nature neuroscience

Constructing scenes from objects in human occipitotemporal cortex

Sean P MacEvoy[1,2] & Russell A Epstein[2]

We used functional magnetic resonance imaging (fMRI) to demonstrate the existence of a mechanism in the human lateral occipital (LO) cortex that supports recognition of real-world visual scenes through parallel analysis of within-scene objects. Neural activity was recorded while subjects viewed four categories of scenes and eight categories of 'signature' objects strongly associated with the scenes in three experiments. Multivoxel patterns evoked by scenes in the LO cortex were well predicted by the average of the patterns elicited by their signature objects. By contrast, there was no relationship between scene and object patterns in the parahippocampal place area (PPA), even though this region responds strongly to scenes and is believed to be crucial for scene identification. By combining information about multiple objects within a scene, the LO cortex may support an object-based channel for scene recognition that complements the processing of global scene properties in the PPA.

Human observers have a remarkable capacity to categorize complex visual scenes, such as 'kitchen' or 'beach', at a single glance[1,2]. Behavioral data and computational models suggest that analysis of global properties, such as spatial layout, texture or image statistics, might provide one route to scene recognition[3–5], and previous neuroimaging work has identified regions of occipitotemporal cortex that are hypothesized to support scene recognition on the basis of whole-scene characteristics[6–8]. At the same time, it is clear that objects can provide important information about scene category; for example, a kitchen and an office are easily distinguished by the objects they contain even if they have similar three-dimensional geometries. The use of object information to support rapid scene recognition presents a substantial challenge, however: scenes usually contain many potentially informative objects, making scene recognition on the basis of serial deployment of attention to each object unacceptably slow. The manner in which the visual system solves this problem is unclear, as are the neural systems involved. Although previous work has identified regions that respond to standalone objects[9] and objects within scenes[10,11], a role for these regions in object-based scene recognition has not been established.

Here we provide evidence for a specific mechanism of object-based scene recognition. Under our hypothesis, the occipitotemporal visual areas that support this mechanism perform parallel analysis of individual objects within scenes and then combine the resulting object codes linearly. The result is a unified scene representation that inherits the neural signatures of the individual constituent objects, thereby uniquely encoding scene categories on the basis of their contents. In essence, we suggest that this mechanism builds 'kitchens' out of 'stoves' and 'refrigerators'; 'bathrooms' out of 'toilets' and 'bathtubs'.

To test this hypothesis, we exploited the fact that different categories of scenes and objects evoke distributed patterns of neural activity that can be distinguished with functional magnetic resonance imaging (fMRI)[12,13]. Participants were scanned with fMRI while they viewed images of four scene categories (kitchen, bathroom, playground and intersection) and eight categories of 'signature objects' strongly associated with the scenes (kitchen: stove and refrigerator; bathroom: toilet and bathtub; playground: swing and slide; and intersection: car and traffic signal; **Fig. 1**). We reasoned that if scene representations in any area were 'built' from their constituent objects, then multivoxel patterns evoked by each scene category should closely resemble combinations of multivoxel patterns evoked by their signature objects when these objects were viewed in isolation. We evaluated this prediction by attempting to decode multivoxel scene patterns on the basis of combinations of multivoxel object patterns in three fMRI experiments.

RESULTS

Multivoxel classification of scenes and objects

In experiment 1, images from the four scene categories and the eight object categories were presented for 1 s followed by a 2-s interstimulus interval, with scenes and objects interleaved in an event-related design. Subjects were asked to press a button and silently name each item. Our analyses focused on the lateral occipital complex (LOC), a region that responds preferentially to objects[9], and the parahippocampal place area (PPA), a region that responds preferentially to scenes[6]. Within the LOC we defined two subregions, the posterior fusiform area (pF) and the more posteriorly situated LO, as previous work suggests that these subregions may support different functions during visual recognition[14,15].

Using multivoxel pattern analysis, we first quantified the amount of information about object and scene category that was reliably present in distributed patterns of activity in these regions of interest (ROIs). Consistent with previous results[8,12,13,16,17], we were able to identify scene categories on the basis of multivoxel patterns that were evoked

[1]Department of Psychology, Boston College, Chestnut Hill, Massachusetts, USA. [2]Department of Psychology, University of Pennsylvania, Philadelphia, Pennsylvania, USA. Correspondence should be addressed to S.P.M. (sean.macevoy.1@bc.edu).

Received 29 March; accepted 6 July; published online 4 September 2011; doi:10.1038/nn.2903

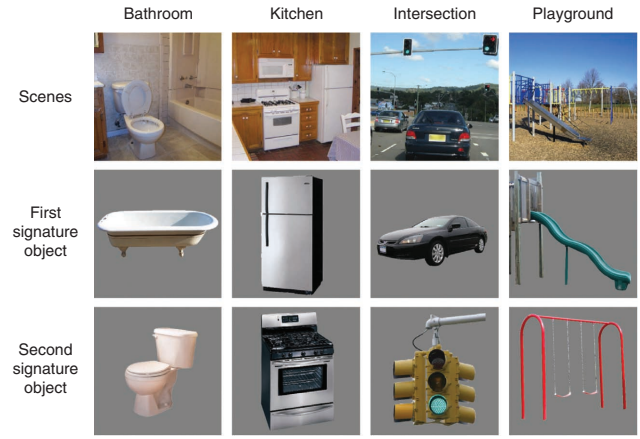


Figure 1 Experimental stimuli. Subjects viewed 104 scene images drawn from four categories (kitchen, bathroom, playground and roadway intersection) and 208 object images drawn from eight categories strongly associated with the scenes (refrigerators and stoves for kitchens, toilets and bathtubs for bathrooms, swings and slides for playgrounds, and traffic signals and cars for intersections). Each scene contained the two corresponding signature objects; however, none of the object exemplars was drawn from any of the scene exemplars.

by scenes and to identify object categories on the basis of multivoxel patterns that were evoked by objects at rates that were significantly above chance in all three regions (two-tailed *t*-test on classification accuracy for objects: LO, $t_{13} = 7.6$, $P < 0.0001$; pF, $t_{13} = 5.7$, $P < 0.0001$; PPA, $t_{13} = 5.3$, $P = 0.0002$; classification accuracy for scenes: LO, $t_{13} = 6.8$, $P < 0.0001$; pF, $t_{13} = 7.4$, $P < 0.0001$; PPA, $t_{13} = 6.2$, $P < 0.0001$).

