



# Statistics of high-level scene context

Michelle R. Greene \*

Department of Computer Science, Stanford University, Stanford, CA, USA

**Edited by:**

Gregory Zelinsky, Stony Brook University, USA

**Reviewed by:**

Ben Tatler, University of Dundee, UK  
Ruth Rosenholtz, Massachusetts Institute of Technology, USA

**\*Correspondence:**

Michelle R. Greene, Department of Computer Science, Stanford University, 353 Serra Mall Rm 241, Stanford, CA 94305, USA  
e-mail: mrgreene@stanford.edu

Context is critical for recognizing environments and for searching for objects within them: contextual associations have been shown to modulate reaction time and object recognition accuracy, as well as influence the distribution of eye movements and patterns of brain activations. However, we have not yet systematically quantified the relationships between objects and their scene environments. Here I seek to fill this gap by providing descriptive statistics of object-scene relationships. A total of 48,167 objects were hand-labeled in 3499 scenes using the LabelMe tool (Russell et al., 2008). From these data, I computed a variety of descriptive statistics at three different levels of analysis: the ensemble statistics that describe the density and spatial distribution of unnamed “things” in the scene; the bag of words level where scenes are described by the list of objects contained within them; and the structural level where the spatial distribution and relationships between the objects are measured. The utility of each level of description for scene categorization was assessed through the use of linear classifiers, and the plausibility of each level for modeling human scene categorization is discussed. Of the three levels, ensemble statistics were found to be the most informative (per feature), and also best explained human patterns of categorization errors. Although a bag of words classifier had similar performance to human observers, it had a markedly different pattern of errors. However, certain objects are more useful than others, and ceiling classification performance could be achieved using only the 64 most informative objects. As object location tends not to vary as a function of category, structural information provided little additional information. Additionally, these data provide valuable information on natural scene redundancy that can be exploited for machine vision, and can help the visual cognition community to design experiments guided by statistics rather than intuition.

**Keywords:** context, scene, ensemble, bag of words, data mining, scene understanding

## INTRODUCTION

Imagine that you are attending a friend’s housewarming party. Although you have never been in this house before, you are not surprised to find a coffee table next to a sofa in the living room, chairs surrounding the dining room table, or framed pictures hanging on the walls. As a considerate guest, you help with the cleanup afterwards; effortlessly finding the trash can under the sink for disposing the waste, and the dishwasher next to the cabinets to wash the dishes. Our interactions in the world are facilitated by virtue of the fact that objects are not randomly strewn about the world but follow some basic laws of where they may be located, how large they are, and what other objects will be found near them. Collectively, these regularities are known as *context*. While context appears to be crucial for human scene recognition and helpful for machine vision (see Bar, 2004; Oliva and Torralba, 2007 for reviews), contextual relations between scenes and their objects have not yet been systematically measured and cataloged.

## WHY MEASURE STATISTICS OF OBJECT CONTEXT?

The last two decades have seen a growing literature on the statistics of natural images. Knowing about the input received by our visual systems allows for a better understanding of visual coding

in the brain. We have a growing understanding of the statistical regularities of natural scenes at the level of basic features such as luminance, contrast, color and Fourier amplitude spectra, as well as the relations between edges and contours (Olshausen and Field, 1996; van Hateren and Ruderman, 1998; Fine and MacLeod, 2001; Geisler et al., 2001; Schwartz and Simoncelli, 2001; Golz and MacLeod, 2002; Torralba and Oliva, 2003; Howe and Purves, 2004). Mid-level regularities have been found for scene textures (Torralba and Oliva, 2003) as well as scene scale and depth (Ruderman, 1994; Torralba and Oliva, 2002). Higher-level statistical regularities, such as object location (Karklin and Lewicki, 2003), scene background to objects (Torralba and Sinha, 2001; Choi et al., 2010) and scene spatial structure (Schyns and Oliva, 1994) have also been measured. The importance of this work lies in the predictive power of image statistics for both behavior and neural responses (Rao et al., 2002; for reviews, see Simoncelli and Olshausen, 2001; Geisler, 2008). It has been hypothesized that the visual system exploits statistical redundancies in order to efficiently code a complex visual world (Attneave, 1954; Zetsche et al., 1993; Barlow, 2001). Thus, knowing the statistical dependencies between objects and scenes can help us to understand the types of compressed visual codes that allow us to rapidly recognize our visual environments.

Despite a large and growing literature on the effects of context on object and scene recognition (Palmer, 1975; Friedman, 1979; Biederman et al., 1982; Boyce et al., 1989; De Graef et al., 1990; Henderson, 1992; Bar and Ullman, 1996; Hollingworth and Henderson, 1998; Henderson et al., 1999; Davenport and Potter, 2004; Eckstein et al., 2006; Neider and Zelinsky, 2006; Auckland et al., 2007; Becker et al., 2007; Davenport, 2007; Joubert et al., 2007; Vö and Henderson, 2009; Mack and Palmeri, 2010), there has yet to be a systematic quantification of scene-object relationships in the world. This is a critical step as recent work has found that principles of attention and perception learned from artificial laboratory stimuli have limited generalizability to real-world stimuli (Neider and Zelinsky, 2008; Wolfe et al., 2011a,b). Here I seek to fill this gap by providing both descriptive statistics of contextual relations and inferential statistics to show how much these types of context can contribute to scene categorization.

Suppose you wanted to know whether object recognition benefits from lawful scene context (e.g., Davenport and Potter, 2004). Traditionally, one would approach the problem by embedding the object of interest in normal a scene context (e.g., a “blender” in a *kitchen*), or an abnormal scene context (e.g., a “blender” in a *bathroom*), and then have human observers perform an object categorization task on both types of stimuli. Similarly, what if you wanted to study the degree to which an object evokes a particular scene context (e.g., Bar and Aminoff, 2003). Or perhaps you are interested in how scenes are formed from diagnostic objects (e.g., MacEvoy and Epstein, 2011). In each of these cases, how do you choose the object and scene contexts that you will use? How do we define diagnosticity for objects, and how do we measure it? Are all abnormal contexts equally bad? In each of these cases, these questions have been answered through introspection and intuition. The aim of this work is to provide baseline statistics of objects in scenes so that these types of questions can be answered with quantitative measures.

### THEORIES OF OBJECT CONTEXT

One of the first theories of object-scene context was known as *frame* or *schema* theory (Bartlett, 1932; Minsky, 1975; Friedman, 1979; Biederman, 1981). According to this theory, scene categories can be represented in a mental structure containing learned associations between the category and objects that are commonly found in it. For example, a *kitchen* schema might activate representations of objects such as “refrigerator,” “blender,” and “cutting board.”

Biederman et al. (1982) argued that there are five object-scene relationships that constitute well-formed visual scenes. Scenes must obey the laws of physics, with objects *supported* by a horizontal surface, and not occupying the same physical space (*interposition*). Furthermore, the objects themselves have a certain *likelihood* of being in a particular scene context, as well as some some probable *position* in it. Finally, every object is constrained to have a particular *size* relative to the other objects in the scene. The first two relationships describe physical constraints on the world, while the last three describe the semantic content of the scene. These authors found that violations in any of these

relationships resulted in reaction time and accuracy deficits for object recognition within a scene, that multiple violations made performance worse, and that both types of relations (physical and semantic) disrupted scene and object processing to similar degrees.

Much of the experimental work on scene-object context has focused on the likelihood contextual relation, often referred to as *consistency*. It is generally accepted that a consistent object in a scene facilitates object and scene recognition (Palmer, 1975; Loftus and Mackworth, 1978; Boyce et al., 1989; De Graef et al., 1990; Bar and Ullman, 1996; Hollingworth and Henderson, 1998; Davenport and Potter, 2004; Eckstein et al., 2006; Becker et al., 2007; Joubert et al., 2007; Vö and Henderson, 2009, 2011; Mack and Palmeri, 2010). However, an open debate still exists over whether this facilitation is perceptually or cognitively based (Hollingworth and Henderson, 1998; Henderson and Hollingworth, 1999; Bar, 2004). The details of this argument are beyond the scope of this paper.

As there are no existing norms for object frequencies in scenes, it is often left to the intuitions of the experimenters to determine which objects are consistent or inconsistent in a scene category. Two salient exceptions include Friedman (1979), who obtained normative rankings by asking participants to brainstorm lists of objects that have various probabilities of being in a particular scene, and Henderson (1992), who provided a pilot experiment where the object-scene pairs were verified by an independent group of observers. However, in the absence of ground truth measurements of object frequency, the notion of object consistency seems to be better capturing object *plausibility* rather than object probability. For example, in Davenport and Potter (2004), a “sand castle” was chosen to be the consistent object for a *beach* scene. While sand castle is a very plausible object in a beach scene, most beaches are unlikely to have sand castles, making “sand castle” a plausible, but low-probability object. By measuring contextual statistics of objects and scenes, we can revisit the consistency effect with experiments reflecting real-world probabilities rather than intuitions.

There is also general agreement that context involves some form of learned associations extracted from interactions in the world. For example, in the phenomenon of contextual cueing (Chun and Jiang, 1998), observers’ reaction times in repeated visual search displays become more efficient over the course of an experiment, suggesting that they implicitly learned the spatial layout of the displays. Brockmole and Henderson (2006) have shown that displaying a letter search array on a real-world scene and consistently pairing the target location with a location in the scene also produces contextual cueing. Furthermore, this result generalizes across different scenes in the same category (Brockmole and Vö, 2010), suggesting that object-scene relationships can be implicitly extracted and used for visual search. However, there remain a number of open questions: how strong are the contextual relations between objects and scenes? Does the strength of the relations differ among different types of scenes (e.g., indoor, natural landscapes, or urban environments)? Understanding and characterizing these relationships will allow the formulation of new experiments examining the extent to which human observers

use various types of context for recognition, search and memory of complex natural scenes.

### USING CONTEXT FOR RAPID SCENE RECOGNITION

The mechanism behind the remarkable rapid scene categorization performance of human observers has been a long-standing mystery. How is a scene recognized as quickly as a single object when scenes contain many objects? Biederman (1981) outlined three paths by which an initial scene representation could be generated: (1) by recognizing a prominent object that is diagnostic of the scene's category; (2) by perceiving and recognizing global, scene-emergent features that were not defined; or (3) by recognizing and spatially integrating a few contextually related objects.

Although global, scene-specific features have been shown to be useful for scene categorization (Greene and Oliva, 2009), observers are able to also report a few objects after a brief glance at a scene (Fei-Fei et al., 2007). The first and third paths outlined by Biederman have been sparsely explored, as what counts as a "diagnostic" or "contextual" object is not immediately obvious. In this work, I operationalize these concepts so we may begin to test these hypotheses.

### SCOPE OF THE CURRENT WORK

In this paper, I introduce a large scene database whose objects and regions have been fully labeled using the LabelMe annotation tool (Russell et al., 2008). The fully labeled data contain names, sizes, locations, and 2D shapes for each object in each scene. In this work, I will provide descriptive statistics on these data at three levels of description: statistical ensembles, bag of words and structural. At the ensemble level, I will examine the overall object density and spatial distribution of unnamed objects and regions across the scene categories. The bag of words level of description uses the object labels to determine which objects occur in which scene categories without regard to the spatial distribution of these objects. The structural description will then examine the spatial relations among objects across scene categories. For each level of description, I will also describe how sufficient these statistics are for predicting scene categories through use of a linear classifier, and discuss how human observers may employ such strategies for rapid scene recognition.

## METHODS

### SCENE DATABASE

The main scene database consists of 3499 full-color scene photographs from 16 basic-level categories. Eight of the basic-level categories are indoor environments (*bathroom, bedroom, conference room, corridor, dining room, kitchen, living room, and office*). These images were downloaded from the web. The remaining scene categories were outdoor environments taken from Oliva and Torralba (2001), with four categories representing urban environments (*skyscrapers, street scenes, city centers, and highways*) and four categories representing natural environments (*coast, open country, mountain, and forest*). There were at least 94 images in each of the 16 basic-level categories. The images varied in size and were selected from a large lab database amassed from the web, personal photographs and books. See **Figure 1** for example images from each basic-level category.

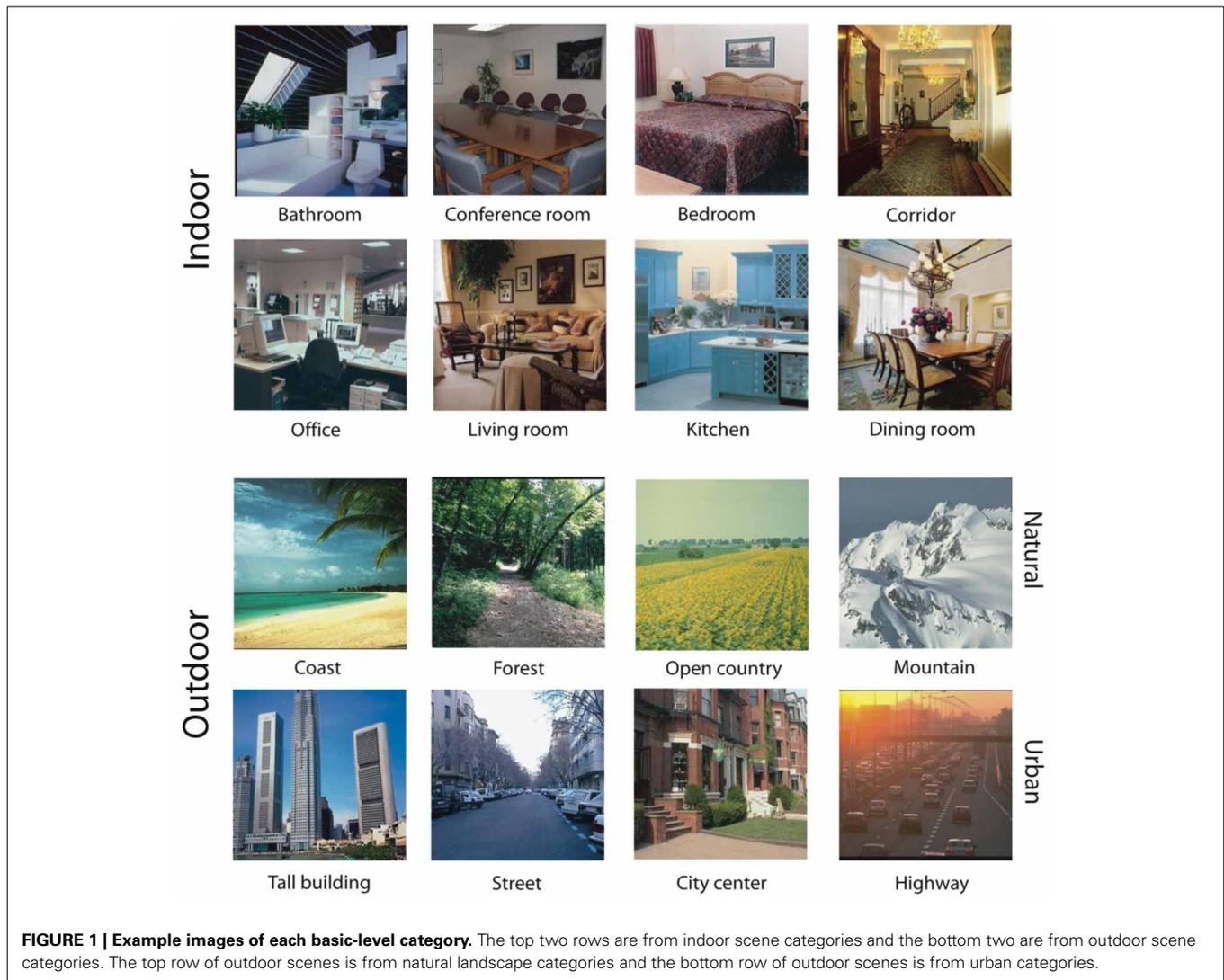
### LABELING PROCEDURE

The image database<sup>1</sup> was hand segmented and labeled using the LabelMe Open Annotation Tool (<http://labelme.csail.mit.edu>, Russell et al., 2008) by four observers (including the author) over the period of several months. Observers were instructed to label all regions and objects in each image and affix the best basic-level name to each region as well as to label objects as individuals, size permitting (e.g., annotate each apple in a bowl of apples except in cases where apples were too small to create an accurate bounding region). It was decided in advance that objects that could be seen through windows would not be annotated, as these objects are not located in the given scene environment. Similarly, objects whose reflections appeared in mirrors were not annotated because this would artificially inflate the count of this object in the scene. Namable parts of objects that are not separable from the object (e.g., the leg of a chair, or headlight of a car) were not labeled. For the labelers, any visual, namable entity counted as an object, so items such as "fog," "cloud," or "sand" were considered objects. Although one typically thinks of "objects" as discrete entities that do not comprise the structure of a scene, regions vary in their "objectness." In order to avoid idiosyncratic labeling strategies, all regions were considered. In cases of occlusion, labelers were instructed to interpolate object boundaries as to do otherwise would increase the count of this type of object. Statistical analysis on these annotations was performed in Matlab using the LabelMe toolbox (Russell et al., 2008).

### CLEANING THE DATABASE

As the LabelMe interface accepts any label typed by an observer, the raw annotations contained numerous typos, misspellings and synonyms. Raw labels were hand-corrected to ameliorate these issues. These changes reduced the number of unique annotations from 1767 to 617. Misspelled items accounted for 21% of the changes (for example "automan" for "ottoman"). Plurals were changed to singular, accounting for 15% of the changes. Labels that were written at the subordinate level, including descriptions of position information ("side view of car" or "slated wooden panel") were changed to the appropriate entry-level category. These accounted for 40% of the changes. Furthermore, items listed at the superordinate level were visually inspected and assigned to the appropriate entry-level category, accounting for 3% of the changes. For example, "art" was a label that referred to a "painting" in one image and a "sculpture" in another, and "island" could refer to either a landmass in water or counter space in the center of a kitchen. In cases where the entry-level category of an object was questionable, attempts were made to group objects by function. For example, "decoration" was chosen as an entry-level as all objects under this label served a common function (e.g., "decorative wall hanging" or "decorative fish"). Object labels that were synonyms according to WordNet (Miller, 1995) were unified under one label (for example, "couch" and "sofa"). Synonyms accounted for 16% of the changes. Labels that encompassed multiple objects (for example, "basket of magazines") were included as the containing, or larger object only

<sup>1</sup>Upon publication, this database and Matlab structures containing statistical data will be publically available for download on the author's website.



(e.g., “basket,” 2% of changes). Labels that referred to object parts that are not independent of the object whole (e.g., “chair leg” is a part that is not removable from a chair without the chair losing its function) were deleted. Parts that could refer to the whole object (e.g., “back of chair” for a chair that was occluded except for the back) were changed to the object’s name. These accounted for 2% of the changes. Finally, there were 288 labels that were simply called “object.” These referred to a variety of small objects that could not be accurately identified from the small images, so the label has not been changed. There were a total of 21 deletions. The list of deletions can be found in Appendix C. A list of raw and final labels can be found in Appendices A and B, respectively.

#### AUXILIARY DATASET

Although ~3500 images is a relatively large database and near the practical limit of what one can hand-annotate, a critical question for the utility of these statistics is the degree to which they generalize to the population of all real-world environments. Indeed, dataset bias is known to limit the knowledge gleaned from this

type of inquiry (Torralba and Efros, 2011). In order to address this question, I compared the contextual statistics from the main database with a completely independent labeled database. As every database has independent bias, the extent to which statistics measured in one database can be successfully applied to another reflects the generalizability of the database.

I created an auxiliary set of images taken from the LabelMe database and annotated by unknown observers. The dataset consisted of 1220 images from the same 16 basic-level scene categories that had at least 85% label coverage. There were 100 images per category for *bathroom*, *kitchen*, *living room*, *city*, *street*, *coast*, and *forest*, and 14–59 in the others, as LabelMe does not have a sufficient number of fully labeled scenes for the other categories. These scenes were labeled by unknown observers without the rules used by the four observers who annotated the main set. As LabelMe allows users to upload their own photographs, this dataset differs from the main dataset in that the depicted environments are less idealized and stylized and seem to come from users snapping views of their own offices, kitchens and streets (see the

**Figure A1**). This set was cleaned as described above. All analyses were repeated on this additional set, and all differences between the two databases are noted in Appendix D.

