Published in final edited form as:

Nat Neurosci. ; 14(10): 1323–1329. doi:10.1038/nn.2903.

Constructing scenes from objects in human occipitotemporal cortex

Sean P. MacEvoy^{1,2}, Russell A. Epstein²

¹Department of Psychology, Boston College, 140 Commonwealth Avenue, Chestnut Hill, Massachusetts 02467, USA

²Department of Psychology, University of Pennsylvania, 3720 Walnut Street, Philadelphia, Pennsylvania 19104, USA

Abstract

We used functional magnetic resonance imaging (fMRI) to demonstrate the existence of a mechanism in 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. Multi-voxel patterns evoked by scenes in LO were well-predicted by the average of the patterns elicited by their signature objects. In 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 play a critical role in scene identification. By combining information about multiple objects within a scene, LO may support an object-based channel for scene recognition that complements 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 based on 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 significant challenge, however: scenes usually contain many potentially informative objects, making scene recognition based upon 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 objects⁹ and objects within scenes^{10, 11}, a role for these regions in object-based scene recognition has not been established.

Author Contributions

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

Here we provide evidence for a specific mechanism for 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 based on their contents. In essence, we hypothesize 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, intersection) and eight categories of “signature objects” strongly associated with the scenes (kitchen: stove and refrigerator; bathroom: toilet and bathtub; playground: swing and slide; intersection: car and traffic signal; Figure 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 based on 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 objects⁹, and the parahippocampal place area (PPA), a region that responds preferentially to scenes⁶. Within LOC we defined two subregions, a posterior fusiform area (pF) and the more posteriorly-situated lateral occipital area (LO), as previous work suggests that these subregions may support different functions during visual recognition^{14, 15}.

Using multi-voxel pattern analysis (MVPA), 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 category based on multi-voxel patterns evoked by scenes, and object category based on multi-voxel patterns evoked by objects, at rates 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 (Figure 2; please see Supplementary Materials for a discussion of this choice of predictors). Even though none of the single-object exemplars were drawn from any of the scenes each object-based predictor type correctly classified scene patterns in LO at a rate significantly above chance (single-object predictor: $t(13) = 3.1$, $p = 0.007$; mean predictor: $t(13) = 3.8$, $p = 0.002$; see Figure 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 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 appeared unrelated to representations of the objects they contained. To eliminate the possible confound presented by stronger overall responses to scenes versus objects in 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; in 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 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 second 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 to the individual objects within the scenes.

Scene-from-object classification results in Exp. 2 were almost identical to those observed in Exp. 1 (Figure 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, $t(12) = 3.8$, $p = 0.002$), with accuracy of the mean predictor 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.

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 likely to contain them¹⁰. To address this, Exp. 3 used a modified version of the Exp. 2 in which a new pool of subjects were first scanned while viewing only scenes and then 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 ($t(13) = 4.0$, $p = 0.0015$; Figure 3) and single object predictors ($t(13) = 3.9$, $p = 0.0017$), with performance for the mean predictor significantly higher than the average of the single object predictors ($t(13) = 2.37$, $p = 0.033$). Thus, success of the object-based scene predictors was not predicated on subjects' being implicitly cued to selectively attend to the signature objects.

Finally, our results could be explained by subjects' alternating 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 to 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 (Exp. 3). However, if subjects did attend to different objects across trials, we 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 by each trial of a given category of scene or object (see Methods), we calculated the Euclidean distances between multivoxel patterns for all possible pairs of trials for that category (e.g., distance between kitchen trial 1 and kitchen trial 2, then between kitchen trial 1 and kitchen trial 3, etc.). 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 subject's median object intertrial distance, had an average value across subjects of -2.47% (bootstrap 95% confidence interval: -9.1% to -0.27%) in Exp 1., -0.06% in Exp. 2 (CI: -0.23% to -0.12%) and 16.1% in Exp. 3 (CI: -5.5% to 50.6%). Although the wide confidence interval in

