What do we perceive in a glance of a real-world scene?

Li Fei-Fei Computer Science Department, Princeton University, Princeton, NJ, USA

Asha Iyer Division of Biological Sciences, California Institute of Technology, Pasadena, CA, USA

Christof Koch Division of Biological Sciences, California Institute of Technology, Pasadena, CA, USA

Pietro Perona Electrical Engineering Department, California Institute of Technology, Pasadena, CA, USA

What do we see when we glance at a natural scene and how does it change as the glance becomes longer? We asked naive subjects to report in a free-form format what they saw when looking at briefly presented real-life photographs. Our subjects received no specific information as to the content of each stimulus. Thus, our paradigm differs from previous studies where subjects were cued before a picture was presented and/or were probed with multiple-choice questions. In the first stage, 90 novel grayscale photographs were foveally shown to a group of 22 native-English-speaking subjects. The presentation time was chosen at random from a set of seven possible times (from 27 to 500 ms). A perceptual mask followed each photograph immediately. After each presentation, subjects reported what they had just seen as completely and truthfully as possible. In the second stage, another group of naive individuals was instructed to score each of the descriptions produced by the subjects in the first stage. Individual scores were assigned to more than a hundred different attributes. We show that within a single glance, much object- and scene-level information is perceived by human subjects. The richness of our perception, though, seems asymmetrical. Subjects tend to have a propensity toward perceiving natural scenes as being outdoor rather than indoor. The reporting of sensory- or feature-level information of a scene (such as shading and shape) consistently precedes the reporting of the semantic-level information. But once subjects recognize more semantic-level components of a scene, there is little evidence suggesting any bias toward either scene-level or object-level recognition.

Keywords: perception, natural scene, real-world scene, indoor, outdoor, sensory-level perception, segmentation, object recognition, subordinate, entry level, superordinate, object categorization, scene categorization, event recognition, free recall

Citation: Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision,* 7(1):10, 1–29, http://journalofvision.org/7/1/10/, doi:10.1167/7.1.10.

Introduction

It is known that humans can understand a real-world scene quickly and accurately, saccading many times per second while scanning a complex scene. Each of these glances carries considerable information. Filmmakers have long exploited this ability through a technique called "flash cut." In a commercial motion picture called *The Pawnbroker* (Lumet, 1965), S. Lumet inserted an unusually brief scene that represented a distant memory. Lumet found that a presentation lasting a third of a second, although unexpected and unrelated to the flow of the main narrative, was sufficient for the audience to capture the meaning of the interposed scene (Biederman, Teitelbaum, & Mezzanotte, 1983).

Pioneering studies extended these anecdotal findings. Potter (1976) and Potter, Staub, Rado, and O'Connor (2002) utilized rapid serial visual presentations of images and revealed that subjects could perceive scene content in less than 200 ms. Furthermore, Potter demonstrated that although the semantic understanding of a scene is quickly extracted, it requires a few hundred milliseconds to be consolidated into memory (Potter, 1976). Later studies documented limits to our perception of a scene. Rensink, O'Regan, and Clark (1997) showed that changes to retinotopically large portions of the scene will sometimes go unobserved. It is likely that this occurs if the regions are not linked to the scene's overall "meaning."

Other hallmark investigations attempted to elucidate the information involved in this "overall meaning"; their conclusions regarding scene perception paralleled concepts in auditory studies of sentence and word comprehension. Biederman et al. found that recognition of objects is impaired when those objects are embedded in a randomly jumbled rather than a coherent scene (Biederman, 1972). They identified several physical (support, interposition) and semantic (probability, position, size) constraints that objects must satisfy within a scene, similar to the syntactic and grammatical rules of language (Biederman, 1982). They investigated how object recognition was modulated

doi: 10.1167/7.1.10 Received December 30, 2005; published January 31, 2007 ISSN 1534-7362

by violating these constraints. They concluded that the *schema* of a scene—or the overall internal representation of a scene that includes objects and object relations—is perceived within a single fixation (Biederman, 1982), regardless of expectation and familiarity (Biederman et al., 1983). Boyce, Pollatsek, and Rayner (1989) also demonstrated that objects are more difficult to identify when located against an inconsistent background given a briefly flashed scene (150 ms), further suggesting that both recognition of objects and global contextual understanding are quickly and deftly accomplished.

