Published in final edited form as:
Cogn Psychol. 2009 March ; 58(2): 137–176. doi:10.1016/j.cogpsych.2008.06.001.

Recognition of natural scenes from global properties: Seeing the forest without representing the trees

Michelle R. Greene[a] and **Aude Oliva**[a]

[a]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue 46-4078, Cambridge, MA 02139, USA

Abstract

Human observers are able to rapidly and accurately categorize natural scenes, but the representation mediating this feat is still unknown. Here we propose a framework of rapid scene categorization that does not segment a scene into objects and instead uses a vocabulary of global, ecological properties that describe spatial and functional aspects of scene space (such as navigability or mean depth). In Experiment 1, we obtained ground truth rankings on global properties for use in Experiments 2–4. To what extent do human observers use global property information when rapidly categorizing natural scenes? In Experiment 2, we found that global property resemblance was a strong predictor of both false alarm rates and reaction times in a rapid scene categorization experiment. To what extent is global property information alone a sufficient predictor of rapid natural scene categorization? In Experiment 3, we found that the performance of a classifier representing only these properties is indistinguishable from human performance in a rapid scene categorization task in terms of both accuracy and false alarms. To what extent is this high predictability unique to a global property representation? In Experiment 4, we compared two models that represent scene object information to human categorization performance and found that these models had lower fidelity at representing the patterns of performance than the global property model. These results provide support for the hypothesis that rapid categorization of natural scenes may not be mediated primarily though objects and parts, but also through global properties of structure and affordance.

1. Introduction

One of the greatest mysteries of vision is the remarkable ability of the human brain to understand novel scenes, places and events rapidly and effortlessly ([Biederman, 1972], [Potter, 1975] and [Thorpe et al., 1996]). Given the ease with which we do this, a central issue in visual cognition is determining the nature of the representation that allows this rapid recognition to take place. Here, we provide the first behavioral evidence that rapid recognition of real-world natural scenes can be predicted from a collection of holistic descriptors of scene structure and function (such as its degree of openness or its potential for navigation), and suggests the possibility that the initial scene representation can be based on such global properties, and not necessarily the objects it contains.

1.1. Rapid basic-level scene categorization

Human observers are able to understand the meaning of a novel image if given only a single fixation (Potter, 1975). During the course of this glance, we perceive and infer a rich collection of information, from surface qualities such as color and texture ([Oliva and Schyns, 2000] and [Rousselet et al., 2005]); objects ([Biederman et al., 1982], [Fei-Fei et al., 2007], [Friedman, 1979], [Rensink, 2000] and [Wolfe, 1998]), and spatial layout ([Biederman et al., 1974], [Oliva and Torralba, 2001], [Sanocki, 2003] and [Schyns and Oliva, 1994]), to functional and conceptual properties of scene space and volume (e.g. wayfinding, [Greene and Oliva, 2006] and [Kaplan, 1992]; emotional valence, Maljkovic & Martini, 2005).

Indeed, from a short conceptual scene description such as “birthday party”, observers are able to detect the presence of an image matching that description when it is embedded in a rapid serial visual presentation (RSVP) stream and viewed for 100 ms ([Potter, 1975] and [Potter et al., 2004]). This short description is also known as the basic-level category for a visual scene ([Rosch, 1978] and [Tversky and Hemenway, 1983]), and refers to the most common label used to describe a place.

The seminal categorization work of Eleanor Rosch and colleagues has shown that human observers prefer to use the basic-level to describe objects, and exhibit shorter reaction times to name objects at the basic-level rather than at subordinant or superordinant (Rosch, 1978). It is hypothesized that the basic-level of categorization is privileged because it maximizes both within-category similarity and between-category variance ([Gosselin and Schyns, 2001] and [Rosch, 1978]). In the domain of visual scenes, members of the same basic-level category tend to have similar spatial structures and afford similar motor actions (Tversky & Hemenway, 1983). For instance, most typical environments categorized as “forests” will represent enclosed places where the observer is surrounded by trees and other foliage. An image of the same place from very close-up might be called “bark” or “moss”, and from very far away might be called “mountain” or “countryside”. Furthermore, the characteristic spatial layout of a scene constrains the actions that can be taken in the space ([Gibson, 1979] and [Tversky and Hemenway, 1983]). A “forest” affords a limited amount of walking, while a “countryside” might afford more options for navigation because the space is open. Although such functional and structural properties are inherent to scene meaning, their role in scene recognition has not yet been addressed.

1.2. The object-centered approach to high-level visual recognition

Many influential models of high-level visual recognition are object-centered, treating objects and parts as the atoms of scene analysis ([Biederman, 1987], [Biederman et al., 1988], [Bülthoff et al., 1995], [Fergus et al., 2003], [Marr, 1982], [Pylyshyn, 1999], [Riesenhuber and Poggio, 1999] and [Ullman, 1999]). In this view, the meaning of a real-world scene emerges from the identities of a set of objects contained within it, learned through the experience of object co-occurrence and spatial arrangement ([Biederman, 1981], [Biederman, 1987], [De Graef et al., 1990] and [Friedman, 1979]). Alternatively, the identification of one or more prominent objects may be sufficient to activate a schema of the scene, and thus facilitate recognition ([Biederman, 1981] and [Friedman, 1979]).

Although the object-centered approach has been the keystone of formal and computational approaches to scene understanding for the past 30 years, research in visual cognition has posed challenges to this view, particularly when it comes to explaining the early stages of visual processing and our ability to recognize novel scenes in a single glance. Under impoverished viewing conditions such as low spatial resolution ([Oliva and Schyns, 1997], [Oliva and Schyns, 2000], [Schyns and Oliva, 1994] and [Torralba et al., in press]); or when only sparse contours are kept, ([Biederman, 1981], [Biederman et al., 1982], [De Graef et al., 1990], [Friedman, 1979], [Hollingworth and Henderson, 1998] and [Palmer, 1975]) human observers are still able to recognize a scene's basic-level category. With these stimuli, object identity information is so degraded that it cannot be recovered locally. These results suggest that scene identity information may be obtained before a more detailed analysis of the objects is complete. Furthermore, research using change blindness paradigms demonstrates that observers are relatively insensitive to detecting changes to local objects and regions in a scene under conditions where the meaning of the scene remains constant ([Henderson and Hollingworth, 2003], [Rensink et al., 1997] and [Simons, 2000]). Last, it is not yet known whether objects that can be identified in a briefly presented scene are perceived, or inferred through the perception of other co-occurring visual information such as low-level features (Oliva &

Torralba, 2001), topological invariants (Chen, 2005) or texture information (Walker Renninger & Malik, 2004).

1.3. A scene-centered approach to high-level visual recognition

An alternative account of scene analysis is a scene-centered approach that treats the entire scene as the atom of high-level recognition. Within this framework, the initial visual representation constructed by the visual system is at the level of the whole scene and not segmented objects, treating each scene as if it has a unique shape (Oliva & Torralba, 2001). Instead of local geometric and part-based visual primitives, this framework posits that global properties reflecting scene structure, layout and function could act as primitives for scene categorization.

Formal work ([Oliva and Torralba, 2001], [Oliva and Torralba, 2002], [Torralba and Oliva, 2002] and [Torralba and Oliva, 2003]) has shown that scenes that share the same basic-level category membership tend to have a similar spatial layout. For example, a corridor is a long, narrow space with a great deal of perspective while a forest is a place with dense texture throughout. Recent modeling work has shown success in identifying complex real-world scenes at both superordinant and basic-levels from relatively low-level features (such as orientation, texture and color), or more complex spatial layout properties such as texture, mean depth and perspective, without the need for first identifying component objects ([Fei-Fei and Perona, 2005], [Oliva and Torralba, 2001], [Oliva and Torralba, 2002], [Oliva and Torralba, 2006], [Torralba and Oliva, 2002], [Torralba and Oliva, 2003], [Vogel and Schiele, 2007] and [Walker Renninger and Malik, 2004]). However, the extent to which human observers use such global features in recognizing scenes is not yet known.

A scene-centered approach involves both global and holistic processing. Processing is global if it builds a representation that is sensitive to the overall layout and structure of a visual scene ([Kimchi, 1992] and [Navon, 1977]). The influential global precedence effect (Navon, 1977, see Kimchi, 1992 for a review) showed that observers were more sensitive to the global shape of hierarchical letter stimuli than their component letters. Interestingly, the global precedence effect is particularly strong for stimuli consisting of many-element patterns (Kimchi, 1998) as is the case in most real-world scenes. A consequence of global processing is the ability to rapidly and accurately extract simple statistics, or summary information, from displays. For example, the mean size of elements in a set is accurately and automatically perceived ([Ariely, 2001], [Chong and Treisman, 2003] and [Chong and Treisman, 2005]), as is the average orientation of peripheral elements (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001); some contrast texture descriptors (Chubb, Nam, Bindman, & Sperling, 2007) as well as the center of mass of a group of objects (Alvarez & Oliva, 2008). Global representations may also be implicitly learned, as observers are able to implicitly use learned global layouts to facilitate visual search ([Chun and Jiang, 1998] and [Torralba et al., 2006]).

While all of these results highlight the importance of global structure and relations, an operational definition of globality for the analysis of real-world scenes has been missing. Many properties of natural environment could be global and holistic in nature. For example, determining the level of clutter of a room, or perceiving the overall symmetry of the space are holistic decisions in that they cannot be taken from local analysis only, but require relational analysis of multiple regions (Kimchi, 1992).

Object and scene-centered computations are likely to be complementary operations that give rise to the perceived richness of scene identity by the end of a glance (200–300 ms). Clearly, as objects are often the entities that are acted on within the scene, their identities are central to scene understanding. However, some studies have indicated that the processing of local object information may require more image exposure (Gordon, 2004) than that needed to identify the scene category ([Oliva and Schyns, 2000], [Potter, 1975] and [Schyns and Oliva, 1994]). In

this study, we examine the extent to which a global scene-centered approach can explain and predict the early stage of human rapid scene categorization performance. Beyond the principle of recognizing the "forest before the trees" (Navon, 1977), this work seeks to operationalize the notion of "globality" for rapid scene categorization, and to provide a novel account of how human observers could identify the place as a "forest", without first having to recognize the "trees".

1.4. Global properties as scene primitives

We propose a set of global properties that tap into different semantic levels of global scene description. Loosely following Gibson (1979) important descriptors of natural environments come from the scene's surface structures and the change of these structures with time (or constancy). These aspects directly govern the possible actions, or affordances of the place. The global properties were therefore chosen to capture information from these three levels of scene surface description, namely structure, constancy and function.

A total of seven properties were chosen for the current study to reflect aspects of scene structure (mean depth, openness and expansion), scene constancy (transience and temperature), and scene function (concealment and navigability). A full description of each property is found in Table 1. These properties were chosen on the basis of literature review (see below) and a pilot scene description study (see Appendix A.1) with the requirement that they reflect as much variation in natural landscape categories as possible while tapping into different levels of scene description in terms of structure, constancy and function. Critically, the set of global properties listed here is not meant to be exhaustive,1 as other properties such as naturalness or roughness (the grain of texture and number and variety of surfaces in the scene) have been shown to be important descriptors of scene content (Oliva & Torralba, 2001). Rather, the goal here is to capture some of the variance in how real-world scenes vary in structure, constancy and function, and to test the extent to which this information is involved in the representation of natural scenes.

1.4.1. Properties of scene structure—Previous computational work has shown that basic-level natural scene categories tend to have a particular spatial structure (or spatial envelope) that is well-captured in the properties of mean depth, openness and expansion ([Oliva and Torralba, 2001] and [Torralba and Oliva, 2002]). In brief, the global property of mean depth corresponds to the scale or size of the space the scene subtends, ranging from a close-up view to panoramic environment. The degree of openness represents the magnitude of spatial enclosure whereas the degree of expansion refers to the perspective of the spatial layout of the scene. Images with similar magnitudes along these properties tend to belong to the same basic-level category: for example, a "path through a forest" scene may be represented using these properties as "an enclosed environment with moderate depth and considerable perspective". Furthermore, these spatial properties may be computed directly from the image using relatively low-level image features (Oliva & Torralba, 2001).

1.4.2. Properties of scene constancy—The degree of scene constancy is an essential attribute of natural surfaces ([Cutting, 2002] and [Gibson, 1979]). Global properties of constancy describe how much and how fast the scene surfaces are changing with time. Here, we evaluated the role of two properties of scene constancy: transience and temperature.

Transience describes the rate at which scene surface changes occur, or alternatively stated, the probability of surface change from one glance to the next. Places with the highest transience would show actual movement such as a storm, or a rushing waterfall. The lowest transience places would change only in geologic time, such as a barren cliff. Although the perception of transience would be more naturalistically studied in a movie rather than a static image, humans

can easily detect implied motion from static images ([Cutting, 2002] and [Freyd, 1983]), and indeed this implied motion activates the same brain regions as continuous motion (Kourtzi & Kanwisher, 2000). Temperature reflects the differences in visual appearance of a place during the changes of daytime and season, ranging from the intense daytime heat of a desert, to a frigid snowy mountain.

1.4.3. Properties of scene function—The structure of scene surfaces and their change over time governs the sorts of actions that a person can execute in an environment (Gibson, 1979). The global properties of navigability and concealment directly measure two types of human–environment interactions deemed to be important to natural scene perception from previous work ([Appelton, 1975], [Gibson, 1958], [Gibson, 1979], [Kaplan, 1992] and [Warren et al., 2001]). Insofar as human perception evolved for goal-directed action in the environment, the rapid visual estimation of possible safe paths through an environment was critical to survival (Gibson, 1958). Likewise, being able to guide search for items camouflaged by the environment (Merilaita, 2003), or to be able to be concealed oneself in the environment (Ramachandran, Tyler, Gregory, Rogers-Ramachandran, Duessing, Pillsbury & Ramachandran, 1996) have high survival value.