Exp. 3 leaves open the possibility that scene patterns may have been more variable than object patterns in that experiment, the narrow CIs spanning negative values near zero in Exp. 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 (Exp. 1: $t(13) = -1.34$, $p = 0.20$; Exp. 2: $t(13) = -0.72$, $p = 0.44$; Exp. 3: $t(13) = -0.61$, $p = 0.55$). Thus we find no evidence to suggest that the classification performance of the mean predictor is owed to alternation of attention across different within-scene objects in different trials. (Please see Supplemental Materials for the results 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 scenes and the objects 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, while 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 (e.g. stoves and refrigerators)¹⁸. Objects from the same context share many semantic attributes; in contrast, their visual similarities are less salient. Thus, we reasoned that semantic coding would be evinced by more similar response patterns between pairs of same-context objects than pairs of different-context objects. We assessed this by attempting to classify each object category based on patterns evoked by the other object category from the same context. Surprisingly, classification accuracies depended upon the length of time that objects were viewed (Figure 4). In Exp. 1, wherein stimuli were presented for 1 s followed by a 2 s interval before the next item, object-from-contextually-related-object classification accuracy 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 PPA ($t(13) = 0.82$, $p = 0.40$). In contrast, accuracy was not above chance in any of these ROIs in Exp. 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: $t(13) = 1.2$, $p = 0.24$), or in Exp. 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 scene and object coding in LO is based mainly on visual features in the short-presentation experiments (Exps. 2 and 3), with greater inclusion of semantic features in the long-presentation experiment (Exp. 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 which 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 based on low-level visual features to a later one based on a high-level

semantic summary¹⁹. Additional analyses related to these points can be found in the Supplementary Results.

Searchlight analysis

To examine responses outside our pre-defined ROIs, we used a hypothesis-free “searchlight” procedure to independently identify regions containing scene patterns that related to patterns evoked by their constituent objects²⁰. For each voxel in the brain we defined a 5mm radius spherical mask centered on that voxel and applied the scene-from-mean classification procedures described above to the multi-voxel patterns defined by that mask. High classification accuracy for scenes using object-average patterns was mainly limited to two voxel clusters: one in medial parietal cortex, the other in LO (Figure 5). (Above-chance accuracy was also observed just anterior to the central sulcus, visible as a ventral cluster in the sagittal slice of Figure 5. This likely corresponds to motor cortex, reflecting the correlation between scene/object category and button presses in the indoor/outdoor task in Exps. 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 Figure 4. For the results of pattern classification in early visual cortex and other ROIs, and data broken down by hemisphere, see Supplemental Materials).

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 answer this, we performed a behavioral study outside the scanner in which a new group of subjects viewed briefly-presented (50 ms) and masked scenes and 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 (Figure 6a). We reasoned that the operation of an object-based system of scene recognition should be evident as a decline in behavioural classification performance when subjects viewed scenes with objects removed compared to intact scenes. 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 ANCOVAs on itemwise accuracy and reaction time (RT) 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 was degraded as more pixels were removed (accuracy: $F(1,380) = 17.50$, $p < 0.0001$; RT: $F(1,380) = 9.57$, $p = 0.002$); furthermore, there was a significant interaction between number of pixels removed and number of objects removed (accuracy: $F(1,380) = 7.0$, $p = 0.009$; RT: $F(1,380) = 9.7$, $p = 0.002$). To characterize this interaction we applied the Johnson-Neyman procedure²¹, which revealed that the number of objects removed (1 vs. 2) had a significant effect on performance, but only when a large enough number of pixels were removed (Figure 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

impact 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 by scenes in LO 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 above chance with knowledge of the patterns evoked by just two of the object categories they contained - even though those objects could be incomplete, occluded, or at peripheral locations, and even though the scenes contained many other objects to which responses were not known. In contrast, no similar relationship between scene and object patterns was observed in the PPA, despite patterns in this region carrying information about the identities of scenes and individual objects at levels of precision comparable to LO. By demonstrating 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 on a scene classification task when objects within scenes are selectively 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. Both multivoxel activity patterns in human LO^{16, 24} and the responses of single inferotemporal neurons in macaques²⁵ evoked by non-scene object arrays resemble the average of patterns evoked by each array element by itself, as long as attention is equally divided between the objects or deployed to none. 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 hypothesis¹⁶: rather than an 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 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 (Exp. 1) and also in the fast-presentation versions (Exps. 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 Exp. 3 viewed scenes before objects to ensure that they would not develop search template for the objects. Nor can our results be explained by 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 an interesting complement to those obtained in recent study in which subjects were pre-cued to search for a single object within a scene; in that case, patterns evoked by scenes resembled those evoked by the target object but not non-target objects that were also present¹⁰. Thus, while 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 objects³². 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 open questions. First, we cannot state with certainty that they will generalize 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, these might defy object-based classification. Similarly, scene recognition may be particularly reliant upon diagnostic objects when the range of scene categories likely to be encountered is relatively narrow, as it was in our experiment and would be in many real-world situations (e.g., having 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 presence of the signature objects in the scene exemplars. It would not be terribly surprising if a predictor based on an object strongly associated with a scene category were to produce correct classifications of scene exemplars in which that object was absent. Indeed, our ability to classify objects from their same-context counterparts in Exp. 1 indicates at least some redundancy in the patterns evoked by objects from the same context, suggesting that scene patterns in LO should be somewhat tolerant to the removal of signature objects. Finally, we have not examined the extent to which scene-evoked patterns in LO are tolerant of identity-preserving object transformations. A number of previous studies have shown that LO is sensitive to object position, size, and viewpoint^{33–37}, all of which could lead to even higher classification performance 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 varying reliably across scene categories, the ensemble of object responses evoked by scenes