These studies show that some comprehension of a scene is rapidly attained. However, in all previous studies of scene perception, the experimenters have a set of predetermined hypotheses to test. Their experiments are hence constructed to illuminate certain parameters relevant to their claims and questions. As a number of questions are left unexplored by this approach, we propose to examine unbiased real-world scene perception as a function of display time. We have designed an experiment in which subjects view one of nearly a hundred novel natural scenes for a brief interval without any priming and pre- or poststimulus cuing, as to its content. We ask them to type freely what they have seen in as much detail as possible. We vary the presentation time (PT) of the image between 27 ms and 500 ms. Through unbiased responses, we hope to uncover new aspects of scene perception that were previously not considered. The following issues arose when we examined the free-recall responses we collected.

1. There has been no commonly accepted definition of the content of "gist." Mandler and Parker (1976) have suggested that three types of information are remembered from a picture: (i) an inventory of objects, (ii) descriptive information of the physical appearance and other details of the objects, and (iii) spatial relations between the objects. In addition to this object information, propositional relationships between objects, spatial layout of the scene, and a general impression of the low-level features that fill the scene (e.g., texture) are speculatively incorporated into the scene gist (Wolfe, 1998a). Finally, Biederman (1982) has proposed that global semantic meaning or context also contributes to the initial surmisal of a scene. Positing the "contents" of a glance as an operational definition of *scene gist*, we would like to ascertain the visual and semantic information comprising scene gist, as revealed by our subjects' responses.
2. Rosch (1978) suggested that one distinguishes between "basic-level," "superordinate-level," and "subordinate-level" object categories. Similarly, Tversky and Hemenway (1983) proposed the same taxonomy for scene categories. These authors motivate their theory with arguments of maximizing the visual and linguistic information conveyed during naming. Does human perception of natural complex scenes reveal a similar hierarchy of objects and scenes?
3. One parameter to vary in examining scene perception is the length of PTs. We are curious to see whether different percepts arise in a given temporal order.

In the Method section, we introduce in detail our experimental paradigm. We first show the images used in our experiments (Stimuli section). The Experimental Stage I section describes how we collected image descriptions. The Experimental Stage II section then explicates how these descriptions are evaluated. Five different observations are presented in the Results and Observations section. We summarize our findings and general discussions in the Conclusion section.

Method

Our subjects were asked to freely recall what they perceive in briefly displayed images of real-world scenes. We explored the evolution of our subjects' reports as a function of the length of PTs. Our data were collected in Stage I and analyzed in Stage II.

In Stage I, subjects viewed briefly a picture of a scene on a computer monitor and were then asked to type what they had seen, using a free-recall method to collect responses. The Experimental Stage I section explains the details of this stage of the experiment.

In Stage II, we asked an independent group of subjects to evaluate and classify the free-recall responses collected in Stage I. The Experimental Stage II section is a detailed account of this evaluation process.

Stimuli

In most previous studies of scene perception or object recognition, line drawings were used as stimuli (Biederman, 1982; Hollingworth & Henderson, 1999). Recently, several studies have used a large commercial database of photographs to study the perception of scenes and categories (Li, VanRullen, Koch, & Perona, 2002; Thorpe, Fize, & Marlot, 1996; Torralba & Oliva, 2003). This data set, unfortunately, is a collection of professionally photographed scenes, mostly shot with the goal of capturing a single type of objects or specific themes of scenes. We are, however, interested in studying images of everyday scenes, as commonly seen by most people in a naturalistic setting.[1] Therefore, we assembled a collection of images trying to minimize this sampling bias.

Figures 1 and 2 show our data set of 44 indoor images and 46 outdoor images collected from the Internet in the following way. We asked a group of 10 naive subjects to randomly call out five names of scenes that first came to their mind. Some of the names overlapped. After pruning, we retained about 25 to 30 different words or word

Figure 1. Forty-six images of outdoor scenes in our data set of 90 grayscale images.

Figure 2. Forty-four images of indoor scenes in our data set of 90 grayscale images.

phrases that corresponded to different environments.[2] We typed each of these words or word phrases in the Google image search engine. From the first few page(s) of search results, we randomly selected 3–6 images that depicted the keyword. The Google image search engine largely returned images found on people's personal websites, most often taken with a snapshot camera. Although everyone has a bias when taking a picture, we believe that the large number of images from different unknown sources would help average out these biases.