1.5. Research questions

The goal of this study is to evaluate the extent to which a global scene-centered representation is predictive of human performance in rapid natural scene categorization. In particular, we sought to investigate the following questions: (1) are global properties utilized by human observers to perform rapid basic-level scene categorization? (2) Is the information from global properties sufficient for the basic-level categorization of natural scenes? (3) How does the predictive power of a global property representation compare to an object-centered one?

In a series of four behavioral and modeling experiments, we test the hypothesis that rapid human basic-level scene categorization can be built from the conjunctive detection of global properties. After obtaining normative ranking data on seven global properties for a large database of natural images (Experiment 1), we test the use of this global information by humans for rapid scene categorization (Experiment 2). Then, using a classifier (Experiment 3), we show that global properties are computationally sufficient to predict human performance in rapid scene categorization. Importantly, we show that the nature of the false alarms made by the classifier when categorizing novel natural scenes is statistically indistinguishable from human false alarms, and that both human observers and the classifier perform similarly under conditions of limited global property information. Critically, in Experiment 4 we compare the global property classifier to two models trained on a local region-based scene representation and observed that the global property classifier has a better fidelity in representing the patterns of performance made by human observers in a rapid categorization task.

Although strict causality between global properties and basic-level scene categorization cannot be provided here, the predictive power of the global property information and the convergence of many separate analyses with both human observers and models support the hypothesis that an initial scene representation may contain considerable global information of scene structure, constancy and function.

2. General method

2.1. Observers

Observers in all experiments were 18–35 years old, with normal or corrected-to-normal vision. All gave informed consent and were given monetary compensation of $10/h.

2.2. Materials

Eight basic-level categories of scenes were chosen to represent a variety of common natural outdoor environments: desert, field, forest, lake, mountain, ocean, river and waterfall. The authors amassed a database of exemplars in these categories from a larger laboratory database of 22,000 (256 × 256 pixel) full-color photographs collected from the web, commercial databases, personal digital images and scanned from books ([Oliva and Torralba, 2001] and [Oliva and Torralba, 2006]). From this large database, we selected 500 images2 chosen to reflect natural environmental variability. To estimate the typicality of each image, independent, naïve observers ranked each of the 500 images on its prototypically for each scene category, using a 1–5 scale (see Appendix A.3 for a description of the typicality norming task). The most prototypical 25 images for each of the eight basic-level category were kept, for a grand total of 200 images which were used in Experiment 1–4 (see details in Appendix A.3). The remaining 300 poly-categorical images were used in Experiment 3, Section 5.2.6. For human psychophysics experiments, we used Matlab and the Psychophysics Toolbox as presentation software ([Brainard, 1997] and [Pelli, 1997]).

3. Experiment 1: Normative rankings of global properties on natural scenes

First, we obtained normative rankings on the 500 natural scenes along the seven global properties. These normative rankings provide a description of each image and basic-level category in terms of their global structural, constancy and functional properties. Namely, each image is described by seven components, each component representing the magnitude along each global property dimension (see examples in Fig. A2 in Appendix A.2).

As Experiments 2–4 involve scene categorization using global property information, robust rankings are essential for selecting images for the human psychophysics in Experiment 2 as well as for training and testing the classifier used in Experiment 3.

3.1. Method

3.1.1. Participants—Fifty-five observers (25 males) ranked the database along at least one global property, and each property was ranked by at least 10 observers.

3.1.2. Procedure—Images were ranked using a hierarchical grouping procedure (Fig. 1, Oliva & Torralba, 2001). This allows the ranking of a large number of images at once, in the context of one another.

For a given global property, each participant ranked two sets of 100 images. The two halves of the database were pre-chosen by the authors to contain roughly equal numbers of images in each semantic category. One hundred picture thumbnails appeared on an Apple 30″ monitor (size of 1.5 × 1.5 deg/thumbnail). The interface allowed participants to drag and drop images around the screen and to view a larger version of the image by double-clicking on the thumbnail.

Participants were given the name and description of a global property at the start of a ranking trial. They were instructed to divide the images into two groups based on a specific global property such that images with a high magnitude along the global property were placed on the right-hand side of the screen while images with a low magnitude were placed on the left (see Fig. 1). In a second step, participants were asked to split each of the two groups into two finer divisions, creating four groups of images that range in magnitudes along the specified global property. Finally, the four groups were split again to form a total of eight groups, ordered from the lowest to the highest magnitude for a given property. At any point during the trial, participants were allowed to move an image to a different subgroup to refine the ranking. Participants repeated this hierarchical sorting process on the remaining 100 pictures in database along the specified global property. Participants had unlimited time to complete the task, but

on average completed a trial in 30 min. As the task was time consuming, not all participants ranked all seven properties, and we are reporting results from 10 observers per property, normalized to fit in the range of 0–1.

3.2. Results

3.2.1. General description—Examples of images that were ranked as low, medium and high for each global property are shown in Fig. 2. Global properties are continuous perceptual dimensions, and therefore image ranks spanned the range of possible values across the database (Scattergrams of rankings by category for all global properties can be seen in Fig. A1, Appendix A.2). It is essential to note in Fig. A1 the high scatter of rankings indicates that the basic-level category label is not the determinant of the global property ranking for any single global property. In other words, concealment is not just another way of saying forestness.

In order to compare the time-unlimited rankings of Experiment 1 to the speeded categorization task of Experiment 2, it is necessary to know that global properties can be rapidly and accurately perceived by human observers. Furthermore, a similar ranking of images along global properties when presentation time is limited ensures that the rankings of Experiment 1 are not due to inferences based on the scene schema. To this end, we ran a control speeded classification task3 (see the description of this experiment in Appendix A.4). Results showed that indeed, global properties could be estimated from limited presentation time. The mean correlation of the speeded classification to the hierarchical rankings was 0.82, ranging from 0.70 for concealment to 0.96 for temperature (all significant), see Appendix A.4 for more details.

3.2.2. Between-observer consistency in ranking images along each global property—The extent to which global properties represent a reasonable basis for natural scene recognition depends critically on the extent to which the global properties can be ranked consistently by human observers.

Here we are using the 200 prototypical images as they give strong ground truth for the purpose of categorization in Experiments 2–4. We computed observers' consistency as a Spearman's rank-order correlation for each possible pairing of observers for all seven global properties. The mean and standard error for these correlation coefficients by global property are shown in Table 2. Between-observer Spearman's rank-order correlations ranged from 0.61 (transience) to 0.83 (openness), and were all statistically significant ($p < 0.01$). This indicates that different observers estimated the degree of these global properties in similar ways (see also [Oliva and Torralba, 2001] and [Vogel and Schiele, 2007] for similar results) and agreed well on which images represented a high, medium and low magnitude for a given global property.

3.2.3. Global property descriptions of semantic categories—The subsequent experiments test the utility of a global property representation for rapid scene categorization. In this representation, images are represented as points in a seven-dimensional space where each axis corresponds to a global property. How are different basic-level categories described in this space?

To visualize the global property signature of each semantic category, we computed the category means and ranking spread for each global property. Fig. 3 shows box-and-whisker plots for the global property rankings for each category, creating a conceptual signature of the category. For example, most deserts were ranked as hot, open and highly navigable environments, with a low magnitude along the transience and concealment dimensions while most waterfalls are closed, highly transient environments that are less navigable. Other categories, such as lakes, have global property ranking averages that were intermediate along each dimension, meaning that most lakes have a medium level of openness and expansion, are neither environments perceived as very cold or very warm, and so on.

Euclidean distance measures between each pair of basic-level categories provided a conceptual distance metric between basic-level categories (see Table A4 and details in Appendix A.6). As expected from intuition, categories like waterfall and river are close to each other, but categories like field and waterfall are very distant.

3.3. Discussion

Here, we have shown that observers can provide normative rankings on global properties with a high degree of consistency. We have also provided a conceptual description of basic-level category prototypes as the mean global property rankings of a category.

To what extent do the scene-centered semantic category descriptions shown in Fig. 3 contribute to human observers' mental representations of scene identity? We test this explicitly in Experiment 2.

4. Experiment 2: Human use of global properties in a rapid scene categorization task

The goal of Experiment 2 was to test the extent to which global property information in natural scenes is utilized by human observers to perform rapid basic-level scene categorization. A global property-based scene representation makes the prediction that scenes from different semantic categories but with similar rankings along a global property (e.g. oceans and fields are both open environments) will be more often confused with each other in a rapid categorization task than scenes that are not similar along a global property (e.g. an open ocean view and a closed waterfall). We tested this hypothesis systematically by recording the false alarm rates for each basic-level category (serving as targets in blocked yes–no forced choice task) when viewed among distractor images that all shared a particular global property pole (such as high concealment or low openness).

4.1. Method

4.1.1. Participants—For a purpose of completeness and replication of our effects, two groups of participants participated in Experiment 2. First, four participants (1 male) completed the entire experimental design. Throughout the experiment, we will refer to this group as the complete-observer group. While having all observers complete all blocks of the experiment is statistically more robust, it could also lead to over-learning of the target images. To eliminate the learning effect, a meta-observer group consisting of 73 individuals (41 male) completed at least 8 blocks (400 trials) of the design, for a total of eight meta-observers (see Appendix A.5 for details on the analysis of meta-observer data). Meta-observer analysis is justified here because the critical analyses are on the image items.

4.1.2. Design—The experimental design consisted of a full matrix of target-distractor blocks where each basic-level category was to be detected amongst distractor images from different semantic categories that shared a similar ranking along one global property. Both high and low magnitudes of each global property were used, yielding 112 blocked conditions (8 target categories × 7 global properties × 2 magnitudes). For example, a block would consist of one semantic category (such as forest) seen among images that were all ranked in Experiment 1 as (for example) high transience. The distractor sets were chosen to reflect a wide variety of semantic categories, and to vary in other global properties while keeping ranks in the manipulated property constant. Therefore, as best as possible, global properties were independently manipulated in this design. Distractor sets for a given global property magnitude were therefore chosen uniquely for each category. High and low rankings were defined as imaged ranked as >0.6 and <0.3 for a given global property.

4.1.3. Procedure—Each of the 112 experimental blocks contained 25 target and 25 distractor images. At the start of each block, participants were given the name of the target category and were instructed to respond as quickly and accurately as possible with a key press ("1" for yes, "0" for no) as to whether each image belonged to the target category. Each trial started with a 250 ms fixation cross followed by an image displayed for 30 ms, immediately followed by a 1/f noise mask presented for 80 ms. Visual feedback (the word "error") followed each incorrect trial for 300 ms.

4.2. Results

For all analyses, we report results for both the complete-observer and the meta-observer groups. Results from the two groups support each other well. In addition to providing a self-replication, examining individuals completing the entire design reduces the noise seen from pooling individual performances. On the other hand, the meta-observer group reduces the problem of over-learning the target images.

In the following, we report four different analyses on both correct detection (hit) and false alarms: 4.21—the general performance of human observers in rapid basic-level scene categorization; 4.22—the power of target–distractor global property resemblance in predicting which particular images will yield false alarms to a basic-level category target.; 4.23—the relation between false alarms made between basic-level categories and the relative distances of those categories in global property space; 4.24—the effect of global property similarity on reaction time.

4.2.1. Basic-level scene categorization: Overall performance—The complete-observers' average hit rate was 0.87 with a mean false alarm rate of 0.19. This level of performance corresponds to an average d′ sensitivity of 2.07. Performance by semantic category is detailed in Table 3. With this short 30 ms presentation time, observers could reliably detect all scene categories (all $d' > 1.0$). However, critical for subsequent analyses, observers made substantial false alarms to each category as well, giving a rich range of performance data to work with.

For the 8 meta-observers, the mean hit rate was 0.78, with a mean false alarm rate of 0.24. This corresponds to a d′ of 1.58. For the complete-observer group, we looked at hit rate across the 14 times they viewed the target images. For each observer, we performed a linear regression on the hit rate over these blocks and found that for 3 of the 4 subjects, there was a positive slope (mean – 0.095, just under 1% per block), indicating that there was learning of the targets over the course of the experiment.

4.2.2. The role of global properties on basic-level categorization performance—A prediction of the scene-centered approach is that distractor images that share a global property ranking with the target prototype should yield more false alarms than images that are less similar to the target prototype. A pictorial representation of sample results is shown in Fig. 4: forests, which tend to be closed (cf. Fig. 3) have more false alarms to closed distractors than to open distractors, and the opposite is true of fields, which tend to be open environments.

A global property-based scene representation would predict that any image's confusability to any target category could be predicted from this image's global property distance to the target category. For example, in general, mountain scenes were ranked as moderately-low navigability (cf. Fig. 3). Therefore, in a block where mountains were to be detected among low navigability distractors, we would expect more false alarms to distractors that are also moderately-low navigability than non-navigable distractors of greater magnitude (such as a very dense forest).

In each of the 112 experimental blocks, a single semantic category was to be detected among distractor images that had a common global property rank. For each distractor image in these blocks, we computed the one-dimensional distance between its global property rank on the manipulated global property to the mean global property rank of the target category for the same property. For example, in a block where deserts were viewed among low-expansion scenes, each distractor would be expressed as the distance between its rank on expansion (given from Experiment 1), and the mean desert rank for expansion (cf. Fig. 3).

Therefore, all of the distractor images in the entire experiment could be ranked from most similar to the target category to least. If global property information is used to help human observers estimate the image category, then global property resemblance should predict the false alarms that are made during the experiment.

To test, we first binned the false alarm data into quartiles based on ascending target–distractor distance. The mean percent correct rejections for each quartile for each data set are shown in Table 4. For both groups, the accuracies increase monotonically with distance, indicating that difficulty of image categorization is in part due to the resemblance of the distractors to the target category prototype. Human categorization performance is not obliterated by this one-dimensional similarity, however as even the most similar 1% of distractors are still classified significantly above chance by the meta-observers: 64% correct, $t(198) = 5.5$, $p < 0.0001$.

Performance suffers with decreasing distance to target prototype, but remains above chance.