**RESULTS**

**GENERAL FINDINGS**

**Quality of annotations**

How much of each image was labeled? Although labelers were instructed to label each pixel, some regions were too small to label, or may have been overlooked. Here I examined the percentage of image pixels assigned to a label. On average, 85.4% of an image’s pixels were assigned to a label (standard deviation: 12.7%). Sixty one percent of images had more than 90% of its pixels labeled. By contrast, only 8.9% of images in the LabelMe database have this level of annotation (Russell et al., 2008) making the main database better suited to describing contextual statistics.

**ENSEMBLE STATISTICS**

Ensemble statistics are statistical summaries of a group of objects, such as mean size (Ariely, 2001; Chong and Treisman, 2003, 2005), center of mass (Alvarez and Oliva, 2008), or mean orientation (Parkes et al., 2001). Although most work in this area has been on laboratory displays of simple shapes, human observers can estimate ensemble statistics over more complicated sets of features as well, such as the average emotion of a crowd of faces (Haberman and Whitney, 2007). Recent work in visual cognition has shown that the human visual system is adept at representing such ensembles both rapidly and outside the focus of attention (Ariely, 2001; Chong and Treisman, 2003, 2005; Alvarez and Oliva, 2008, 2009; for a review see Alvarez, 2011). Although the use of statistical ensembles has been posited as a potential mechanism for scene gist recognition (Haberman and Whitney, 2012), there has been little work on what statistical ensembles might exist in real-world images.

Here, I examined several summary statistics representing the general density and location of objects in scene categories. The utility of each measure for scene categorization is assessed both individually and as a group using a linear classifier.

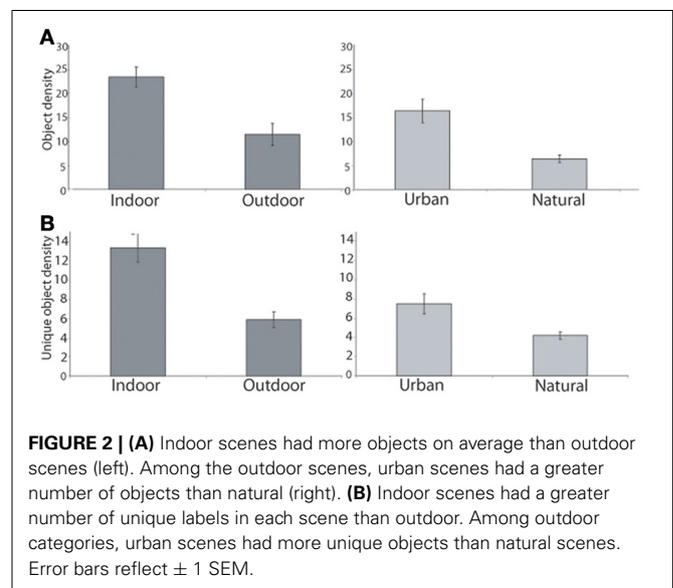
**Object density and variety**

The first ensemble statistic is simply the density of labeled objects in each scene. The number of objects in a scene ranged from 1 to 88 (median: 11). Do all scene categories have a similar number of objects? To answer this question, I examined the number

of objects per scene as a function of basic- and superordinate-level scene category labels. While human observers have no problems recognizing scenes with a variety of object densities (Potter, 1976; Wolfe et al., 2011a,b), classical visual search experiments show clear performance costs as the number of objects in a display increases (Biederman, 1988; Vickery et al., 2005). In order to better understand how the number of objects in a scene affects categorization performance in that scene, it is important to first understand how scenes vary in terms of object density.

In this database, the mean object density ranged from 5.1 objects per *mountain* scene to 33.1 objects per *kitchen* (see **Table 1A**). As shown in **Figures 2A, 3**, indoor scenes had a significantly higher mean object density than outdoor scenes [23.45 and 11.44 objects per scene, respectively,  $t_{(14)} = 4.14, p < 0.01$ ], and among the outdoor scenes, urban environments had a significantly higher average density than natural [16.43 vs. 6.46 objects per scene, respectively,  $t_{(6)} = 4.52, p < 0.01$ ]. This indicates that the degree of human intervention in an environment results in more labeled regions and objects.

I also examined the number of unique objects in each scene: the larger this number, the greater the heterogeneity and possibly the complexity of the scene. In the database, the number of unique items in a scene ranged from 1 to 42 (median: 6).

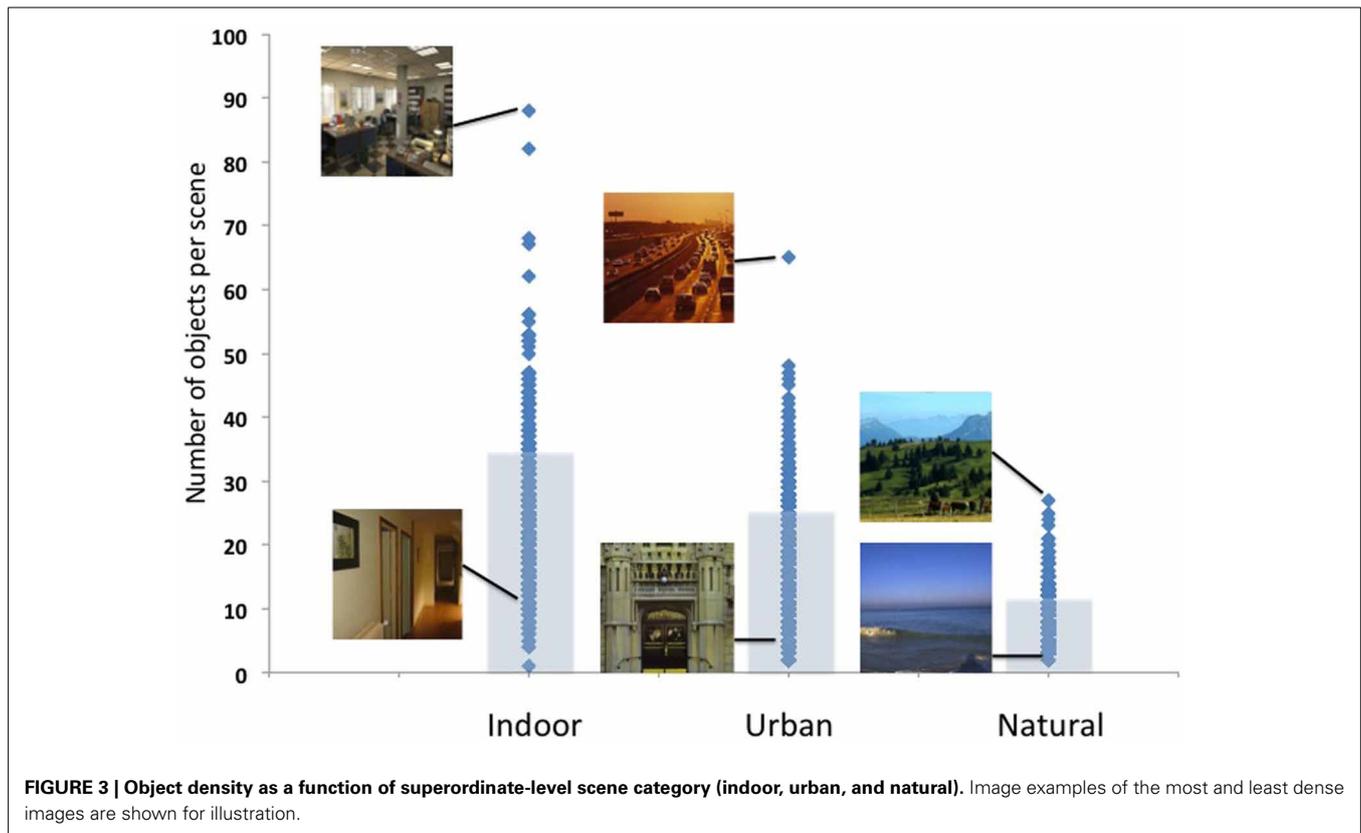


**FIGURE 2 | (A)** Indoor scenes had more objects on average than outdoor scenes (left). Among the outdoor scenes, urban scenes had a greater number of objects than natural (right). **(B)** Indoor scenes had a greater number of unique labels in each scene than outdoor. Among outdoor categories, urban scenes had more unique objects than natural scenes. Error bars reflect  $\pm 1$  SEM.

**Table 1 | (A) The mean number of total labeled regions per scene for each of the basic-level scene categories; (B) the mean number of uniquely labeled regions per scene.**

	Indoor							Urban				Natural				
	Bath	Bed	Conf.	Corr.	Dine	Kit' n.	Liv.	Off.	Tall	City	Strt.	High	Cst.	OpC	Mntn.	Frst.
(A)	20.2	18.3	25.8	15.6	22.6	33.1	24.8	27.1	12.3	20.0	20.1	13.3	5.5	7.4	5.1	7.8
(B)	14.7	12.5	8.6	7.9	12.2	19.5	15.2	15.9	5.0	9.0	8.4	7.7	4.4	7.7	3.2	4.5

Scene category abbreviations (left to right) are: bathroom, bedroom, conference room, corridor, dining room, kitchen, living room, office, tall building (skyscraper), city, street, highway, coast, open country, mountain and forest. This convention will be followed for all tables in this article.



The number of unique regions in a scene varied from 19.5 in a *kitchen*, to 3.2 in *mountain* scenes (see **Table 1B**). As with total object density, there were more unique objects in indoor scenes when compared with outdoor scenes [13.3 and 5.7 unique items per scene,  $t_{(14)} = 4.78$ ,  $p < 0.001$ ], and among outdoor scene categories, more unique objects in urban scenes compared with natural scenes [7.55 and 3.93 unique items per scene,  $t_{(6)} = 3.76$ ,  $p < 0.01$ , see **Figure 2B**). Manufactured environments therefore have both a greater number and greater variety of objects than natural environments.

#### Mean and variance of object size

Human observers are able to quickly and accurately compute the mean size of objects in laboratory displays (Ariely, 2001; Chong and Treisman, 2003, 2005). Are statistical properties of object size diagnostic of scene category? Although object size and density are related, it is important to consider that the two-dimensional labeling of a three-dimensional scene results in overlapping polygons. For example, a pillow on a bed will overlap with the bed, or a chair in front of a table with overlap with the table. Thus, mean object size is not trivially the inverse of object density.

For each scene, the size of each object was expressed as percent of total image area. In general, labeled regions were relatively small (the median mean object size was 17% of the total image area). There was considerable range in mean object size in our database, from a minuscule 0.05% of image area to a nearly all-encompassing 99.4%. Among basic-level categories, *living rooms* had the smallest mean object size (5% of

image area) and *mountains* had the largest (43%), see **Table 2**. Predictably, indoor scenes had a smaller mean object size compared to outdoor scenes [7.5 vs. 24.4%,  $t_{(14)} = 4.94$ ,  $p < 0.001$ ]. Among the outdoor superordinate-level categories, natural scenes trended toward having a larger mean object size compared to urban [30.2 and 18.6%, respectively,  $t_{(6)} = 2.28$ ,  $p = 0.063$ ].

Next, I examined object size variance across basic- and superordinate-level scene categories. For each scene, the size variance of all objects in the scene was computed. For each basic-level category, I computed the mean of object size variance, finding that *living rooms* had the smallest variance of object size, and *mountains* had the largest. Overall, indoor scenes had smaller variance of mean object size compared to outdoor [ $t_{(14)} = 5.84$ ,  $p < 0.001$ ], but no reliable difference was found between natural and urban scenes [ $t_{(6)} = 1.07$ , n.s.].

#### Center of mass

The previous ensemble statistics have shown us that, relative to outdoor environments, indoor scenes have a higher density of objects, and lower variance of object size. However, these do not tell us anything about where these objects are located in the scene. Previous work has shown that human observers are sensitive to the center of mass of a group of objects and can accurately compute this location even when attention is diverted elsewhere (Alvarez and Oliva, 2008). Are there robust differences in the locations of objects in different basic- and superordinate-level scene categories?

**Table 2 | Mean (top) and standard deviation (bottom) of mean object size in percentage of total image area.**

Indoor								Urban				Natural			
Bath	Bed	Conf.	Corr	Dine	KiLn.	Liv.	Off.	Tall	City	Strt	High	Cst.	OpC	Mn tn.	Frst.
5.7	6.9	10.3	14.0	7.1	5.4	5.0	5.3	24.9	17.4	20.0	12.0	23.9	25.2	42.9	29.0
2.7	3.3	5.5	7.4	3.8	3.7	2.1	2.7	15.4	11.6	9.5	6.5	11.6	12.1	15.3	22.4

For each scene, the center of each object was computed as (xMax-xMin, yMax-yMin) of the polygon vertices. The center of mass for the scene was then computed as the mean of these values, weighted by the size of the object (as computed above). As expected, there was a strong tendency for the objects to center along the vertical axis (basic-level category centroids were located between 46 and 53% of total horizontal extent), indicating that objects were located with equal probabilities in the left and right sides of a scene. I observed a certain degree of diversity in position in the vertical axis, with basic-level category centroids occupying 35–75% of the vertical axis. This makes sense, as vertical location is a possible cue for scene depth. In particular, outdoor environments had a higher center of mass in the image plane (65% of vertical axis) than indoor environments [47% of vertical axis,  $t_{(14)} = 4.09$ ,  $p < 0.01$ ], reflecting the presence of objects such as skyscrapers, buildings and sky. However, no systematic difference was found between the natural and urban outdoor scenes [ $t_{(6)} < 1$ , n.s.]. Therefore, vertical center of object mass may contain diagnostic information for scene category.

### Object spacing regularity

The center of object mass tells us about the general location of objects in an image, but this statistic does not tell us about the spacing of these objects. Objects that cluster together can be perceptually grouped (Gestalt law of proximity), and may have functional relations in a scene. Do scene categories vary in their object spacing regularity?

For each scene, pairwise distances between each of the objects in the scene were computed, using the (x,y) locations of each object's center of mass. Then for each scene, I computed the variability of object spacing as the standard deviation of distances, normalized by the mean distance between each object. Normalizing by the mean allows us to compare images that were not the same size. A low degree of variability indicates a very regular, grid-like spacing of objects while a high degree of variability suggests a more clustered spatial layout. While basic-level categories varied in their degrees of spacing variability, no systematic differences were found between indoor and outdoor scenes [ $t_{(14)} = 1.04$ , n.s.] nor between the natural and urban outdoor scenes [ $t_{(6)} = 1.12$ , n.s.]. This seems to be partially due to the fact that indoor scene categories were themselves quite variable: *bathroom* scenes displayed the highest degree of object spacing regularity of all 16 categories while *living rooms* displayed the lowest.

### Scene classification with an ensemble statistics model

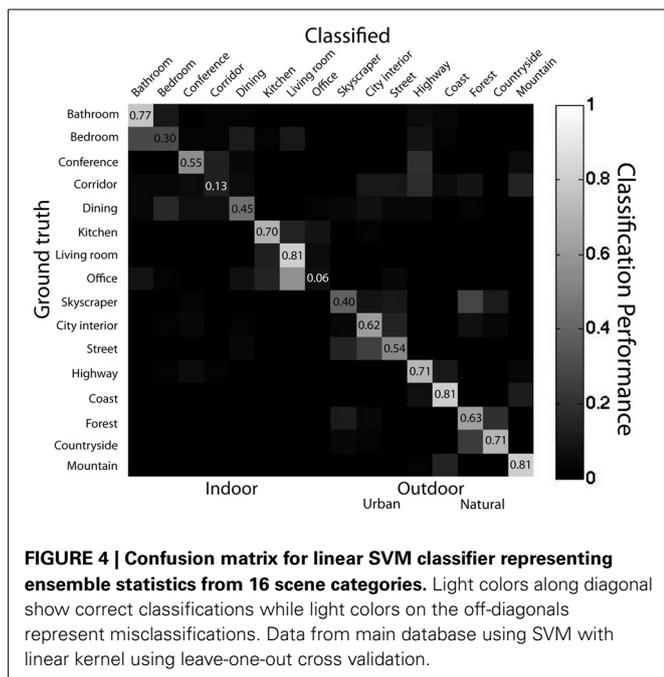
To what extent do these ensemble statistics provide information about the basic- and superordinate-level scene categories? To examine this question, I expressed each image in the database

according to its object density, unique object density, mean object size, object size variance, center of mass and variability of object spacing. Using a support vector machine (SVM) classifier [linear kernel, using LIBSVM, Chang and Lin (2011)], I tested basic- and superordinate-level scene categorization. Each image was separately used as a test image after training with the remaining 3498 images in the database. This procedure was the same for all SVM analyses in this manuscript. LIBSVM uses a one-against-one multi-class classification, with all parameters remaining the same for each classification task. All default parameters for LIBSVM were employed. For the superordinate-level categorization task, the classifier achieved an accuracy of 91% correct for natural, 63.4% for urban and 76.5% for indoor scenes (overall AUC: 0.83). This overall level of performance is well above the chance level of 33% (binomial test, all  $p < 0.0001$ ).

For the basic-level categorization task, mean performance was 61% correct (AUC = 0.77), well above the chance level of 6.25%. Performance on each basic-level category ranged from 6% for *offices* to 81% for *living rooms*. Binomial tests on the performance of each basic-level category indicated that all categories except for *office* were classified above chance ( $p < 0.01$ ). Basic-level classification performance did not differ significantly between outdoor and indoor scene categories [65 and 47% correct, respectively,  $t_{(14)} = 1.7$ ,  $p = 0.10$ ], nor did basic-level categorization performance differ among natural and urban scene categories [73 and 55% correct, respectively,  $t_{(6)} = 1.8$ ,  $p = 0.12$ ]. Incorrectly classified *offices* were frequently classified as *living rooms* (67% of mistakes), *kitchens* (13%), or *bathrooms* (10%), see **Figure 4** for full confusion matrix.

What was the nature of the misclassifications? Twenty seven percent of misclassified scenes were misclassified across the indoor-outdoor superordinate distinction (i.e., a scene was indoor and was classified as one of the outdoor categories). This pattern is unlike human scene categorization, where mistakes are nearly always within the same superordinate-level category (Renninger and Malik, 2004; Kadar and Ben-Shahar, 2012). Although this classifier has remarkably high performance given its simplicity, the pattern of performance suggests that human observers use different or additional information for performing rapid scene categorization.

How does each of the ensemble statistics contribute to classification performance? To address this question, I performed the same SVM analysis as described above, using only one ensemble statistic at a time. Basic-level categorization performance was above chance for each of the ensemble statistics (binomial test, all  $p < 0.001$ ), and ranged from 10.2% for center of mass to 31.6% for spacing regularity. A one-way ANOVA on the accuracy of each classifier revealed significant differences in performance ( $p < 0.001$ ), suggesting that certain ensemble statistics



are more useful than others for categorization. As shown in section **Object Spacing Regularity**, the regularity of object spacing did not differ reliably among superordinate level scene categories, even though it has the highest basic-level categorization performance when tested alone, indicating that this feature carries information about a scene's basic-level category, but not superordinate-level category.

In order to understand how the dimensionality of these features affects classification performance, I ran classifiers trained and tested on each combination of 2–6 ensemble statistics. Classification performance grew linearly in this range (slope: 8.3% per feature,  $r^2 = 0.98$ ). Extrapolating, ceiling performance could be expected with 11 ensemble features.

How well do ensemble statistics from the main database generalize to the auxiliary database? I trained a linear SVM on ensemble statistics from the main database, and tested categorization performance on the auxiliary database. Above-chance performance of this classifier indicates shared information because bias between the two databases should not be strongly correlated. Indeed, basic-level categorization performance for a model trained on the main database and tested on the auxiliary set was 17% ( $AUC = 0.52$ ), significantly above chance level (binomial test,  $p < 0.001$ ), indicating that ensemble statistics measured from one database contain information about the pattern of ensemble statistics in an independent database.

How does classifier performance compare to human performance on a rapid scene categorization task? Here, I compared the classifier to the human categorization data of Kadar and Ben-Shahar (2012) who tested observers on 12 of the 16 categories in the current database. In their experiment, participants were briefly shown two images and were then asked to determine whether the images were in the same category. Images were presented for 27–1000 ms and followed by a  $1/f$  noise mask.