A number of authors have suggested that color information is not critical for the rapid categorization of scenes (Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001; Fei-Fei et al., 2005). While color could be diagnostic in a later stage of recognition (Oliva & Schyns, 2000), and uncommon colors might even hinder rapid scene categorization (Goffaux, Jacques, Mauraux, Oliva, Schynsand, & Rossion, 2005), we are mostly concerned with the initial evolution of scene perception. Thus, we decided to use only grayscale versions of our images for our experiments. It will be, however, interesting to compare our results with a future study using colored images.

Experimental Stage I: Free recall

Subjects

Twenty-two highly motivated California Institute of Technology students (from 18 to 35 years old) who were proficient in English served as subjects in Experiment Stage I. One author (A.I.) was among the subjects. All subjects (including A.I.) were naive about the purpose of the experiments until all data were collected.

Apparatus

Subjects were seated in a dark room especially designed for psychophysics experiments. The seat was approximately 100 cm from a computer screen, connected to a Macintosh (OS9) computer. The refresh rate of the monitor was 75 Hz. All experimental software was programmed using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) and MATLAB.

Procedure

Figure 3 illustrates a single trial of Stage I. An image from our data set was presented for one of seven different possible PTs: 27, 40, 53, 67, 80, 107, and 500 ms. For each trial, the particular PT was randomly selected with equal probability from these choices. The image was then masked by one of eight natural image perceptual masks, constructed by superposing white noise band-passed at different spatial frequencies (Li et al., 2002; VanRullen, & Koch, 2003). The subject was then shown a screen with the words:

> Please describe in detail what you see in the picture. Two sample responses are: 1. City scene. I see a big building on the right, and some people walking by

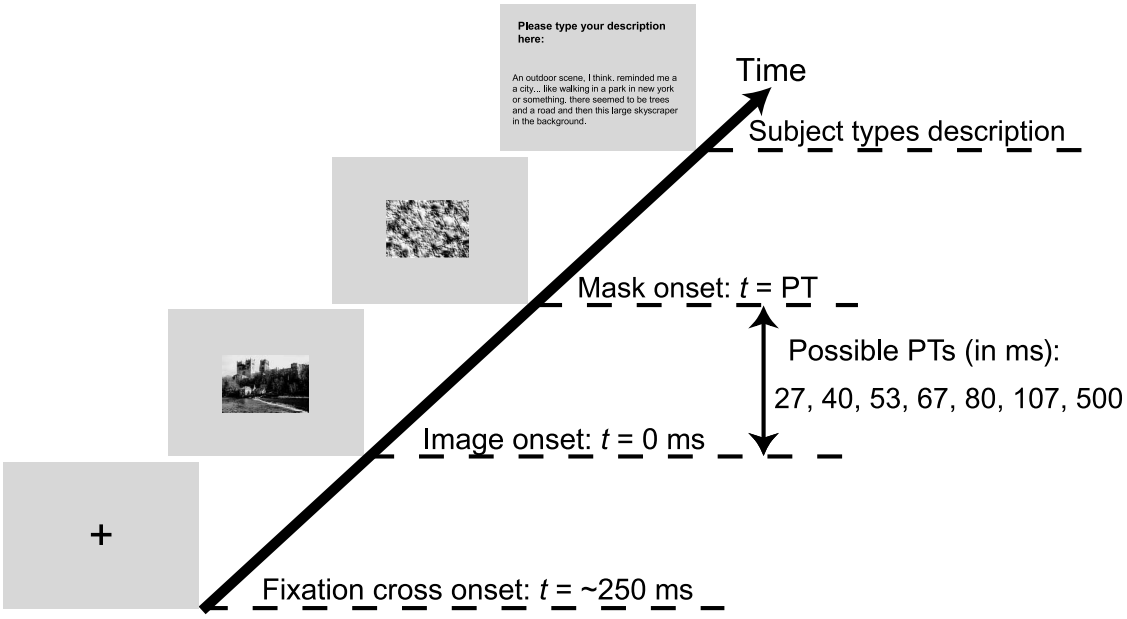

Figure 3. A single trial in Stage I: A fixation cross appeared for about 250 ms. An image from our data set was then presented at the center, subtending 6° × 8° in visual angle. After a variable PT, the image was masked by one of eight natural image perceptual masks (for details of the mask, see Li et al., 2002). The time between the onset of the image and the onset of the mask is called the PT. The mask was presented for 500 ms. Afterward, subjects were prompted to a screen in which they were asked to type in what they had seen of the image. Subjects were given an unlimited amount of time to write down their responses. When they were ready to continue, they could initiate the next trial by pressing the space bar.

shops. There are also trees. Most of the trees are on the left of the picture, against some background buildings. 2. Possibly outdoor. I really cannot tell much. Probably some animals, maybe mammals…

Subjects were given an unlimited amount of time to write down their responses.