in LO might form a “shorthand” code -- somewhat noisy perhaps, owing to differences in the exact object sets contained in different exemplars of each scene category, but robust enough to facilitate recognition of the scene based on these objects.

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 chance classification accuracy of scenes from object-based predictors. Furthermore, we did not observe a relationship between the patterns evoked by contextually-related objects in the PPA, contrary to what one might have expected based on previous work¹⁸. 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. Consistent with results recent neuroimaging studies, we hypothesize that the underlying representation might consist either of a statistical summary of the visual properties of the stimulus, or geometric information about the layout of the most salient spatial axes^{38,39}. With regards to the last hypothesis, it is worth noting that most of the objects in the current study were large fixed items that would play a role in determining the geometry of local navigable space. In contrast, an earlier study that compared PPA response patterns across smaller, moveable objects found no reliable differences⁴⁰. It is also noteworthy that object-based predictors failed to classify scenes in pF, despite above chance object and scene classification and the results of previous studies demonstrating even greater tolerance of identity-preserving object transformations in pF than in LO. The reasons for the failure 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 sum, our results demonstrate 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 appear to be chiefly concerned with large-scale spatial features such as spatial layout rather than the coding of within-scene objects^{6, 41, 42}. In contrast, scene-evoked patterns in LO appear to be “built” from those of their constituent objects. These results suggest that the PPA and LO can be seen as nodes along parallel pathways supporting complementary modes of scene recognition⁸, with the PPA supporting recognition based principally upon global scene properties^{3–5} and LO supporting recognition based upon the objects they contain.

Methods

fMRI

Subjects.—In Experiments 1 and 2, 28 subjects (14 subjects each; Exp. 1: 6 female, aged 19–25 years; Exp. 2: 6 female, aged 21–28 years) with normal or corrected-to-normal vision were recruited from the University of Pennsylvania community and gave written informed consent in compliance with procedures approved by the University of Pennsylvania Institutional Review Board. For Experiment 3, 14 subjects (11 female, aged 18–23 years) subject to the same inclusion criteria were recruited from the Brown University community

and 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.

MRI acquisition.—Subjects participating in Exps. 1 and 2 were scanned at the Center for Functional Neuroimaging at the University of Pennsylvania on a 3T Siemens Trio scanner equipped with an eight-channel multiple-array head coil. Subjects participating in Exp. 3 were scanned at the Brown University MRI Research Facility on a 3T 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 a 3D MPRAGE pulse sequences (TR = 1620 ms, TE = 3 ms, TI = 950 ms, voxel size = $0.9766 \times 0.9766 \times 1\text{ mm}$, matrix size = $192 \times 256 \times 160$). T2* weighted scans sensitive to blood oxygenation level-dependent (BOLD) contrasts were acquired using a gradient-echo echo-planar pulse sequence (TR = 3000ms, TE = 30ms, voxel size = $3 \times 3 \times 3\text{ mm}$, matrix size = $64 \times 64 \times 45$). At the University of Pennsylvania, the entire projected field subtended $22.9 \times 17.4^\circ$ and was viewed at 1024×768 pixel resolution; at Brown University the field subtended $24 \times 18^\circ$ 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 (Exp. 1), fourteen participated in a “short-presentation” version in which they viewed scene and object stimuli interleaved in the same scan runs (Exp. 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; Exp. 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, 8 object categories). Scene images were scaled to 9° by 9° . Object images were edited in Photoshop to remove any background information and were scaled and cropped so that the longest dimension of the object spanned 9° .