The authors published confusion matrices for the scene categories averaged over presentation time. Overall, sensitivity of the ensemble statistics classifier was lower than that of the human observers [mean  $A'$  = 0.65 for classifier, 0.85 for participants  $t_{(22)} = 7.7$ ,  $p < 0.0001$ ]. When comparing the confusion matrices of the classifier to those of the human observers, I found that although the patterns of classifier confusion were not well correlated with human error patterns at the basic-level ( $r = 0.04$ ), error patterns were quite similar when averaged over superordinate-level categories ( $r = 0.79$ ). Therefore, the ensemble classifier can predict human performance at rapid scene categorization at a coarse level, adding support for the plausibility of such a coding scheme as a mechanism for scene gist perception.

Together, these analyses show that simple ensemble statistics, such as the number and location of nameless objects, are sufficient for above-chance scene categorization at both the basic and superordinate levels, and that the pattern of performance mimics human categorization performance at a coarse level.

### Ensemble statistics discussion

In this section, I have described real-world images in terms of very simple statistics that express the quantity and coarse spatial distributions of “things” in a scene. These are of interest because they are rapidly computed by human observers on laboratory displays (Parkes et al., 2001; Chong and Treisman, 2005; Haberman and Whitney, 2007) and may explain aspects visual representations outside the fovea (Balas et al., 2009) or outside the focus of attention (Alvarez and Oliva, 2008, 2009).

Descriptively, these results show that statistical ensembles vary considerably with the degree of human manufacturing of an environment. In particular, indoor scenes have more total objects, a greater variety of objects, and a smaller average object size when compared to outdoor scenes. This same trend holds for urban scenes when compared to natural scenes, as urban scenes have a higher degree of manufacture. Spatially, indoor scenes had a lower center of mass compared to outdoor scenes. There are two reasons for this. Outdoor scenes have more objects further off the ground than indoor scenes (“sky,” “cloud,” “skyscraper,” “bird,” “tree”). Also, outdoor scenes also have a larger mean depth than indoor scenes (Torralba and Oliva, 2002). Objects receding in depth tend to be located higher in the x-y image plane, leading to a higher center of mass for outdoor scenes. Therefore, although scenes are treated as a single class in the literature, this result suggests that scenes are a heterogeneous set of entities, leaving open the possibility that different environments may be processed differently by the visual system.

Through the use of a linear classifier, I have shown that such simple statistics carry sufficient information to categorize a scene at both the basic- and superordinate- levels significantly above chance, demonstrating for the first time that ensemble statistics could be a plausible mechanism for scene gist recognition in human observers. Although this classifier had lower performance than the human observers from Kadar and Ben-Shahar (2012), the patterns of errors made by this model were similar to those made by the human observers when averaged over superordinate-level categories, suggesting that human observers

may build an initial scene representation using ensemble-like features. Of course, the majority of work on ensemble statistics has been on very sparse laboratory displays. It remains to be seen whether observers can accurately report statistical information from complex, real-world images.

### BAG OF WORDS MODELS

The statistical ensemble model considered all annotations to be nameless “things.” However, the identity of these “things” is critical to scene identity. In linguistics, models that consider statistical patterns of word use independent of syntactical relations (so-called “bag of words” models) have been successful in document classification and spam detection (Deerwester et al., 1990; Blei et al., 2003). In computer vision, growing bodies of models perform similar operations on visual “words” given by interest-point or object detectors (Sivic and Zisserman, 2003; Fei-Fei and Perona, 2005). Visual bag of words models have been very successful for scene classification in recent years (Bosch et al., 2006; Lazebnik et al., 2006; Li et al., 2010).

In the model considered here, a scene is represented as a list of the objects contained in it. Measures such as object frequency (overall as well as conditioned on scene category) and mutual information between objects and scenes will be employed while still ignoring the spatial relations existing between these objects and regions. As before, I will examine the fidelity of a bag of words model for predicting basic- and superordinate-level scene categories through the use of a linear classifier, and evaluate proposed schemes by which human rapid scene categorization might occur via object recognition.

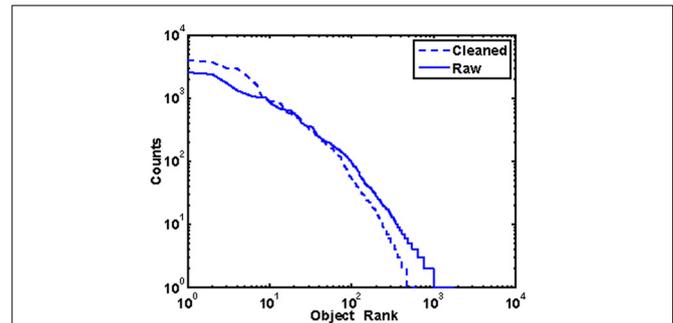
### Overall object frequency

Which objects are most common in the world? Just as certain words are more common than others in written text (“the” is more common than “aardvark”), certain objects appear in the world with greater frequency than others. Each of the 617 uniquely labeled regions in the database appeared between 1 and 3994 times in 1–2312 images. Nearly a quarter of the labels (22.7%) appeared only once in the database while eight objects (0.23%) appeared more than 1000 times. Overall, the frequency of objects in the database is inversely proportional to the frequency rank of the object, a relationship known in the linguistics literature as Zipf’s law (Li, 1992; see Figure 5).

The 10 most common objects are listed in Table 3 where I list both the total counts for an object (right column), and the number of scenes that contain at least one instance of that object (middle column). It should be noted that these counts represent a lower bound for the number of objects in the scenes. In scenes where several exemplars of a small object were grouped together, but too small to individuate (e.g., “apples” in a “bowl,” or “books” on a “shelf”), it was typical for annotators to list these as a group using the plural.

### Object frequency

What are the most frequent objects in each scene category? Knowing object frequency will allow us to find out how sensitive human observers are to these frequencies, and thus better understand the role of expectation in scene perception.



**FIGURE 5 | Object frequency is inversely proportional to frequency rank.** This pattern does not strongly depend on how the database was cleaned.

**Table 3 | The 10 most common objects in the database.**

Object name	Nb scenes (%)	Total counts
Sky	2312 (66)	2393
Tree	1377 (39)	3680
Building	1139 (33)	3994
Mountain	963 (28)	1615
Road	871 (25)	1064
Window	821 (23)	2981
Car	635 (18)	2943
Door	488 (14)	901
Ceiling	484 (14)	513
Plant	465 (13)	859

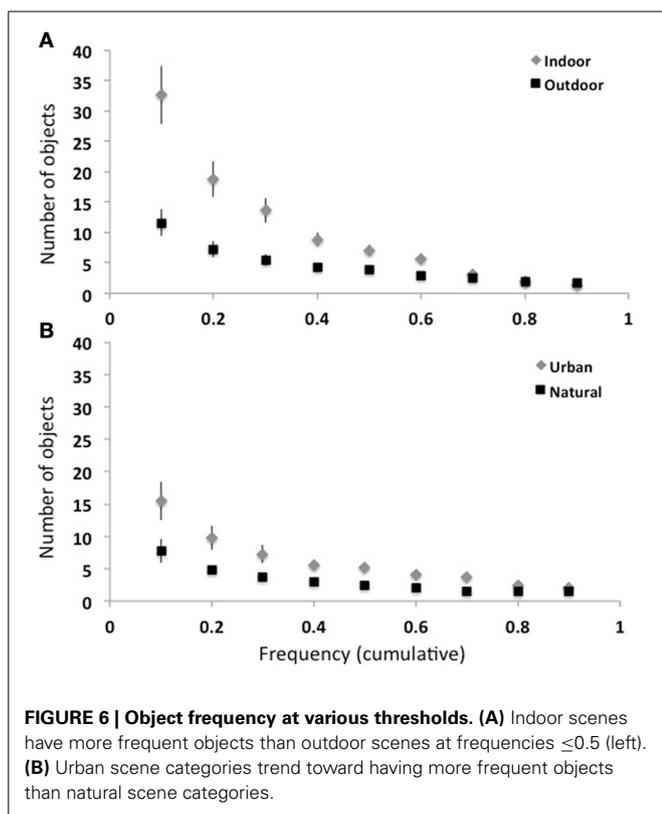
The middle column shows the number of scenes containing at least one exemplar of this object. The percentage of scenes with this object is shown in parentheses. The right column shows the total number of these objects in the database.

Table 4 shows the 10 most frequent objects in each basic-level scene category. It is of note that there are relatively large differences between basic-level scene categories in terms of the frequency of the most typical objects: while “sofa” is an intuitively important object for *living rooms*, it was present in only 86% of living room scenes, while “faucet” was labeled in over 99% of *bathrooms*.

What is the overall frequency-rank relationship for each of the 16 scene categories? For each basic-level scene category, I computed the number of objects that were present in at least half of the images. Indoor scenes had a greater number of frequent objects compared to outdoor scenes [7.1 vs. 3.9 objects,  $t_{(14)} = 3.1$ ,  $p < 0.01$ ]. Among the outdoor scenes, urban scenes had a greater number of frequent objects compared to natural [5.3 vs. 2.5,  $t_{(6)} = 4.0$ ,  $p < 0.01$ ]. Again, this pattern shows that the degree of human manufacture affects the distribution of object frequencies. To probe at a finer level of detail, I computed the number of objects at frequency levels between 0.1 and 0.9. Figure 6 shows the average of outdoor and indoor scenes (top, A) and the average of natural and urban scenes (bottom, B). *T*-tests performed at each threshold level showed that no statistical difference exists between the number of objects in outdoor and indoor

**Table 4 | The 10 most frequent objects in each basic-level category along with the proportion of scenes in each category that contain at least one exemplar of that object.**

	Indoor										Urban					Natural				
	Bed	Conf.	Corr.	Dine	Kit'in.	Liv.	Off.	Tall	City	Strt.	High	Cst.	OpC	Mntn.	Frst.					
Faucet	0.90	0.99	0.99	0.98	0.96	0.86	0.94	0.95	0.94	1.00	1.00	0.99	0.96	1.00	0.96					
Towel	0.79	0.95	0.95	0.96	0.94	0.83	0.89	0.94	Window	Road	Sky	0.99	Tree	Sky	0.51					
Sink	0.77	0.61	0.91	0.70	0.67	0.80	Monitor	Skyscraper	Door	Sky	Car	0.44	Field	Tree	0.31					
Bath	0.69	0.61	0.76	0.66	0.63	0.68	Window	Tree	Road	Car	Tree	0.42	Mountain	River	0.25					
Mirror	0.68	0.43	0.49	0.56	0.62	0.67	Keyboard	Road	Sky	Sidewalk	Sign	0.33	River	Snow	0.23					
Floor	0.63	0.35	0.29	0.55	0.62	0.64	Book	River	Sidewalk	Person	Fence	0.17	Hill	Rock	0.17					
Toilet	0.60	0.33	0.21	0.53	0.60	0.64	Lamp	Car	Car	Tree	Mountain	0.15	Building	Road	0.14					
Window	0.59	0.30	0.21	0.41	0.55	0.61	Paper	Sidewalk	Tree	Van	Median	0.14	Bush	Ground	0.13					
Bottle	0.45	0.27	0.15	0.41	0.55	0.51	Plant	Streetlight	Person	Streetlight	Building	0.14	Path	Mtn pass	0.13					
Wall	0.43	0.13	0.13	0.36	0.54	0.47	Ceiling	Antenna	Plant	Sign	Streetlight	0.11	Rock	Person	0.12					
		0.19																		



scenes for frequency thresholds above 0.5 (Bonferroni corrected), suggesting that although indoor scenes have more objects than outdoor scenes, all scenes have similar numbers of very frequent objects. Among the outdoor scene categories, natural and urban scenes did not reliably differ.

### Object diagnosticity

How important is an object to scene identity? Important objects may frequently occur in scenes, but not all frequent objects provide information about scene category. For example, some objects, such as “tree” can occur in many environments, while other objects such as “toilet” can only occur in a specific context, such as a *bathroom*. To formalize this notion, I introduce *diagnosticity*, which is the probability of a scene belonging to a particular scene category conditioned on the presence of a particular object [ $p(\text{scene}|\text{object})$ ]. Although “chair” is a frequent object in *dining rooms*, chairs are not diagnostic of dining rooms because they are also found in *bedrooms*, *conference rooms*, *offices*, etc. Similarly, there may be objects that are diagnostic that are not frequent, and these might reflect object-scene pairs that have been used in the object consistency literature (recall the example of the “sand castle” on the *beach*). This measure is of particular interest as some models of human rapid scene categorization posit that categorization can be mediated through the recognition of one or more diagnostic objects (Friedman, 1979; Biederman, 1981).

Diagnosticity was measured for every object and scene category in the database. However, this metric over-represents rare objects. As nearly one quarter of labeled objects occurred only

once in the database, all of these objects have full diagnosticity for the scene category they were found in. However, because they are rare, these objects may not be informative. Therefore, I am reporting the diagnosticity of objects with at least 10 instances in the database. The most diagnostic objects for each scene category are listed in **Table 5**.

In addition, I examined the number of completely diagnostic objects (diagnosticity = 1) across scene categories. All objects were included in this analysis. I found that indoor scenes tended to have a higher number of completely diagnostic objects compared to outdoor scenes [25.3 vs. 11.8,  $t_{(14)} = 1.93$ ,  $p = 0.07$ ], although both urban and natural scene categories had the same number of completely diagnostic objects on average (11.8). Again, this is not surprising as indoor scenes had more objects overall, as well as more infrequent objects.

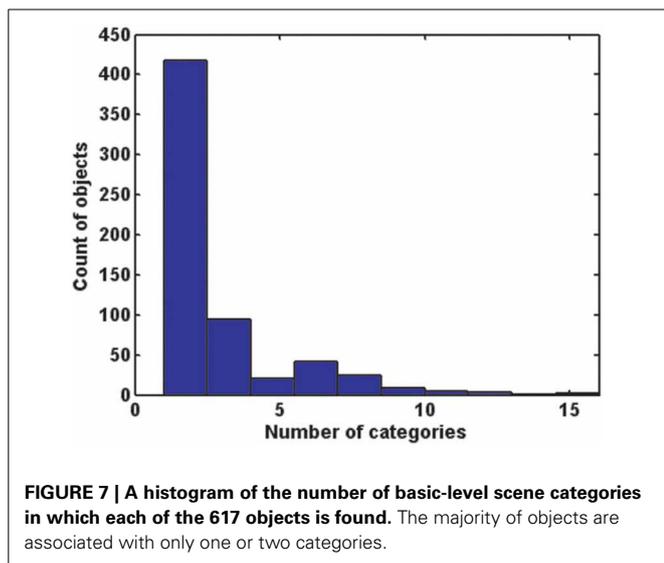
As noted in section Using Context for Rapid Scene Recognition, the notion of diagnosticity can be used to test hypotheses on the mechanisms of rapid scene categorization. Biederman (1981) first posited that a scene might be recognized through the recognition of a prominent, diagnostic object. How diagnostic are the largest objects in the scene? For each of the 3499 scenes, I examined the diagnosticity of the largest object in that scene for the scene’s category. On average, the largest object has a diagnosticity of 0.32 for the scene category it is in (95% CI: 0.04–0.99). Thus, although knowing the identity of the largest object in the scene will allow you to guess the scene category at an above-chance level, it does not reflect the outstanding performance that human observers have with rapid scene categorization. What if you know the identity of the object nearest the center of the image? The mean diagnosticity of the center object was 0.33 (95% CI: 0.03–1.00). Although this is a little better than knowing the largest object [ $t_{(6996)} = 2.3$ ,  $p < 0.05$ ], it seems unlikely that human scene gist performance can be explained from recognizing the center object alone.

### Scene-object specificity

How many scene categories contain a particular object? Here, I investigated the question by computing the number of scene categories in which each object is found. This measure is useful in the design of experiments in object and scene perception, as it allows experimenters to choose objects that are strongly tied to only one scene category (for example, to study response bias, e.g., Castelano and Henderson, 2008) or to use objects found in a variety of scenes to de-couple object recognition from inferential effects.

As shown in **Figure 7**, the majority of objects are closely tied to one or two scene categories. The median number of scene categories containing an object was two. Forty eight percent of objects were only found in one scene category, and of these, 53% had at least two instances in the database, suggesting that this effect was not solely driven by infrequent objects. In fact, 31 of the objects found in only one scene category (5% of the total) had 10 or more instances. These are listed in **Table 6**. On the other hand, there was only one object present in all 16 categories (“wall”), and 19 (3% of total) were present in at least nine of the 16 categories. These are also listed in **Table 6**.





**Table 6 | Objects with 10 or more database instances found in only scene category (left), and objects found in at least nine of the 16 basic-level scene categories (right).**

Objects in 1 category		Objects in 9+ categories
Attic	Pan	Bench
Bath mat	Podium	Box
Bedspread	Porch	Chair
Cow	Projection screen	Clock
Cutting board	Seagull	Column
Desert	Shower	Decoration
Dish towel	Shower curtain	Door
File organizer	Soap dish	Lamp
Goose	Spotlight	Light
Hay bale	Stove	Person
Headboard	Stove hood	Plant
Kettle	Toothbrush	Poster
Keyboard		Rock
Microphone		Staircase
Mountain pass		Statue
Mouse		Table
Mouse pad		Tree
Nightstand		Wall
Oven		Window

**Object pairs and groups**

While the previous statistics have examined relationships between the scenes and single objects in them, it is also important to examine the relationships between multiple objects in a scene. Object co-occurrence has been shown to guide visual search in naturalistic scenes (Mack and Eckstein, 2011); interacting objects tend to be perceptually grouped (Green and Hummel, 2006); object interactions have been shown to increase activity in object-selective cortex (Kim and Biederman, 2010); and scene identity can be predicted from pairs of objects in object-selective cortex (MacEvoy and Epstein, 2011). How informative are groups of objects, and

how many objects do you need to be able to predict the scene’s category?

First, I examined the frequency of co-occurrence of object pairs in each basic-level scene category. The 10 most frequent object pairs for each basic-level category are shown in Table 7.

As shown in Table 7, some object pairs are functionally related (such as “faucet” and “sink” for *bathroom*), while many are not (e.g., “sky” and “building” in *skyscraper* scenes). There are 20 object pairs in this table that are listed in multiple basic-level categories. In fact, *conference rooms* and *dining rooms* share 8 of the 10 most frequent object pairs. However, of these 20 object pairs, only two are shared across superordinate-level categories (“sky” + “building” and “tree” + “sky”). Both of these pairs are shared across natural and urban scene categories. No object pair in this group was observed in both indoor and outdoor scenes. Therefore, although single objects may be found across all superordinate categories, pairs of objects do not share this property.

Next, I examined the 617 by 617 object co-occurrence matrix collapsed over all scene categories. Overall, the object co-occurrence matrix was sparse, with only 9% of possible object pairings having been observed. Of the observed object pairings, 8% had a co-occurrence probability of 1, indicating that these pairs of objects were always found together, and of these, 9% ( $n = 254$ , 0.73% of total pairings) were for objects with more than one instance in the database. Thus, requisite object pairs are relatively rare in the world, and arbitrary pairs of objects are generally not seen together.

What are the most frequent groups of  $n$  objects in each of the basic-level scene categories? Table 8 shows the most frequent groups of three, four, and five objects for each of the basic-level scene categories. Larger groups are not shown because many natural landscape images have fewer than 6 total objects.

How much information do these object groups provide about scene categories? More specifically, are these groups of multiple objects more diagnostic of a scene category than single objects? Here, I computed the diagnosticity [ $p(\text{category}|\text{object})$ ] of the most frequent groups of one to five objects. As shown in Figure 8, although the most common object in a scene category has an average diagnosticity of only 0.35, diagnosticity increases with increasing group size up to 0.78 for groups of five. The diagnosticity of object groups did not reliably differ across superordinate categories. This result gives some insight into the third path to scene recognition proposed by Biederman (1981), that scene recognition can arise through the spatial integration of a few contextually-related objects. Although this bag-of-words approach neglects the spatial relationships between objects, this analysis places a lower bound on the categorization performance that can be achieved by knowing the identities of a few objects. In section Structural Statistics, we will examine the effect of knowing coarse spatial relationships.