Each subject was shown all 90 images in the database, broken into five 22-trial sessions. The images were presented in random order. At the beginning of each session, 4 images outside of the database were used to familiarize the subject with the responses and PTs. Free-recall responses for these 20 (4 × 5) images were excluded from all data analysis. Order of image presentation, as well as the choice of PT for each image, was randomized among all subjects. Each subject thus contributed one description for each image at one of the PTs. Overall, our 22 subjects provided 1,980 descriptions; that is, we obtained between 3 and 4 descriptions for each image and each PT.

Experimental Stage II: Description evaluation

Subjects

Five paid volunteer undergraduate students from different schools in the Los Angeles area (from 18 to 35 years old) served as scorers in Experiment Stage II. As scorers needed to analyze and interpret unstructured written responses, they were required to be native English speakers. All scorers were naive about the purpose of the experiments until all response evaluation was finished.

Apparatus

The scorers' task was to evaluate and classify the image descriptions obtained in the previous stage. For this purpose, they used Response Analysis software that we designed and implemented for this purpose (Figure 5). Subjects were seated in a lighted office room. The seat was approximately 100 cm from a computer screen, connected to a Macintosh (OS9) computer. The refresh rate of the monitor was 75 Hz. All Response Analysis user interface software was programmed using MATLAB and the GUI toolbox.

Procedure

Our aim was to evaluate free-recall responses in a consistent and uniform manner for all subjects. To do this, we assessed the content of all responses with respect to a standardized list of attributes.

The list of attributes was constructed by the experimenters, who examined the entire set of free-recall responses/descriptions to extract a comprehensive inventory of terms referred to in these descriptions. Most