In Exp. 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 minutes 24 seconds or 6 minutes 6 seconds long, 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 once⁴³. Six repetitions of each image category were contained in the shorter scan of each pair, and 7 repetitions in the longer. Each subject was scanned with a unique continuous carryover sequence, which was repeated either two or four times.

In Exps. 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 via 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 seconds. In Exp. 2, scene and objects were interleaved in 4 continuous carryover sequences, each filling a single scan lasting 5 minutes 27 seconds. These 4 sequences were uniquely generated for each subject, and were each shown twice. In Exp. 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 minutes 20 seconds each, followed by scans in which they viewed objects only, arranged into 6 continuous carryover sequences spread across two scans of 7 minutes each.

Functional localizer scans were 6 minutes 15 seconds 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 previously⁴⁴.

MRI Analysis.—Functional images were corrected for differences in slice timing by resampling slices in time to match the first slice of each volume, 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 FWHM Gaussian filter; all other data were left unsmoothed. Data were analyzed using a general linear model as implemented in VoxBo (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 (beta values) associated with each stimulus condition. In Exps. 1 and 2, the resulting activity patterns were grouped into halves (e.g., 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 object 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 (e.g., kitchen) were constructed by taking the voxelwise average of the patterns evoked by the two objects associated with that scene (e.g., refrigerator and stove). 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 (e.g., refrigerator, tub, car, slide), and reducing the object set in the other half to include only the remaining object from each context (stove, toilet, traffic signal, swing). Patterns in each half were then labelled by context (kitchen, bathroom, playground, intersection), and the accuracy with which patterns from one half predicted the context label of the other half was assessed.

Analysis of activity patterns in Exp. 3 was similar, except that accuracy was accumulated across all four possible pairwise comparisons between the two scene and two object scans (e.g., first scene scan versus first object scan, first scene scan versus second object scan, etc). This scheme served to improve our estimates of classification accuracy by increasing the total number of unique classification decisions.

In addition to the pattern classification analyses performed within preset ROIs, we also used a “searchlight” analysis approach to identify regions of high classification accuracy throughout the brain²⁰. For each brain voxel we defined a spherical 5mm surrounding region (the searchlight cluster), performing 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 second resolution in Matlab using a low-pass interpolating filter (cutoff at 0.167 Hz) sampling symmetrically from the nearest 8 original volumes.. Response vectors for each stimulus trial were defined from the MR signal in each voxel averaged across the 4 time points from 3 to 7.5 seconds following stimulus onset.

Regions of Interest.—Functional ROIs were defined based on data from a separate set of functional localizer scans. The lateral occipital complex (LOC) was defined as the set of voxels in the lateral/ventral occipitotemporal region exhibiting stronger responses ($t > 3.5$) to objects than to scrambled objects. We divided LOC into anterior and posterior segments associated with the posterior fusiform sulcus (pF) and area LO, respectively. The

parahippocampal place area (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. Prior to any analysis, LO and pF 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 early visual cortex (EVC). 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. EVC was defined by significantly higher responses to scrambled objects than intact objects ($t > 3.5$) in the posterior occipital lobe.

Behavioral

Subjects.—Sixteen subjects (12 female, aged 19 to 28 years) with normal or corrected-to-normal vision were recruited from the University of Pennsylvania community and 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 sized 400×400 pixels. Critically, the photographs each contained two strongly diagnostic objects (e.g., a toilet and a bathtub for bathroom; a slide and swings for playground); these 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 the majority 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 and classify it as a bathroom, intersection, kitchen, or playground by pressing one of four different buttons. 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, implemented in MATLAB. ANCOVAs were implemented in SPSS.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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