We also performed a correlation on the distractor distance data, using the mean false alarm rate for each distractor to its distance from target prototype mean. For the complete-observer group, we found a striking relation with correlation coefficients ranging from 0.98 to 0.91, when binning the data, respectively, in 8 bins and 25 bins (for all correlations, $p < 0.0001$). For the meta-observers, correlations ranged from 0.95 for 8 bins, to 0.81 for 25 bins, all correlations were significant ($p < 0.001$).

This strong relation shows that images that resemble the category global property prototype are more often mistaken with the target category than other images, and suggests that with a short presentation time, global property information is used by human observers to categorize natural scenes into basic-level categories.

4.2.3. Distance in global property space predicts pairwise category false alarms

—Are some semantic categories confused with each other more often than others? Can such asymmetries be understood through a scene-centered global property representation? Ashby and Lee (1991) showed that false alarms increase with increasing similarity between targets and distractors in a detection task. Therefore, if our global properties are incorporated into the human scene representation, we would expect false alarms made between semantic categories in the rapid categorization task to follow from the categories' similarity in global property space (from Experiment 1, see Fig. 3).

As the experimental task was a yes–no forced choice within a block of uniform target categories, the false alarms made in a given block provide insight into which category observers believed the image to be. For example, a false alarm to a forest image while looking for river targets indicates that the observer believed the picture of the forest to be a river. False alarm rates between each pair of categories were thus computed (see Appendix A.6 for more details).

We then computed the Euclidean distance between each category in the global property space ($n * (n - 1)/2 = 28$ pairwise comparisons for the $n = 8$ categories). This is a dissimilarity metric: larger values indicate more differences between two categories (see Appendix A.5 for more details).

For the complete-observer group, we found a strong negative correlation between category dissimilarity and false alarm rates ($r = -0.76$, $p < 0.001$), indicating that pairs of categories that are similar in global property space (such as river and waterfall) are more often confused by human observers than pairs of categories that are more distant, such as field and waterfall. The same pattern held for the meta-observers: ($r = -0.78$, $p < 0.001$).

4.2.4. The reaction time effects of global property similarity—The previous analyses have shown that the probability of human observers mis-categorizing images given a brief presentation is strongly related to how similar a given distractor is to the target category in global property space. Is there evidence of global property similarity for the images that are correctly categorized? In particular, is the speed at which an image can be correctly categorized inversely related to how similar it is to the category prototype? One can imagine that a distractor sharing very few global properties with the target category might be more quickly rejected than a distractor that more closely resembles the target category.

For this analysis, we report data from the complete-observer group as individual differences in reaction time from the meta-observer group are confounded in the blocked design. For all correctly rejected distractors, we correlated the participants' reaction time to the Euclidean distance of that distractor to the target category in global property space. We found that there was a strong inverse relation between target–distractor resemblance and reaction time ($r = -0.82$, $p < 0.0001$), indicating that distractors that are more dissimilar to the target category are more quickly rejected than distractors that are more similar. In other words, similarity in global property space predicts the mistakes that human observers tend to make as well as which images will take longer to categorize.

4.3. Discussion

The previous analyses have shown that with a very brief image exposure, human observers are able to detect the basic-level category of a natural scene substantially above chance (Section 4.2.1; see also [Joubert et al., 2007], [Oliva and Schyns, 2000], [Potter, 1975] and [Rousselet et al., 2005]). However, participants' performances were far below ceiling, suggesting that the scene representation afforded by this amount of image exposure was incomplete, providing a rich array of false alarms that are useful for understanding the initial representation.

In this experiment, we have shown converging evidence from different analyses indicating that human observers are sensitive to global property information during very brief exposures to natural images, and that global information appears to inform basic-level categorization.

First, we have shown that the probability of false alarm to a given image can be very well predicted from the one-dimensional distance of this image's rank along a global property to the target category prototype for that same property (Section 4.2.2). We have also shown that semantic categories that are more often confused by human observers are more similar to one another in global property space (Section 4.2.3). As distractor images varied in semantic categories, other global properties and objects, this implies that global property information makes up a substantial part of the initial scene representation. Last, we have shown that the reaction times for correctly rejected distractors were also related to the distractors' resemblance to the target category (Section 4.2.4). Altogether, these results support a scene-centered view of scene understanding that asserts that spatial and functional global properties are potential primitives of scene recognition.

5. Experiment 3: The computational sufficiency of global properties for basic-level scene categorization

We have shown so far that global property information strongly modulates human performance in a rapid scene categorization task. To what extent is a global property representation sufficient to predict human rapid scene categorization performance? To answer this question, we built a conceptual naïve Bayes classifier whose only information about each scene image was from the normative ranking data of Experiment 1. Therefore, the classifier is agnosic to any other visual information (color, texture, objects) that human observers could have used to perform the task. Here, we compare the performance of this classifier (correct and false alarms) to the human scene categorization performance of Experiment 2.

5.1. Method

The training input to the classifier consisted of the ranks that each image received for each of the seven global properties along with a label indicating which semantic category the image belonged to. From this input, the classifier estimated Gaussian distributions for each category along each global property. Then, given a test image (not used in training), the classifier computed the most likely semantic category for the set of global properties given to it

$$C_{\text{est}} = \arg\max_{c \in C} \sum_{k=1}^{k} \ln \frac{1}{\sqrt{2\pi\sigma_{jk}^2}} - \frac{1}{2\sigma_{jk}^2}\left(x - \mu_{jk}\right)^2$$

where the log likelihood of each category j is estimated from the distributions of each property dimension k (for background, see Mitchell, 1997). For a discussion on the assumptions of such a classifier, see Appendix A.6.

The classifier was run 25 times, testing each image in turn using a leave-one-out design. In each run, 24 images from each semantic category (192 total) served as training, and the last eight (one from each category) were used for testing.

It is of note that the naïve Bayes classifier was chosen to be the simplest classifier for testing this global property representation. All reported results were also done with a linear discriminant analysis with no significant performance differences (see Appendix A.7).

5.2. Results

In comparing a classifier's performance to human performance for the goal of gaining insight into the human representation, it is necessary to examine classifier performance at several levels. Similar overall performance is not enough since any psychophysical task can be made arbitrarily harder or easier by changing presentation time, for example. The errors made by a classifier are more informative than the overall correct performance because similarities in errors make a stronger argument for a similar representation. Conversely, dissimilarities in the patterns of errors are informative in refining hypotheses. We report here four distinct types of analyses using data from Experiments 1, 2 and 3: Section 5.2.1—the overall performance of the classifier relative to human scene categorization performance from Experiment 2; Section 5.2.2—an examination of the types of classification errors made by both humans and classifier; Section 5.2.3—an examination of the distances between categories in our scene-centered space (Experiment 1) and how this predicts errors made by both classifier and human observers; and Sections Sections 5.2.4 and 5.2.5—a comparison of how the classifier and human observers perform under conditions where a complete global property representation cannot be used for scene categorization. As a last test of this model (Section 5.2.6), we compare the classifier's

responses to non-prototypical images to that of the human norming data of Experiment 1 (see Appendix A.3).

5.2.1. Classifier performance: Percent correct and correlation to human basic-level category performance—Overall, the performance of the classifier was remarkably similar to that of human meta-observers: the overall proportion correct for the classifier was 0.77 (0.77 for human meta-observers, $t(7) < 1$). The performance for the complete-observer group was higher (proportion correct was 0.86), in part because of the over-learning of the stimuli.

To get an idea of how classifier performance compared to human performance by basic-level category, we correlated meta-observer's correct performance and classifier correct performance and found a striking similarity: the by-category correlation was $r = 0.88$, $p < 0.01$ (see Fig. 5). This level of agreement did not differ from meta-observer agreements ($r = 0.78$: $t(7) = 1.72$, $p = 0.13$), indicating that the classifier's overall correct performance and correct performance by category were indistinguishable from human performances. Similarly, the correlation between the classifier and the mean correct performance of the complete-observer group was similarly high ($r = 0.75$, $p < 0.01$).

5.2.2. Error analysis: Easy and difficult images—Do human observers and the classifier have difficulty classifying the same images? We looked at the errors that both humans and classifier made in a by-image item analysis, comparing the probability of classifier failure (average performance in 4, 10 and 25 bins, due to the binary classification of the 200 images by the classifier) to human false alarm rates (over the same bins).

We found a significant correlation between the classifier and the meta-observers (for 10 bins, $r = 0.89$, $p < 0.0001$) indicating that indeed humans and classifier have trouble categorizing the same images. Bin size did not affect the nature of the result: using bin sizes of 4 and 25, the correlation coefficients were 0.97 and 0.76, respectively (all significant). Similarly, the correlation between the classifier and participants from the complete-observers design were all significant ($p < 0.001$, $r = 0.96$, $r = 0.81$, and $r = 0.64$ for the same bin sizes).

5.2.3. Qualitative error analysis: Distribution of error types—Next, we sought to determine the qualitative similarity of the false alarms made by both classifier and human observers. The yes–no forced choice task of the human observers allowed insights into which category observers believed an image to be given a false alarm, and this can be compared directly to the output of the classifier. In other words, in a block where the target image was river, and an observer made a false alarm to an image of a forest, does the classifier also call this forest a river?

Given an error made by the classifier, we found that at least one human observer in the meta-observer group made the same false alarm in 87% of the images (88% for the complete-observer group). However, human observers are also prone to finger errors, attentional lapses and other possible mistakes, so when we include only the false alarms that at least five of the eight meta-observers made; there was human-classifier correspondence on 66% of the images (59% for at least 3 of the 4 participants who completed the entire experiment).

Examples of the correct responses and the false alarms made by the classifier and human observers (meta-observer group) are shown in Fig. 6. This indicates that the scene categorization performance of a classifier knowing only about global property rankings is highly similar to that of human observers when given a 30 ms exposure to a scene image.

We have shown so far that the overall performance of the global property classifier as well as the types of errors it made is highly similar to the performance of human observers in a rapid scene categorization task. To further compare classifier to human performance, we created a category-by-category confusion matrix for the global property classifier (see false alarms Table A6 in Appendix A.6) and human observers (human matrix of false alarms from Experiment 2, see Table A5 in Appendix A.6). We found that the between-category confusions made by the classifier were highly correlated with those made by human observers ($r = 0.77$, $p < 0.0001$ for complete-observers and $r = 0.73$ for the meta-observers, $p < 0.0001$). It is of note that the diagonals of the confusion matrices (the correct detections) were taken out for both as it would have led to a spuriously high correlation. This analysis further suggests that a scene representation containing only global property information predicts rapid human scene categorization, a result which strengthens the hypothesis that a global scene-centered representation may be formed by human observers at the beginning of the glance.

5.2.4. "Knocking-out" a global property I: Missing properties—A stronger case for a global scene representation in human observers would be if the classifier and humans are similarly impaired under degraded conditions. We have shown so far that these global properties are sufficient to predict human performance in rapid scene categorization. From Experiment 2, we found that human observers are remarkably flexible in scene categorization under conditions where target–distractor similarity along a global property dimension decreases the utility of that dimension for categorization—performance suffers but remains above chance with such incomplete information. How does the classifier perform when similarly impaired? To test, we compared human false alarms in Experiment 2 to runs of the classifier trained with all global properties but one in turn. Experiment 2 "knocked-out" global properties for human observers by matching the target and distractors on that property, reducing the utility of the property for categorization. For example, assuming high transience is a diagnostic property of oceans, classifying oceans among high transience scene distractors will render transience useless for the task. Likewise, training the classifier without a property "knocks-out" that property because there is no representation of the property at all.

All training and testing procedures were identical to the previously presented method in Section 5.1 except that all images were represented by six global properties instead of the full set of seven, which served as a performance baseline. For the human comparison, for each global property we used the pole (high or low rank) that yielded the most false alarms. For each category, we compared these false alarm rates to the average performance of that category over all distractor conditions.

For each basic-level category we compare the increase in false alarms for the classifier to the increase in false alarms for human observers. Interestingly, "knocking-out" the use of a global property decreased performance to a similar degree: overall increase in false alarms by category was an average of 5.2% more for the classifier and 3.2% more for the complete-observer group (3.1% for meta-observers, difference between humans and model were not significant, $t(7) < 1$) indicating that the loss of global property information affected both human observers and the model to a similar degree, and that the classifier's representation was similarly robust to incomplete global property information. Furthermore, the correlation between classifier and human correct performance by category remains strong in this manipulation ($r = 0.81$, $p < 0.0001$ for the complete-observers, and $r = 0.83$ for meta-observers), indicating that the absence of each global property is similarly disruptive to categorization, and suggesting that both observer types are using similar diagnostic global property information to perform the categorization task. Again, the correlation existing between the classifier and mean human performance was not different from the agreement between meta-observers ($t(7) < 1$), indicating that the classifier's performance is indistinguishable from human observers.

5.2.5. "Knocking-out" a global property II: The role of all properties—What is the limit of the classifier's ability to deal with incomplete information and to what extent are all of the global properties necessary to predict human categorization performance? To address this question, we ran the classifier on exhaustive combinations of incomplete global property data, from one to six global properties.

The average performance of the classifier for each number of global properties used is shown in Fig. 7A. Interestingly, when the classifier is trained on only one of the global properties, categorization performance is still significantly above chance (30%, chance being 12.5%, $t(6) = 7.93$, $p < 0.0001$) and reaches a plateau when combinations of six global properties are used (74%).

Next, we looked at which combinations of global properties lead to maximum performance for all eight basic-level categories. We tabulated the average performance of global property combinations containing each global property. If the maximum classifier performance is carried by one or two properties, one would expect maximum performance when these properties are present and diminished performance with other combinations. Instead, Fig. 7B shows that all properties were represented in these combinations with similar frequency (between 54% and 61% correct). Although global property combinations containing transience are slightly higher than the mean performance ($t(6) = 2.0$, $p < 0.05$), and combinations containing expansion trend toward lower performance ($t(6) = 1.8$, $p = 0.12$), this result suggests that overall categorization performance is not carried by one or two global properties, but rather that each global property provides essential information for classifying all eight basic-level categories. This result is conferred by the multi-dimensional scaling solution on the rankings as described in Appendix A.2 (showing that there is no obvious asymptote in the stress of a six-dimensional solution over a seven-dimensional solution).