We reasoned that if scene representations in any of these areas were built from those of their constituent objects, we should be able to classify scene-evoked patterns using combinations of object-evoked patterns. To test this idea, we attempted to classify scenes using a set of object-based predictors: two 'single-object' predictors that were simply the patterns evoked by that scene category's two associated objects, and a 'mean' predictor that was the average of the two associated object patterns and was representative of a linear combination rule (**Fig. 2**; please see **Supplementary Results** for a discussion of this choice of predictors). Even though none of the single-object exemplars was drawn from any of the scenes, each object-based predictor type correctly classified scene patterns in LO at a rate that was significantly above chance (single-object predictor: $t_{13} = 3.1$, $P = 0.007$; mean predictor: $t_{13} = 3.8$, $P = 0.002$; see **Fig. 3**). Performance of the mean predictor was significantly higher than the average performance of the single-object predictors ($t_{13} = 2.7$, $P = 0.019$). Neither of the object-based predictors produced performance above chance in pF (single: $t_{13} = 0.78$, $P = 0.45$; mean: $t_{13} = 0.71$, $P = 0.5$). These results indicate that patterns of activity evoked by scenes in LO, but not in pF, carry information about the identities of multiple objects within them, even in the absence of any requirement of subjects to attend to those objects individually.

The success of object-based predictors in LO stands in distinct contrast to their poor performance in the PPA, where scene classification using the predictors did not differ from chance (single: $t_{13} = 1.2$, $P = 0.27$; mean: $t_{13} = 0.46$, $P = 0.65$). In other words, even though activity patterns in the PPA contained information about both scenes and standalone objects, neural representations of scenes seemed to be unrelated to representations of the objects they contained. To eliminate the possible confound presented by stronger overall responses to scenes versus objects in the PPA, we repeated our classification procedure after independently normalizing each scene and predictor pattern by converting it to a vector with unit magnitude. Even after this step, classification of scenes from object-based predictors did not significantly differ from chance (50%) for any of the predictors; by contrast, accuracy for the predictor models in LO improved slightly.

Role of attentional shifts

Our results suggest that scene representations in LO are linear combinations of the representations elicited by their constituent objects. However, given the slow time course of the fMRI signal, which effectively integrates neural activity over a lengthy temporal window, we could have obtained the same results if subjects directed their attention serially to the individual objects within the scenes during the relatively long, 1-s presentation time. To address this possibility, experiment 2 repeated our basic design in a new set of subjects using a faster stimulus sequence in which scenes and objects were shown for only 150 ms each, followed immediately by a phase-scrambled mask. Subjects performed an indoor–outdoor discrimination task. Although this presentation time was sufficient to allow subjects to interpret scenes (evinced by greater than 95% accuracy on the behavioral task), it reduced subjects' ability to direct attention sequentially to the individual objects within the scenes.

Scene-from-object classification results in experiment 2 were almost identical to those observed in experiment 1 (**Fig. 3**). In LO, scene classification accuracy was significantly above chance for both of the object-based predictors (single: $t_{12} = 3.3$, $P = 0.006$; mean: $t_{12} = 3.8$, $P = 0.002$), and the accuracy of the mean predictor was significantly higher than for the single-object predictors ($t_{12} = 3.31$, $P = 0.006$). Accuracy for both predictors was only marginally above chance in pF (single: $t_{13} = 1.9$, $P = 0.08$; mean: $t_{13} = 2.1$, $P = 0.06$) and not significantly above chance in the PPA (single: $t_{13} = 0.38$, $P = 0.7$; mean: $t_{13} = 0.42$, $P = 0.68$). Thus, scene patterns in LO resemble averages of object patterns even when subjects have little time to move attention between the objects in the scene.

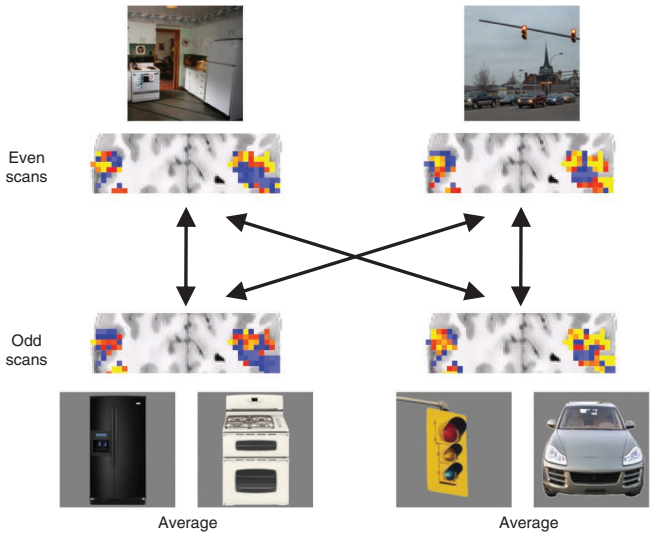


Figure 2 Logic of scene classification analysis. Scene patterns evoked by actual scenes in one half of scans were compared to predictor patterns derived from object-evoked patterns from the opposite half. Activity maps shown are actual scene-evoked patterns (top) and the averages of object-evoked patterns (bottom) for one subject. Correct scene-from-object classification decisions occurred when actual scene patterns were more similar to predictors that are based on their own associated objects than to the predictors that are based on objects from other scene contexts.

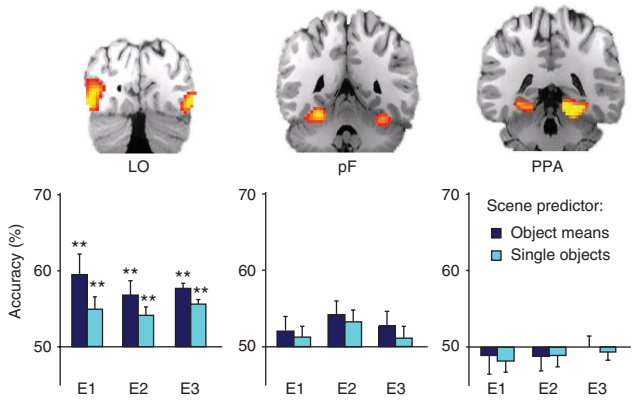


Figure 3 Multivoxel classification of scenes using object-based predictors. Results for each experiment are plotted separately for each ROI. Scene classification accuracy using the mean and single-object predictors was significantly above chance in LO in experiments 1, 2 and 3 (E1, E2 and E3, respectively). Accuracy was not above chance in any experiment in either pF or PPA. Furthermore, performance in LO was higher for the mean predictor than for the single-object predictor in each experiment ($P < 0.05$; see Results). Error bars indicate s.e.m. $^{**}P < 0.01$.