**Scene combinations**

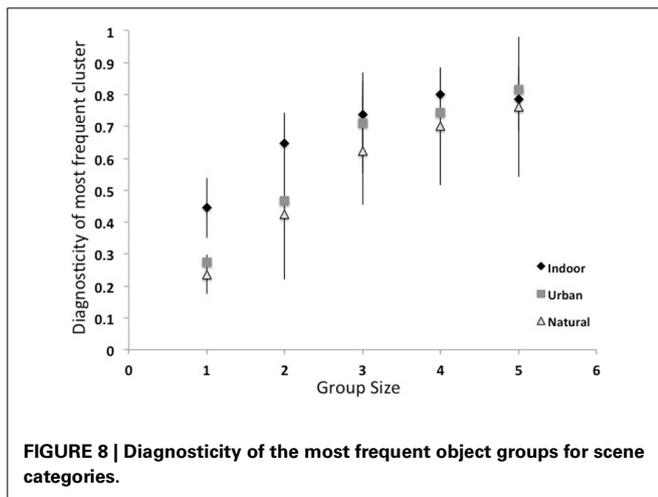
How many unique combinations of objects were observed in the database? Do certain scene categories have more object combinations than others? Let us first examine the theoretical limit:

**Table 7 | The 10 most frequent pairs of objects in each basic-level scene category.**

	Indoor										Urban										Natural													
	Bed	Conf.	Corr	Dine	KiIn.	Liv.	iff.	Tall	City	Strt.	High	Cst.	OpC	Mntn.	Frst.	Bed	Conf.	Corr	Dine	KiIn.	Liv.	iff.	Tall	City	Strt.	High	Cst.	OpC	Mntn.	Frst.				
Sink	Table	Floor	Table	Counter	Sofa	Desk	Sky	Building	Road	Road	Sky	Sky	Building	Road	Sky	Sky	Sky	Ocean	Trees	Sky	Sky	Tree	Sky	Building	Road	Road	Sky	Ocean	Trees	Sky	Sky	Mountain	Tree	
Faucet	Chair	Wall	Chair	Cabinet	Pillow	Chair	Building	Window	Building	Sky	Building	Building	Window	Building	Sky	Ocean	Mountain	Tree	Field	Sky	Mountain	Bush	Sky	Skyscraper	Road	Car	Mountain	Field	Mountain	Tree	Mountain	Tree		
Faucet	Ceiling	Wall	Window	Window	Table	Monitor	Sky	Building	Building	Sky	Building	Sky	Building	Car	Sky	Sky	Sky	Sky	Field	Sky	Mountain	Tree	Sky	Skyscraper	Road	Car	Mountain	Field	Mountain	Tree	Mountain	Tree		
Towel	Window	Chair	Chair	Cabinet	Sofa	Chair	Sky	Building	Building	Sky	Building	Sky	Building	Car	Sky	Sky	Sky	Sky	Field	Sky	Mountain	Tree	Sky	Skyscraper	Road	Car	Mountain	Field	Mountain	Tree	Mountain	Tree		
Mirror	Bed	Floor	Window	Sink	Table	Monitor	Building	Door	Road	Tree	Tree	Building	Door	Tree	Tree	Ocean	Mountain	Tree	Sky	Sky	Tree	Rock	Sky	Building	Tree	Tree	Ocean	Sky	Tree	Tree	Rock	Tree		
Faucet	Lamp	Ceiling	Table	Cabinet	Pillow	Chair	Skyscraper	Window	Sky	Skyscraper	Chair	Skyscraper	Window	Sky	Road	Mountain	Tree	Tree	Tree	Tree	Sky	Tree	Rock	Sky	Building	Tree	Mountain	Tree	Sky	Sky	Tree	Tree		
Bath	Bed	Wall	Bouquet	Window	Sofa	Window	Building	Road	Road	Building	Window	Building	Road	Car	Tree	Sky	Sky	Sand	Mountain	Sky	Sky	River	Sky	Building	Tree	Sky	Sand	Sky	River	Mountain	Tree	Tree		
Faucet	Ceiling	Door	Chair	Counter	Lamp	Chair	Tree	Building	Car	Tree	Chair	Tree	Building	Car	Sky	Sand	Sand	Mountain	Trees	Field	Snow	Grass	Sky	Building	Car	Sky	Ocean	Field	Snow	Mountain	Tree	Tree		
Sink	Lamp	Floor	Table	Faucet	Window	Monitor	Tree	Building	Road	Sky	Keyboard	Tree	Building	Road	Sky	Ocean	Sand	Trees	Trees	Field	Snow	Grass	Sky	Building	Car	Sky	Ocean	Field	Snow	Mountain	Tree	Tree		
Towel	Pillow	Door	Bouquet	Cabinet	Pillow	Monitor	Sky	Window	Car	Car	Monitor	Sky	Window	Car	Sky	Sand	Sand	Trees	Tree	Tree	Sky	Ground	Tree	Sky	Building	Car	Rock	Tree	River	Tree	Tree	Tree		
Sink	Bed	Ceiling	Ceiling	Plant	Sofa	Keyboard	Tree	Sky	Car	Car	Keyboard	Tree	Building	Car	Sky	Sand	Rock	Trees	Trees	Field	Sky	Ground	Tree	Sky	Building	Car	Rock	Tree	River	Tree	Tree	Tree		
Mirror	Curtain	Door	Chair	Cabinet	Plant	Desk	Skyscraper	Bunding	Car	Car	Desk	Skyscraper	Bunding	Car	Sky	Rock	Tree	Tree	Tree	Tree	Sky	Ground	Tree	Sky	Building	Car	Rock	Tree	River	Tree	Tree	Tree		
Floor	Pillow	Light	Painting	Stove	Lamp	Window	Road	Sky	Building	Sign	Window	Road	Sky	Building	Sky	Sky	Mountain	Trees	Mountain	Rock	Mountain	Bush	Sky	Building	Sign	Road	Rock	Mountain	Mountain	Sky	Sky	Sky	Sky	
Faucet	Window	Wall	Chair	Cabinet	Pillow	Desk	Building	Window	Sidewalk	Road	Desk	Building	Window	Sidewalk	Sign	Rock	Rock	Trees	Trees	Rock	Mountain	Bush	Sky	Building	Sidewalk	Road	Rock	Mountain	Mountain	Sky	Sky	Sky	Sky	
Towel	Bed	Light	Table	Sink	Table	Chair	Road	Sidewalk	Road	Sign	Keyboard	Road	Sidewalk	Sign	Sand	Sand	Field	Field	Field	Snow	Rock	Rock	Sky	Building	Sign	Sky	Sand	Field	Snow	Sky	Sky	Sky	Sky	
Bath	Painting	Ceiling	Ceiling	Counter	Lamp	Chair	Sky	Building	Road	Sign	Chair	Road	Building	Sky	Mountain	Mountain	Tree	Tree	Tree	Tree	Sky	Rock	Rock	Sky	Building	Sign	Sky	Mountain	Tree	Sky	Sky	Sky	Sky	Sky
Toilet	Ceiling	Floor	Light	Cabinet	Sofa	Book	Road	Sidewalk	Sky	Sign	Book	Road	Sidewalk	Sign	Ocean	Ocean	River	Trees	River	Sky	Mountain	Mountain	Sky	Building	Sign	Car	Boat	River	Sky	Rock	Rock	Rock	Rock	
Faucet	Pillow	Light	Chair	Stove	Painting	Desk	Skyscraper	Window	Sidewalk	Car	Desk	Skyscraper	Window	Sidewalk	Boat	Boat	Trees	Trees	Trees	Sky	Rock	Tree	Tree	Sky	Building	Car	Boat	Trees	Rock	Rock	Rock	Rock	Rock	
Towel	Bed	Door	Light	Plant	Table	Chair	River	Road	Sidewalk	Tree	Chair	River	Road	Sidewalk	Sky	Sky	Mountain	Trees	Mountain	Sky	Mountain	Mountain	Sky	Building	Sign	Tree	Sky	Mountain						
Mirror	Nightstand	Light	Table	Counter	Plant	Book	Building	Door	Sidewalk	Car	Book	Building	Door	Sidewalk	Boat	Boat	Tree	Tree	Tree	Sky	Road	River	Tree	Building	Sign	Tree	Boat	Mountain	Mountain	Road	Road	Road	Bush	

**Table 8 | The most frequent groups of three, four, and five objects found in each of the 16 basic-level scene categories.**

Category	Three-objects	Four-objects	Five-objects
Bathroom	Faucet, sink, towel	Faucet, mirror, sink, towel	Bath, faucet, mirror, sink, towel
Bedroom	Bed, pillow, window	Bed, ceiling, pillow, window	Bed, ceiling, painting, pillow, window
Conference	Ceiling, chair, table	Ceiling, chair, light, table	Ceiling, chair, light, table, window
Corridor	Ceiling, floor, wall	Ceiling, door, floor, wall	Ceiling, door, floor, light, wall
Dining room	Chair, table, window	Bouquet, chair, table, window	Bouquet, ceiling, chair, curtain, wine glass
Kitchen	Cabinet, counter, plant	Cabinet, counter, faucet, sink	Cabinet, counter, faucet, sink, window
Living room	Pillow, sofa, table	Pillow, sofa, table, window	Lamp, pillow, sofa, table, window
Office	Book, chair, desk	Chair, monitor, desk, window	Chair, monitor, desk, whiteboard, window
Tall building	Building, sky, skyscraper	Building, sky, skyscraper, tree	Building, road, sky, skyscraper, tree
Inside city	Building, door, window	Building, door, sky, window	Building, door, road, sidewalk, window
Street	Building, car, road	Building, car, road, sky	Building, car, road, sidewalk, sky
Highway	Car, road, sky	Car, road, sky, tree	Car, road, sign, sky, tree
Coast	Ocean, rock, sky	Mountain, ocean, rock, sky	Mountain, ocean, rock, sand, sky
Open country	Field, sky, tree	Field, mountain, sky, tree	Building, field, mountain, river bank, sky
Mountain	Mountain, sky, tree	Ground, mountain, sky, tree	Ground, mountain, road, sky, tree
Forest	Bush, sky, tree	Bush, river, rock, tree	Bush, river, rock, sky, tree



if all 617 objects in the database were independent, and could occur with equal probability in all scenes, then there would be  $2^{617}$  possible combinations of objects. Even if we examine only the possible combinations of 6 objects (the median number of unique objects in a scene from our database), this leaves us with an astounding  $7.5 \times 10^{13}$  combinations!

In contrast, I observed only 2552 unique object combinations in the 3499-scene database. In other words, 26% of scenes had the exact same combination of objects as at least one other scene in the database. However, this redundancy was not evenly distributed among the different basic-level scene categories. Ninety nine percent of indoor scenes had unique object combinations compared to only 68.6% of outdoor scenes [ $t_{(14)} = 3.71$ ,  $p < 0.01$ ]. Among the outdoor scenes, 85.1% of urban scenes had a unique object combination vs. 52.1% of natural scenes [ $t_{(6)} = 2.89$ ,  $p < 0.05$ ]. *Mountain* scenes in particular had very high redundancy in terms of unique object combinations, as only 33.7% of these scenes had a unique combination of objects.

### Entropy

Information theory provides a formal means of expressing redundancy between objects and scene categories. If all objects in the database were independent and equally probable, then the redundancy of the database could be expressed as  $\log_2(617) = 9.27$  bits per object. However, object frequencies are not uniformly distributed: objects such as “chair” and “sky” are much more frequent than others such as “scaffolding” or “zebra” (section Object Diagnosticity). Relative object frequencies can be accounted for by computing the entropy of the database:

$$N = \sum_p(o) \log p(o)$$

Where  $p(o)$  refers to the observed probability of each object in the database. In this instance, taking relative frequencies into account reduces the number of bits per object needed to encode the database to 6.25. Imagine that you are trying to guess an object’s identity by playing the game “20 questions.” The rules of this game stipulate that you may only ask questions whose answer is “yes” or “no.” This entropy result tells us that you would be able to correctly guess the object by asking, on average, 6 binary questions.

### Mutual information

How much information do objects and scenes provide about one another? For example, how much evidence do you have about the category *dining room* from the presence or absence of an object such as a “chair?” To formalize this notion, I computed the mutual information between all objects and their scene categories. While diagnosticity tells us how likely an image is to belong to a particular scene category given the presence of a particular object, it does not easily tell us which objects are important, as objects occurring only once in the database are by definition completely diagnostic of that category. Mutual information measures the degree of dependence between objects and scenes and is therefore more immune to the problem of small numbers. Formally, mutual



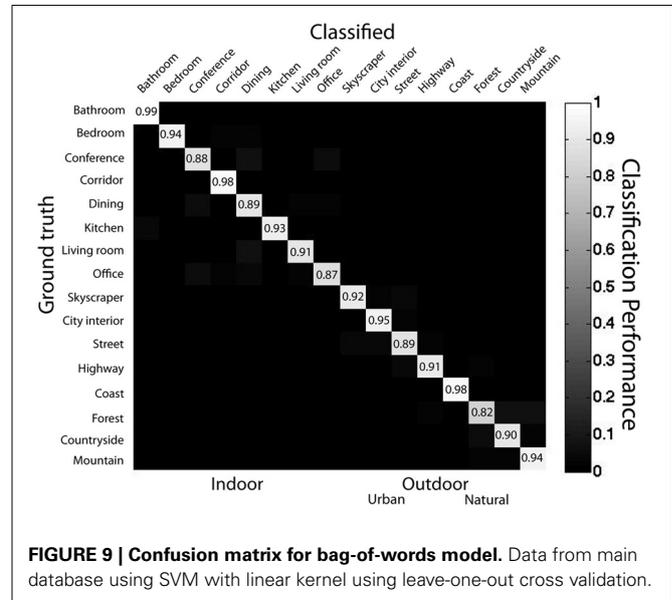
**Table 10 | The 10 objects with the highest mutual information for each of the 16 basic-level categories.**

	Indoor										Urban					Natural				
	Bed	Bed	Conf.	Corr.	Dine	Kit'in	Liv.	Off.	Tall	City	Strt.	High	Cst.	OpC	Mntn.	Frst.				
Towel	Bed	Chair	Wall	Chair	Counter	Sofa	Desk	Skyscrpr	Building	Road	Road	Ocean	Field	Mountain	Tree					
Bath	Nightstnd	Table	Floor	Table	Stove	Pillow	Monitor	Building	Door	Car	Car	Sand	Sky	Sky	Bush					
Faucet	Pillow	Prijtn Scrn	Ceiling	Bouquet	Cabinet	Armchair	Keybrd	Sky	Sidewalk	Building	Sign	Sky	Window	Window	Rock					
Toilet	Dresser	Light	Door	Buffet	Pot	Coffee table	Cmptr	Window	Window	Sidewalk	Median	Rock	Hill	Plant	River					
Sink	Lamp	Podium	Column	Plate	Oven	Lamp	Chair	Antenna	Road	Person	Fence	Sun	Tree	Snow	Building					
Mirror	Curtain	Whitebrd	Tree	Napkin	Sink	Ottoman	Mouse	Door	Balcony	Sky	Sky	Boat	Mountain	Wall	Land					
Soap	Painting	Tree	Exit sign	Candle	Dishwshr	Painting	Book	Tree	Shp wdvw	Van	Tree	Chair	River	Door	Branch					
Twl rack	Ceiling	Ceiling	Light	Painting	Bowl	Fireplace	Paper	Plant	Staircase	Crosswalk	Window	Cloud	Rock	Building	Grass					
Bottle	Carpet	Projector	Arch	Placemat	Faucet	Plant	Phone	Dock	Terrace	Chair	Bridge	Mountain	Desert	Mtn pass	Road					
Floor	Bedsprd	Microphone	Fire alrm	Light	Plate	Decor	Printer	Mountain	Chair	Stright	Streetlight	Road	Door	Car	Wall					

Only objects making an appearance in the scene category are listed.

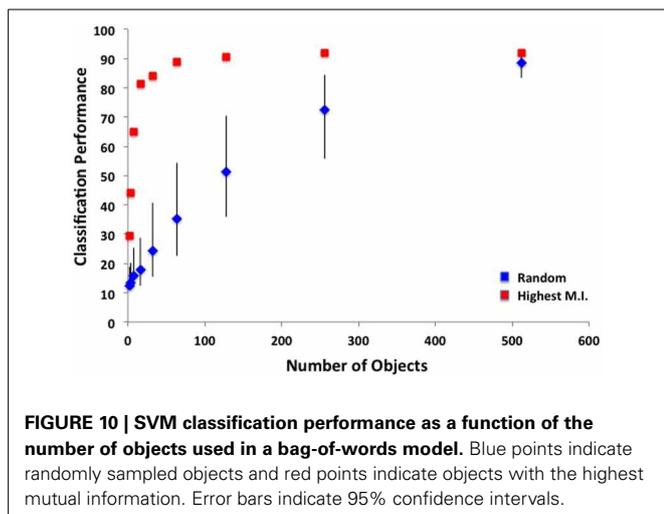
**Table 11 | Objects with the highest mutual information for all scene categories.**

- Object name
- Sky
- Building
- Chair
- Road
- Table
- Ceiling
- Tree
- Car
- Window
- Pillow



is sufficient to categorize scene images at both basic- and superordinate-levels.

To what extent is the higher performance of the bag-of-words model compared to the ensemble statistics model due to the higher dimensionality of this model? To answer this question, I ran SVM analyses on sets of six objects, either by randomly sampling from the 617 total objects, or by taking the objects with the highest overall mutual information. For sets of randomly selected objects, mean classification performance was 15.1% correct (95% CI: 13.9–16.1%), well below the 61% achieved by the same number of ensemble features. When taking the six objects with the highest overall mutual information (see **Table 11**), classification performance was 51.4%, only marginally worse than that of the ensemble statistic model (binomial test,  $p = 0.051$ ). How many objects are necessary to reach ceiling performance? I ran additional SVM analyses on sets of 2–512 objects, either by randomly sampling objects or selecting objects with the highest mutual information. As shown in **Figure 10**, ceiling performance is reached with the 64 best objects. Therefore, although higher performance was achieved using a bag-of-words approach, this



performance can be attributed to the larger dimensionality as the features contained in the ensemble statistics model contained at least as much information as a similar number of object features.

How does the performance of the bag of words model compare to the human rapid scene categorization performance reported in Kadar and Ben-Shahar (2012)? Overall sensitivity was similar between the bag of words classifier and human observers [ $A' = 0.85$  for both,  $t_{(22)} < 1$ ]. However, the patterns of errors for the classifier and human observers were markedly dissimilar. As with the ensemble statistics classifier, error patterns were not well correlated at the basic-level ( $r = 0.04$ ). However, error patterns at the superordinate level actually showed opposite trends from the human observers ( $r = -0.88$ ), suggesting that the bag of words representation, although similar in performance to the human observers, is not similar to the human scene gist representation.

### Bag of words discussion

Here I have examined statistical regularities between object identities and scene categories, ignoring the spatial relationships between these objects. The measures include object frequency, object diagnosticity, the mutual information between an object and its scene category and the number of scene categories each object is found in. At this level of analysis, the relationships that objects have to one another was also considered by examining the co-occurrence frequencies of two or more objects.

Object frequencies are not equivalent to object “consistency” as used in the visual cognition literature, which tends to be a Boolean variable (a “blender” in a *kitchen* is consistent, a “fire hydrant” in a *kitchen* is inconsistent). Here, object frequencies are continuous and range from 0 (no observed instances of this object for this scene category) to 1 (each scene in this category contains this object). This continuous scale allows the design of new experiments, allowing researchers to ask questions about the perceptual processing or memory differences that might exist for objects that are present in nearly all scene exemplars (frequency =  $\sim 1$ ) vs. objects that are present in only about half of the exemplars (frequency = 0.5), vs. objects that are plausible but rare (frequency  $< 0.2$ ).