References

1. Potter MC Meaning in visual search. *Science* 187, 965–966 (1975). [PubMed: 1145183]
2. Biederman I, Rabinowitz JC, Glass AL & Stacy EW On the Information extracted from a glance at a scene. *J Exp Psych* 103 597–600 (1974)
3. Schyns PG & Oliva A From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science* 5, 195–200 (1994).
4. Renninger LW & Malik J When is scene identification just texture recognition? *Vision Research* 44, 2301–2311 (2004). [PubMed: 15208015]
5. Greene MR & Oliva A Recognition of natural scenes from global properties: seeing the forest without representing the trees. *Cognit Psychol* 58, 137–176 (2009). [PubMed: 18762289]
6. Epstein R & Kanwisher N A cortical representation of the local visual environment. *Nature* 392, 598–601 (1998). [PubMed: 9560155]
7. Aguirre GK, Zarahn E & D’Esposito M An area within human ventral cortex sensitive to “building” stimuli: Evidence and implications. *Neuron* 21, 373–383 (1998). [PubMed: 9728918]
8. Park S, Brady TF, Greene MR & 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). [PubMed: 21273418]
9. Malach R, et al. Object-Related Activity Revealed by Functional Magnetic-Resonance-Imaging in Human Occipital Cortex. *Proc Natl Acad Sci U S A* 92, 8135–8139 (1995). [PubMed: 7667258]
10. Peelen MV, Fei-Fei L & Kastner S Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature* 460, 94–97 (2009). [PubMed: 19506558]
11. Epstein R, Graham KS & Downing PE Viewpoint-specific scene representations in human parahippocampal cortex. *Neuron* 37, 865–876 (2003). [PubMed: 12628176]
12. Haxby JV, et al. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2430 (2001). [PubMed: 11577229]
13. Walther DB, Caddigan E, Fei-Fei L & Beck DM Natural scene categories revealed in distributed patterns of activity in the human brain. *J Neurosci* 29, 10573–10581 (2009). [PubMed: 19710310]
14. Drucker DM & Aguirre GK Different spatial scales of shape similarity representation in lateral and ventral LOC. *Cereb Cortex* 19, 2269–2280 (2009). [PubMed: 19176637]
15. Haushofer J, Livingstone MS & Kanwisher N Multivariate patterns in object-selective cortex dissociate perceptual and physical shape similarity. *PLoS Biol* 6, e187 (2008). [PubMed: 18666833]
16. MacEvoy SP & Epstein RA Decoding the representation of multiple simultaneous objects in human occipitotemporal cortex. *Curr Biol* 19, 943–947 (2009). [PubMed: 19446454]
17. Diana RA, Yonelinas AP & Ranganath C High-resolution multi-voxel pattern analysis of category selectivity in the medial temporal lobes. *Hippocampus* 18, 536–541 (2008). [PubMed: 18446830]
18. Bar M & Aminoff E Cortical analysis of visual context. *Neuron* 38, 347–358 (2003). [PubMed: 12718867]
19. Potter MC, Staub A & O’Connor DH Pictorial and Conceptual Representation of Glimpsed Pictures. *Journal of Experimental Psychology* 30, 478–489 (2004). [PubMed: 15161380]
20. Kriegeskorte N, Goebel R & Bandettini P Information-based functional brain mapping. *Proc Natl Acad Sci U S A* 103, 3863–3868 (2006). [PubMed: 16537458]
21. Rogosa D Comparing nonparallel regression lines. *Psychological Bulletin* 88, 307–321 (1980).