We also considered the possibility that the relationship we observed between scene- and object-evoked patterns might reflect the development of templates for object search. That is, repeated exposure to signature objects presented alone may have led subjects to automatically search for those objects when they were presented within scenes that were likely to contain them[10]. To address this, experiment 3 used a modified version of experiment 2 in which a new group of subjects was scanned while viewing only scenes and then scanned again while viewing only objects. While viewing scenes, subjects were unaware that they would subsequently view objects associated with those scenes. Replicating the results from the first two experiments, scene-from-object classification accuracy in LO was significantly above chance for the mean predictor ($t_{13} = 4.0$, $P = 0.0015$; **Fig. 3**) and for the single-object predictors ($t_{13} = 3.9$, $P = 0.0017$); furthermore, accuracy of the mean predictor was significantly higher than the average accuracy of the single-object predictors ($t_{13} = 2.37$, $P = 0.033$). Thus, the success of the object-based scene predictors was not predicated on the subjects being implicitly cued to selectively attend to the signature objects.

Finally, our results could be explained by the subjects alternating their attention between signature objects within scenes across trials. That is, subjects could have attended to refrigerators in one half of the trials during which they saw kitchens and on stoves in the other half, producing a trial-averaged kitchen pattern that resembled a linear combination of the stove and refrigerator patterns. We consider this behavioral pattern unlikely, as the tasks would tend to induce subjects to attend to the entire scene. Moreover, we have already shown that scene-evoked patterns resembled linear combinations of object-evoked patterns even when subjects had no motivation to attend to any particular objects within scenes (experiment 3). However, if subjects did attend to different objects across trials, we would have expected scene-evoked patterns to show greater trial-to-trial variability than object-evoked patterns, reflecting alternation between the activated object representations.

We examined this issue by analyzing the multivoxel response patterns evoked by scenes and objects on individual trials in LO. After extracting activity patterns evoked on each trial for a given category of scene or object (see Online Methods), we calculated the Euclidean distances between multivoxel patterns for all possible pairs of trials for that category (for example, the distance between kitchen trial 1 and kitchen trial 2, then between kitchen trial 1 and kitchen trial 3, and so on). These distances provide a measure of intertrial variability for scene and object patterns; in particular, because distances must always be positive, consistently greater variability should be reflected in larger median intertrial distances. After pooling within-category intertrial distances for each subject across all scenes and, separately, all objects, we computed the difference between each subject's median scene intertrial distance and median object intertrial distance. The resulting variable, expressed as a percentage of each subject's median object intertrial distance, had an average value across subjects of −2.47% in experiment 1 (bootstrap 95% confidence interval, −9.1% to −0.27%), −0.06% in experiment 2 (bootstrap 95% confidence interval, −0.23% to 0.12%) and 16.1% in experiment 3 (bootstrap 95% confidence interval, −5.5% to 50.6%). Although the wide confidence interval in experiment 3 leaves open the possibility that scene patterns may have been more variable than object patterns in that experiment, the narrow confidence intervals spanning negative values near zero in experiments 1 and 2 are inconsistent with generally greater variability for scenes than objects. Traditional statistical testing revealed no significant differences between scene and object variability in any of the three experiments (experiment 1: $t_{13} = -1.34$, $P = 0.20$; experiment 2: $t_{13} = -0.72$, $P = 0.44$.; experiment 3: $t_{13} = 0.61$, $P = 0.55$). Thus, we find no evidence to suggest that the classification performance of the mean predictor is a result of alternation of attention across different within-scene objects in different trials. (See **Supplementary Results** and **Supplementary Figs. 1–3** for descriptions of several additional control analyses.)

Visual versus semantic similarities

The existence of an ordered relationship between scene and object patterns in LO suggests that this region encodes features that are common to both the scenes and the objects that they contain. What are these features? There are at least two possibilities. First, the common features could be visual: stoves have flat tops and knobs, which are visible both when the stoves appear within a scene and when they are presented alone. Second, the common features could be semantic: both kitchens and stoves are associated with cooking, whereas both playgrounds and swings are associated with play.

We attempted to partially distinguish these possibilities by examining the relationship between response patterns evoked by objects drawn from the same context (for example, stoves and refrigerators)[18]. Objects from the same context share many semantic attributes; by contrast, their visual similarities are less salient. Thus, we reasoned that semantic coding would be evinced by more similar response

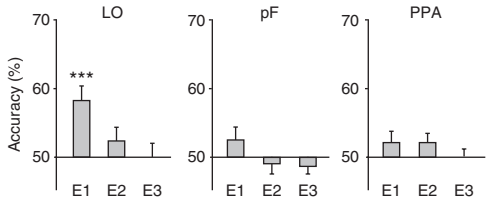


Figure 4 Classification of objects on the basis of the patterns elicited by their same-context counterpart objects (for example, the accuracy of discriminating refrigerators from bathtubs on the basis of patterns evoked by stoves and toilets). Accuracy was significantly above chance in LO in experiment 1 (E1), but not above chance in any ROI in experiment 2 (E2) or experiment 3 (E3). Error bars indicate s.e.m. $^{***}P < 0.001$.

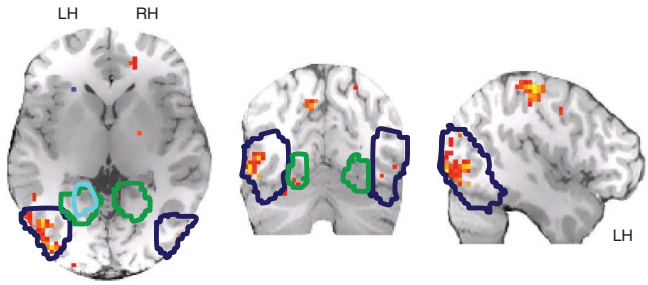
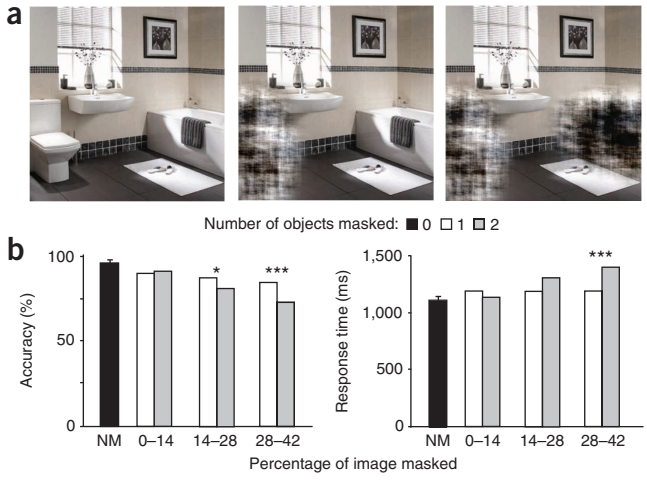