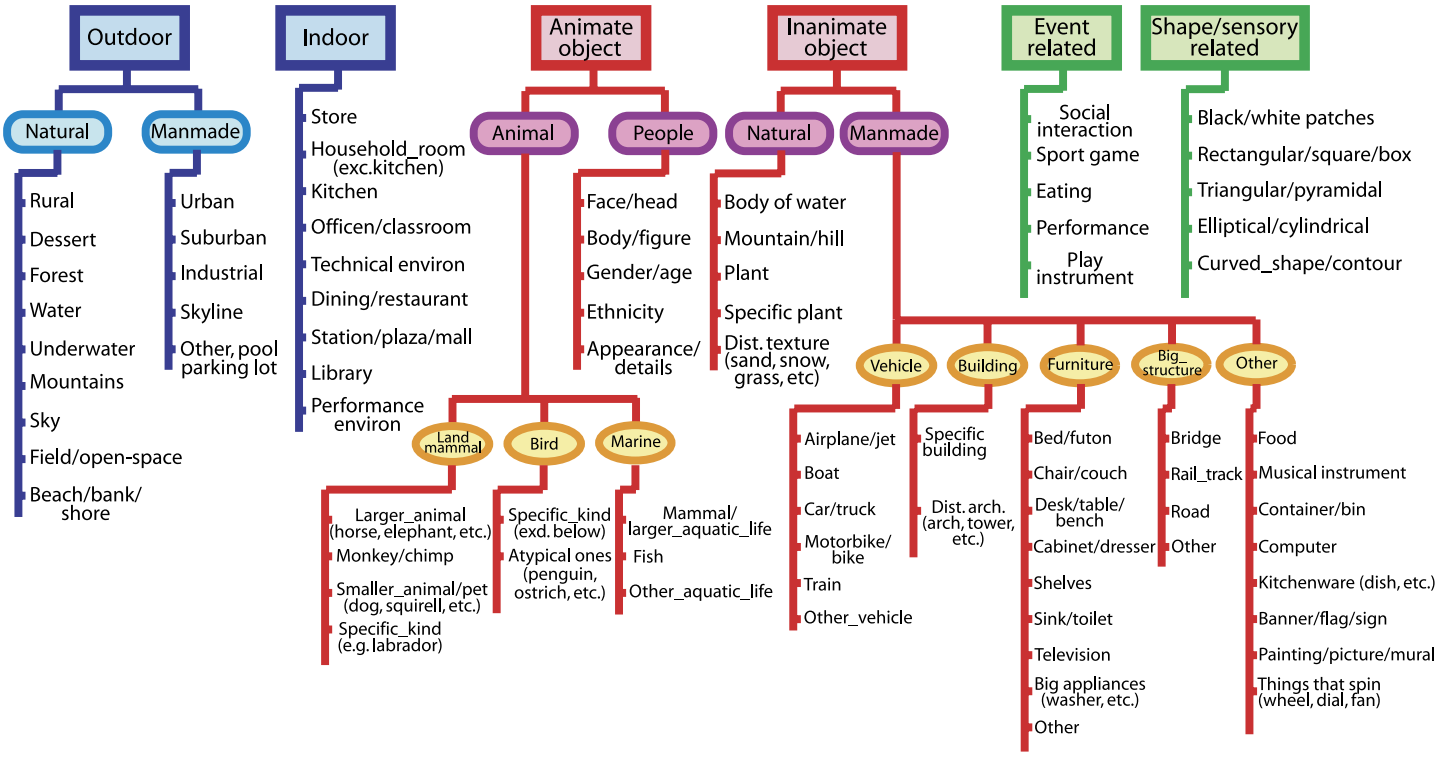

Figure 4. Attribute tree. The list of attributes was constructed by examining the entire set of free-recall responses/descriptions to extract a comprehensive inventory of terms referred to in these descriptions.

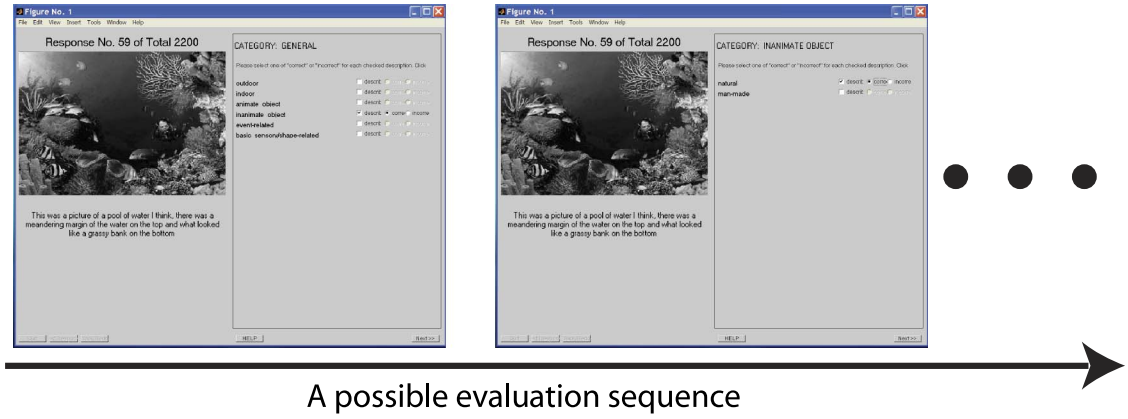

Figure 5. Experiment Stage II: Evaluating the free-recall responses. This figure is best viewed with magnification.

attributes described fell into one of six categories: inanimate objects, animate objects, outdoor scenes, indoor scenes, visual/perceptual features (i.e., shapes, lines), or event related (this category comprised a more cognitive understanding of the picture, in which human behavior related to the scene was inferred, i.e., social interaction, sports/games, performances, concert; see Figure 4 for the entire list of attributes). It goes without saying that this attribute tree is not a complete reflection of *everything* reported by subjects. We chose to focus on descriptions of sensory information (e.g., shape, shading), objects and scenes, rather than on more cognitive functionalities such as emotions. In addition, explicit verbal reports are likely to indicate a lower bound of perceived information limited by both short-term memory as well as a natural tendency of reporting more abstractly than what has been seen.