5.2.6. Global property classifier generalizes to less prototypical images—Up until this point, all scene images we have used have been ranked as being very prototypical for a basic-level scene category. However, scenes, unlike objects can often be members of more than one basic-level category (Tversky & Hemenway, 1983). A candidate scene representation is not complete without being able to generalize to and deal with images that span category boundaries. Many of the images in the natural world contain elements from multiple categories (poly-categorical). Take, for example the bottom image in Fig. 8. This scene contains elements that could reasonably be assigned to forest, mountain, river or lake scenes. What assignment will the global property classifier give to such a scene?

Recall that the 200 typical scene images used so far were chosen from a larger pool of 500 images that had been ranked by human observers by how prototypical they were for these eight scene categories (Appendix A.3). Recall also that the global property classifier is a maximum likelihood estimator, who computes the probability of an image being in each of the eight basic-level categories. Therefore, we can directly compare the order of category membership given by the human observers to the order of category probability given by the classifier (see examples in Fig. 8).

First, for the 300 poly-categorical images, we compared the top-ranked choice from a category-ranking experiment (see Appendix A.3) to the most likely category given by the classifier when trained on the 200 prototypical images. We found that the classifier's top category choice matched human observers' top category choice in 56% of images. It is of note that we would not expect the classifier performance on poly-categorical images to exceed its percent correct on prototype images (77%, Section 5.2.1). It is also unreasonable to expect the model to agree better with human observers than these observers agree with each other about an image's category (Spearman's correlation 0.73, see Appendix A.3).

A further complexity is that an image might be ranked equally prototypical for multiple categories, have possibility to be ranked in most categories, or have low overall prototypicality for all of the categories used in this experiment. In order to account for these, we then only analyzed images that received a score of at least 3 out of 5 for a category on the prototypicality scale (see Appendix A.3 for method details), and those without a close second-place category rank. For these images, the model's top category choice matched the human observers' top category choice in 62% of the images. It is also notable that the top two category choices for the model match the top choice for the human observers in 92% of the images.

5.3. Discussion

Given that Experiment 2 showed that human observers were sensitive to global property information while rapidly categorizing natural scenes, in Experiment 3 we investigated the extent to which a scene-centered global description is sufficient to predict human rapid scene categorization performances. To do this, we employed a simple classifier whose only image information was the global property ranking data from Experiment 1. In terms of overall accuracy, the classifier is comparable to human performance (Section 5.2.1), and has a similar performance by semantic category (Section 5.2.1), indicating that the same semantic categories that are easier for human observers are also easier for the classifier. We have also shown that the errors made by the classifier are similar to the false alarms made by human observers (Sections Sections 5.2.2 and 5.2.3). Critically, the exact errors are often repeatable (in other words, if a human observer makes a false alarm to a particular mountain as a forest, the classifier will most often make the same mistake). We have shown that the classifier, when trained on incomplete global property data, replicates the false alarms made by human observers in Experiment 2 when certain global properties were rendered less diagnostic for the classification task (Sections Sections 5.2.4 and 5.2.5). Finally, we have shown that the global property representation can deal with non-prototypical images as well as it deals with prototypical images (Section 5.2.6). Altogether, we have shown that in terms of accuracy and errors, a representation that only contains global property information has high predictive value for human performance at rapid basic-level scene categorization.

6. Experiment 4: An alternative hypothesis—comparing a global property representation to a local region representation

The global property-based classifier shows remarkable human-like performance, in terms of both quantity and fidelity, in a rapid scene categorization task. Could any reasonably informative representation achieve such high fidelity? Basic-level scene categories are also defined by the objects and regions that they contain. Here, we test the utility of a local representation for predicting human rapid natural scene categorization by creating an alternative representation of our database that explicitly represents all of the local regions and objects in each scene. In order to fairly test the local representation, we employed two different models using these data, based on implementations of proposals in the literature: the local semantic concept model (Vogel & Schiele, 2007) and the prominent object model ([Biederman, 1981] and [Friedman, 1979]).

The *local semantic concept* model presents the case where an exhaustive list of scene regions and objects is created, and that scene recognition takes place from this list. Vogel and Schiele (2007) showed that very good machine scene classification could be done by representing a natural landscape image as a collection of local region names drawn from a small vocabulary of semantic concepts: an image could be represented as 9% sky, 25% rock, and 66% water, for example. Here, we implement a similar idea, using the names of all regions and objects using a set of basic-level and superordinant region concepts along with their percent image area in a scene (see Section 6.1 and Appendix A.8 for details).

The prominent object model represents the case where scene recognition proceeds from a single, prominent object or region rather than an exhaustive list. This has been a popular potential mechanism for scene understanding proposed in the literature ([Biederman, 1981] and [Friedman, 1979]). Our implementation calculates the predictability of a scene category given the identity of the largest annotated object in the image. For example, we would predict that an image whose largest object is “trees” to be a forest, or an image whose largest region is “grass” is likely a field. Of course, objects can be prominent without necessarily being the largest objects, and a related literature is devoted to determining the image features that make an object prominent, or salient (for a review, see Itti & Koch, 2001). As the nature of these features is still relatively open, here we are limiting our definition of “prominent” to only include size.

It is important to note that both local region models present two conceptually different views about how scene recognition might proceed from local region and object information. The local semantic concept model categorizes a scene based on the co-occurrence of regions from an exhaustive list of scene regions, assuming that in a glance all objects can be segmented, perceived and abstracted into concepts. This model represents the best-case scenario for the local approach, in which the identities of all of the objects and regions in the scene are known, as well as their relative sizes.

By contrast, the prominent object model assumes that not all regions have equal diagnostic information for the scene category, and that in particular, if an object is prominent in the scene, it will contain more information about the scene's category. Scene categorization is therefore an inference based on the recognition of this prominent (and informative) object. However, it is important to note that as size information is also included in the local semantic concept model, all of the information in the prominent object model is contained in the local semantic concept model. Therefore, the essential difference in the two models is in the relative importance of one object verses the importance of all objects.

6.1. Method

Two independent observers (one author, and one naïve observer) hand-segmented and labeled all regions and objects in the 200 image database. The labeling was done using the online annotation tool LabelMe (Russell, Torralba, Murphy, & Freeman, 2008). Example annotations are found in Fig. 9. There were a total of 199 uniquely labeled regions in the database. All of the labels were pared down to 16 basic and superordinant level region names by removing typos, misspellings, synonyms and subordinant-level concept names (for example “red sand” instead of “sand”). We used the following region concepts for the local semantic concept model: sky, water, foliage, mountains, snow, rock, sand, animals, hills, fog, clouds, grass, dirt, manmade objects, canyon and road. This list includes the nine semantic concepts used by Vogel and Schiele (2007) as well as others that were needed to fully explain our natural image database. In Appendix A.8, we report that the performance of this 16 concept model is not different from a model using the raw data (199 concepts), or a model using 50 basic-level region concepts.

Each image's list of regions (along with their image area) was used to train and test a naïve Bayes classifier using the same leave-one-out procedure as described in Experiment 3. As with the global property classifier of Experiment 3 results are compared to the human psychophysical performance of Experiment 2.

For the *prominent object* model, the naïve Bayes classifier was not needed because the relevant information could be calculated directly from the statistics of the LabelMe annotations. For each image, we calculated the probability of the image being from each basic-level category based on the identity of the scene's largest object. For this analysis, we used the 50 local concept

list (see Appendix A.8) as it had the best balance between distinctiveness and representation of the object concepts.

For each image, we computed a 50 region by 8 category matrix of object predictability from the 199 remaining scenes where each entry (i, j) was the probability of the region (i) being in the basic-level category (j). Taking the row representing the largest region in the test image, we selected the category for maximum probability for that region. For example, if a scene's largest region was sky, the probabilities of the scene being from each of the eight categories are as follows: 0.20 desert; 0.14 field; 0.04 forest; 0.16 lake; 0.17 mountain; 0.15 ocean; 0.05 river; 0.09 waterfall. Therefore, the scene is most likely a desert.

6.2. Results

A summary of the classification results of the two local region models, along with a comparison to the global property model of Experiment 3, can be found in Table 5.

The local semantic concept model refers to a model in which a scene is represented as a co-occurrence vector of all labeled regions and objects along with their relative sizes. The prominent object model refers to the predictability of the scene category conditioned on the presence of its largest object. The by-category correlation (cf. Section 6.2.1 for the local models and Section 5.2.1 for global model) shows the extent to which the models are similar to the pattern of human correct performance rate by category for the eight basic-level categories. The item analysis (Sections Sections 6.2.2 and 5.2.2 for local and global models, respectively, bins of 25) shows the extent to which the models tend to misclassify the same images as humans do. The between-category confusion correlation (Sections Sections 3.1 and 5.2.3 for local and global models, respectively) shows the extent to which the patterns of confusability between pairs of basic-level categories for the models were similar to those of human observers.

*Indicates significant correlations ($p < 0.05$).

6.2.1. Local models' performance: Percent correct and correlation to human basic-level category performance—The local semantic concept model averaged an overall 60% correct categorizations (vs. 77% for the global property classifier, Section 5.2.1), which was significantly lower than the percent correct of human meta-observers (77%, $t(7) = 2.88$, $p < 0.05$, also recall 86% for complete-observers). To ensure that the local semantic concepts were not too general, we compared this performance to the performance on a larger list of 50 basic-level region concepts, finding no significant performance difference to the semantic concept model ($t(398) < 1$, see Appendix A.8 for details, including the percent of correct classifications per semantic category). The prominent object model performed well overall. The overall percent correct for this model was 52% (chance being 12.5%), but still under the rate of human observers ($t(7) = -9.4$, $p < 0.0001$).

To evaluate how the local models compared to human performance by category, we correlated meta-observer correct performance and object models' correct performance for the eight basic-level categories (as in Section 5.2.1 and Fig. 5 for the global property model): none were significant ($r = 0.64$, $p = 0.09$, for the local semantic model, and $r = 0.55$, $p = .16$, for prominent object model).

These results suggest that the scene categories that are easy or hard for human observers to classify at a short presentation time are not necessarily the same for the object models. In fact, the categories field, forest and mountain are classified by all three models at human performance levels, whereas the object models' classifications drop for desert, lake, ocean and river. Indeed, field, forest and mountain are environments that are mostly composed of one or two prominent regions or objects (e.g. grass for field, trees for forest, and mountain for

mountain), whereas other scene categories share more objects between them, putting local models at a disadvantage.

6.2.2. Error analysis: Easy and difficult images—As we did in Experiment 3 (Section 5.2.2), we performed an item analysis to determine if the local region models would have trouble classifying the same images that human observers do. This analysis quantifies whether an error is made on an image, but not the type of error made.

Both the local semantic concept model and the prominent object model reflected the level of difficulty of the images for humans as well as the global property model did (for bins of 25, $r = 0.69$ for both object models, both correlations significant $p < 0.001$, see Table 5. Bins of 10 yielded higher coefficients, $r = 0.89$ for local semantic concept model and $r = 0.85$ for the prominent object model). These correlations indicate that both global and local representations have a tendency to perform well or poorly on the same images. However, this analysis does not give information about the type of errors made. In other words, the local models and human observers tend to misclassify the same images, but do they misclassify these images as being the same category? We explore this issue below.

6.2.3. Qualitative error analysis: Distribution of error types—In order to evaluate further the types of errors made by the local models, we analyzed the extent to which the distribution of errors made by the object models was similar to the distributions of false alarms made by human observers. For instance, in the rapid scene categorization task (Experiment 2), humans often confused river and waterfall, as well as desert with field (Table A5). However, they almost never mistake a forest for an ocean. Are the pairs of categories often confused by human observers also often confused by the local region models? As in Section 5.2.3, we compared the pairwise basic-level category confusions made by the local region models to the distribution of false alarms made by the human observers for each pair of categories. For both local models, there was no significant relation between their patterns of category confusability and those of the human observers: $r = 0.23$ ($p = .25$) for the local semantic concept model, and $r = 0.06$ ($p = 0.75$) for the prominent object model (the global property model gave $r = 0.77$ for comparison). This indicates that there is limited similarity between the local models and human observers in terms of the pairs of categories confused, and suggests that these local models do not capture the richness of the representation built by human observers in a 30 ms presentation time.

6.3. Discussion

The high performance of the global property model begs the question of whether any reasonably rich and informative representation could predict human rapid scene categorization performance.

Here we have explored two distinct alternative hypotheses to the global property scene representation. In particular, our results suggest that a local, region-based approach, based on suggestions from the literature does not have the same capacity to explain human rapid scene categorization as the global property model does. It is of note that the local semantic concept model represents one of the best-case scenario for the local approach, in which the identities of all of the objects and regions in the scene are known, as well as their relative sizes.

While the local semantic concept model shows relatively good percent correct performance at basic-level scene categorization (60%, chance being 12.5%), it does not have the fidelity to predict the types of false alarms made by human observers in a rapid scene categorization task (cf. Table 5). For instance, Fig. 10 shows example false alarms made by the global property classifier of Experiment 3 with the local semantic concept model of Experiment 4. Strikingly, the top desert and river are classified by the global property classifier as being field and forest,

respectively. This mirrors the pattern of false alarms made to the same images by human observers in Experiment 2. However, the lake and river shown at the bottom of Fig. 10 were classified as ocean and field, respectively, by the local semantic concept model; errors that were not made often by the human observers in Experiment 2. At first glance, it seems strange that such a prototypical river (bottom right of Fig. 10) would be classified as a field at all. However, as fields in our database have large amounts of sky, trees and rock (similar to rivers), this image was classified as a field by the local semantic concept model.