Figure 5 Group random-effects analysis of local searchlight accuracy maps for the classification of scenes from object averages, including subjects from all three experiments. Painted voxels represent centers of searchlight clusters with above-chance classification accuracies ($P < 0.005$, uncorrected). Displayed slices are cardinal planes containing the occipitotemporal voxel of peak significance, which was found in the left hemisphere (LH). Outlined regions are LO (dark blue), pF (light blue) and the PPA (green), which are defined from random-effects analysis of volumes across subjects ($P < 0.00001$, uncorrected). Although pF and PPA overlap when defined using these group data, they did not overlap when defined at the individual subject level. The apparent bias toward higher performance in left LO is addressed in the **Supplementary Results** and **Supplementary Figure 7.** RH, right hemisphere.

patterns between pairs of same-context objects than between pairs of different-context objects. We assessed this by attempting to classify each object category on the basis of patterns evoked by the other object category from the same context. Unexpectedly, classification accuracies depended upon the length of time that objects were viewed (**Fig. 4**). In experiment 1, wherein stimuli were presented for 1 s followed by a 2 s interval before the next item, the accuracy of discriminating objects from contextually related objects was significantly above chance in LO ($t_{13} = 5.7$, $P < 0.0001$), but not above chance in pF ($t_{13} = 1.2$, $P = 0.24$) or the PPA ($t_{13} = 0.82$, $P = 0.40$). By contrast, accuracy was not above chance in any of these ROIs in experiment 2, wherein stimuli were presented for 150 ms followed by a 350 ms mask and then a 1 s interval before the next trial (LO: $t_{12} = 1.4$, $P = 0.18$; pF: $t_{13} = 0.5$, $P = 0.60$; PPA: $t_{13} = 1.2$, $P = 0.24$). Nor was accuracy above chance in experiment 3, which used the same temporal parameters (LO: $t_{13} = 0.08$, $P = 0.94$; pF: $t_{13} = -0.49$, $P = 0.62$; PPA: $t_{13} = 0.19$, $P = 0.85$).

These results suggest that LO primarily encoded visual features of scenes and objects in the short-presentation experiments (experiments 2 and 3), but encoded semantic features in addition to visual features in the long-presentation experiment (experiment 1). The reason for the differences is unclear, but may relate to the fact that subjects in the first experiment covertly named each item—a task that may have activated abstract representations tied to language—whereas subjects in the other two experiments did not. Alternatively, the faster presentation rate in the second and third experiments may have interrupted a transition between an initial representation that was based on low-level visual features to a later one that was based on a high-level semantic summary[19]. Additional analyses related to these points can be found in the **Supplementary Results** and **Supplementary Figure 4**.



Searchlight analysis

To examine responses outside our pre-defined ROIs, we used a whole-brain 'searchlight' procedure to independently identify regions containing scene patterns that related to patterns evoked by their constituent objects[20]. For each voxel in the brain, we defined a 5-mm-radius spherical mask centered on that voxel and applied the scene-from-mean classification procedures described above to the multivoxel patterns defined by that mask. High classification accuracy for scenes using object-average patterns was mainly limited to two voxel clusters: one in the medial parietal cortex and the other in LO (**Fig. 5**). (Above-chance accuracy was also observed in a dorsal cluster, visible the sagittal slice in **Fig. 5**. This cluster is likely to correspond to motor cortex, reflecting the correlation between scene/object categories and button presses in the indoor/outdoor task in experiments 2 and 3.) These results suggest that LO is unique among occipitotemporal visual areas in possessing neural representations of scenes that are constructed from the representations of their constituent objects. (For whole-brain searchlight analyses of the object-from-contextual-counterpart classification, see **Supplementary Fig. 5**. For the results of pattern classification in the early visual cortex and other ROIs and for data broken down by hemisphere, see **Supplementary Results** and **Supplementary Fig. 6**).

Behavioral evidence for object-based scene recognition

Our fMRI results suggest that, by preserving information about individual objects within scenes, LO houses a potentially rich resource in support of scene recognition. But is this information actually used for this purpose? To address this, we conducted a behavioral study outside the scanner in which a new group of subjects viewed scenes that were briefly presented (50 ms) and masked and then performed a four-alternative forced-choice classification task. Each scene had zero, one or two of its signature objects obscured by a visual noise pattern (**Fig. 6a**). We reasoned that the operation of an object-based system of scene recognition should be evident as a decline in behavioral

Figure 6 Behavioral evidence for object-based scene recognition. (**a**) Subjects saw briefly presented exemplars of scenes from each of the four categories used in the fMRI studies and performed a four-alternative forced-choice scene identification task. Each exemplar was shown intact (left) or with one (middle) or both (right) of its signature objects obscured. (**b**) Average accuracy (left) and response time (right) are shown for images with zero, one or two objects removed. NM, no mask (zero objects removed). Data for conditions with objects removed are shown for three different ranges of the percentage of image pixels removed. For matched percentages of pixel deletion, accuracy and reaction time were significantly degraded when both signature objects were removed compared to when just one was removed; this effect was only significant when a high percentage of the scene pixels were deleted, which is likely to correspond to the range in which image-based identification falters. Accuracy and response time estimates are from application of the Johnson–Neyman procedure, which does not produce error bars. $^*P < 0.05$ and $^{***}P < 0.001$.

classification performance when subjects viewed scenes with objects removed, compared to intact-scene performance. We observed a significant effect of the number of objects removed on both classification accuracy ($F_{2,28} = 33.7$, $P < 0.0001$) and reaction time ($F_{2,28} = 35.3$, $P < 0.0001$). To determine whether this effect was simply a consequence of image degradation, we performed one-way analyses of covariance (ANCOVAs) on itemwise accuracy and reaction time with the number of objects removed (either one or two) as a factor and the number of pixels removed as covariate. This analysis revealed that performance degraded as more pixels were removed (accuracy: $F_{1,380} = 17.50$, $P < 0.0001$; reaction time: $F_{1,380} = 9.57$, $P = 0.002$). Furthermore, there was a significant interaction between the number of pixels removed and the number of objects removed (accuracy: $F_{1,380} = 7.0$, $P = 0.009$; reaction time: $F_{1,380} = 9.7$, $P = 0.002$). To characterize this interaction, we applied the Johnson–Neyman procedure[21], which revealed that the number of objects removed (one versus two) had a significant effect ($P < 0.01$) on performance, but only when a large enough number of pixels were removed (**Fig. 6b**). These findings are consistent with the operation of parallel object- and image-based systems of scene recognition: scene identification falters when both of the signature objects are removed, but only if enough of the image is obscured to simultaneously affect the image-based recognition system. Conversely, even when large portions of scenes are obscured, the presence of a single diagnostic object is sufficient to rescue recognition.

