a way that might not occur typically. Our first answer to this argument is that a strong link between the output of object recognition pathways and scene categorization is generally assumed by extant theories of object-based scene categorization, which require extraction of objects’ identities as precursor to activation of scene schemata or context frames (Bar, 2004; Biederman, 1972; Davenport & Potter, 2004; Henderson & Hollingworth, 1999; Mudrik, Lamy, & Deouell, 2010; Potter, 1975, 1976). Our second answer is that cues to attend to one scene feature or another are frequently present during real-world scene categorization, simply in different forms than they took in our study. In most real-world contexts, the range of scene categories an observer may need to choose among is
quite small, and prior knowledge of the scenes in that range provides a strong cue to attend to the scene feature (object or spatial properties) that best differentiates among them. For instance, someone navigating through an unfamiliar house knows that a room she is about to peer into can only plausibly belong to one of perhaps a handful of categories (kitchen, bathroom, bedroom, etc.). Based on experience with those categories that tells her that spatial properties will be unlikely to differentiate conclusively among them (Linsley & MacEvoy, 2014), she therefore also knows that the optimal strategy for identifying the room is to attend to the objects it contains. There is no reason to believe that the fact that this context cue is nonlexical makes it any less capable of linking scene judgments of object recognition than the block cues in our study. Indeed, the absence of an average reaction time difference between object and size blocks suggests that the mechanism of scene categorization engaged by object blocks was not out of the ordinary.

One surprising feature of our results was the right-preferring laterality of the ability of LO to predict decisions, which appears to conflict with results suggesting a greater involvement of the left hemisphere in categorical processing (Grossman et al., 2002; Laeng, Zarrinpar, & Kosslyn, 2003; Li, Ostwald, Giese, & Kourtzi, 2007; Seger et al., 2000). However, because our results do not address where categorical decisions are made, but instead where the information supporting those decisions is encoded, they do not necessarily conflict with left lateralized decision mechanisms. The absence of evidence for a contribution by pF to scene categorization is consistent with the prior observation that pF encodes the identity of objects in scenes with considerably less fidelity than does LO (MacEvoy & Epstein, 2011).

Showing a decoding profile opposite that of LO, activity patterns in PPA were significantly better at decoding judgments of ambiguous scenes in size blocks than in object blocks. The implication of this result, that pF encodes the identity of objects in scenes with considerably less fidelity than does LO (MacEvoy & Epstein, 2011). (Spatial variability was added to ambiguous scenes to provide participants greater impetus to follow directions in size blocks. It increased the odds that their attention would be diverted from scenes’ object contents and consequently maximized our ability to detect a cue dependent contribution to categorization by LOC.) More important for our study, therefore, was the absence of a significant predictive ability for PPA patterns during object blocks, indicating that PPA did not mediate the influence of object recognition networks on scene judgments.

It is noteworthy that among all ROIs, the highest overall decoding accuracy (i.e., pooled across block type) was observed in RSC, a region that has also been linked to scenes’ global properties (Epstein, 2008; Harel et al., 2012; Park & Chun, 2009). However the absence of a significant accuracy difference between block types makes this observation uninterpretable. It is possible that this reflects an equal dependence of scene judgments on RSC activity in object and size blocks, although this account is difficult to square with its even greater bias towards spatial property-based scene processing than PPA (Harel et al., 2012). More likely is that RSC patterns simply followed categorical judgments regardless of the scene feature they were based on. Indeed, though RSC has been shown to be sensitive to scenes’ spatial properties, this appears less related to scene categorization per se than to spatial navigation in general (Epstein, 2008). Inasmuch as our task had no navigation-relevant component, it is perhaps not surprising to find no causal link between RSC patterns and behavior. Given the task-dependence of LOC and PPA contributions, however, we suspect an RSC link could emerge from a task related to observers’ perceived whereabouts in the world.

**Keywords**: lateral occipital complex, parahippocampal place area, scene categorization, fMRI, MVPA

**Acknowledgments**

This work was funded by Boston College.

Commercial relationships: none.

Corresponding author: Drew Linsley.

Email: linsleyd@bc.edu.

Address: Department of Psychology, Boston College, Chestnut Hill, MA, USA.

**References**

Aguirre, G. K., & D’Esposito, M. (1999). Topograph-