The bag of words level of analysis shows additional ways that scene categories differ. The ensemble level of analysis showed large differences between superordinate-level categories in terms of the amount of unnamed objects in scenes: indoor scenes having more than outdoor, and urban having more than natural. At this level of analysis, I found that objects strongly segregate themselves into different basic level scene categories—any given object was only found in a small number of scene categories, and when an object is found in multiple basic-level categories, these categories do not cross superordinate classes. A classifier given all object identities achieved near-ceiling performance at both superordinate- and basic-level scene classifications. Thus, knowledge of either a scene’s category or an object’s identity gives a great deal of information about the other, and full knowledge of all objects in a scene is sufficient for scene categorization.

Additionally, ceiling performance can be achieved with fewer objects, provided you have the “best” objects (i.e., the objects with the highest mutual information for distinguishing scene categories). Here, I demonstrated that ceiling performance could be reached with the 64 most informative objects. This is of use to those in the computer vision community who perform scene classification using hundreds of off-the-shelf object detectors (e.g., Li et al., 2010). By choosing objects that are informative, rather than frequent, these systems could be made far more efficient.

The results of the linear SVM classifier suggest that if one knows the identities of all of the objects in a scene, one will know the category of the scene. Although this has been posited as a possible route to scene understanding (Biederman, 1981), behavioral evidence suggests that human observers do not apprehend all of a scene’s objects in a single glance (Fei-Fei et al., 2007; Greene and Oliva, 2009). Similarly, although the bag of words classifier had similar overall performance to human observers, it had markedly different patterns of errors, suggesting a representation different from humans. How many objects do people understand in a glance at a scene? This is a notoriously difficult problem as conceptual short term memory is relatively fragile (Potter, 1976), human observers can inflate performance through elaborate rehearsal or guessing strategies (Liu and Jiang, 2005), and observers can demonstrate sensitivity (in the form of negative priming) to objects that they cannot overtly name (VanRullen and Koch, 2003). The most stringent tests estimate that observers can only accurately report one object from a scene after a 250 ms masked display (Liu and Jiang, 2005).

Can scene recognition proceed from the recognition of just one object? When examining some plausible scenarios, such as perceiving the largest, or the most centered object, diagnosticity for the scene category is around 0.33, far below the performance of human observers in rapid scene classification. Of course, diagnosticity increases with increasing numbers of objects (section Object Pairs and Groups). However, classification performance for smaller numbers of objects, even the most informative objects, lagged behind that of the ensemble statistics model, suggesting that individual objects may not make the best features for human scene understanding and categorization.

While the bag of words level of analysis is a powerful and popular computer vision model of objects in scenes, the spatial

relationships between objects and regions are also critical to scene identity. I explore this level of analysis in the next session.

### STRUCTURAL STATISTICS

The third level of object-scene relationships I will explore is aimed toward obtaining a “syntax” of visual scenes that includes the nature of the spatial relations between objects. Just as the relations between object parts are key to the object identity (e.g., a key difference between a pail and a mug is the placement of the handle, Biederman, 1987), the relations between objects may provide additional information into scene identity. The spatial layout of a scene is created in part by the relative positioning of the objects within it, and regularities in layout allow a scene to be identified under highly degraded conditions, such as under sparse contours (Biederman, 1981) or blur (Oliva and Torralba, 2007) where object identities cannot be recovered. Indeed, two of the three pathways to scene gist outlined by Biederman (1981) can come from structural relations.

Here, I will examine the locations of objects in scenes, as well as the distances between objects and the spatial distributions of the important diagnostic and informative objects. As with the other two levels of analysis, I will examine the extent to which these structural statistics can be used to classify scenes at the basic- and superordinate- levels.

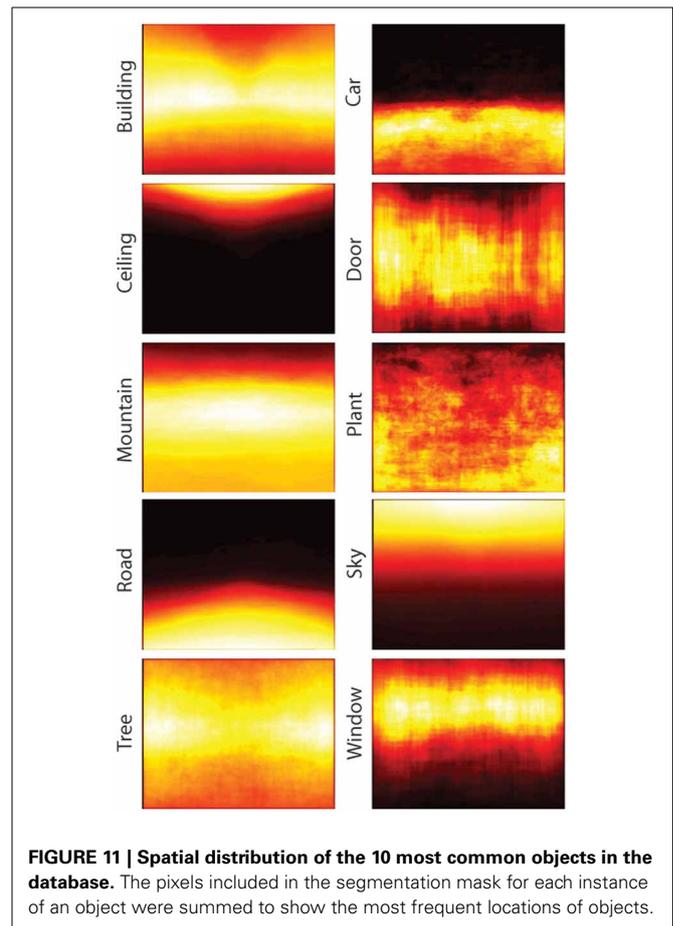
#### Object position specificity

One basic structural description is the position specificity of individual objects. In other words, how stereotyped are the x-y locations of the objects in the database? **Figure 11** shows a heat map of the spatial locations of the 10 most common objects in the database. Some regions, such as “ceiling,” are tightly bound to a particular image location while others, such as “plant” or “building,” can be found throughout the image plane. To quantify this notion, I examined the variance in x-y position for each object center across the database as well as the position variance of objects in each of the basic-level scene categories.

**Table 12** shows the 10 objects with the most position variance as well as the 10 objects with the least position variance in the database. Unsurprisingly, objects with a great deal of position specificity (low variance in x-y position) are often objects that make up the spatial boundaries of a scene (such as “carpet” and “sky”).

For basic-level scene categories, *bedrooms* had the most position variance while *open country* scenes had the least. Overall, indoor scenes tended to have more position variance compared to outdoor scenes [ $t_{(14)} = 2.98, p < 0.01$ ]. However, among the outdoor scenes, no distinct pattern emerged [ $t_{(6)} < 1$ ].

Are objects found in different locations when they are found in different scene categories? If this is the case, then position can provide diagnostic scene information. Here, I took the 17 objects that had at least 10 instances in indoor categories and at least 10 instances in outdoor scene categories (“bench,” “box,” “chair,” “clock,” “column,” “door,” “light,” “person,” “plant,” “poster,” “railing,” “sign,” “staircase,” “statue,” “trash can,” “wall,” and “window”) and examined image locations for the object when found outdoors and compared it to the locations where the object was found indoors. For each of these objects at each pixel location,



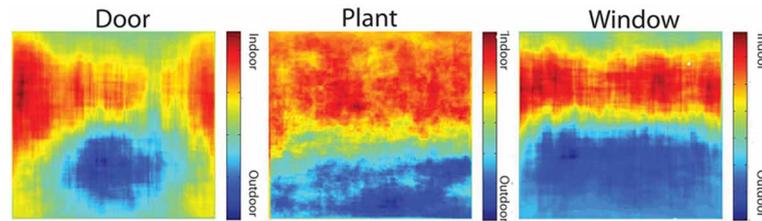
**FIGURE 11 | Spatial distribution of the 10 most common objects in the database.** The pixels included in the segmentation mask for each instance of an object were summed to show the most frequent locations of objects.

**Table 12 | The 10 objects with the least position variance (most static) and with the most position variance (least static).**

Most static	Least static
Carpet	Molding
Desert	Leaves
Ceiling	Dock
Exit sign	Dome
Bath mat	Pan
Bedsread	Basket
Ocean	Grill
Fan	Lighthouse
Bed	Calendar
Sky	Toy

The search for these objects was constrained to objects with at least 10 instances in the database.

I subtracted the number of instances the object was found in that location in an outdoor scene from the number of times the object was found in that location in an indoor scene. Significance was determined by Bonferroni corrected *t*-tests. Only three objects (“door,” “window,” and “plant”) had different location patterns in indoor scenes compared to outdoor scenes. **Figure 12** shows that these objects are found in higher positions in the image



**FIGURE 12 | Heat maps of the locations of doors (left), plants (center), and windows (right), conditioned on whether the object was found in an outdoor or indoor scene.** Warmer colors indicate

locations that are more probable in an indoor scene while cooler colors indicate locations that are more probable for the object in an outdoor scene.

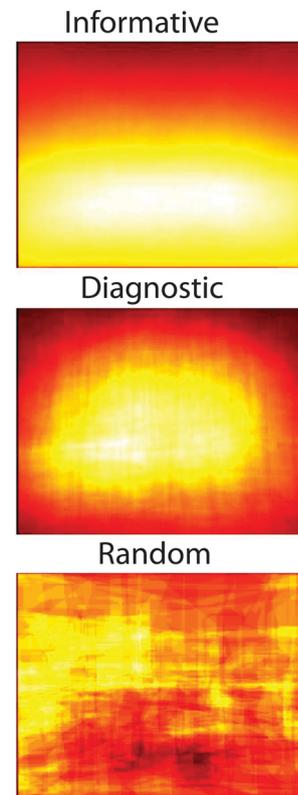
plane when found indoors compared to where they are found outdoors. Therefore, most objects are found in similar scene locations regardless of category, so position information generally does not add additional information beyond that of object identity. Additionally, these small differences may reflect both differences in the structure of these environments (such as depth differences, as discussed in section Center of Mass), as well as differing strategies of photographers for capturing the relevant information in different environments. Our knowledge of the three-dimensional world tells us that a “door” is located in a “wall,” and just above the “floor.” Therefore, these differences reflect statistics of photographs, as well as statistics of the external world.

#### ***Spatial distribution of diagnostic and informative objects***

Where are the most informative regions of a scene? Photographers tend to center pictures on objects of interest (Tatler et al., 2005), and objects in LabelMe tend to be labeled from the center out (Elazary and Itti, 2008). Do these centered objects have the highest diagnosticity or mutual information for their scene category?

For each of the 16 scene categories, I plotted all pixels associated with that most informative object or the object with the highest diagnosticity for the scene category. As shown in **Figure 13**, diagnostic objects tend to be centered overall, while informative objects tend to be centered lower in the image. This is not just due to spatial regression to the center, as random selections of objects do not display this behavior.

This analysis formalizes the notion of center-bias in photographs, demonstrating that photographs tend to be centered on scene regions that contain highly diagnostic objects. Highly informative regions, on the other hand, tend to cluster near the bottom of the image. This analysis also shows key differences between the notions of mutual information and diagnosticity. Many of the most informative objects are structural or boundary elements of a scene that can coarsely distinguish between categories, but are not necessarily the most important or interesting objects in a scene (see section Mutual Information and **Table 11**). For example, although “carpet” is highly informative because it distinguishes between outdoor and indoor environments, it is not a terribly interesting region. As it is known that the central fixation bias of human observers persists even when important features are moved to the periphery (Tatler, 2007), this finding is unlikely to provide additional insight into human scene perception mechanisms. However, researchers in computer



**FIGURE 13 | Spatial distribution of the most informative objects for all scene categories (top), most diagnostic object for each scene category (middle) and a random object for each scene (bottom).**

vision might find greater scene classification efficiency in applying object detectors from the center out rather than in a sliding window.

#### ***Scene classification with a structural model***

How much information do object locations provide about scene categories? To answer this question, I divided all of the 3499 scene images into quadrants and computed the number of times each object was found in each quadrant for each image. Thus, compared to the bag of words model, each object is represented four times, once in each of the four quadrant locations. This matrix

was fed into a linear SVM classifier with the same training and testing procedures outlined earlier. Superordinate and basic-level categorizations were assessed. Any increase in performance above the bag of words level can be taken as evidence for the utility of spatial object information for scene categorization.

Overall, this classifier had 98% accuracy ( $AUC = 0.99$ ) at superordinate-level scene categorization (not significantly different from the 98% correct performance of the bag-of-words model,  $Z < 1$ ) and 89.6% accuracy ( $AUC = 0.95$ ) at basic-level categorization (significantly lower than the 92% correct from the bag-of-words model,  $Z = 3.7$ ,  $p < 0.001$ ). There was no reliable difference in basic-level categorization accuracy for indoor (90.8% correct) vs. outdoor (89.4% correct) scenes [ $t_{(14)} < 1$ ], nor between urban and natural scene categories [ $t_{(6)} < 1$ ]. Performance by category was similar to the bag of words classifier—best performance was achieved by *bathroom* (98%, tied with *corridor* and *coast*), while the classifier had the poorest performance on *open country* images (75%). These *open country* images were frequently confused with *mountains* (38.2%), *forests* (34.3%), and *highways* (14.7%).

Altogether, adding coarse spatial information to the bag of words classifier did not result in higher classification performance. This is not very surprising as the bag of words classifier was at near ceiling performance, as the majority of objects were only found in one or two scene categories (section Scene-Object Specificity), and even objects found in multiple scene categories were generally found in similar locations regardless of category (section Object Position Specificity). The lower performance for basic-level classification is likely due to an increased number of features (617 vs. 2468) with the same number of training examples.

### Structural discussion

In this section, I have described scenes in terms of objects and their locations in the image plane. First, I described the location variability of each object, showing that objects that describe a scene's boundaries, such as "floor" or "sky" show less position variance than non-structural objects. Interestingly, most objects are found in similar locations in all scene categories. Of the objects found frequently in both outdoor and indoor scene environments, only "door," "window," and "plant" showed different patterns. For each of these cases, the object is found higher in the image plane in indoor scenes relative to outdoor scenes. This makes sense as the spatial enclosure of indoor scenes allows objects to be found in these locations. However, knowing an object's position in the x-y image plane does not provide much additional information over knowing its identity.

Next, I demonstrated that the center bias of photographs shows up in this database as a tendency for the most diagnostic and informative objects to be located near the center of the image. This may reflect the photographer's inherent sensitivity to object diagnosticity, and desire to convey the maximum amount of information about an environment in a single viewpoint. However, as informative objects tend to be large structural areas of a scene, diagnostic objects were more centered in the image.

Of course, both of these measures reflect statistical regularities of photographs rather than statistical regularities of the world. Although I have shown a tendency of photographers to photograph a "door" higher in the image plane in an indoor environment, we know that doors in the world are located above the "ground," and within "walls" in all environments. Similarly, "center bias" has no meaning in the immersive, three-dimensional real world. Despite these limitations, statistics of photographs provide insight into how human observers choose to represent information from the real world when forced to choose a single view.

A linear classifier trained on the bag of words model with coarse spatial location information did not outperform the pure bag of words model, and in fact, fared a little worse in basic-level categorization. There are two reasons for this: (1) most objects are only found in one or two scene categories (section Scene-Object Specificity), so the position of these objects is not going to provide additional category-related information; and (2) of the objects that are found in several scene categories, the majority are found in similar locations regardless of category (section Object Position Specificity).

This does not mean that structural information does not contribute unique scene information, however. One limitation of measuring structural relationships on scene photographs is that we lose the three spatial dimensions that are available in the world. The third dimension would allow the disambiguation of a variety of object relationships, including containment, support and adjacency. Indeed, these types of object relations can be easily extracted using 3D models (Fisher and Hanrahan, 2010). Additionally, object pairs and groups may have spatial arrangements that are diagnostic for scene category and a more sophisticated learning approach could glean these from the data. For example, although both *dining rooms* and *conference rooms* tend to have centrally located "table" and "chairs," and may also contain a "telephone," the presence of telephone *on top of* the table is diagnostic of *conference room*. On the other hand, a structural description on a scene may not be a good model for human scene gist as it has been shown that human scene classification performance can be well explained as the perception of a set of unbound features (Evans and Treisman, 2005). Similarly, electrophysiological markers structural scene processing occur later than markers of semantic processing (Võ and Wolfe, 2013). Taken together, these suggest that the first scene representation may include little structural information.

As ensemble statistics had better classification performance, feature-for-feature, compared to individual objects, a structural model that coarsely localizes these types of features may prove to be more fruitful for future work.

## GENERAL DISCUSSION

In this work, I have provided a set of real world image statistics at the level of labeled objects, and assessed the utility of these measurements for scene categorization. By understanding the regularities of natural images, we can design experiments to understand how these redundancies are exploited by the human visual system to efficiently recognize environments and search for objects in those environments.

## CATEGORY INFORMATION COMES FROM DIFFERENT LEVELS OF ANALYSIS

I have examined scene-object relationships at three levels of analysis: the ensemble level, the bag of words level, and the structural level. Statistics measured at each level of analysis contained sufficient information to categorize scene environments into basic- and superordinate-level categories. Although we intuitively know that *kitchens* and *offices* differ in terms of the objects found in them, this work also demonstrates that scene categories differ in terms of the amount and types of “things” found in them (ensemble statistics), and to a certain degree in the spatial distribution of their objects (structural statistics).

Additionally, quantitative analysis of objects in scenes allows us to test the plausibility of hypotheses on the role of object perception in rapid scene categorization. Biederman (1981) suggested that scenes might be recognized by first recognizing a single, prominent object in the scene. In section Object Pairs and Groups, I demonstrated that knowledge of either the largest object or the most centered object was insufficient to reproduce the high classification performance of human observers. Adding additional objects increases the diagnosticity for the scene, so a path for future work will be to examine how small groups of objects might be rapidly perceived to give rise to scene gist. Classification performance using a few objects as features lagged behind classification performance of ensemble statistics, suggesting that the coarse object information provided by the ensembles was more informative about scene category than individual objects.

## NOT ALL SCENE CATEGORIES ARE CREATED EQUALLY

Similarly, scene categories in different superordinate categories (indoor vs. outdoor, or natural vs. urban) differ markedly from one another at each level of analysis. Compared to outdoor scene categories, indoor environments have a higher object density as well as greater object variety.

The identities of the objects found in scenes also differs between superordinates, as very few objects were found in both indoor and outdoor scenes. The majority of objects in the database were found in only one or two scene categories, so knowing that an object is present in a scene provides considerable information about the scene environment. However, when considering the few objects that are found in many scene categories (such as “door,” “window,” or “trash can”), object position in the image can (but tends not to) differ by superordinate category, thus giving little additional predictive information about the scene category above that of the object identity.

Why do these indoor scene categories differ from the outdoor scene categories? One limitation of this database is that the indoor scene categories reflect small-scale indoor environments in the home and workplace. Perhaps larger indoor environments such as *department store* or *warehouse* would show patterns more similar to the outdoor environments, as larger environments mean that more objects will be too small to individually label, leading to a smaller number of measured objects.

Interestingly, categorization accuracy for all superordinate-level categories was found to be similar for each of the classifiers considered here. This was unexpected, as indoor scene

categorization is often considered to be a harder problem than outdoor scene categorization (Quattoni and Torralba, 2009). This result suggests that machine vision systems performing indoor scene categorization can be improved in at least two ways: first, the use of “objectness” detectors (Alexe et al., 2012) could be employed to understand object density and other ensemble statistics that are somewhat diagnostic of scene categories, and second, to use object detectors for the objects that provide the most mutual information for distinguishing scene categories.

## ARE ALL OBJECT TYPES CREATED EQUALLY?

Throughout this paper, I have treated each annotated label equally for the purposes of statistical analysis. “Sky” is just as much of an object as “book” in the database, even though it is not tangible and has no clear boundaries in the world. Although defining what counts as an object is a notoriously difficult problem (for a review see Feldman, 2003), one might want to consider sub-types of objects. For example, one might distinguish between object labels that refer to count nouns vs. mass nouns (Burge, 1972; Adelson, 2001; Huntley-Fenner et al., 2002; Prasada et al., 2002). Count nouns are labeled objects that are discrete and countable (“mug,” “building,” “car,” “book”) while mass nouns are regions with no fixed units or boundaries (“field,” “water,” “smoke,” “sky”). This distinction appears to be a fundamental difference in object representation that is present from a very early age (Huntley-Fenner et al., 2002). Alternatively, some of the annotated labels reflect background or boundary elements of a scene, such as “ground,” “sky,” “wall,” or “ceiling.” As a well-accepted definition of a visual scene includes the lawful arrangement of objects on a background (Henderson and Hollingworth, 1999), it is possible that these labeled regions have a different perceptual status than other labels such as “bowl” or “book.” Indeed, objects that make up scene boundaries have the highest mutual information for distinguishing between scene categories (see Table 11). However, a glance at the labels in Appendix B will convince the reader that it is very easy to find unclear cases.