DISCUSSION

The principal finding of this study is that patterns of activity evoked in LO by scenes are well predicted by linear combinations of the patterns evoked by their constituent objects. Despite the complexity of the real-world scenes used, we were able to classify the patterns they evoked at rates that were above chance. Furthermore, we could do this with knowledge of the patterns evoked by just two of the object categories the scenes contained, even though the objects in the scenes could be incomplete, occluded or at peripheral locations, and even though the scenes contained many other objects for which the response patterns were not known. By contrast, no similar relationship between scene and object patterns was observed in the PPA, even though patterns in this region carried information about the identities of scenes and individual objects at levels of precision that were comparable to those in LO. By demonstrating the neural construction of scenes from their constituent objects in LO, our results suggest the existence of a previously undescribed channel supporting object-based scene recognition. The existence of such a channel is further supported by behavioral results demonstrating degraded performance in a scene classification task when objects within scenes are obscured.

In contrast to previous studies showing that patterns evoked by complex scenes can be predicted from a comprehensive inventory of the responses of individual voxels to other scenes[22,23], our results show that patterns evoked by scenes can be predicted by a stimulus class—objects—that occupies a different categorical space. By doing so, our findings provide an important extension of previous work examining neural responses to multiple-object arrays. When objects are shown in non-scene arrays, both the multivoxel activity patterns in human LO[16,24] and the responses of single inferotemporal neurons in macaques[25] resemble the average of those evoked by each array element by itself, as long as attention is equally divided among the objects or is directed away from all of them. Although these and similar phenomena[26,27] are often explained in terms of competition between stimuli for limited neural resources[24,28–31], we have previously advocated an alternative hypothesis[16]: rather than reflecting the outcome of an indeterminate attentional state, response averaging reflects a strategy for low-loss encoding of information about multiple simultaneous objects in populations of broadly tuned neurons. This coding scheme would be particularly useful during scene recognition: encoding scenes as linear combinations of their constituent objects would ensure that scene-evoked patterns varied reliably across scene categories while retaining information that could be useful for identifying the objects themselves should they be individually attended. The current results demonstrate that the combination rules previously observed to mediate neural representations of non-scene object arrays also apply to the representations of real-world scenes, even though scenes are highly complex and contain many varieties of information (for example, spatial layout and perspective) that are not present in non-scene arrays.

The relationship between scene and object patterns did not appear to result from the subjects paying attention to individual objects in scenes either within or across trials. The same relationship between scene and object patterns was observed both in the slow-presentation version of the experiment (experiment 1) and in the fast-presentation versions (experiments 2 and 3), even though subjects viewed stimuli in the latter two experiments for only 150 ms followed by a mask, and even though subjects in experiment 3 viewed scenes before objects to ensure that they would not develop a search template for the objects. Furthermore, our results cannot be explained by the subjects paying attention to different signature objects within scenes across different trials, as this would predict greater trial-by-trial variability for scene patterns than for object patterns, which was not observed. Rather, our results seem to reflect the outcome of an averaging mechanism that operates on object representations when subjects direct attention not to these objects as individual items but to the scene as a whole. As such, these results provide a complement to those obtained in a recent study in which subjects were pre-cued to search for a single object within a scene. In that case, the patterns evoked by scenes resembled those evoked by the target object, but did not resemble patterns evoked by non-target objects that were also present[10]. Thus, although attention to one object in a scene can bias the scene-evoked response to more closely match the pattern evoked by that object, from our results we argue that directed attention is not a prerequisite to scene–object links. Indeed, the absence of an attentional requirement in the generation of object-based scene representations is consistent with the phenomenology of scene recognition, which can occur 'at a glance', without serial deployment of attention to individual objects[32]. Instead of producing a loss of information, our results show that the absence of attentional bias allows information about multiple objects to be represented simultaneously, expanding the precision with which scenes can be encoded.

The current results leave several unresolved issues. First, we cannot state with certainty that they will apply to every scene category under all circumstances. Certain scenes—for instance, a stadium interior—have few salient objects and may require a heavier reliance on global features for recognition. In an experiment such as ours, such scenes might defy object-based classification. Conversely, scene recognition might be especially reliant on diagnostic objects when the range of scene categories that is likely to be encountered is relatively narrow, as it was in our experiment and would be in many real-world situations. (For example, when one has already entered a house, the set of plausible scenes is fairly small.) Second, we do not know whether all objects in a scene contribute to scene-evoked patterns in LO; contributions may be limited to the most visually salient objects or to the most diagnostic objects. Third, we do not know whether the success of object-based scene classification in our study depended on the actual presence of the signature objects in the scene exemplars. It would not be surprising if a predictor linked to an object that is strongly

associated with a scene category were to produce correct classifications of scenes in which that object was absent. Indeed, the ability to classify objects from their same-context counterparts in experiment 1 indicates at least some redundancy in the patterns evoked by objects from the same context, suggesting that scene patterns in LO should be at least somewhat tolerant to removal of signature objects. Finally, we have not examined the extent to which scene-evoked patterns in LO are sensitive or invariant to identity-preserving object transformations. Several previous studies have shown that responses in LO depend on object position, size and viewpoint[33–37]; this suggests that even higher classification performance could be obtained if these quantities were conserved. Nevertheless, our results indicate that even when these quantities vary across stimuli, enough information is preserved about object identity in LO response patterns to allow scene discrimination. By differing reliably between scene categories, the ensemble of object-based responses evoked in LO can be seen as a robust, if somewhat 'noisy', shorthand code facilitating scene recognition.

The findings in LO stand in sharp contrast to those observed in the PPA. Even though PPA activity patterns in our study contained information about object category when objects were presented singly, this information was absent when objects were embedded in scenes, as evinced by the failure of patterns evoked by objects to predict patterns evoked by scenes containing them. Furthermore, we did not observe a relationship between the patterns evoked by contextually related objects in the PPA, which is contrary to what one might have expected on the basis of previous work[18]. These results suggest that the PPA encodes either visual or spatial information that is unique to each scene and object category but does not allow scenes to be related to their component objects or objects to be related to their contextual associates. We suggest that, consistent with the results of recent neuroimaging studies, the underlying representation might consist either of a statistical summary of the visual properties of the stimulus or of geometric information about the layout of the most salient spatial axes[38,39]. With regards to the geometric hypothesis, it is worth noting that most of the objects in the current study were large, fixed items that would help determine the geometry of local navigable space. By contrast, an earlier study that compared PPA response patterns across smaller, moveable objects found no reliable differences[40]. It is also noteworthy that object-based predictors did not classify scenes in pF in our study, despite above-chance object and scene classification and despite previous studies showing that pF has an even greater tolerance of identity-preserving object transformations than LO. The reasons for the low classification accuracies in pF are unclear, but as in the PPA, the results suggest that scenes may be considered to be distinct items unto themselves in pF, rather than combinations of objects.