The prominent object model, while having the lowest overall correct categorization performance of the models, still performed substantially above chance. This is because some categories, such as field and forest were very well categorized by this model. This makes intuitive sense, as typical prominent objects for these categories were grass and trees, respectively, which were very diagnostic for these categories. However, these categories which were easy for the model to classify had limited similarity to the categories that were easy for the human observers to classify, which is why the by-category correlation was modest. While the prominent object model had a tendency to correctly categorize the same images as human observers, it could not predict the types of errors that the human observers would make. For example, if water was the largest object in a scene, the prominent object model could not distinguish whether the scene was a lake, ocean, river or waterfall because water is equally diagnostic for these categories.

Likewise, the local semantic concept model was able to correctly classify the majority of the images in the database. This is because there is a considerable amount of redundancy in image categories that allowed the model to learn that a scene with cliffs, water and sky is likely to be a waterfall while a scene with sand, rock and sky is likely to be a desert. However, the pattern of correct category classification of this model showed only modest similarity to that of the observers. For example, field was very well classified by the model while it was on average, one of the more difficult categories for the human observers in the rapid categorization task. This is likely because the model was relying heavily on the presence of objects such as grass or flowers that are unique to this category. Like the prominent object model, the local semantic concept model tended to correctly classify the same images as human observers, but could not predict the types of false alarms made by humans. In particular, categories such as lake and river have very similar sets of objects (typical objects include sky, water, trees and grass), so it was difficult for the local semantic concept model to distinguish between these categories, even though human observers did not have such a difficulty.

In contrast, the global property model of Experiment 3 had higher correct classification performance than the local models, and was very similar to human observers' performance. Also in contrast to the local models, its pattern of performance by category significantly correlated with that of the human observers'. Like both of the local models, it also tended to correctly classify the same images that human observers did. However, unlike the local models, it has the power to predict the types of false alarms made by the human observers. To go back to the lake and river example, the local models made errors in these categories because the objects in them are very similar. However, the global property model can distinguish between them because they have different layout and surface properties: lakes are more open, and less transient, for example (see Fig. 3). To the human observers, few errors are made between these categories, perhaps because the observers are using the structural differences between these categories to distinguish them.

Clearly, more sophisticated object models that incorporate structure and layout information should be able to capture more of the essence of a natural scene ([Grossberg and Huang, in press] and [Murphy et al., 2003]). Our point here is that object models testing simple instantiations of valid propositions from the visual cognition literature do not have the same

explanatory power as our global property model for predicting human rapid scene categorization performance.

Importantly, we do not mean to imply that local objects are regions are not represented in early processing of the visual scene. Instead we have shown that the remarkable fidelity of a global property representation for predicting human rapid scene categorization performance cannot be achieved with any reasonably informative description of the visual scene.

While local region and object information most certainly make up an important part of a scene's identity, our results suggest that the representation formed by human observers after a very brief glance at a scene is not dominated by local object information (see also Fei-Fei et al., 2007). Our results suggest the possibility that our qualia of object perception in a brief glance might be based upon inference of these objects given global scene structure and schema activation.

7. General discussion

In this paper, we have shown that a global scene-centered approach to natural scene understanding closely predicts human performance and errors in a rapid basic-level scene categorization task. This approach uses a small vocabulary of global and ecologically relevant scene primitives that describe the structural, constancy and functional aspects of scene surfaces without representing objects and parts. Beyond the principle of recognizing the “forest before the trees” (Navon, 1977), here we propose an operational definition of the notion of “globality” for natural scene recognition, and provide a novel account of how human observers could identify a place as a “forest”, without first having to recognize the “trees”.

Several independent analyses, on human performance alone (Experiments 1 and 2), and on human performance compared to a classifier (Experiments 3 and 4), were undertaken to finely probe the relation between a global scene representation and human rapid natural scene categorization performance. Although strict causation cannot be inferred from these correlational results alone, all results taken together are suggestive of the view that a scene-centered approach can be used by human observers for basic-level scene categorization. Strengthening this view is the fact that performance of a classifier representing the local objects and regions of the images (Experiment 4) does not have the same explanatory power as the global property representation (Experiment 3) for predicting human performance and false alarms (Experiment 2).

We have shown that human performance at a rapid scene categorization task can be dramatically influenced by varying the distractor set to contain more global property similarities to a target category (cf. Fig. 4, Section 4.2.2). Moreover, the item analysis which calculates the probability of a false alarm occurring to single distractor images was very well predicted from each distractor's distance from the target-category mean for a global property, suggesting that rapid image categorization performance follows the statistical regularities of global properties' distributions in basic-level categories. Last, the relative confusability of basic-level categories (Section 4.2.3, Table A5 and Table A6) to one another is also well-explained by the basic-level categories' similarity in global property space.

To determine how computationally sufficient the global properties are for explaining the human rapid scene categorization data in Experiment 2, we compared a simple classifier to human performance on several metrics (Experiment 3). First, the overall categorization performance of the classifier was similar to humans', and the relative performance of the classifier by category was also well-correlated with human observers.

However, similar levels of performance are not enough: if the global property representation is a plausible human scene representation, then the classifier should also predict the false alarms made by human observers. We have shown that image difficulty for the classifier is very similar to image difficulty for human observers, and that the same qualitative errors are made by both (e.g. false alarming to a particular river image as a waterfall) the majority of the time (Section 5.2.3). Furthermore, we have shown that when a global property is not available for use in categorization, either because it is not explicitly represented (classifier), or because the distractors make it non-diagnostic of the target category (humans), performance suffers similarly (Sections Sections 5.2.4 and 5.2.5). Furthermore, we have shown in Section 5.2.6 that the high fidelity of categorization performance in the global property model can generalize beyond prototypical images. In particular, the level of agreement between the classifier and human observers is not different from the agreement between the human observers. Lastly, the striking predictability of the global property model for human scene categorization performance is not found in two local object models that we tested (Experiment 4).

It has been known that visual perception tends to proceed in a global-to-local manner (Navon, 1977), but for stimuli as complex as a natural scene, it is not obvious what the global level might be. Computational models have shown that basic-level scene categories can emerge from a combination of global layout properties ([Oliva and Torralba, 2001], [Oliva and Torralba, 2002] and [Oliva and Torralba, 2006]), or from a collection of regions ([Fei-Fei and Perona, 2005], [Grossberg and Huang, in press], [Vogel and Schiele, 2007] and [Vogel et al., 2006]) but no psychological foundation has yet been established between global scene properties and basic-level scene categorization performance. This work has tried to make this link. By grounding our search in the principles of environmental affordance ([Gibson, 1979] and [Rosch, 1978]), we found a collection of global properties that are sufficient to capture the essence of many natural scene categories.

Our result is also in the spirit of seminal scene understanding studies from the 1970s and 1980s. Biederman and collaborators have shown that coherent scene context aided the search for an object within the scene, even when the identity and location of the object were known in advance (Biederman, 1972). Furthermore, lack of coherent spatial context seemed particularly disruptive on negative trials where the object was not in the scene, but had a high probability of being in the scene (Biederman, Glass, & Stacy, 1973). Together, this suggests that scene identity information may be accessed before object identity information is complete. Biederman (1981) outlined three paths by which such scene information could be computed: (1) a path through the recognition of a prominent object; (2) a global path through scene-emergent features that were not defined at this time; (3) the spatial integration of a few context related objects.

Our results offer positive evidence for path 2 (the global path suggested by Navon, 1977, but never operationalized) and non-conclusive evidence for path 1 (the prominent object). Path 3 supposes that the co-occurrence of a few objects in a stereotypical spatial arrangement would be predictive of the scene category. The semi-localized local model of Vogel and Schiele (2007) along with the studies of relation processing by Hummel and colleagues (e.g. Saiki & Hummel, 1998) has started to find evidence for this path. However, there is also reason to believe that path 3 may not be the only approach for capturing the type of representation built over a brief glance at a novel scene. This view requires that several objects be segmented, recognized and relationally organized for scene categorization to occur. However, it is still not clear that humans can segment, identify and remember several objects in a scene at a glance. Potter, Staub, and O' Connor (2004) demonstrated that, in a memory test following an RSVP sequence of images, a large number of false alarms were made to images that were conceptually similar to an image presented in the sequence, but did not necessarily have the same objects and regions, suggesting that what is encoded and stored from a brief glance at a scene is a more

general description of the image than an exhaustive list of its objects. This view is corroborated with the facts that human observers also make systematic errors in remembering the location of objects from a briefly glimpsed display (Evans & Treisman, 2005), and are relatively insensitive to changes in single objects in a scene (change blindness, [Rensink et al., 1997] and [Simons, 2000]).

A consequence of our global precedence finding could be that the perceptual entry-level for visual scenes is not the basic-level category, but rather an image's global property descriptions, at a superordinate level ([Joubert et al., 2007], [Oliva and Torralba, 2001] and [Oliva and Torralba, 2002]). This idea is not necessarily contradictory of the behavioral findings of Rosch and colleagues. We argue that the basic-level category is the entry-level for communication about objects and places because it represents a compromise between within-category similarity and between-category distinctiveness. However, under the constraints of a rapid categorization task, perhaps the initial scene representation would benefit from processing distinctiveness first, making a superordinate description an ideal level, particularly if the visual features used to get this superordinate description do not require a segmentation stage, known to be computationally more expensive than an holistic analysis (Oliva & Torralba, 2001).

Finding the image-level features that mediate such rapid visual categorizations is a fascinating, yet rather open question that is beyond the scope of the current work (cf. McCotter, Gosselin, Sowden, & Schyns, 2005). Indeed, previous work has shown that certain spatial layout properties, such as openness and mean depth can be well-described from a set of low-level image features corresponding to spatially localized second-order image statistics ([Oliva and Torralba, 2001], [Oliva and Torralba, 2002], [Torralba and Oliva, 2002] and [Torralba and Oliva, 2003]). Some properties, such as temperature, might even be represented by simpler images features, such as the color distribution. However, functional properties such as navigability and concealment may be more complex to represent, as their spatial structures might not co-vary in a simple way with first or second-order image statistics. For instance, if a scene is very open, it is open because it has a very salient horizon line somewhere near the vertical center, and all scenes that are consistently ranked as highly open share this feature. A navigable scene however, might be navigable because the scene is open and free of clutter, or it could be navigable because it has a very obvious path through an otherwise dense environment. Therefore, image features of a higher complexity might be needed to fully represent these global properties, a question that future research will investigate.

A global scene-centered representation is a plausible coding of visual scenes in the brain and a complementary approach to object-based scene analysis. This present work suggests that rapid scene recognition can be performed by global scene-centered mechanisms and need not be built on top of object recognition. Indeed, work in functional imaging has shown a dissociation between brain areas that represent scenes (the parahippocampal place area, or PPA, Epstein & Kanwisher, 1998) and those that represent individual objects ([Bar, 2004] and [Grill-Spector et al., 2001]). Furthermore, the PPA seems to be sensitive to holistic properties of the scene layout, but not to its complexity in terms of quantity of objects (Epstein & Kanwisher, 1998). The neural independence between scenes and object recognition mechanisms was recently strengthened by Goh, Siong, Park, Gutchess, Hebrank and Chee (2004). They observed activation of different parahippocampal regions when pictures of scenes were processed alone compared to pictures containing a prominent object, consistent within that scene. Steeves, Humphreys, Culham, Menon, Milner and Goodale (2004) have shown that an individual with profound visual form agnosia could still identify pictures of real-world places from color and texture information only. These findings are consistent with the hypothesis that whole scene recognition may be dissociated from object identification.

What is the mechanism by which a scene-centered pathway could arise in the brain? Although we are far from a definitive answer, an examination of the time course of visual processing yields critical insights. Thorpe and colleagues (1996) have made a case that the speed of high-level visual processing necessitates a single feed-forward wave of spikes through the ventral visual system. Furthermore, biologically inspired models of this architecture yield high performances in detection tasks ([Delorme and Thorpe, 2003] and [Serre et al., 2007]). However, very rapid feedback might also mediate this performance. Physiological evidence shows that there is considerable overlap in time between spikes arriving in progressive areas of the ventral visual stream (Schmolesky et al., 1998), suggesting that feedback from higher visual areas can feed back to early visual areas to build a simple yet global initial scene representation. Furthermore, a combined EEG/MEG and fMRI study has shown a V1 feedback signal as early as 140 ms after stimulus presentation (Noesselt et al., 2002) furthering the idea that scene recognition may be mediated through rapid feedback. Strikingly, there is evidence of the global pattern from a contextual cueing display being processed 100 msec after stimulus presentation (Chaumon, Drouet, & Tallon-Baudry, 2008). These results confer with behavioral evidence which suggest that global properties such as concealment or naturalness are available for report with less exposure time than basic-level categories (Greene & Oliva, in preparation; [Joubert et al., 2005], [Joubert et al., 2007] and [Kaplan, 1992]). Although this does not necessarily imply that they are processed first by the brain, it is consistent with the view that global properties are reasonable scene primitives for basic-level categorization.

Emphasizing the importance of a scene-centered view does not imply that objects are not an important part of rapid scene recognition. Surely, as objects can make up the identity of the scene and are the entities acted on by agents in a scene, they are of critical importance for scene understanding with longer image exposures. However, it appears that objects might not necessarily be the atoms of high-level recognition especially under degraded conditions of blur or at the very beginning of visual analysis ([Oliva and Schyns, 2000] and [Schyns and Oliva, 1994]). But given longer image exposures, objects become increasingly important in our representations of scenes during the course of the first fixation ([Fei-Fei et al., 2007] and [Gordon, 2004]) and a framework that would combine objects and their spatial relationships with global properties would capture more of the richness of scene identity.

In this paper, we have demonstrated that global property information is more diagnostic of natural scene categories than local region and object information. A natural question is then what roles both types of information play in other types of environments, such as indoor scenes? Intuitively, the prominent object model from Experiment 4 seems like it would do a good job at categorizing some indoor categories such as bedrooms or living rooms because the largest object (bed or sofa) is not typically found in other scene categories. However, it does not seem that all indoor categories are so strongly object-driven. A corridor, for example, is unique among indoor scene categories as having a great deal of perspective. A conference room and a dining room might also be confused by a prominent object model as they both have prominent tables surrounded by chairs. Part of our ongoing effort is characterizing the relative use of global and local diagnostic information for scene categorization for a greater variety of scene categories.