## GENERALIZABILITY AND DATABASE BIAS

How generalizable are these findings? In other words, how much do they say about the distribution of objects in the world, and how much do they say about the idiosyncrasies of this particular database? Although the eight-category database from Oliva and Torralba (2001) used in the main database is a standard scene classification set in computer vision, more modern work has criticized this database for relying too heavily on the Corel Stock Photo collection (Torralba and Efros, 2011), which may represent only over-stylized representations of scenes. Similarly, the indoor images largely consist of highly idealized environments from real estate websites. Do these generalize to more everyday environments? In order to address this question, I have computed all statistics on a separate auxiliary database, and I have shown the similarities and differences between the two datasets whenever possible in Appendix D. Assuming that the bias in these datasets is independent, their degree of overlap reflects the generalizability of these statistics (Torralba and Efros, 2011; Khosla et al., 2012). This assumption is likely to be optimistic, however, as both datasets are

part of the larger set of scenes that people find remarkable enough to photograph and share on the web in the first place.

The two datasets examined in this work showed considerable but not perfect overlap. It is likely that the noted differences between natural landscapes and indoor environments are robust to dataset bias, but perhaps not the differences between urban and indoor scenes. The auxiliary set showed that the differences in these superordinates is driven primarily by natural landscape images, as this database contained very complex urban environments whose images had object density similar to indoor environments. Both datasets showed remarkable overlap in object frequency and mutual information, making these measures generally useful for the design of new experiments on object-scene context. Similarly, the measured entropy was very similar between the two datasets, suggesting that this statistic is robust to any dataset bias. The specificity of objects to a particular scene category was also observed in both sets. However, other measurements should be taken with more caution. The main dataset showed more redundancy (scenes having the same combination of objects) than the auxiliary set, and this manifested itself in higher classifier performance across the board. Appendix D contains more details on the specific differences in the findings between the two data sets.

Future investigations will continue to validate the generalizability of these data via comparison to other annotated databases such as SUn (Xiao et al., 2010), or through modeling the bias directly (Khosla et al., 2012). Separately, one can see how these statistics match the intuitions of human observers, although observers' intuitions should not necessarily be counted as ground truth, because we are insensitive to statistical base rates in some domains (Tversky and Kahneman, 1974).

### THE UTILITY OF OBJECT CONTEXT STATISTICS

Although it is generally recognized that lawful contextual relationships facilitate scene and object recognition, work in this area

has been limited because these contextual relationships have not been fully characterized and quantified. Previous work has characterized contextual relationships as merely being the intuitive plausibility of an object for a given scene environment. Many of the scene-object pairs in these experiments include informative but rare objects, such as a "moose" in a *forest*. Although a moose is more likely to be found in a forest when compared to other types of environments, the vast majority of *forest* images will not include a "moose." By measuring object frequency, diagnosticity and mutual information, experimenters will be able to determine the perceptual and memory consequences of these relationships individually. Furthermore, current experiments treat contextual relationships as binary—an object is either contextually related to an environment or it is not. However, the statistics measured here are continuous, allowing for more subtle questions to be asked.

More broadly, it has been argued that we cannot yet perform well-controlled studies on natural scene images because it is too difficult to understand or control the stimuli (Rust and Movshon, 2005). The results presented here take a necessary step toward this goal by characterizing complex scene stimuli in terms of quantified object-scene relationships. At all levels of analysis, real-world scene images show remarkable redundancy that can be utilized by the brain to represent the world efficiently. Therefore, measuring these statistics allows us to better understand and control our stimuli and to move forward into more real-world vision research.

### ACKNOWLEDGMENTS

Thanks to Antonio Torralba, Aude Oliva, and Jeremy Wolfe for providing advice and guidance for this project. Thanks to Karla Evans, Melissa Vo, Talia Konkle, Fei-Fei Li Adam Katz, Marius C. Iordan, Chris Baldassano, and Olga Russakovsky for useful conversations about this project. This work was supported by an NSF-GRF to Michelle R. Greene, as well as NRSA F32 EY19815 to Michelle R. Greene.

### REFERENCES

- Adelson, E. H. (2001). On seeing stuff: the perception of materials by humans and machines. *Proc. SPIE* 4299, 1–12. doi: 10.1117/12.429489
- Alexe, B., Deselaers, T., and Ferrari, V. (2012). Measuring the objectness of image windows. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 2189–2202. doi: 10.1109/TPAMI.2012.28
- Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends Cogn. Sci.* 15, 122–131. doi: 10.1016/j.tics.2011.01.003
- Alvarez, G., and Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychol. Sci.* 19, 392–398. doi: 10.1111/j.1467-9280.2008.02098.x
- Alvarez, G. A., and Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proc. Natl. Acad. Sci.* 106, 7345–7350. doi: 10.1073/pnas.0808981106
- Ariely, D. (2001). Seeing sets: representation by statistical properties. *Psychol. Sci.* 12, 157–162. doi: 10.1111/1467-9280.00327
- Attneave, F. (1954). Some informational aspects of visual perception. *Psychol. Rev.* 61, 183–193. doi: 10.1037/h0054663
- Auckland, M. E., Cave, K., and Donnelly, N. (2007). Nontarget objects can influence perceptual processes during object recognition. *Psychon. Bull. Rev.* 14, 332–337. doi: 10.3758/BF03194073
- Balas, B., Nakano, L., and Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *J. Vis.* 9, 1–18. doi: 10.1167/9.12.13
- Bar, M. (2004). Visual objects in context. *Nat. Rev. Neurosci.* 5, 617–625. doi: 10.1038/nrn1476
- Bar, M., and Aminoff, E. (2003). Cortical analysis of visual context. *Neuron* 38, 347–358. doi: 10.1016/S0896-6273(03)00167-3
- Bar, M., and Ullman, S. (1996). Spatial context in recognition. *Perception* 25, 343–352. doi: 10.1068/p250343
- Barlow, H. (2001). The exploitation of regularities in the environment by the brain. *Behav. Brain Sci.* 24, 602–607. doi: 10.1017/S0140525X01000024
- Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge: Cambridge University Press.
- Becker, M., Pashler, H., and Lubin, J. (2007). Object-intrinsic oddities draw early saccades. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 20–30. doi: 10.1037/0096-1523.33.1.20
- Biederman, I. (1981). "On the semantics of a glance at a scene," in *Perceptual Organization* (Hillsdale, NJ: Erlbaum), 213–253.
- Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115–147. doi: 10.1037/0033-295X.94.2.115
- Biederman, I. (1988). "Aspects and extensions of a theory of human image understanding," in *Computational Processes in Human Vision: An Interdisciplinary Perspective*, ed Z. W. Pylyshyn (Norwood, NJ: Ablex), 370–428.
- Biederman, I., Mezzanotte, R., and Rabinowitz, J. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cogn. Psychol.* 14, 143–177. doi: 10.1016/0010-0285(82)90007-X
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet

- allocation. *J. Mach. Learn. Res.* 3, 993–1022.
- Bosch, A., Zisserman, A., and Muñoz, X. (2006). “Scene classification via pLSA,” in *Computer Vision—ECCV 2006*, eds A. Leonardis, H. Bischof, and A. Pinz (Graz: Springer Berlin Heidelberg), 517–530.
- Boyce, S. J., Pollatsek, A., and Rayner, K. (1989). Effect of background information on object identification. *J. Exp. Psychol. Hum. Percept. Perform.* 15, 556–566. doi: 10.1037/0096-1523.15.3.556
- Brockmole, J. R., and Henderson, J. M. (2006). Using real-world scenes as contextual cues for search. *Vis. Cogn.* 13, 99–108. doi: 10.1080/13506280500165188
- Brockmole, J. R., and Võ, M. (2010). Semantic memory for contextual regularities within and across scene categories: evidence from eye movements. *Attent. Percept. Psychophys.* 72, 1803–1813. doi: 10.3758/APP.72.7.1803
- Burge, T. (1972). Truth and mass terms. *J. Philos.* 69, 263–282. doi: 10.2307/2024729
- Castelhano, M. S., and Henderson, J. (2008). The influence of color on the perception of scene gist. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 660–675. doi: 10.1037/0096-1523.34.3.660
- Chang, C. C., and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 1–27. doi: 10.1145/1961189.1961199
- Choi, M., Lim, J., Torralba, A., and Willsky, A. (2010). “Exploiting hierarchical context on a large database of object categories.” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, (San Francisco, CA), 129–136.
- Chong, S. C., and Treisman, A. (2003). Representation of statistical properties. *Vision Res.* 43, 393–404. doi: 10.1016/S0042-6989(02)00596-5
- Chong, S. C., and Treisman, A. (2005). Attentional spread in the statistical processing of visual displays. *Percept. Psychophys.* 67, 1–13. doi: 10.3758/BF03195009
- Chun, M. M., and Jiang, Y. (1998). Contextual cueing: implicit learning and memory of visual context guides spatial attention\*1, \*2. *Cogn. Psychol.* 36, 28–71. doi: 10.1006/cogp.1998.0681
- Davenport, J. (2007). Consistency effects between objects and scenes. *Mem. Cognit.* 35, 393–401. doi: 10.3758/BF03193280
- Davenport, J., and Potter, M. C. (2004). Scene consistency in object and background perception. *Psychol. Sci.* 15, 559–564. doi: 10.1111/j.0956-7976.2004.00719.x
- Deerwester, S., Dumais, S., Landauer, T., Furnas, G., and Harshman, R. (1990). Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* 41, 391–407.
- De Graef, P., Christiaens, D., and D’Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychol. Res.* 52, 317–329. doi: 10.1007/BF00868064
- Eckstein, M., Drescher, B., and Shimozaki, S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. *Psychol. Sci.* 17, 973–980. doi: 10.1111/j.1467-9280.2006.01815.x
- Elazary, L., and Itti, L. (2008). Interesting objects are visually salient. *J. Vis.* 8, 1–15. doi: 10.1167/8.3.3
- Evans, K. K., and Treisman, A. (2005). Perception of objects in natural scenes: is it really attention free? *J. Exp. Psychol. Hum. Percept. Perform.* 31, 1476–1492. doi: 10.1037/0096-1523.31.6.1476
- Fei-Fei, L., Iyer, A., Koch, C., and Perona, P. (2007). What do we perceive in a glance of a real-world scene? *J. Vis.* 7, 1–29. doi: 10.1167/7.1.10
- Fei-Fei, L., and Perona, P. (2005). “A Bayesian hierarchical model for learning natural scene categories,” in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, *IEEE Computer Society*, Vol. 2, (San Diego, CA), 524–531. doi: 10.1109/CVPR.2005.16
- Feldman, J. (2003). What is a visual object? *Trends Cogn. Sci.* 7, 252–256. doi: 10.1016/S1364-6613(03)00111-6
- Fine, I., and MacLeod, D. I. A. (2001). Visual segmentation based on the luminance and chromaticity statistics of natural scenes. *J. Vis.* 1, 63. doi: 10.1167/1.3.63
- Fisher, M., and Hanrahan, P. (2010). Context-based search for 3D models. *ACM Trans. Graph.* 29, 182:1–182:10. doi: 10.1145/1882261.1866204
- Friedman, A. (1979). Framing pictures: the role of knowledge in automatized encoding and memory for gist. *J. Exp. Psychol. Gen.* 108, 316–355. doi: 10.1037/0096-3445.108.3.316
- Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annu. Rev. Psychol.* 59, 167–192. doi: 10.1146/annurev.psych.58.110405.085632
- Geisler, W. S., Perry, J. S., Super, B. J., and Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Res.* 41, 711–724. doi: 10.1016/S0042-6989(00)00277-7
- Golz, J., and MacLeod, D. I. A. (2002). Influence of scene statistics on colour constancy. *Nature* 415, 637–640. doi: 10.1038/415637a
- Green, C., and Hummel, J. (2006). Familiar interacting object pairs are perceptually grouped. *J. Exp. Psychol. Hum. Percept. Perform.* 32, 1107–1119. doi: 10.1037/0096-1523.32.5.1107
- Greene, M. R., and Oliva, A. (2009). Recognition of natural scenes from global properties: seeing the forest without representing the trees. *Cogn. Psychol.* 58, 137–176. doi: 10.1016/j.cogpsych.2008.06.001
- Haberman, J., and Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Curr. Biol.* 17, R751–R753. doi: 10.1016/j.cub.2007.06.039
- Haberman, J., and Whitney, D. (2012). “Ensemble perception: summarizing the scene and broadening the limits of visual processing,” in *From Perception to Consciousness: Searching with Anne Treisman*, eds J. Wolfe and L. Robertson (Oxford, UK: Oxford University Press), 339–349.
- Henderson, J. M. (1992). Object identification in context: the visual processing of natural scenes. *Can. J. Psychol.* 46, 319–341. doi: 10.1037/h0084325
- Henderson, J. M., and Hollingworth, A. (1999). High-level scene perception. *Annu. Rev. Psychol.* 50, 243–271. doi: 10.1146/annurev.psych.50.1.243
- Henderson, J. M., Weeks, P., and Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *J. Exp. Psychol. Hum. Percept. Perform.* 25, 210–228. doi: 10.1037/0096-1523.25.1.210
- Hollingworth, A., and Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *J. Exp. Psychol. Gen.* 127, 398–415. doi: 10.1037/0096-3445.127.4.398
- Howe, C. Q., and Purves, D. (2004). Size contrast and assimilation explained by the statistics of natural scene geometry. *J. Cogn. Neurosci.* 16, 90–102. doi: 10.1162/089892904322755584
- Huntley-Fenner, G., Carey, S., and Solimando, A. (2002). Objects are individuals but stuff doesn’t count: perceived rigidity and cohesiveness influence infants’ representations of small groups of discrete entities. *Cognition* 85, 203–221. doi: 10.1016/S0010-0277(02)00088-4
- Joubert, O. R., Rousselet, G. A., Fize, D., and Fabre-Thorpe, M. (2007). Processing scene context: fast categorization and object interference. *Vision Res.* 47, 3286–3297. doi: 10.1016/j.visres.2007.09.013
- Kadar, I., and Ben-Shahar, O. (2012). A perceptual paradigm and psychophysical evidence for hierarchy in scene gist processing. *J. Vis.* 12:pii:16 doi: 10.1167/12.13.16
- Karklin, Y., and Lewicki, M. S. (2003). Learning higher-order structures in natural images. *Network* 14, 483–499. doi: 10.1088/0954-898X/14/3/306
- Khosla, A., Zhou, T., Malisiewicz, T., Efros, A., and Torralba, A. (2012). “Undoing the damage of database bias,” in *European Conference on Computer Vision (ECCV)*, (Florence).
- Kim, J. G., and Biederman, I. (2010). Where do objects become scenes? *Cereb. Cortex* 21, 1738–1746. doi: 10.1093/cercor/bhq240
- Lazebnik, S., Schmid, C., and Ponce, J. (2006). “Beyond bags of features: spatial pyramid matching for recognizing natural scene categories,” in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society*, Vol. 2, (New York, NY), 2169–2178.
- Li, L.-J., Su, H., Lim, Y., and Fei-Fei, L. (2010). “Objects as attributes for scene classification,” in *Trends and Topics in Computer Vision*, ed K. N. Kutulakos (Crete: Springer Berlin Heidelberg), 57–69.
- Li, W. (1992). Random texts exhibit Zipf’s-law-like word frequency distribution. *IEEE Trans. Inf. Theory* 38, 1842–1845. doi: 10.1109/18.165464
- Liu, K., and Jiang, Y. (2005). Visual working memory for briefly presented scenes. *J. Vis.* 5, 650–658. doi: 10.1167/5.7.5
- Loftus, G. R., and Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *J. Exp. Psychol. Hum. Percept. Perform.* 4, 565–572. doi: 10.1037/0096-1523.4.4.565
- MacEvoy, S. P., and Epstein, R. A. (2011). Constructing scenes from objects in human occipitotemporal cortex. *Nat. Neurosci.* 14, 1323–1329. doi: 10.1038/nn.2903
- Mack, M. L., and Palmeri, T. J. (2010). Modeling categorization of scenes containing consistent versus inconsistent objects. *J. Vis.* 10, 1–11. doi: 10.1167/10.3.11

- Mack, S. C., and Eckstein, M. P. (2011). Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *J. Vis.* 11, 1–16. doi: 10.1167/11.9.9
- Miller, G. A. (1995). WordNet: a lexical database for English. *Commun. ACM* 38, 39–41. doi: 10.1145/219717.219748
- Minsky, M. (1975). “A framework for representing knowledge,” in *The Psychology of Computer Vision*, ed P. H. Winston (New York NY: McGraw Hill), 211–277.
- Neider, M. B., and Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Res.* 46, 614–621. doi: 10.1016/j.visres.2005.08.025
- Neider, M. B., and Zelinsky, G. J. (2008). Exploring set size effects in scenes: identifying the objects of search. *Vis. Cogn.* 16, 1. doi: 10.1080/13506280701381691
- Oliva, A., and Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.* 42, 145–175. doi: 10.1023/A:1011139631724
- Oliva, A., and Torralba, A. (2007). The role of context in object recognition. *Trends Cogn. Sci.* 11, 520–527. doi: 10.1016/j.tics.2007.09.009
- Olshausen, B. A., and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609. doi: 10.1038/381607a0
- Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Mem. Cogn.* 3, 519–526. doi: 10.3758/BF03197524
- Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., and Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nat. Neurosci.* 4, 739–744. doi: 10.1038/89532
- Potter, M. C. (1976). Short-term conceptual memory for pictures. *J. Exp. Psychol. Hum. Learn. Mem.* 2, 509–522. doi: 10.1037/0278-7393.2.5.509
- Prasada, S., Ferenz, K., and Haskell, T. (2002). Conceiving of entities as objects and as stuff. *Cognition* 83, 141–165. doi: 10.1016/S0010-0277(01)00173-1
- Quattoni, A., and Torralba, A. (2009). “Recognizing indoor scenes.” in *IEEE Transactions on Computer Vision and Pattern Recognition (CVPR)*, (Miami, FL).
- Rao, R., Olshausen, B., and Lewicki, M. (eds.). (2002). *Probabilistic Models of the Brain: Perception and Neural Function*. Cambridge, MA: MIT Press.
- Renninger, L. W., and Malik, J. (2004). When is scene identification just texture recognition? *Vision Res.* 44, 2301–2311. doi: 10.1016/j.visres.2004.04.006
- Ruderman, D. (1994). The statistics of natural images. *Network* 5, 517–548. doi: 10.1088/0954-898X/5/4/006
- Russell, B., Torralba, A., Murphy, K., and Freeman, W. (2008). LabelMe: a database and web-based tool for image annotation. *Int. J. Comput. Vis.* 77, 157–173. doi: 10.1007/s11263-007-0090-8
- Rust, N. C., and Movshon, T. (2005). In praise of artifice. *Nat. Neurosci.* 8, 1647–1650. doi: 10.1038/nn1606
- Schwartz, O., and Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nat. Neurosci.* 4, 819–825. doi: 10.1038/90526
- Schyns, P. G., and Oliva, A. (1994). From blobs to boundary edges: evidence for time- and spatial-scale-dependent scene recognition. *Psychol. Sci.* 5, 195–200. doi: 10.1111/j.1467-9280.1994.tb00500.x
- Simoncelli, E. P., and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annu. Rev. Neurosci.* 24, 1193–1216. doi: 10.1146/annurev.neuro.24.1.1193
- Sivic, J., and Zisserman, A. (2003). “Video Google: a text retrieval approach to object matching in videos,” in *Proceedings of the Ninth IEEE International Conference on Computer Vision*, Vol. 2 (Washington, DC), 1470–1477. doi:10.1109/ICCV.2003.1238663
- Tatler, B. W. (2007). The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *J. Vis.* 7, 1–17. doi: 10.1167/7.14.4
- Tatler, B. W., Baddeley, R. J., and Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Res.* 45, 643–659. doi: 10.1016/j.visres.2004.09.017
- Torralba, A., and Efros, A. (2011). “An unbiased look at dataset bias.” in *IEEE Transactions on Computer Vision and Pattern Recognition (CVPR)*, (Colorado Springs, CO) 1521–1528.
- Torralba, A., and Oliva, A. (2002). Depth estimation from image structure. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 1–13. doi: 10.1109/TPAMI.2002.10.33214
- Torralba, A., and Oliva, A. (2003). Statistics of natural image categories. *Network* 14, 391–412. doi: 10.1088/0954-898X/14/3/302
- Torralba, A., and Sinha, P. (2001). “Statistical context priming for object detection.” in *IEEE Proceedings of the International Conference in Computer Vision*, Vol. 1, (Vancouver, CA), 763–770.
- Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131. doi: 10.1126/science.185.4157.1124
- Ullman, S., Vidal-Naquet, M., and Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nat. Neurosci.* 5, 682–687. doi: 10.1038/nn870
- van Hateren, J. H., and Ruderman, D. L. (1998). Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc. R. Soc. B Biol. Sci.* 265, 2315–2320. doi: 10.1098/rspb.1998.0577
- VanRullen, R., and Koch, C. (2003). Competition and selection during visual processing of natural scenes and objects. *J. Vis.* 3, 75–85. doi: 10.1167/3.1.8
- Vickery, T. J., King, L., and Jiang, Y. (2005). Setting up the target template in visual search. *J. Vis.* 5, 81–92. doi: 10.1167/5.1.8
- Võ, M. L., and Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *J. Vis.* 9, 1–15. doi: 10.1167/9.3.24
- Võ, M. L.-H., and Henderson, J. M. (2011). Object-scene inconsistencies do not capture gaze: evidence from the flash-preview moving-window paradigm. *Attent. Percept. Psychophys.* 73, 1742–1753. doi: 10.3758/s13414-011-0150-6
- Võ, M. L.-H., and Wolfe, J. M. (2013). Differential electrophysiological signatures of semantic and syntactic scene processing. *Psychol. Sci.* 24, 1816–1823. doi: 10.1177/0956797613476955
- Wolfe, J. M., Alvarez, G. A., Rosenholtz, R., Kuzmova, Y. I., and Sherman, A. M. (2011a). Visual search for arbitrary objects in real scenes. *Attent. Percept. Psychophys.* 73, 1650–1671. doi: 10.3758/s13414-011-0153-3
- Wolfe, J. M., Võ, M. L.-H., Evans, K. K., and Greene, M. R. (2011b). Visual search in scenes involves selective and nonselective pathways. *Trends Cogn. Sci.* 15, 77–84. doi: 10.1016/j.tics.2010.12.001
- Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2010). “SUN database: Large-scale scene recognition from abbey to zoo,” in *Presented at the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (San Francisco, CA), 3485–3492. doi: 10.1109/CVPR.2010.5539970
- Zetzsche, C., Barth, E., and Wegmann, B. (1993). “The importance of intrinsically two-dimensional image features in biological vision and picture coding,” in *Digital Images and Human Vision*, ed A. B. Watson (Cambridge, MA: MIT Press), 109–138.