In summary, our results show the existence of an object-based channel for scene recognition in LO. By doing so, they address a long-standing challenge to our understanding of the neural mechanisms of scene recognition: even though the identities of objects in a scene can greatly aid its recognition, brain regions strongly activated by scenes such as the PPA seem to be chiefly concerned with large-scale spatial features, such as spatial layout, rather than the coding of within-scene objects[6,41,42]. By contrast, scene-evoked patterns in LO seem to be 'built' from the individual patterns of the objects within a scene. These results suggest that the PPA and LO can be seen as nodes along parallel pathways supporting complementary modes of scene recognition[8], with the PPA supporting recognition based principally on global scene properties[3–5] and LO supporting recognition based on the objects the scenes contain.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/natureneuroscience/.

Note: Supplementary information is available on the Nature Neuroscience website.

ACKNOWLEDGMENTS

The authors wish to thank E. Ward, A. Stigliani and Z. Yang for assistance with data collection. This work was supported by US National Eye Institute grant EY-016464 to R.A.E.

AUTHOR CONTRIBUTIONS

S.P.M. and R.A.E. designed the experiments. S.P.M. collected fMRI data and R.A.E. collected behavioral data. S.P.M. analyzed data with input from R.A.E. S.P.M. and R.A.E. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/natureneuroscience/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Potter, M.C. Meaning in visual search. *Science* **187**, 965–966 (1975).
2. Biederman, I., Rabinowitz, J.C., Glass, A.L. & Stacy, E.W. On the Information extracted from a glance at a scene. *J. Exp. Psychol.* **103**, 597–600 (1974).
3. Schyns, P.G. & Oliva, A. From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychol. Sci.* **5**, 195–200 (1994).
4. Renninger, L.W. & Malik, J. When is scene identification just texture recognition? *Vision Res.* **44**, 2301–2311 (2004).
5. Greene, M.R. & Oliva, A. Recognition of natural scenes from global properties: seeing the forest without representing the trees. *Cognit. Psychol.* **58**, 137–176 (2009).
6. Epstein, R. & Kanwisher, N. A cortical representation of the local visual environment. *Nature* **392**, 598–601 (1998).
7. Aguirre, G.K., Zarahn, E. & D'Esposito, M. An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. *Neuron* **21**, 373–383 (1998).
8. Park, S., Brady, T.F., Greene, M.R. & Oliva, A. Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *J. Neurosci.* **31**, 1333–1340 (2011).
9. Malach, R. *et al.* Object-related activity revealed by functional magnetic-resonance-imaging in human occipital cortex. *Proc. Natl. Acad. Sci. USA* **92**, 8135–8139 (1995).
10. Peelen, M.V., Fei-Fei, L. & Kastner, S. Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature* **460**, 94–97 (2009).
11. Epstein, R., Graham, K.S. & Downing, P.E. Viewpoint-specific scene representations in human parahippocampal cortex. *Neuron* **37**, 865–876 (2003).
12. Haxby, J.V. *et al.* Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).
13. Walther, D.B., Caddigan, E., Fei-Fei, L. & Beck, D.M. Natural scene categories revealed in distributed patterns of activity in the human brain. *J. Neurosci.* **29**, 10573–10581 (2009).
14. Drucker, D.M. & Aguirre, G.K. Different spatial scales of shape similarity representation in lateral and ventral LOC. *Cereb. Cortex* **19**, 2269–2280 (2009).
15. Haushofer, J., Livingstone, M.S. & Kanwisher, N. Multivariate patterns in object-selective cortex dissociate perceptual and physical shape similarity. *PLoS Biol.* **6**, e187 (2008).
16. MacEvoy, S.P. & Epstein, R.A. Decoding the representation of multiple simultaneous objects in human occipitotemporal cortex. *Curr. Biol.* **19**, 943–947 (2009).
17. Diana, R.A., Yonelinas, A.P. & Ranganath, C. High-resolution multi-voxel pattern analysis of category selectivity in the medial temporal lobes. *Hippocampus* **18**, 536–541 (2008).
18. Bar, M. & Aminoff, E. Cortical analysis of visual context. *Neuron* **38**, 347–358 (2003).
19. Potter, M.C., Staub, A. & O'Connor, D.H. Pictorial and conceptual representation of glimpsed pictures. *J. Exp. Psychol. Hum. Percept. Perform.* **30**, 478–489 (2004).
20. Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc. Natl. Acad. Sci. USA* **103**, 3863–3868 (2006).
21. Rogosa, D. Comparing nonparallel regression lines. *Psychol. Bull.* **88**, 307–321 (1980).
22. Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M. & Gallant, J.L. Bayesian reconstruction of natural images from human brain activity. *Neuron* **63**, 902–915 (2009).
23. Kay, K.N., Naselaris, T., Prenger, R.J. & Gallant, J.L. Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008).

24. Reddy, L., Kanwisher, N.G. & VanRullen, R. Attention and biased competition in multi-voxel object representations. *Proc. Natl. Acad. Sci. USA* **106**, 21447–21452 (2009).
25. Zoccolan, D., Cox, D.D. & DiCarlo, J.J. Multiple object response normalization in monkey inferotemporal cortex. *J. Neurosci.* **25**, 8150–8164 (2005).
26. Moran, J. & Desimone, R. Selective attention gates visual processing in the extrastriate cortex. *Science* **229**, 782–784 (1985).
27. MacEvoy, S.P., Tucker, T.R. & Fitzpatrick, D. A precise form of divisive suppression supports population coding in the primary visual cortex. *Nat. Neurosci.* **12**, 637–645 (2009).
28. Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* **18**, 193–222 (1995).
29. Reynolds, J.H. & Heeger, D.J. The normalization model of attention. *Neuron* **61**, 168–185 (2009).
30. Beck, D.M. & Kastner, S. Stimulus similarity modulates competitive interactions in human visual cortex. *J. Vis.* **7**(19), 11–12 (2007).
31. Beck, D.M. & Kastner, S. Stimulus context modulates competition in human extrastriate cortex. *Nat. Neurosci.* **8**, 1110–1116 (2005).
32. Fei-Fei, L., Iyer, A., Koch, C. & Perona, P. What do we perceive in a glance of a real-world scene? *J. Vis.* **7**(1), 10 (2007).
33. Schwarzlose, R.F., Swisher, J.D., Dang, S. & Kanwisher, N. The distribution of category and location information across object-selective regions in human visual cortex. *Proc. Natl. Acad. Sci. USA* **105**, 4447–4452 (2008).
34. Sayres, R. & Grill-Spector, K. Relating retinotopic and object-selective responses in human lateral occipital cortex. *J. Neurophysiol.* **100**, 249–267 (2008).
35. Kravitz, D.J., Kriegeskorte, N. & Baker, C.I. High-level visual object representations are constrained by position. *Cereb. Cortex* **20**, 2916–2925 (2010).
36. Andresen, D.R., Vinberg, J. & Grill-Spector, K. The representation of object viewpoint in human visual cortex. *Neuroimage* **45**, 522–536 (2009).
37. Eger, E., Kell, C.A. & Kleinschmidt, A. Graded size sensitivity of object-exemplar-evoked activity patterns within human LOC subregions. *J. Neurophysiol.* **100**, 2038–2047 (2008).
38. Epstein, R.A. Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends Cogn. Sci.* **12**, 388–396 (2008).
39. Kravitz, D.J., Peng, C.S. & Baker, C.I. Real-world scene representations in high-level visual cortex: it's the spaces more than the places. *J. Neurosci.* **31**, 7322–7333 (2011).
40. Spiridon, M. & Kanwisher, N. How distributed is visual category information in human occipito-temporal cortex? An fMRI study. *Neuron* **35**, 1157–1165 (2002).
41. Habib, M. & Sirigu, A. Pure topographical disorientation: a definition and anatomical basis. *Cortex* **23**, 73–85 (1987).
42. Aguirre, G.K. & D'Esposito, M. Topographical disorientation: a synthesis and taxonomy. *Brain* **122**, 1613–1628 (1999).