An extension of the present work that could indirectly probe the neural representation of visual scenes is to measure if global properties are adaptable (Greene & Oliva, 2008). A ubiquitous property of neural systems is that repeated presentation of a represented property leads to a temporary decrease in sensitivity to that property, a phenomenon known as adaptation. This phenomenon is seen at all levels of visual processing for entities that seem to have dedicated processing, from basic properties such as color, motion, orientation and spatial frequency (for a review, see Wade & Verstraten, 2005) to complex features such as facial emotion and identity ([Leopold et al., 2001] and [Webster et al., 2004]). Furthermore, adapting to low-level image

features can modulate higher level perceptual judgments for surface glossiness (Motoyoshi, Nishida, Sharan, & Adelson, 2007) or the naturalness of real-world scenes (Kaping, Tzvetanov, & Treue, 2007).

7.1. Concluding remarks

The present work was designed to operationalize the notion of globality in the domain of natural real-world images. We have shown that global properties capture much of the variance in how real-world scenes vary in structure, constancy and function, and are involved in the representation of natural scenes that allows rapid categorization.

All together, our results provide support for an initial scene-centered visual representation built on conjunctions of global properties that explicitly represent scene function and spatial layout.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to thank George Alvarez, Timothy Brady, Barbara Hidalgo-Sotelo, Todd Horowitz, Talia Konkle, Mary Potter, Joshua Tenenbaum, Antonio Torralba, Jeremy Wolfe and three anonymous reviewers for very helpful comments and discussion. This research is supported by a NSF graduate research fellowship awarded to M.R.G. and by a NEC Research Support in Computer and Communication and a National Science Foundation Career Award (0546262) and an NSF Grant (0705677) to A.O.

References

Alvarez GA, Oliva A. The representation of simple ensemble visual features outside the focus of attention. Psychological Science 2008;19(4):392–398. [PubMed: 18399893]

Appelton, J. The experience of landscape. Wiley; London: 1975.

Ariely D. Seeing sets: Representation by statistical properties. Psychological Science 2001;12:157–162. [PubMed: 11340926]

Ashby F, Lee W. Predicting similarity and categorization from identification. Journal of Experimental Psychology: General 1991;120(2):150–172. [PubMed: 1830609]

Bar M. Visual objects in context. Nature Reviews: Neuroscience 2004;5:617–629.

Biederman I. Perceiving real-world scenes. Science 1972;177:77–80. [PubMed: 5041781]

Biederman, I. Perceptual organization. Lawrence Erlbaum; Hillsdale, New Jersey: 1981. On the semantics of a glance at a scene; p. 213-263.

Biederman I. Recognition by components: A theory of human image understanding. Psychological Review 1987;94(2):115–147. [PubMed: 3575582]

Biederman I, Glass A, Stacy E. Searching for objects in real-world scenes. Journal of Experimental Psychology 1973;97(1):22–27. [PubMed: 4704195]

Biederman I, Rabinowitz JC, Glass AL, Stacy EW. On the information extracted from a glance at a scene. Journal of Experimental Psychology 1974;103:597–600. [PubMed: 4448962]

Biederman, et al. Biederman I, Mezzanotte RJ, Rabinowitz JC. Scene perception: Detecting and judging objects undergoing relational violations. Cognitive Psychology 1982;14:143–177. [PubMed: 7083801]1982

Biederman, et al. Biederman I, Blickle TW, Teitelbaum RC, Klatsky GJ, Mezzanotte RJ. Object identification in nonscene displays. Journal of Experimental Psychology: Human Learning, Memory, and Cognition 1988;14:456–467.1988

Brainard DH. The psychophysics toolbox. Spatial Vision 1997;10:443–446. [PubMed: 9176954]

Bülthoff H, Edelman S, Tarr M. How are three-dimensional objects represented in the brain? Cerebral Cortex 1995;3:247–260.

Chaumon M, Drouet V, Tallon-Baudry C. Unconscious associative memory affects visual processing before 100 ms. Journal of Vision 2008;8(3):1–10. [PubMed: 18484816]

Chen L. The topological approach to perceptual organization. Visual Cognition 2005;12(4):553–637.

Chong SC, Treisman A. Representation of statistical properties. Vision Research 2003;43:393–404. [PubMed: 12535996]

Chong S, Treisman A. Statistical processing: Computing the average size in perceptual groups. Vision Research 2005;45(7):891–900. [PubMed: 15644229]

Chubb C, Nam J, Bindman D, Sperling G. The three dimensions of human visual sensitivity to first-order contrast statistics. Vision Research 2007;47(17):2237–2248. [PubMed: 17619044]

Chun MM, Jiang Y. Contextual cueing: Implicit learning and memory of visual context guides spatial attention. Cognitive Psychology 1998;36:28–71. [PubMed: 9679076]

Cutting J. Representing motion in a static image: Constraints and parallels in art, science and popular culture. Perception 2002;31(10):1165–1193. [PubMed: 12430945]

De Graef P, Christaens D, d'Ydewalle G. Perceptual effects of scene context on object identification. Psychological Research 1990;52:317–329. [PubMed: 2287695]

Delorme A, Thorpe S. SpikeNET: An event-driven simulation package for modeling large networks of spiking neurons. Network: Computation in Neural Systems 2003;14:613–627.

Epstein R, Kanwisher N. A cortical representation of the local environment. Nature 1998;392:598–601. [PubMed: 9560155]

Evans K, Treisman A. Perception of objects in natural scenes: Is it really attention free? Journal of Experimental Psychology: Human Perception and Performance 2005;31(6):1476–1492. [PubMed: 16366803]

Fei-Fei L, Perona P. A Bayesian Hierarchical model for learning natural scene categories. IEEE Proceedings in Computer Vision and Pattern Recognition 2005;2:524–531.

Fei-Fei L, Iyer A, Koch C, Perona P. What do we perceive in a glance of a real-world scene? Journal of Vision 2007;7(1):1–29. [PubMed: 17997664]

Fergus R, Perona P, Zisserman A. Object class recognition by unsupervised scale-invariant learning. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2003;2:II-264–II-271.

Freyd J. The mental representation of movement when static stimuli are viewed. Perception and Psychophysics 1983;33:575–581. [PubMed: 6622194]

Friedman A. Framing pictures: The role of knowledge in automatized encoding and memory of scene gist. Journal of Experimental Psychology: General 1979;108:316–355. [PubMed: 528908]

Gibson J. Visually controlled locomotion and visual orientation in animals. British Journal of Psychology 1958;49(3):182–194. [PubMed: 13572790]

Gibson, JJ. The ecological approach to visual perception. Houghton-Mifflin; Boston: 1979.

Goh JOS, Siong SC, Park D, Gutchess A, Hebrank A, Chee MWL. Cortical areas involved in object, background, and object-background processing revealed with functional magnetic resonance adaptation. Journal of Neuroscience 2004;24:10223–10228. [PubMed: 15537894]

Gordon R. Attentional allocation during the perception of scenes. Journal of Experimental Psychology: Human Perception and Performance 2004;30(4):760–777. [PubMed: 15301623]

Gosselin F, Schyns P. Why do we SLIP to the basic level? Computational constraints and their implementation. Psychological Review 2001;108(4):735–758. [PubMed: 11699115]

Greene, Oliva, Greene MR, Oliva A. The briefest of glances: The time course of natural scene understanding. in preparationin preparation

Greene, MR.; Oliva, A. Natural scene categorization from conjunctions of ecological global properties; Proceedings of the 28th annual conference of the cognitive science society; Vancouver, Canada. 2006. p. 291-296.

Greene MR, Oliva A. High-level aftereffects to natural scenes. Journal of Vision 2008;8(6):1104–1104a.

Grill-Spector K, Kourtzi Z, Kanwisher N. The lateral occipital complex and its role in object recognition. Vision Research 2001;41:1409–1422. [PubMed: 11322983]

Grossberg S, Huang T-R. ARTSCENE: A neural system for natural scene classification. Journal of Vision. in press

Henderson J, Hollingworth A. Global transsaccadic change blindness during scene perception. Psychological Science 2003;14(5):493–497. [PubMed: 12930482]

Hollingworth A, Henderson J. Does consistent scene context facilitate object perception? Journal of Experimental Psychology: General 1998;127(4):398–415. [PubMed: 9857494]

Itti L, Koch C. Computational modeling of visual attention. Nature Reviews Neuroscience 2001;2(3): 194–203.

Joubert O, Fize D, Rousselet G, Fabre-Thorpe M. Categorization of natural scenes: Global context is extracted as fast as objects. Perception 2005;34s:140.

Joubert O, Rousselet G, Fize D, Fabre-Thorpe M. Processing scene context: Fast categorization and object interference. Vision Research 2007;47:3286–3297. [PubMed: 17967472]

Kaping D, Tzvetanov T, Treue S. Adaptation to statistical properties of visual scenes biases rapid categorization. Visual Cognition 2007;15(1):12–19.

Kaplan, S. Environmental preference in a knowledge-seeking, knowledge-using organism. In: Barkow, JH.; Cosmides, L.; Tooby, J., editors. The adaptive mind. Oxford University Press; New York: 1992. p. 535-552.

Kimchi R. Primacy of wholistic processing and global/local paradigm: A critical review. Psychological Bulletin 1992;112:24–38. [PubMed: 1529037]

Kimchi R. Uniform connectedness and grouping in the perceptual organization of hierarchical patterns. Journal of Experimental Psychology: Human Perception and Performance 1998;24:1105–1118. [PubMed: 9706711]

Kourtzi Z, Kanwisher N. Activation of human MT/MST by static images with implied motion. Journal of Cognitive Neuroscience 2000;12(1):48–55. [PubMed: 10769305]

Leopold D, O'Toole A, Vetter T, Blanz V. Prototype-referenced shape encoding revealed by high-level aftereffects. Nature Neuroscience 2001;4:89–94.

Maljkovic V, Martini P. Short-term memory for scenes with affective content. Journal of Vision 2005;5 (3):215–229. [PubMed: 15929647]

Marr, D. Vision: a computational investigation into the human representation and processing of visual information. Henry Holt and Co., Inc.; New York: 1982.

McCotter M, Gosselin F, Sowden P, Schyns P. The use of visual information in natural scenes. Visual Cognition 2005;12:938–953.

Merilaita S. Visual background complexity facilitates the evolution of camouflage. Evolution 2003;57 (6):1248–1254. [PubMed: 12894933]

Mitchell; Mitchell, TM. Machine learning. McGraw Hill; New York: 1997. 1997

Motoyoshi I, Nishida S, Sharan L, Adelson E. Image statistics and the perception of surface qualities. Nature 2007;447:206–209. [PubMed: 17443193]

Murphy, K.; Torralba, A.; Freeman, W. Advances in neural information processing systems 16 (NIPS). MIT Press; Vancouver, BC: 2003. Using the forest to see the trees: A graphical model relating features, objects and scenes.

Navon D. Forest before trees: The precedence of global features in visual perception. Cognitive Psychology 1977;9:353–383.

Noesselt T, Hillyard S, Woldorff M, Schoenfeld A, Hagner T, Jäncke L, et al. Delayed striate cortical activation during spatial attention. Neuron 2002;35(3):575–587. [PubMed: 12165478]

Oliva A, Schyns P. Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. Cognitive Psychology 1997;34:72–107. [PubMed: 9325010]

Oliva A, Schyns P. Diagnostic colors mediate scene recognition. Cognitive Psychology 2000;41:176–210. [PubMed: 10968925]

Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 2001;42:145–175.

Oliva, A.; Torralba, A. Scene-centered description from spatial envelope properties. In: Bulthoff, H.; Lee, SW.; Poggio, T.; Wallraven, C., editors. Proceedings of 2nd international workshop on biologically motivated computer vision; Springer-Verlag; Tuebingen, Germany. 2002. p. 263-272.

Oliva A, Torralba A. Building the gist of a scene: The role of global image features in recognition. Progress in Brain Research: Visual Perception 2006;155:23–36.

Palmer, SE. Visual perception and world knowledge: Notes on a model of sensory-cognitive interaction. In: Norman, D.; Rumelhart, D., editors. Explorations in cognition. Erlbaum; Hillsdale, NJ: 1975. p. 279-307.

Parkes L, Lund J, Angelucci A, Solomon J, Morgan M. Compulsory averaging of crowded orientation signals in human vision. Nature Neuroscience 2001;4:739–744.

Pelli DG. The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision 1997;10:437–442. [PubMed: 9176953]

Potter MC. Meaning in visual scenes. Science 1975;187:965–966. [PubMed: 1145183]

Potter MC, Staub A, O' Connor DH. Pictorial and conceptual representation of glimpsed pictures. Journal of Experimental Psychology: Human Perception and Performance 2004;30:478–489. [PubMed: 15161380]

Pylyshyn Z. Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. Behavior and Brain Sciences 1999;22:341–423.

Ramachandran V, Tyler C, Gregory R, Rogers-Ramachandran D, Duessing S, Pillsbury C, et al. Rapid adaptive camouflage in tropical flounders. Nature 1996;379:815–818. [PubMed: 8587602]

Rensink, Rensink RA. The dynamic perception of visual scenes. Visual Cognition 2000;7(13):17–42.2000

Rensink, et al. Rensink RA, O'Regan JK, Clark JJ. To see or not to see: The need for attention to perceive changes in scenes. Psychological Science 1997;8(5):367–373.1997

Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. Nature Neuroscience 1999;2(11):1019–1025.

Rosch, E. Principles of categorization. In: Rosch, E.; Lloyd, B., editors. Cognition and categorization. Lawrence Erlbaum; Hilldale, NJ: 1978.

Russell BC, Torralba A, Murphy KP, Freeman WT. LabelMe: A database and web-based tool for image annotation. International Journal of Computer Vision 2008;77(1–3):157–173.

Rousselet GA, Joubert OR, Fabre-Thorpe M. How long to get to the “gist” of real-world natural scenes? Visual Cognition 2005;12(6):852–877.