**Conflict of Interest Statement:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 25 April 2013; accepted: 03 October 2013; published online: 29 October 2013.

Citation: Greene MR (2013) Statistics of high-level scene context. *Front. Psychol.* 4:777. doi: 10.3389/fpsyg.2013.00777

This article was submitted to *Perception Science*, a section of the journal *Frontiers in Psychology*.

Copyright © 2013 Greene. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

## APPENDIX A: RAW LABELS (SIC)

<blank>, “?object,” abinet, adding machine, advert, aerial, aerial building, air conditioning, air conditioner, air conditioning, air conditioning occluded, air filter, air vent, airco, airplane, alarm, alarm clock, alarmclock, alcove, all, American flag, animal, animal seal, animals, antique bureau, apple, apples, arcade, arcade occluded, arch, arches, archway, arearug, arid ground, arm chair, armchair, armchair back, armchair part, amchair top, armoire, art, art piece, art?, artichokes, ashtray ash tray, attic, automan, avenue, awning, awning crop, awning occluded, awning tidy up, back of chair, back wall, backpack, badge, bag, balcony occluded, balcony, balcony crop, balcony occluded, bale, balustrade, bandstand, banister, bank, bar, barbeque, barrier occluded, base board, basket, basket food, basket fruits, basket of fruit, basket of magazines, basket of towels, basket table, basket of brushes, baskets, bath, bath faucet, bath mat, bath tub, bath robe, bathtub, bathtub plug, beach, beam, bed, bed part, bed posts, bed skirt, bedskirt, bedspread, bell, bell pepper, bell tower, bench, bench occluded, bib, biclist occluded, bicycle, bicycle occluded, bicyclist, bicyclists, bidet, bidet faucet, bilding, bin, bin occluded, binder binders, bird, bird figurine, bird occluded, birdcade, blackboard, blanket, blanket, blender, blind, blinds, block, block of cheese, blocks, bluildings, boad, board, board games, boarder, boat, boat crop, boat cropped, boat decoration, boat occluded, boats, book, book case, book shelf, book shelves, bookcase, books, books on shelf, bookshelf, bookshelves, boooks, boot, boots, border, bottle, bottle, bottle top, bottles, bottom bunk, bottom of chair, bottom of chairs, bouldings, bouldings occluded, bouquet, bouquet flowers, bowel, bowl, bowl of apples, bowl of fruit, bowl of popcorn, bowl of strawberries, bowl of vegetables, bowl with food, bowl with fruit, bowl with vegis, bowls, box, box contents, box occluded, box?, boxes, braided garlic, branch, branches, branches, brand name, brand name crop, brand name occluded, bread, bread tray, brick fireplace, brick wall, bridge, bridge crop, bridge handrail, bridge occluded, briefcase, broom, brushes, brush, brushes, bucket, buffet, buggy, buiding, buiding occluded, buidings occluded, buildig occluded, buildin, buildin occluded, building, building aerial, building crop, building façade, building occluded, building occluded, building skyscraper, buildingl occluded, buildings, buildings crop, buildings occluded, buildings occludeds, buildingsoccluded, buildins, buildins occluded, buildintgs occluded, building, buildings occluded, bulletin board, buoy, bus, bus stop, bus occluded, bus stop, bus stop occluded, bushes, business card, bus occluded, cabin, cabin occluded, cabinet, cabinet door, cabinet dresser, cabinets, cabinets, cabinetet shelf, cake, cake dish, cake stand, calander, calendar, calender, can, canal water, candel, candle, candle holder, candle stand, candle stick, candles, canister, canister with brushes, canister with utensils, canisters, cannon, canoe, canopy, cans, captus, captus crop, car, car crop, car side, car az90deg, car crop, car frontal, car frontal occluded, car occluded, car occluded, car rear, car rear az90deg, car rear occluded, car rear side, car rear side crop, car rear side crop, car side, car side az0deg, car side az180deg, car side crop, car side occlude, car side occluded, car side rear, car side rear crop, car sides, car\_back, car\_front, car\_left, car\_right, car\_top\_back, car\_top\_fornt, car\_top\_front, card, carpet, carpet floor, carrots, cars, cars occluded, cars side, cars side crop, cars

side occluded, cars sides, cart, case, cassettes, castle, cat, cd, cd’s, cdoor, cds, ceiiing light, ceiiing, ceiling, ceiling fan, ceiling lamp, ceiling lamps, ceiling light, ceiling molding, ceiling tile, ceiling vent, ceiling lamp, ceiling light, celing, cell phone, centerpiece, centra reservation, central reservation, ceramic rooster, certain, certificate, chair, chair back, chair bottom, chair crop, chair leg, chair legs, chair occluded, chair part, chair seat, chair top, chair wheel, chairs, chaise lounge, chandelier, chandelier, chandelier, chandler, changing table, channel, char, cheese, chess board, chessboard, chest, chiar, child walking, chimney, china, china cabinet, china hutch, china plate, cigarettes, city, cliff, clip board, clock, clock occluded, closet, closets, cloth, clothes, clothes hamper, cloths, cloud, clouds, coaster, coat, coat rack, coatrack, coffee maker, coffee pot, coffee table, column, column, column occluded, columns, comfoter, computer, computer monitor, computer screen, computer tower, computers, conference table, cooler, copier, cords, couch, counter, counter top, counter, cove, covered balcony, covered balcony occluded, cow, cows, cpboard, crane, crane occluded, crib, cross, crosswalk, cubbord, cubby, cubical, cubicle, culumn, cup, cup and saucer, cup holders, cup with flower, cupbard, cupboard, cups, curb, curtain, curtain rod, curtains, cushion, cutting board, cutting board island, cutting board with vegetables, cyclist, dam, decoration, decorations, decorative ball, decorative balls, decorative boat, decorative bowl, decorative box, decorative fish, decorative kimono, decorative mask, decorative mirror frame, decorative molding, decorative object, decorative objects, decorative pillow, decorative plat, decorative plate, decorative pots, decorative tree, decorative urn, decorative wall hanging, decoratvie plate, deoderant, derer ground, desert, desert field, desert ground, desk, desk calendar, desk calender, desk divider, desk lamp, desk mat, desk organizer, desk separator, desks, digital watch, dining table, disc, dish, dish rack, dish towel, dish towels, dishes, dishrack, dishwahaer, dishwasher, disks, display case, display stand, dock, dog, dog crop, doily, doll, dolphin, dome, dome crop, dome occluded, door, door crop, door entrance, door frame, door knob, door mat, door occluded, door\_, doorframe, doorpart, doors, doorway, double door, double door crop, double door occluded, double window occluded, drape, drapes, drawer, drawer nob, drawers, drawyers, dresser, dressing screen, dressor, dried flowers, dried plant, drinking fountain, driver, dry earase board, dry earaser, dry erase board, dryer, dune, eagle, egg crate, eggplant, eilling lamp, electric mixer, electrical outlet, elevator door, elf, embankment, emplem, enclave, end board, end table, entertainment center, entrance, entrance occluded, entry, entry-phone, entry-way, envelope, envelopes, equipement, esplanade, estate, exit sign, external driver, external drivers, eye, falg, fall branch, fan, faucet, faucit, faucst, fax machine, fence, fence crop, fence occluded, fences crops, fern, ferns, field, field desert, field flowers, field grass, figurine, figurines, file, file box, file cabinet, file cabinets, file organizer, files, filing cabinet, fining cabinet\_, filing cabinets, fire, fire alarm, fire escape, fire extinguisher, fire hydrant, fire place, fire sprinkler, firehose, fireplace, fireplace screen, firewood, fish tank, flag, fip flop, fllor, flock, floor, floor carpet, floor marble, floor, carpet; floor\_, floral centerpiece, flower, flower arrangement, flower in vase, flower pot, flowers, flowers in pot, flowers in vase, flute, flyer, flyers, fog banck, fog bank, folder,

folders, food, food on plate, foot board, footboard, footbridge, footrest, forest, forest mountain, fork, forks, fountain, fountain crop, fountain occluded, fp screen, frame, framed mirror, framed picture, frozen over river, fruit, fruit bowl, fruit in bowl, fruit plate, fruit stand, fruits, fruits bowl, fruits plate, frying pan, funicular (railway), garage, garage door, garage door occluded, garages, garden, garland, garlic, garment bag, gas pump, gas station, gate occluded, glass, glass bowl, glass cupboard, glass door, glass doors, glass piece, glass shelf, glass table, glass wall, glasses, globe, goblet, goblets, goose, goose occluded, geese, grandfather clock, grapes, grass, grass field, green field, green peppers, greenhouse, grill, grille, grille occluded, ground, ground grass, ground grass crop, ground snow-covered, ground grass, hair, hair brush, hall, hand rail, hand towel, hand vacuum, handle, handrail, hanger, hanging bird, hanging fish, hanging hat, hanging instrument, hanging lamp, hanging light, hanging monitor, hanging pan, hanging pans, hanging patches, hanging pitcher, hanging plant, hanging plate, hanging pot, hanging pot rack, hanging rack, hanging rug, hanging toy, hanging utensils, hanging utensils, hanging utensil rack, hanging wall flower, harbor, hat, hay bale, head board, head stand, headboard, headphones, hearth, heater, hedge, highway, hill, hill urban, hills, hold back, hood, horse, hot pad, house, house occluded, house crop, house in ruins, house occluded, houses, houses crop, houses crops, houses occluded, houses occludeds, human, hung out, hutch, hydrant, ice bucket, idol, silver ware, image frame, in box, industry, inset ceiling light, inset ceiling lights, instrument, intercom, ipod, iron, island, isle, jar, jars, jet, jetski, jetty, jewelry box, joist, joists, jug, junk, kettle, key board, key board shelf, key pad, keyboard, kiosk, kite, Kleenex, Kleenex box, knife, knife holder, knife holder, knife set, knives, knobs, ladder, ladle, lake, lake water, laminating machine, lamp, lamp part, lamp shade, lamps, lampshade, land, lap top, laptop, large bowl, large window, lava, leaf, leaves, leaves tree, ledge, lemon, letter bin, lettuce, lichen, light, light fixture, light switch, lighter, lighthouse, lighthouse, lights, lighthouse, line persons, liquor bottle, litter bin, litter bin crop, little bear, lodge, loft, logs, lotion, lots of chairs, lower cabinets, machine, machines, magazine, magazine holder, magazine rack, magazines, magnifier, magnifying glass, mailbox, mailboxes, make up case, man, manhole, mantle, map, markers, massage table, massager, mast, mat, mattress, message board, message board\_, metal door, metal door occluded, metal shutters crop, microphone, microwave, night stand table, mill, mini blinds, mirror, mirror extender, misc., mixer, molding, monitor, monitor stand, monitor\_display, monitors, monitor, monkey, monkey occluded, monolith, monolith crop, monument, moon, motorbike occluded, motorbike, motorbike crop, motorbike occluded, motorbike side, motorbike side crop, motorbike side occluded, motorbikes side, motorcyclist, motorcyclist, motorcyclist crop, mountain, mountain pass, mountains, mountainside, mountain, mouse, mouse pad, mouse stand, mountain, mouth, movie screen, muffin, mug, mugs, napkin, napkin in glass, napkins, neck pillow, news stand, newspaper, night stand, night stand cabinet, night stand dresser, night stand part, night stand table, nightstand, notebook, notebooks, nothing, notice, nozzle, nutcracker, obelisk, object, object, object, objects, occluded, occluded sky, ocean, ocean water, object, onion, orange, ottoman, ottoman, ounter

top, outlet, overhead projector, pad, paint, painting, paintings, pallet, palm tree, palm pilot, palm tree, palm tree crop, palm tree cropped palm tree occlude, palm tree occluded, palm tree trunk, palm trees, palms trees, pan, pane, pane crop, pane occluded, panel, panelling, panels, pants, paper, paper filer, paper roll, paper sheet, paper towel, paper towel holder, paper towels, paper weight, papers, park, parking lot, parking meter, parking place, parking door, parking meter, parking meter occluded, pass window, passway, pasta, path, object, pear, pears, pedestal, pedestal, pedestrian street, pen, pen box, pen holder, pen holder, pen set, pencil, pencil cup, pencils, pens, people, people riding, people sitting, people walking, pepper mill, pepper shaker, person, person standing occluded, person, person standing, person walking, person boy, person boy standing, person crop, person cyclist, person in line, person man back crop, person man sitting, person man standing, person man walking, person man walking occluded, person occluded, person riding, person standing, person sitting, person sitting cropped, person sitting occluded, person skiing, person skiing crop, person skiing occluded, person standing, person standing, person standing crop, person standing kid, person standing occluded, person sweeping, person swimming, person swimming, person wading, person walking, person walking crop, person walking occluded, person walking occluded, person walking, person wise, person wise back, person wise back occluded, person wise back occluded, person wise crop, person wise in profile, person wise occluded, person woman, person woman sitting, person woman standing, person woman walking, person working, persons rowing, persons standing, persons walking, person standing, person walking, person walking, pestle and mortar, pet bowl, phone, photocopier, piano, picture, picture frame, pictures, pie, pig statue, pillar, pillow, pillows, pipe, pitcher, placard, place mat, place mat, placemat, plain, plants trees, plan pot, plank, plank occluded, plank, plant, plant box, plant box occluded, plant crop, plant fern, plant grapevine, plant in stand, plant in vase, plant occluded, plant pot, plant pot occluded, plant stand, plant tree, plants, plants fern, plants ferns, plaque, plate, plate and bowl, plate of fruit, plates, platform, platter, plug, plumbing, pneumatic tire, podium, pole, pole crop, pole occluded, poles, pond, porch, porch occluded, portfolio, portico, post it note, post it notes, poster, poster board, poster crop, poster occluded, posters pole, pot, pot holder, pot plant, pot plant crop, pot plant occluded, pots, potted flowers, potted plant, power cord, person driving occluded, precipice, printer, projection screen, projector, projection screen, projector, person standing, person walking, puddle, puffy object, pulley, purse, pylon, pylon occluded, quail, quay, quilt, rack, rack of fruit, radiator, radio, radio alarm clock, rail, railing, railings, rail, railroad track, railway, rainbow, range hood, recycling bins, red light, reflection, refrigerator, refrigerator, refuge, remote, remote control, revolving doors, rig, river, river bed, river side, river water, riverside, river water, road, road highway, road traffic, roads, room shade, roasting pan, robe, rock, rock cropped, rock wall, rocks, rocky ground, rocky hill, rocky mountain, rocky mountains, rocky mountain, rocky plain, rocks, rolling pin, roman shade, roof, roof occluded, room label, rooster, rooster figurine, round table, row of chairs, row of desks, rubber duckie, rug, ruin, runner, sack, sailboat, sailing boat, sailing railing,

salt shaker, sand, sand beach, sands, caofa, satellite dish, sauce pot, saucer and bowl, scaffolding, scaffolding crop, scale, scanner, sconce, screen, scrub brush, scrubland, sculpture, sculpture crop, sculpture occluded, sculpyure, sea, sea water, sea beach, sea water, seagull, seashell, serving tray, sewer, shade, shades, shadow, shampoo, shedders, sheep, shelf, shell-bowl of popuri, shelves, ship, ship occluded, shop window, shop window crop, shop window occluded, shower, shower curtain, shower curtain rod, shower door, shower faucet, shower head, shower nozzle, shrub, shrub cropped, shrubs, shrubs crop, shrubs occluded, shudder, shutter, side, side rail, side tabel, side table, side walk, sideboard, sidetable, sidewalk, sidewalk café, sig, sign, sign crop, sign occluded, signes, signs, siky, silver ware, silverware, sing, sink, sink faucet, site office, sk, skier, skier crop, skly, sky, sky light, sky occluded, sky sunset, skyscraper occluded, skyscrape, skyscraper, skyscraper building, skyscraper crop, skyscraper occluded, skyscraperoccluded, skyscrapers, skyscrapers occluded, skyscrapr occluded, skyscrapre occluded, skyscrpaer occluded, skyscrper occluded, skyscraper occluded, slated wooden panel, sled, sleepers, sleeping robe, sliding door, sliding door crop, sliding glass door, slipper, slippers, slope, sloped ceiling, small bowl, small plate, small rug, small table, small table part, small vase, smoke alarm, smoke detector, snow, snow covered, snow covered ground, snow covered mountain pass, snow covered plain, snow covered road, snow covered valley, snow land, snowly mountain, snowy covered field, snowy ground, snowy hill, snowy mountain, snowy mountain pass, snowy plain, snowy road, snowy trees, soap, soap bars, soap bottle, soap box, soap dish, soap dispenser, soap holder, soap on a rope, soaps, socket, sofa, sofa bed, sofa part, soil, soldier, sonwy mountain, sown field, space heater, spatulas, speaker, speaker, spice jar, spice rack, sping, sponge, spoon, spotlight, spray bottle, sprinkler, ssnowy mountain, stacj of papers, stack of books, stack of papers, stack of plates, staiscase, stained-glass window, stainless steel back splash, stainless steel splash guard, stainless steel wall, stair board, stair railing, staircase, staircase occluded, stairs, stairway, stake, stand, stand occluded, stands, stapler, starage rack, starfish, station, statue, streetlight, step, steps, stero, stick, sticker, sticks, stone, stone ball, stone ball crop, stone occluded, stone vase with flowers, stones, stones wall, stool, storage box, storage rack, stove, stove, stove gaurd, stove hood, stove nob, stove sheild, stove top, streelight, street, street light, street lighting, street market, streetcar, streetlamp, streetlight, streetlight crop, streetlight occluded, streetlights, streetlilght, string, stump of tree, subway, suitcase, sun, sun occluded, sunflower field, sunflowers field, sunflowers field, sunset, supermarket, supplies, support beam, tabel occluded, table, table cloth, table lamp, table leg, table occluded, table part, table runner, table top, table with tablecloth, tablecloth runner, tableland, tables, tank, tanker occluded, tapestry, tea cup and saucer, tea kettle, tea pot, teapot, teddy bear, telephone, telephone booth, telephone box, telephone box occluded, television, television cabinet, television case, television screen, television stand, tence, terotauuko, terrace, terrace occluded, text, thermos, thermostat, tile, tile ?, tiled wall, tissue, tissue box, tissues, toast, toaster, toaster oven, toilet, toilet brush, toilet brush stand, toilet paper, toilet paper holder, toilet paper roll, toll gate, tomatoes, toolbox, tooth brush holder, tooth brushes, toothbruch holder, toothbrush, toothbrush in cup, toothbrush in jar, toothbrushes,