ONLINE METHODS

Functional magnetic resonance imaging. *Subjects.* In experiments 1 and 2, 28 subjects (14 subjects each; experiment 1: 6 females, 19–25 years old; experiment 2: 6 females, 21–28 years old) with normal or corrected-to-normal vision were recruited from the University of Pennsylvania community. The subjects gave written informed consent in compliance with procedures approved by the University of Pennsylvania Institutional Review Board. For experiment 3, 14 subjects (11 female, 18–23 years old) were recruited from the Brown University community using the same inclusion criteria. They gave written informed consent in compliance with procedures approved by the Institutional Review Boards of Boston College and Brown University. Subjects received payment for their participation.

Magnetic resonance imaging acquisition. Subjects participating in experiments 1 and 2 were scanned at the Center for Functional Neuroimaging at the University of Pennsylvania on a 3-T Siemens Trio scanner equipped with an eight-channel multiple-array head coil. Subjects participating in experiment 3 were scanned at the Brown University MRI Research Facility on a 3-T Siemens Trio scanner using a 32-channel head coil. Identical scanner parameters were used at the two facilities. Specifically, structural T1*-weighted images for anatomical localization were acquired using three-dimensional magnetization-prepared rapid-acquisition gradient echo (MPRAGE) pulse sequences (TR = 1,620 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 contrast were acquired using a gradient-echo echo-planar pulse sequence (TR = 3,000 ms, TE = 30 ms, voxel size = 3 × 3 × 3mm, matrix size = 64 × 64 × 45). At the University of Pennsylvania, the entire projected field subtended 22.9 × 17.4° and was viewed at 1,024 × 768 pixel resolution; at Brown University the field subtended 24 × 18° at the same resolution.

Experimental procedure. Scan sessions comprised two functional localizer scans followed by either four or eight experimental scans. Fourteen subjects participated in a 'long-presentation' version of the main experiment (experiment 1), 14 participated in a 'short-presentation' version in which they viewed scene and object stimuli interleaved in the same scan runs (experiment 2), and the remaining 14 subjects participated in a 'short-presentation' version in which they viewed object and scene stimuli in different scan runs (scenes always preceding objects; experiment 3). The stimulus set for all three experiments consisted of 312 photographic images drawn in equal numbers from the 12 image categories (4 scene categories and 8 object categories). Scene images were scaled to 9° by 9°. Object images were edited in Adobe Photoshop to remove any background information and were scaled and cropped so that the longest dimension of the object spanned 9°.

In Experiment 1, stimuli were presented one at a time for 1 s each followed by a 2-s interstimulus interval during which subjects fixated on a central cross. For each image, subjects were asked to press a button and covertly name the object or scene. Subject sessions were split into an equal number of scans running either 5 min 24 s or 6 min 6 s long, and arranged in pairs. Each of these scan pairs contained 13 repetitions of each image category interspersed with 13 6-s fixation-only null trials. These were arranged in a continuous carryover sequence, a serially balanced design ensuring that each image category followed every other image category and itself exactly once[43]. Six repetitions of each image category were contained in the shorter scan of each pair and seven repetitions in the longer. Each subject was scanned with a unique continuous carryover sequence, which was repeated either two or four times.

In experiments 2 and 3, scenes and objects were shown for 150 ms each followed by a 350-ms phase-scrambled mask and then a 1-s interstimulus interval. Subjects indicated by a button press whether the scenes were indoor or outdoor, or whether the objects were typically found indoors or outdoors. Performance on this task was very high (mean = 95.6%, averaged over both experiments), indicating that subjects could recognize the stimuli after brief presentations. Fixation-only null trials lasted 3 s. In experiment 2, scenes and objects were interleaved in four continuous carryover sequences, each filling a single scan lasting 5 min 27 s. These four sequences were uniquely generated for each subject and were each shown twice. In experiment 3, scenes and objects were not interleaved but presented in different scans. Subjects first viewed scenes only, arranged into 12 non-repeating continuous carryover sequences spread across two scans lasting 6 min 20 s each, followed by scans in which they viewed objects only, arranged into six continuous carryover sequences spread across two scans of 7 min each.

Functional localizer scans were 6 min 15 s long and were divided into blocks during which subjects viewed color photographs of scenes, faces, common objects and scrambled objects presented at a rate of 1.33 pictures per second, as described previously[44].

Magnetic resonance imaging analysis. Functional images were corrected for differences in slice timing by resampling slices in time to match the first slice of each volume; they were then realigned with respect to the first image of the scan and spatially normalized to the Montreal Neurological Institute (MNI) template. Data for localizer scans were spatially smoothed with a 9-mm full-width half-maximum Gaussian filter; all other data were left unsmoothed. Data were analyzed using a general linear model as implemented in VoxBo (http://www.voxbo.org/), including an empirically derived $1/f$ noise model, filters that removed high and low temporal frequencies, and nuisance regressors to account for global signal variations and between-scan signal differences.