Saiki J, Hummel J. Connectedness and the integration of parts with relations in shape perception. Journal of Experimental Psychology: Human Perception and Performance 1998;24(1):227–251. [PubMed: 9483827]

Sanocki T. Representation and perception of spatial layout. Cognitive Psychology 2003;47:43–86. [PubMed: 12852935]

Schmolesky MT, Wang Y, Hanes DP, Thompson KG, Leutgeb S, Schall JD, et al. Signal timing across the macaque visual system. Journal of Neurophysiology 1998;79(6):3272–3278. [PubMed: 9636126]

Schyns PG, Oliva A. From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. Psychological Science 1994;5:195–200.

Serre T, Oliva A, Poggio TA. A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Sciences 2007;104(15):6424–6429.

Simons D. Current approaches to change blindness. Visual Cognition 2000;7(1):1–15.

Steeves JKE, Humphreys GK, Culham JC, Menon RS, Milner AD, Goodale MA. Behavioral and neuroimaging evidence for a contribution of color and texture information to scene classification in a patient with Visual Form Agnosia. Journal of Cognitive Neuroscience 2004;16:955–965. [PubMed: 15298783]

Thorpe S, Fize D, Marlot C. Speed of processing in the human visual system. Nature 1996;381:520–522. [PubMed: 8632824]

Torralba A, Oliva A. Depth estimation from image structure. IEEE Pattern Analysis and Machine Intelligence 2002;24:1226–1238.

Torralba A, Oliva A. Statistics of natural images categories. Network: Computation in Neural Systems 2003;14:391–412.

Torralba A, Oliva A, Castelhano M, Henderson J. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. Psychological Review 2006;113:766–786. [PubMed: 17014302]

Torralba A, Fergus R, Freeman W. 80 million tiny images: A large dataset for non-parametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. in press

Tversky. Features of similarity. Psychological Review 1977;84(4):327–352.

Tversky B, Hemenway K. Categories of environmental scenes. Cognitive Psychology 1983;15(1):121–149.

Ullman; Ullman, S. High-level vision: Object recognition and visual cognition. MIT Press; Cambridge: 1999. 1999

Vogel J, Schwaninger A, Wallraven C, Bülthoff H. Categorization of natural scenes: Global vs. local information. Symposium on Applied Perception in Graphics and Visualization APGV 2006;153:33–40.

Vogel J, Schiele B. Semantic scene modeling and retrieval for content-based image retrieval. International Journal of Computer Vision 2007;72(2):133–157.

Wade, N.; Verstraten, F. Fitting the Mind to the World: Adaptation and after-effects in high-level vision. Oxford University Press; New York: 2005. Accommodating the past: A selective history of adaptation; p. 83-102.

Renninger, Walker; Malik; Renninger, L. Walker; Malik, J. When is scene identification just texture recognition? Vision Research 2004;44:2301–2311. [PubMed: 15208015]2004

Warren W, Kay B, Zosh W, Duchon A, Sahuc S. Optic flow is used to control human walking. Nature Neuroscience 2001;4(2):213–216.

Webster M, Kaping D, Mizokami Y, Duhamel P. Adaptation to natural face categories. Nature 2004;428 (6982):557–561. [PubMed: 15058304]

Wolfe J. Visual memory: What do you know about what you saw? Current Biology 1998;8:R303–R304. [PubMed: 9560330]

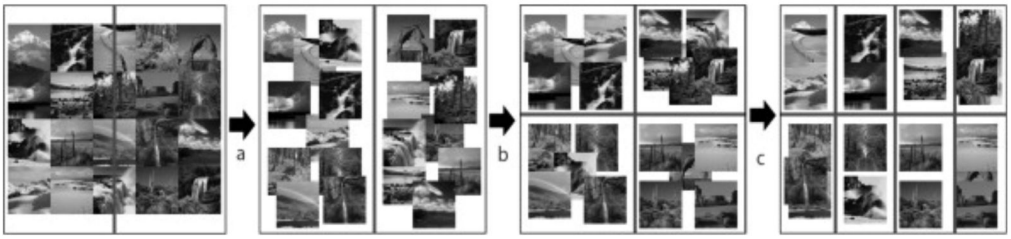

Figure 1.
A schematic illustration of the hierarchical grouping task of Experiment 1. Here, a ranking along the global property *temperature* is portrayed. (a) The images are divided into two groups with the "colder" scenes on the left and the "warmer" scenes on the right. (b) Finer rankings are created by dividing the two initial groups into two subgroups. (c) Images in each quadrant are again divided into two subgroups to create a total of eight groups, ranked from the "coldest" scenes to the "hottest" scenes.

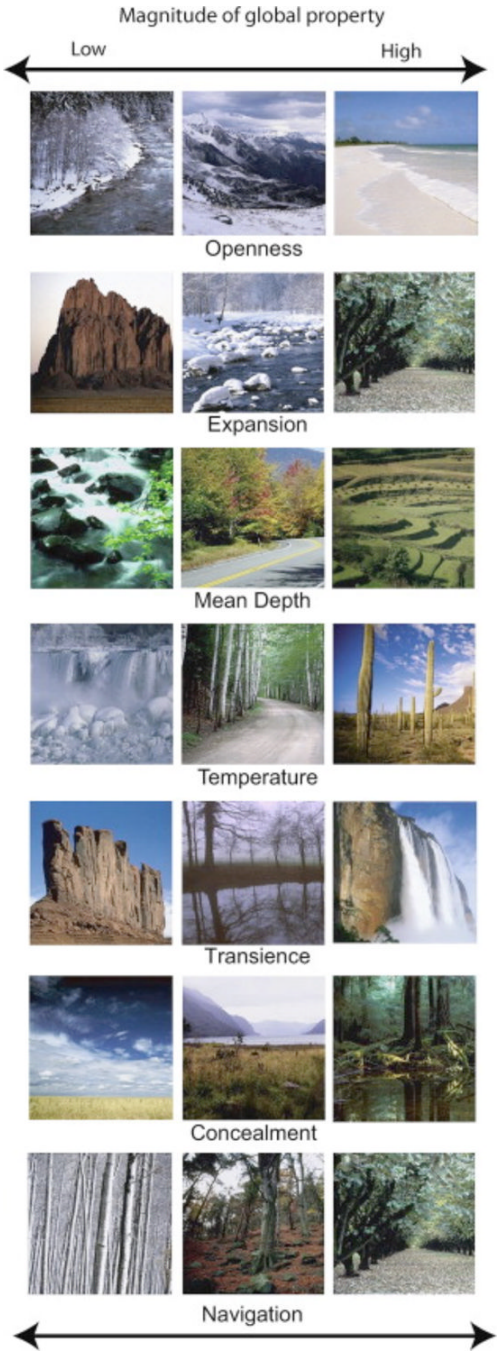

Figure 2.
Examples of scenes with low, medium and high rankings from Experiment 1 along each global property.

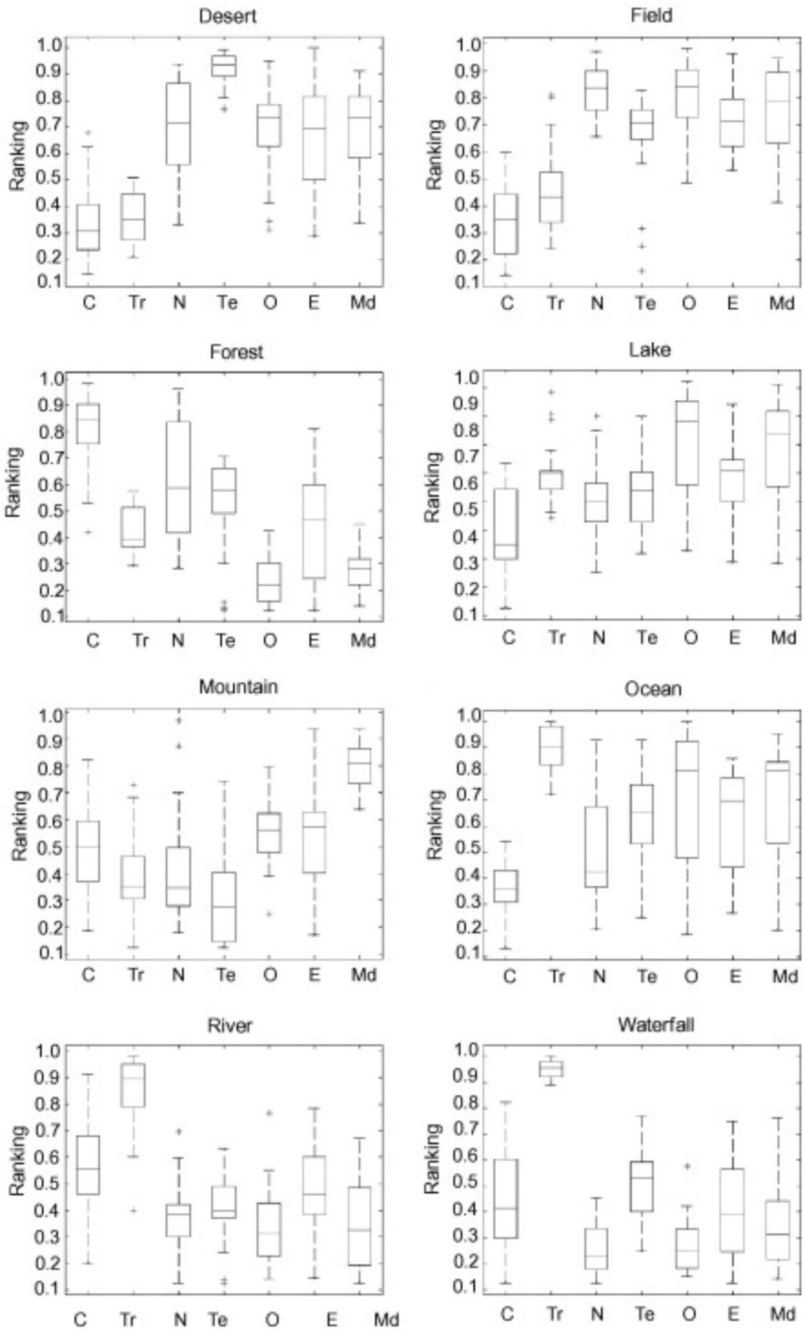

Figure 3.

Box-and-whisker plots of global property rankings for each semantic category, calculated from the ranking data in Experiment 1. Properties are, right to left, C, concealment; Tr, transience; N, navigability; Te, temperature; O, openness; E, expansion and Md, mean depth. Lines indicate median rankings, boxes indicate quartiles and whiskers indicate range. Significant outlier images are shown as crosses.

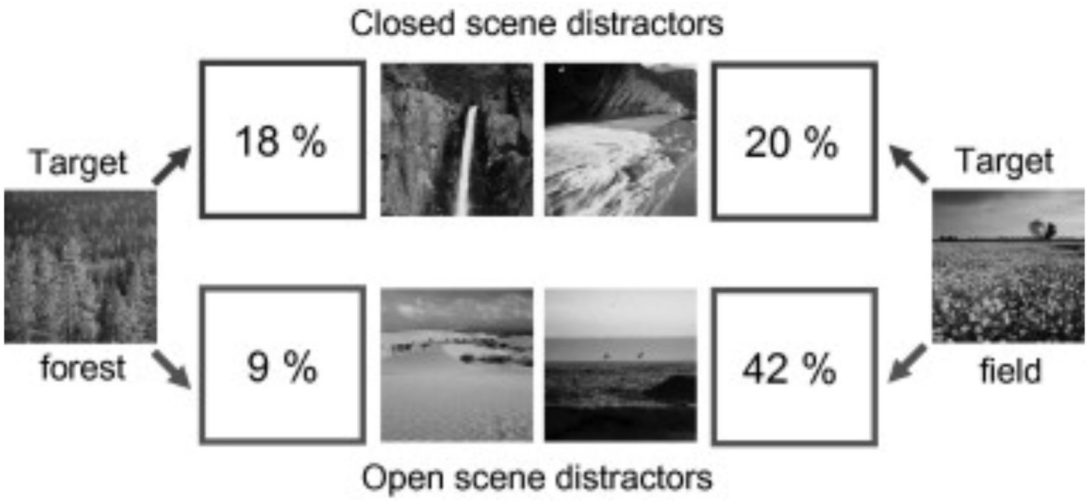

Figure 4.
Illustration of human performance along different distractor sets in Experiment 2. Distractor sets that share a global property with the target category (*closed* is a property of forests and *open* is a property of fields) yield more false alarms than distractor sets that do not. Representative numbers taken from meta-observers' data.

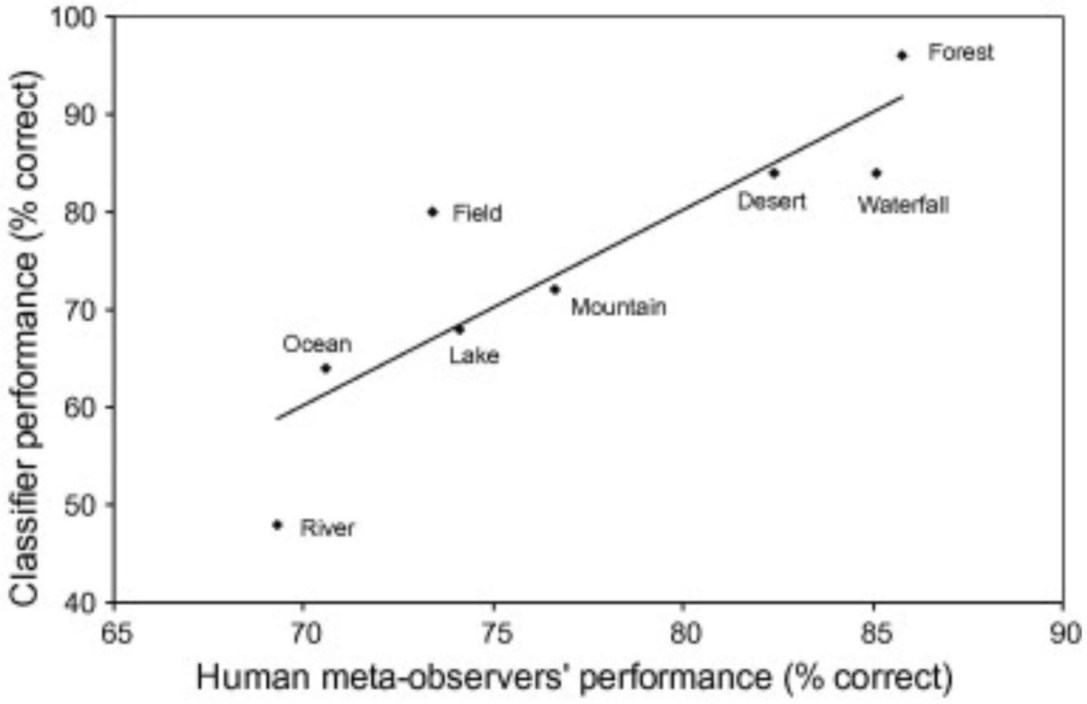

Figure 5.
Categorization performance (percent correct) of naïve Bayes classifier in Experiment 3 is well-correlated with human rapid categorization performance from Experiment 2 (meta-observer data).