toothpaste, toothbrush, top bunk, top chair, top of chair, top wall, torch, towel, towel hanger, towel rack, towels, tower, tower crop, tower occluded, town, toy, toys, tp roll, track, track lighting, tractor, traffic lights, traffic light, traffic light frontal, traffic light side, traffic lights, traffic lights crop, traffic sign, trafficlighs, trash, trash can, trash compactor, tray, tray table, tray with supplies, trays, tree, tree crop, tree cropped, tree cut down, tree in pot, tree leaves, tree occluded, tree top, tree trunk, tree trunk crop, tree trunk cropped, tree trunk fallen, tree trunk occluded, tree trunks, trees, trees cropped, trees occluded, trees occludeds, trees top, trophy, trres, truck, truck crop, truck frontal, truck frontal occluded, truck occluded, truck rear, truck rear occluded, truck side, truck side occluded, trucks occluded, truk, truk occluded, truk side occluded, trumpet, trunk, trunk tree fall, tub, tube, tuck occluded, tumbledown building, tunnel, tv, tv stand, type-writer, umbrella occluded, umbrellas, undergroth, undergrowth, urban plain, urban valley, urbanization, urinal, urn, utencil, utencil holder, utencil rack, utencil, utensils, valance, valley, van, van crop, van frontal, van occluded, van rear, van rear side, van side, van side crop, van side occluded, vanity, varehouse, vase, vase with flowers, vase with leaves, vase with plant, vases, vault, VCR, vegetable, vegetables, vegetation, vegitable, vegetables, vegetables in bowl, vegetables on a plate, vehicles, vent, vents, verge, vertical blinds, verticle blinds, viewpoint, vineyard, volcano, votive, wakk, wal, wall, wall boarder, wall border, wall clock, wall hang, wall hanging, wall lamp, wall light, wall mount, wall occluded, wall outlet, wall outlets, wall stitch, walls, wardrobe, wardrobe part, wash cloth, washcloths, washing machine, waste basket, water, water bottle, water bottles, water cooler, water fall, water fountain, water ocean, water pond, water river, water sea, waterfall, watering can, watermelon, wathervane, wave splash, waves, web cam, webcam, wheat field, wheelbarrow, wheelchair, wheels, white board, white out, whiteboard, whte board, widnow ledge, wind chime, windex, window, window ceiling, window crop, window frame, window ledge, window occluded, window pane, window seat, window shade, window shades, window shop, window shudders, window shutter, window sill, windows, windwo, wine bottle, wine cupboard, wine glass, wine glass, wine glasses, wine rack, wineglasses, wineglass, wineglasses, wire, wire rack, woman, woman walking, wood, wood beam, wood post, wooden ship, workstation, x, xx, xxx, zebra, zebra crossing, zebras.

## APPENDIX B: FINAL REGION LABELS

Adding machine, advert, air conditioner, air filter, airplane, alarmclock, alcove, antenna, apple, arcade, arch, armchair, artichokes, ashtray, attic, awning, backpack, bag, balcony, bandstand, barbecue, base board, basket, bath, bath mat, bath plug, bathrobe, beach, beam, bed, bed posts, bedskirt, bedspread, bell, bell pepper, bench, bicycle, bidet, binder, bird, birdcage, blanket, blender, block, board games, boat, book, bookshelf, boot, bottle, bouquet, bowl, box, braided garlic, branch, bread, bridge, briefcase, broom, bucket, buffalo, buffet, buggy, building, bulletin board, buoy, bus, bus stop, bush, business cards, cabin, cabinet, cactus, cake, cake dish, cake stand, calendar, can, canal water, candle, candle holder, canister, cannon, canoe, canopy, car, card, carpet, carrots, case, cassettes, castle, cat, cave, cds, ceiling, ceiling fan, ceiling tile, cell

telephone, centerpiece, certificate, chair, chaise lounge, chandelier, changing table, chart, cheese, chessboard, chest, chimney, china, china hutch, cigarettes, city, cliff, clip board, clock, closet, cloth, clothes, clothes hamper, cloud, coaster, coat, coat rack, coffee maker, coffee table, column, computer, computer monitor, cooler, counter, cow, crane, crib, cross, crosswalk, cubicle, cup, curb, curtain, curtain rod, cutting board, dam, decoration, deer, deodorant, desert, desk, desk divider, desk mat, desk organizer, dish, dish towel, dishrack, dishwasher, display case, dock, dog, doily, doll, dolphin, dome, door, door mat, drawer, dresser, dressing screen, drinking fountain, dryer, dune, eagle, egg crate, eggplant, electric mixer, electrical outlet, elevator door, emblem, end table, entertainment center, entrance, envelope, equipment, eraser, esplanade, estate, exit sign, external hard drive, fan, faucet, fax machine, fence, fern, field, file, file organizer, filing cabinet, fire, fire alarm, fire escape, fire extinguisher, fire hydrant, fire sprinkler, firehose, fireplace, fireplace screen, firewood, fish tank, flag, flock, floor, flower pot, flowers, fog, food, footboard, footrest, forest, fork, fountain, frame, fruit, fruit bowl, fruit stand, frying pan, garage, garage door, garden, garden, garland, garlic, garment bag, gas pump, gas station, gate, glass, glasses, globe, goblet, goose, grandfather clock, grapes, grass, greenhouse, grill, ground, hair brush, hall, hanger, harbor, hat, hay bale, headboard, headphones, hearth, heater, hedge, hill, horse, house, instrument, intercom, ipod, iron, island, jar, jetski, jetty, jewelry box, jug, kettle, keyboard, kiosk, kite, knife, knife set, knob, ladder, ladle, lake, laminating machine, lamp, lamp shade, land, laptop, lava, leaves, ledge, lemon, lettuce, lichen, light, light switch, lighter, lighthouse, lodge, loft, lotion, machine, magazine, magazine rack, magnifying glass, mailbox, make up case, manhole, mantle, map, markers, massage table, massager, mast, mat, mattress, median, microphone, microwave, mirror, mirror extender, molding, monitor stand, monkey, monolith, monument, moon, motorcycle, mountain, mountain pass, mouse, mouse pad, muffin, mug, napkin, news stand, newspaper, nightstand, notebook, nozzle, nutcracker, obelisk, object, ocean, onion, orange, ottoman, oven, overhead projector, painting, pallet, palm pilot, pan, panel, pants, paper, paper, roll, paper towels, paper towels holder, paper weight, park, parking lot, parking meter, parking place, pasta, path, pear, pedestal, pen, pen holder, pen set, pencil, pencil cup, pepper shaker, person, pestle and mortar, pet bowl, photocopier, piano, picture, pie, pillow, pipe, pitcher, placard, placemat, plank, plant, plant pot, plaque, plate, plateau, platform, platter, plumbing, podium, pole, pond, porch, portfolio, post it note, poster, pot, pot holder, pot rack, power cord, printer, projection screen, projector, puddle, pulley, purse, pylon, quay, quilt, rack, radiator, radio, railing, railroad track, railway, rainbow, recycling bins, reflection, refrigerator, remote control, river, river bank, road, rock, rolling pin, roof, room label, roots, rubber duckie, rug, ruin, salt shaker, sand satellite dish, scaffolding, scale, scanner, sconce, scrub brush, seagull, sealion, seashell, sewer, shadow, shampoo, sheep, shelf, shelves, shoe, shop window, shower, shower curtain, shower door, shower head, shutters, sideboard, sidewalk, sidewalk café, sign, silverware, sink, site office, sky, sky light, skyscraper, sled, sliding door, slippers, slope, smoke detector, snow, soap, soap dish, soap dispenser, soap on a rope, sofa, soil, space heater, spatulas, speaker, splash guard, sponge, spoon, spotlight, spray bottle,

sprinkler, stained glass window, staircase, stake, stand, stapler, starfish, station, statue, step, stereo, stick, sticker, stool, stove, stove hood, stove top, street market, streetcar, streetlight, string, stump of tree, subway, suitcase, sun, supermarket, table, table runner, tablecloth, tank, tapestry, teapot, teddy bear, telephone, telephone booth, television, television stand, terrace, thermos, thermostat, tile, tire, tissue, toast, toaster, toilet, toilet brush, toilet brush stand, toilet paper, toilet paper holder, toll gate, tomatoes, toolbox, toothbrush, toothbrush holder, toothpaste, torch, towel, towel rack, tower, toy, tractor, traffic light, trash, trash can, trash compactor, tray, tray table, tree, trophy, truck, trumpet, trunk, tunnel, typewriter, umbrella, urinal, urn, utensils, utensils rack, vacuum, valley, van, vanity, vase, vault, VCR, vegetable, vent, verge, vineyard, volcano, wall, wall hanging, wall mount, wardrobe, warehouse, washing machine, watch, water, water bottle, water buffalo, water cooler, water fountain, waterfall, watering can, watering can, water cooler, water fountain, waterfall, watering can, watermelon, wave, weathervane, webcam, wheel, wheelbarrow, wheelchair, white out, whiteboard, wind chime, windex, windmill, window, window frame, window ledge, window seat, window shade, window sill, wine bottle, wine glass, wine rack, wire, wood, wood post, zebra.

#### APPENDIX C: LABELS THAT WERE DELETED (SIC)

Bottom of chair, box contents, chair leg, chair wheel, eye, hold back, junk, lamp part, misc., mouth, nothing, occluded, precipice, side, spinning, supplies, table leg, terotaukuo, viewpoint, xx, xxx.

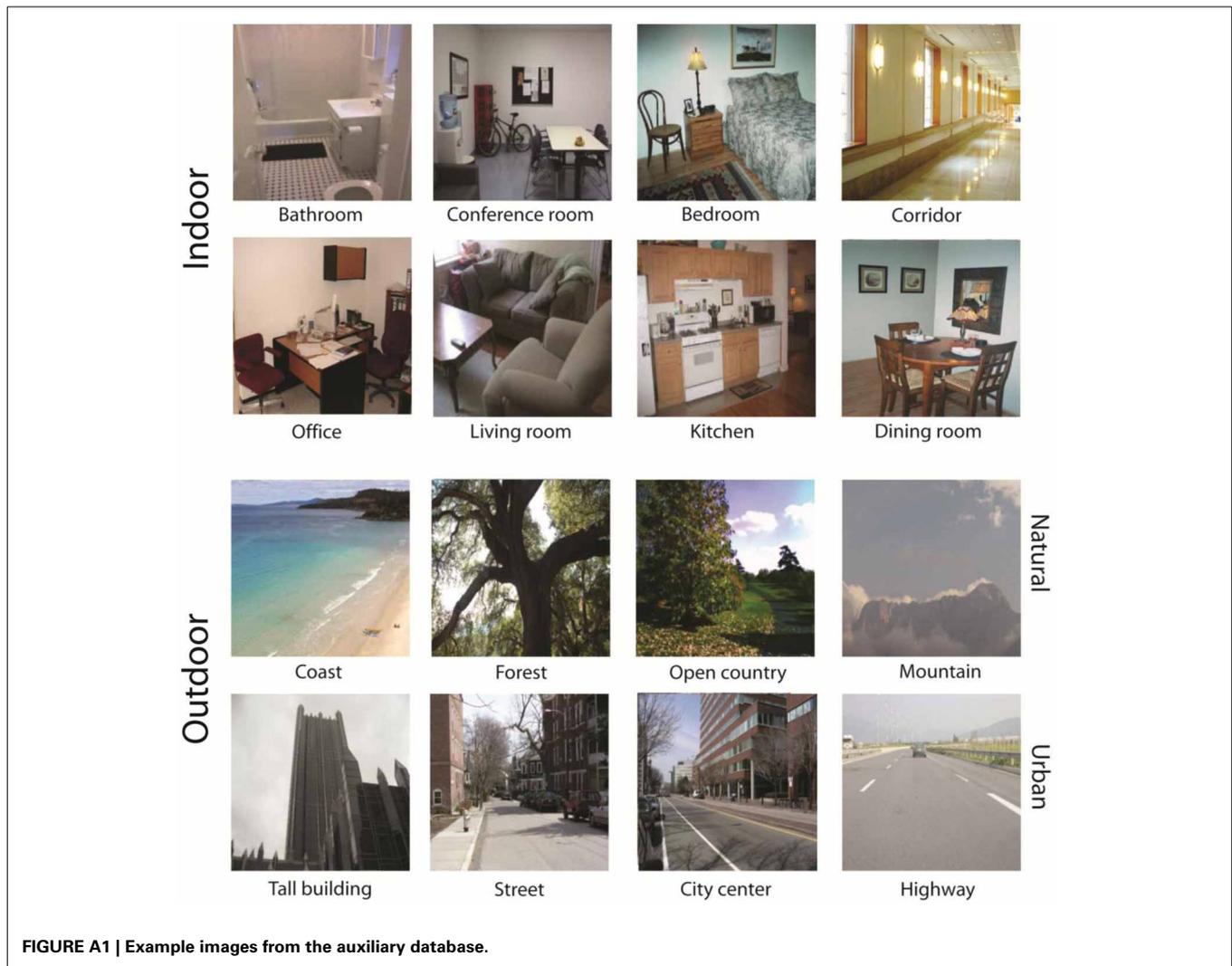
#### APPENDIX D: AUXILIARY DATA SET

##### (a) Object density

Unique object density for the auxiliary set was highly correlated with the main data set ( $r = 0.82$ ). While natural landscapes in the auxiliary set had fewer objects when compared to urban scenes [ $t_{(6)} = 3.46, p < 0.05$ ], there was no reliable difference in object density in this database between indoor and outdoor environments [ $t_{(14)} < 1$ ], as urban scene images in this database were more complex than urban images in the main set.

##### (b) Classification using ensemble statistics

For the auxiliary set, images could be classified at the superordinate level with 85% accuracy (AUC = 0.86). Likewise, basic-level categorization could be done on this data set with 33% accuracy (AUC = 0.61). Although the categorization performance is significantly above the chance level of 6.3%, it is lower than the 61% achieved by the main database ( $Z = 13.8, p < 0.0001$ ). To what extent is this lower performance due to the smaller size of the database? I sampled 5000 sets of 1220 images from the main database and tested classification performance on these sets, and found a median basic-level classification accuracy of 51% (95% confidence interval: 48–54%). Thus, although database size explains some of the performance difference, the higher classification performance on the main dataset may reflect homogeneity in the images that might not be reflected in the real world. This does not diminish from the main point of this analysis: that simple ensemble statistics of objects in scenes provides above-chance information about scene categories.



- (c) Ten most frequent objects in auxiliary set (In descending order of frequency). Wall, sky, floor, window, tree, road, building, lamp, door, table.
- (d) Overlap between frequent objects in main database vs. auxiliary database  
How did the object frequencies measured in the main database compare to the object frequencies in the auxiliary set? For each basic-level scene category, I computed the overlap between the 10 most frequent objects in both databases. Overall, there was 69% overlap, ranging from 50% (*conference room*) to 90% (*bathroom*).
- (e) Object diagnosticity  
To compare object diagnosticity between the main and auxiliary datasets, I computed the overlap of the 10 most diagnostic objects in both sets. 34% of the 10 most diagnostic objects in the main set were in the 10 most diagnostic objects in the auxiliary set.  
The 10 most diagnostic objects in each basic-level category: (In descending order of diagnosticity). *Bathroom*: sink, shower, shower door, soap, toilet lid, toilet paper, towel rack,

toilet, towel, bath. *Bedroom*: bed, nightstand, dresser, pillow, painting, telephone, closet, fan, clock, picture. *Conference room*: board, object, laptop, bottle, ground, bench, bicycle, speaker, chair, table. *Corridor*: entryway, light, light switch, bench, door handle, door, poster, ceiling, trash can, umbrella. *Dining room*: plate, dish, flower, vase, fireplace, tray, radiator, armchair, knife, bowl. *Kitchen*: cutlery, dish towel, stove hood, oven, coffee maker, stove, dishwasher, refrigerator, microwave, blender. *Living room*: sofa, fireplace, magazine, television stand, television, pillow, carpet, armchair, furniture, vase. *Office*: mousepad, keyboard, computer monitor, computer, desk, mug, mouse, pen, speaker, can. *Tallbuilding*: skyscraper, river, bus, building, statue, roof, fire hydrant, streetlight, pole, chimney. *City*: umbrella, house, balcony, parking meter, street, motorcycle, traffic light, wheel, van, awning. *Street*: hedge, wire, manhole, headlight, shutters, windshield, tail light, truck, van, license plate. *Highway*: bridge, median, tail light, fence, grass, road, cloud, truck, car, skyscraper. *Coast*: beach, ocean, sand, boat, cliff, rock, water, sky, cloud, mountain. *Open country*: field, hill, grass, cloud,

bush, bison, house, sky, mountain, snow. *Mountain*: bison, snow, mountain, forest, hill, cloud, water, house, rock, sky. *Forest*: path, tree trunk, forest, foliage, river, snow, ground, tree, rock, grass.

The smaller degree of overlap for diagnosticity, compared to frequency (section Object Frequency) or mutual information (section Mutual Information) is likely due to the fact that the auxiliary set is much smaller, so there are fewer objects with at least 10 instances in this dataset.

- (f) Scene-object specificity  
Forty six percent of objects were found in only one category, and of these, 43% had more than one instance in the database. A total of 19 objects (3.6%) were found in nine or more scene categories. Comparing these 19 objects to the 19 objects from the main database found in nine or more categories, 13 objects were found in both. These are: “box,” “chair,” “clock,” “door,” “lamp,” “light,” “person,” “plant,” “poster,” “table,” “tree,” “wall,” “window.”
- (g) Number of unique combinations  
In the 1220 image auxiliary set, 1108 images had unique object combinations (9% redundancy).
- (h) Entropy

The entropy of the auxiliary set was 6.15 bits per object.

- (i) Mutual information  
Fifty four percent of the 10 most informative objects for each scene category were shared between the main database and the auxiliary set, and six of the 10 objects with the highest overall mutual information for all scene categories were shared between the two data sets.
- (j) Classification using bag of words model  
Ninety seven percent accuracy ( $AUC = 0.97$ ) at superordinate-level categorization (not significantly different from main set performance:  $Z = 1.2, p = 0.23$ ) and 80.2% basic-level categorization accuracy ( $AUC=0.88$ ). This was significantly lower performance compared to the main dataset ( $Z = 14, p < 0.001$ ). As the two data sets contained different objects, it is not possible to train on one and test on the other as in the previous section.
- (k) Classification using structural model  
The auxiliary set achieved 77.5% correct at basic-level categorization (significantly lower than main dataset:  $Z = 13, p < 0.001$ ) and 97% correct at superordinate-level categorization (similar to performance on main data set:  $Z < 1$ ).