22. Naselaris T, Prenger RJ, Kay KN, Oliver M & Gallant JL Bayesian reconstruction of natural images from human brain activity. *Neuron* 63, 902–915 (2009). [PubMed: 19778517]
23. Kay KN, Naselaris T, Prenger RJ & Gallant JL Identifying natural images from human brain activity. *Nature* 452, 352–355 (2008). [PubMed: 18322462]
24. Reddy L, Kanwisher NG & VanRullen R Attention and biased competition in multi-voxel object representations. *Proc Natl Acad Sci U S A* 106, 21447–21452 (2009). [PubMed: 19955434]
25. Zoccolan D, Cox DD & DiCarlo JJ Multiple object response normalization in monkey inferotemporal cortex. *Journal of Neuroscience* 25, 8150–8164 (2005). [PubMed: 16148223]
26. Moran J & Desimone R Selective attention gates visual processing in the extrastriate cortex. *Science* 229, 782–784 (1985). [PubMed: 4023713]
27. MacEvoy SP, Tucker TR & Fitzpatrick D A precise form of divisive suppression supports population coding in the primary visual cortex. *Nat Neurosci* 12, 637–645 (2009). [PubMed: 19396165]
28. Desimone R & Duncan J Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* 18, 193–222 (1995).
29. Reynolds JH & Heeger DJ The normalization model of attention. *Neuron* 61, 168–185 (2009). [PubMed: 19186161]
30. Beck DM & Kastner S Stimulus similarity modulates competitive interactions in human visual cortex. *J Vis* 7, 19 11–12 (2007).
31. Beck DM & Kastner S Stimulus context modulates competition in human extrastriate cortex. *Nat Neurosci* 8, 1110–1116 (2005). [PubMed: 16007082]
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, 10 (2007).
33. Schwarzlose RF, Swisher JD, Dang S & Kanwisher N The distribution of category and location information across object-selective regions in human visual cortex. *Proc Natl Acad Sci U S A* 105, 4447–4452 (2008). [PubMed: 18326624]
34. Sayres R & Grill-Spector K Relating retinotopic and object-selective responses in human lateral occipital cortex. *J Neurophysiol* 100, 249–267 (2008). [PubMed: 18463186]
35. Kravitz DJ, Kriegeskorte N & Baker CI High-level visual object representations are constrained by position. *Cereb Cortex* 20, 2916–2925 (2010). [PubMed: 20351021]
36. Andresen DR, Vinberg J & Grill-Spector K The representation of object viewpoint in human visual cortex. *Neuroimage* 45, 522–536 (2009). [PubMed: 19100844]
37. Eger E, Kell CA & Kleinschmidt A Graded size sensitivity of object-exemplar-evoked activity patterns within human LOC subregions. *J Neurophysiol* 100, 2038–2047 (2008). [PubMed: 18632884]
38. Epstein RA Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends Cogn Sci* 12, 388–396 (2008). [PubMed: 18760955]
39. Kravitz DJ, Peng CS & Baker CI Real-World Scene Representations in High-Level Visual Cortex: It's the Spaces More Than the Places. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 31, 7322–7333 (2011). [PubMed: 21593316]
40. Spiridon M & Kanwisher N How distributed is visual category information in human occipito-temporal cortex? An fMRI study. *Neuron* 35, 1157–1165 (2002). [PubMed: 12354404]
41. Habib M & Sirigu A Pure topographical disorientation: a definition and anatomical basis. *Cortex* 23, 73–85 (1987). [PubMed: 3568707]
42. Aguirre GK & D'Esposito M Topographical disorientation: a synthesis and taxonomy. *Brain* 122, 1613–1628 (1999). [PubMed: 10468502]
43. Aguirre GK Continuous carry-over designs for fMRI. *Neuroimage* 35, 1480–1494 (2007). [PubMed: 17376705]
44. Epstein RA & Higgins JS Differential parahippocampal and retrosplenial involvement in three types of visual scene recognition. *Cereb. Cortex* (2006).
45. Epstein RA, Parker WE & Feiler AM Where am I now? Distinct roles for parahippocampal and retrosplenial cortices in place recognition. *J Neurosci* 27, 6141–6149 (2007). [PubMed: 17553986]

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 were drawn from any of the scene exemplars.

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). Correct scene-from-object classification decisions occurred when actual scene patterns were more similar to predictors based on their own associated objects than to the predictors 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 each of the three experiments (E1, E2, and E3). 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 are s.e.m. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Figure 4. Classification of objects based on the patterns elicited by their same-context counterpart objects (e.g. stoves from refrigerators). Accuracy was significantly above chance in LO in Exp. 1, (E1), but not above chance in any ROI in Exps. 2 or 3 (E2, E3). Error bars are s.e.m. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Figure 5.

Group random-effects analysis of local searchlight accuracy maps for classification of scenes from object averages. Painted voxels represent centers of searchlight clusters with above-chance classification accuracies ($p < 0.005$, uncorrected). Displayed slices are cardinal planes containing occipitotemporal voxel of peak significance, which was found in the left hemisphere. Outlined regions are LO (blue), pF (cyan) and PPA (green), defined from random effects analysis of volumes across subjects ($p < 0.00001$, uncorrected). Note that although pF and PPA overlap when defined on these group data, they were non-overlapping when defined at the individual subject level. The apparent bias towards higher performance in left LO is addressed in the Supplementary Results. Ventral cluster of high accuracy in sagittal slice corresponds to motor cortex.

Figure 6.

Behavioral evidence for object-based scene recognition. (a) Subjects viewed 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 total reaction time (right) are shown for images with zero, one, or two objects masked. Data for conditions with objects removed are shown for three different ranges of the percentage of image pixels masked with object removal. For matched percentages of pixel deletion, accuracy and reaction time were significantly degraded when both signature objects were removed compared to just one; this effect was only significant when a high percentage of the scene was masked, likely corresponding to the range where image-based identification falters. Accuracy and RT estimates are from application of the Johnson-Neyman procedure, which does not produce error bars. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.