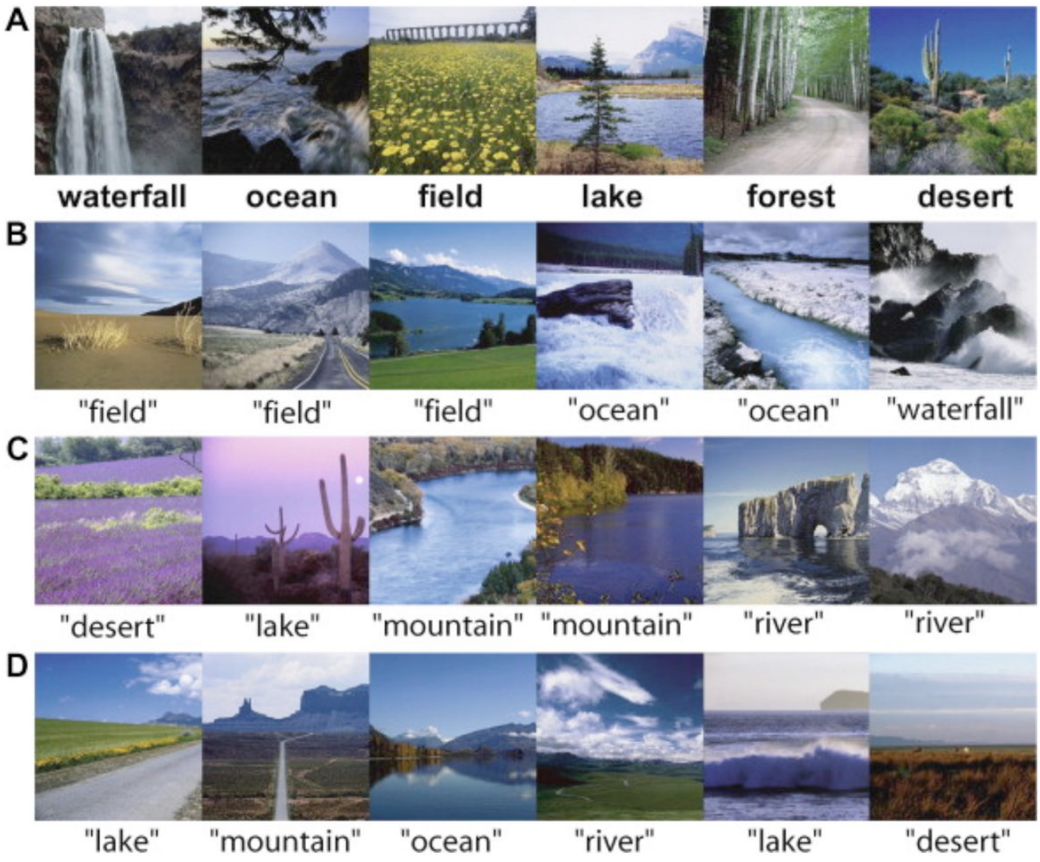

Figure 6.
Examples of human and model performances. (A) (bold titles) corresponds to the correct responses made by both humans (Experiment 2) and the global property classifier (Experiment 3) for the above scene pictures. The other rows (with titles in quotes) represent categorization errors made, respectively, by both humans and the model (B); by the model only (C); by the humans only (D), for the respective scene pictures.

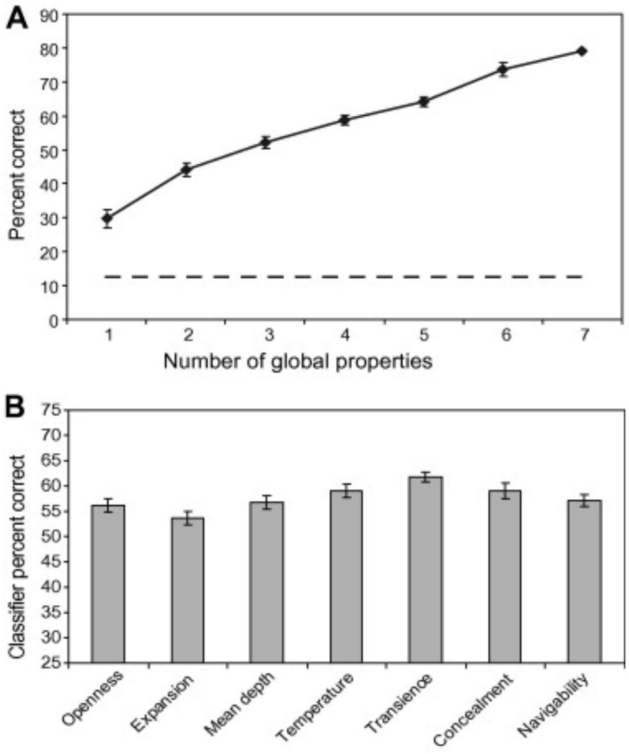

Figure 7.

A) Classifier's performance in Experiment 3 when trained with incomplete data, using from 1 to 7 global properties. The classifier can perform above chance with only one global property (30%), and performance linearly increases with additional properties. Chance level is indicated with the dotted line. (B) Mean classifier performance when trained with incomplete data that contained a particular global property. Classifier performed similarly when any particular global property was present.

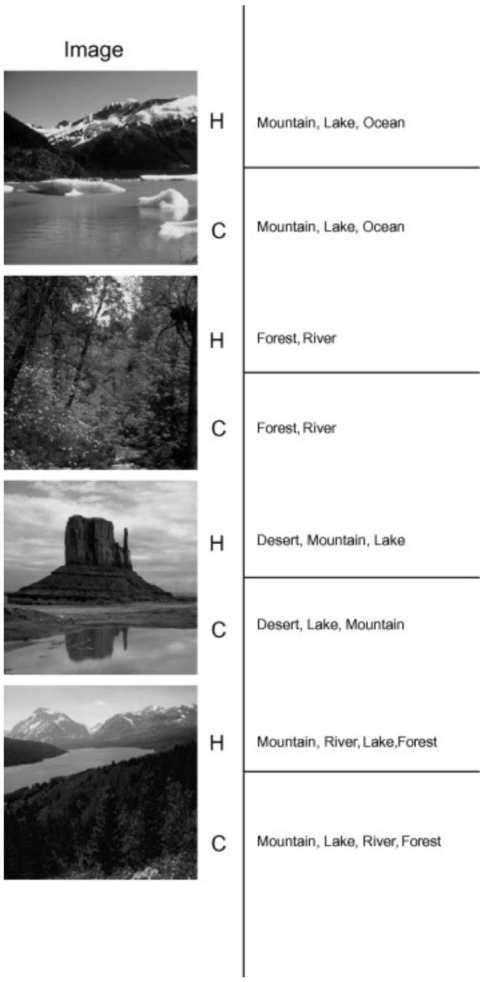

Figure 8.

Examples of non-prototypical images. Human observers ranked the images according to their prototypicality along one or more categories (Appendix A.3). For all examples (H) indicates the order of prototypicality given by the human observers and (C) is the order of classification given by the global property classifier. Although the classifier rates the probability of the image being in each category, we show only the top choices for the same number of categories ranked by the human observers. In other words, if the human observers gave prototypicality rankings for two categories, we show the top two choices of the classifier.

Fig. 9.
Examples of segmentations and annotations made using the LabelMe annotation tool, and used as the basis for the local scene representation in Experiment 4.

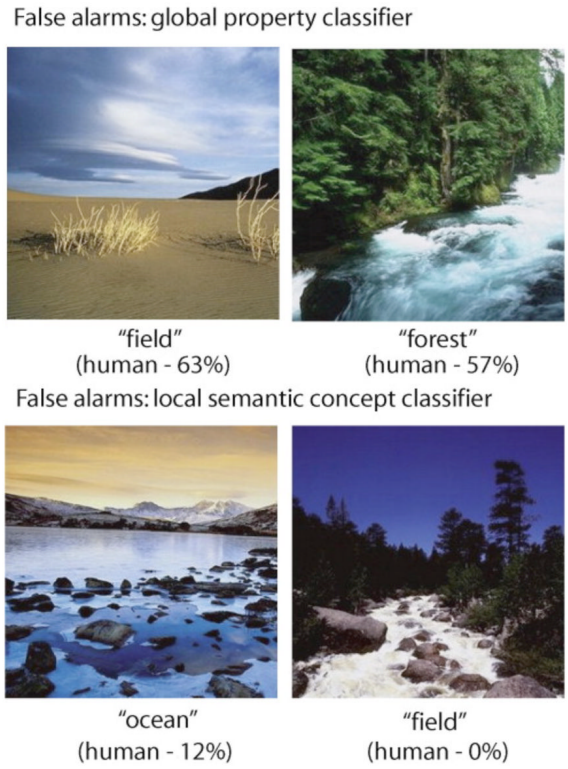

Figure 10.
Examples of false alarms made by the global property classifier of Experiment 3 and the local semantic concept classifier of Experiment 4. Underneath, we report the percent of human false alarms made on that image. The global property classifier captures the majority of false alarms made by human observers while the local semantic concept classifier captures less (see Table 5).

Table 1

Description of the seven global properties of natural scenes used in Experiments 1,2 and 3

Structural properties
Openness [1,2,3,4] represents the magnitude of spatial enclosure. At one pole, there is a clear horizon and no occluders. At the other pole, the scene is enclosed and bound by surfaces, textures and objects. Openness decreases when the number of boundary elements increases
Expansion [1] refers to the degree of linear perspective in the scene. It ranges from a flat view on a surface to an environment with strong parallel lines converging on a vanishing point
Mean depth [1,3] corresponds to the scale or size of the space, ranging from a close-up view on single surfaces or object to panoramic scenes
Constancy properties
Temperature [2,4] refers to the physical temperature of the environment if the observer was immersed in the scene. In orther words, it refers to how hot or cold an observer would feel inside the depicted place.
Temperature [4,5,7] refers to the rate at which the environment depicted in the image is changing. This can be related to physical movement, such as running water or rustling leaves. It can also refer to the transience of the scene itself (fog is lifing, sun is setting). At one extreme, the scent identify is changing only in geological time, and at the other, the identity depends on the photograph being taken at the exact moment.
Functional properties
Concealment [4,6] refers to how efficiently and completely a human would be able to hide in a space, or the probability of hidden elements in the scene that would be difficult to search for. It ranges from complete exposures in a sparse space to complete concealment due to dense and variable surfaces and objects.
Navigability [2,4,5] corresponds to the ease of self-propelled movement through the scene. This ranges from complete impenetrability of the space due to clutter, obstacles or treacherous conditions to free movement in any direction without obstacle.

The numbers refer to additional references describing the properties ([1] Oliva and Torralba (2001); [2] Gibson (1979); [3] Torralba and Oliva (2002); [4] Greene and Oliva (2006); [5] Kaplan 91992); [6] Appleton (1975)).

Table 2

Spearman's rank-order correlations along with standard error of the mean between observers for each global property from the rankings given in Experiment 1

	Openness	Expansion	Mean depth	Temperature	Transience	Concealment	Navigability
r	0.83	0.64 ± 0	0.76	0.73	0.61	0.65	0.69
sem	0.01	0.01 ± 0	0.01	0.01	0.01	0.01	0.01

Table 3

Overall human performance in rapid categorization task of Experiment 2

	Hit	False alarm	*d′*
Desert	0.83 (0.88)	0.18 (0.17)	1.88 (2.13)
Field	0.77 (0.88)	0.30 (0.20)	1.27 (2.02)
Forest	0.88 (0.96)	0.17 (0.11)	2.23 (2.97)
Lake	0.74 (0.91)	0.26 (0.18)	1.32 (2.28)
Mountain	0.78 (0.88)	0.25 (0.17)	1.50 (2.15)
Ocean	0.68 (0.87)	0.27 (0.25)	1.11 (1.79)
River	0.69 (0.89)	0.30 (0.23)	1.03 (1.97)
Waterfall	0.91 (0.95)	0.20 (0.16)	2.29 (2.67)

Table 4

Average human correct rejection performance for both experimental groups in Experiment 2 on distractor images arranged from smallest distance to target category prototype to largest

Quartile	25	50	75	100
% Correct rejection (meta-observers)	71.9	74.5	78.1	82.1
% Correct rejection (complete-observer)	75.9	78.1	84.4	88.5

Performance suffers with decreasing distance to target prototype, but remains above chance.

Table 5

A summary of performance of local region-based models tested in Experiment 4 with the global property model of Experiment 3

	Percent correct (%)	By-category correlation	Item analysis correlation	Between-category confusion correlation
Prominent object model	52	0.55	0.69	0.06
Local semantic concept model	60	0.64	0.69	0.23
Global property model	77	0.88	0.76	0.77