For each scan, functional volumes without spatial smoothing were passed to a general linear model, which allowed the calculation of voxelwise response levels (β values) associated with each stimulus condition. In experiments 1 and 2, the resulting activity patterns were grouped into halves (for example, even sequences versus odd sequences) and patterns within each half were averaged; this and the following steps were repeated for each possible half-and-half grouping of the data. A 'cocktail' average pattern across all stimuli was calculated separately for each half of the data and then subtracted from each of the individual stimulus patterns. Separate cocktails were computed for objects and scenes. Our first set of analyses examined scene patterns and object patterns separately, without considering the relationship between them. Following the logic of previous experiments, we attempted to classify single objects from single objects and scenes from scenes. Pattern classification proceeded as a series of pairwise comparisons among objects and, separately, scenes. For each pairwise comparison, we calculated the Euclidean distances between patterns evoked by the same category in the two halves of the data and between different categories in the two halves. Correct classification decisions were registered when the distance between same-category patterns was shorter than between different-category patterns. For each pair of conditions, there were four such decisions, corresponding to each possible pairing of one vertical and one diagonal arrow in **Figure 2**. Pattern classification accuracies for each ROI were computed as the average of the accuracies from each hemisphere, measured separately. We observed the same classification results when we used correlation, rather than Euclidean distance, as the measure of pattern similarity.

We then performed a separate set of analyses that examined the relationship between patterns evoked by scenes and patterns evoked by their constituent objects. Specifically, we assessed how well predictor patterns constructed from object data in one half of scans classified scene-evoked patterns in the remaining half of scans. Mean predictors for each scene category (for example, kitchen) were constructed by taking the voxelwise average of the patterns evoked by the two objects (for example, refrigerator and stove) associated with that scene. To assess classification accuracy, these predictor patterns were simply substituted for scene-evoked patterns before executing the classification procedure. As part of this analysis, we also measured classification accuracy for scenes using the individual patterns for their constituent objects (without combining these single-object patterns together). Finally, we assessed how well the pattern evoked by one object from a given scene context could predict the pattern evoked by the other object from the same context. To do so, we repeated the object classification procedure after reducing the object set in one half of the data to include just one object from each scene context (for example, refrigerator, tub, car or slide) and reducing the object set in the other half to include only the remaining object from each context (for example, stove, toilet, traffic signal or swing). Patterns in each half were then labeled by context (kitchen, bathroom, playground or intersection), and the accuracy with which patterns from one half predicted the context label of the other half was assessed.

The analysis of activity patterns in experiment 3 was similar, except that accuracy was accumulated across all four possible pairwise comparisons between the two scene and two object scans (for example, first scene scan versus first object scan, first scene scan versus second object scan and so on). This scheme improved our estimates of classification accuracy by increasing the total number of unique classification decisions.

doi:10.1038/nn.2903

In addition to the pattern classification analyses performed within preset ROIs, we used a 'searchlight' analysis approach to identify regions of high classification accuracy throughout the brain[20]. For each brain voxel, we defined a spherical, 5 mm surrounding region (the searchlight cluster) and performed the same pattern classification steps outlined in the previous two paragraphs for each possible searchlight position. Classification accuracy for each cluster was assigned to the voxel at its center, producing whole-brain maps of local accuracy. These maps were combined across participants and subjected to random-effects group analysis to identify regions of above-chance performance.

To extract single-trial response vectors from LO to measure trial-by-trial response variability, we upsampled functional volumes to 1.5 s resolution in MATLAB using a low-pass interpolating filter (cutoff at 0.167 Hz) sampling symmetrically from the nearest eight original volumes. Response vectors for each stimulus trial were defined from the magnetic resonance signal in each voxel averaged across the four time points from 3 to 7.5 s following stimulus onset.

Regions of interest. Functional ROIs were defined on the basis of data from a separate set of functional localizer scans. The LOC was defined as the set of voxels in the lateral–ventral occipitotemporal region that showed stronger responses ($t > 3.5$) to objects than to scrambled objects. We divided the LOC into anterior and posterior segments associated with the posterior fusiform sulcus (pF) and lateral occipital area (LO), respectively. The PPA was defined as the set of voxels in the posterior parahippocampal–collateral sulcus region that responded more strongly ($t > 3.5$) to scenes than to objects. Before any analysis, LO and pF imaging segments were trimmed to exclude any voxels of overlap with the PPA. Supplementary analyses examined three additional ROIs: the scene-responsive retrosplenial complex (RSC), a scene-responsive focus in the transverse occipital sulcus (TOS) and the early visual cortex (EVC). The RSC and TOS were defined using the same scene–object contrast used to define the PPA, except that scene-responsive voxels were selected in this case from the retrosplenial–parietal-occipital sulcus region (RSC) or the transverse occipital sulcus region (TOS)[45]. The EVC was defined by significantly higher responses to scrambled objects than to intact objects ($t > 3.5$) in the posterior occipital lobe.

Behavioral analyses. *Subjects.* Sixteen subjects (12 female, 19–28 years old) with normal or corrected-to-normal vision were recruited from the University of Pennsylvania community. They gave written informed consent in compliance with procedures approved by the University of Pennsylvania Institutional Review Board. Subjects received course credit for participation.

Experimental procedure. Participants performed a four-alternative forced-choice task in which they categorized images of bathrooms, intersections, kitchens and playgrounds. Stimuli were unmodified and modified versions of 32 color photographs from each scene category and were 400 × 400 pixels in size. The photographs each contained two strongly diagnostic objects (for example, a toilet and a bathtub for the bathroom scene, and a slide and swings for the playground scene). The photographs could either be shown in their original form or with one or two signature objects obscured by a noise mask with feathered edges that did not reveal the object's contour and left most of the image intact. Noise masks were drawn from phase-scrambled versions of the original image, thus preserving global image statistics to the best extent possible.

After completing practice trials, each participant categorized one version of each of the 128 photographs. Assignment of the four versions of each scene (intact, object A removed, object B removed, both objects removed) was counterbalanced across subjects. Each stimulus was presented for 50 ms followed by a mask and participants were instructed to press a button when they felt they could categorize the scene as a bathroom, intersection, kitchen or playground, and to then indicate the category of the scene by making a second button press. Stimuli were presented in one run, with a 2-s fixation screen between trials. Masks were jumbled scenes constructed by first dividing each image in the stimulus set into 400 equally sized image fragments and then drawing 400 fragments at random from the complete set; a unique mask was used on each trial.

Statistical analysis. Unless otherwise noted, all *t*-tests were paired and two-sided, and implemented in MATLAB (MathWorks). ANCOVAs were implemented in SPSS (IBM). Bootstrap parameter estimates were generated in MATLAB and were based on 10,000 samples.

43. Aguirre, G.K. Continuous carry-over designs for fMRI. *Neuroimage* **35**, 1480–1494 (2007).
44. Epstein, R.A. & Higgins, J.S. Differential parahippocampal and retrosplenial involvement in three types of visual scene recognition. *Cereb. Cortex* **17**, 1680–1693 (2007).
45. Epstein, R.A., Parker, W.E. & Feiler, A.M. Where am I now? Distinct roles for parahippocampal and retrosplenial cortices in place recognition. *J. Neurosci.* **27**, 6141–6149 (2007).