The attribute list consisted of 105 terms. We organized these attributes into a hierarchical tree structure, where the highest level represented the most general level of description (e.g., inanimate object); the intermediate stages exhibited a greater degree of specificity (e.g., manmade inanimate object, building); and the lowest level corresponded to the most detailed level of description (e.g., Capitol building). This taxonomy schema stems from conventional notions of object and scene categorization, as originally developed by Rosch (1978) and Tversky and Hemenway (1983), predicated on the superordinate level, the entry (or basic) level, and the subordinate level. The findings of these authors formed the basis of our hierarchical classification for the animate object, inanimate object, indoor, and outdoor branches of the tree. The last two branches—sensory related and event related—have received less investigation and, thus, were classified parsimoniously with only two levels: more general (e.g., sensory related) and more detailed (e.g., lines, shapes).

Each of the five scorers read every response (22 subjects who each responded to the same 90 images, for a total of 1,980 responses) and assayed them for mention or description of each attribute as well as correctness. The scorer was guided through this task with the Response Analysis interface tool (Figure 5). For each response, the scorer proceeded as follows: the first screen contained the text of one of the responses, the image described in the response, and a box with labels for the most general attributes: *indoor*, *outdoor*, *animate object*, *inanimate object*, *event related*, and *shape related*. Next to each attribute, a button allowed the scorer to indicate whether the attribute had been described in the written response. If an attribute was checked as "described," the scorer was additionally required to indicate whether the description of the attribute was either an "accurate" or "inaccurate" depiction of the corresponding image. This completed the first screen. For any attribute checked, a successive screen was displayed, which, again, comprised the text of the response and the image, but instead of the general attributes, the next level of more detailed attributes was used; for example, if *inanimate object* had been checked in the first screen, a following screen would have contained the labels *manmade* and *natural* (Figure 4), for which the user would again be prompted to indicate whether these attributes were described in the response, and if so, whether they were accurately or inaccurately described. If the user had then checked *natural*, a following screen would have contained the text of the response, the image, and the next level of attributes: *body of water*, *plant*, *specific plant*, *mountain/hill*, and *distinctive texture*. The entire branch was thus traversed.

If, on the first screen, the scorer had also checked *indoor*, then subsequent screens would have also displayed the text of the response, the image, and the next level of attributes: *store*, *household room*, *kitchen*, *office/classroom*, *technical environment*, *dining/restaurant*, *station/plaza*, *library*, and *performance venue*. In this manner, the relevant portions of the tree were traversed, one branch at a time. This process was repeated for each response.

As explicated earlier, three to four responses were provided for a given image at a given PT. For a given attribute, each scorer judged whether each of these three to four responses accurately described the attribute in the respective image. The percentage of responses rated as

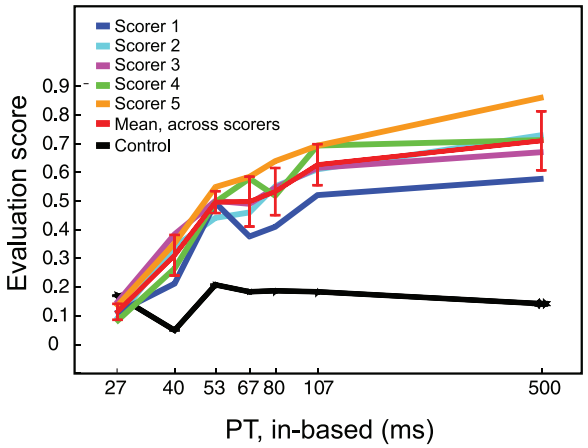

Figure 6. A sample score plot for the *building* attribute.

accurate measured the "degree" to which the attribute was perceived in this image. This initial score thus reflected a particular image, PT, and scorer. The scores were then normalized: The seven scores for a given image (one for each PT) were divided by the highest score achieved for that image (across all PTs). All evaluation scores were therefore between 0 and 1. Due to this "within-image" normalization, inherent differences in "difficulty" of perceiving or understanding scenes between different images were eliminated.

These scores were then utilized in three general kinds of analyses, depending on the issues we were interested in exploring. Most questions we asked fall into the realm of characterizing the content of subject's perception. Hence, in the first type of analysis, the evaluation scores were further averaged over images so that the averaged evaluation score represented the degree to which the attribute was perceived at a given PT across the entire image set. Finally, the scores were averaged over all five scorers.

Because this is the primary analysis employed, we will focus on the evaluation of one attribute, *building*, to better illustrate the parameters just discussed (depicted in Figure 6).

On the *x*-axis are the seven PTs for which images were displayed. The *y*-axis reflects normalized accuracy evaluation score. For the PT of 80 ms, for example, each scorer sees roughly three responses for each image. For each response, the scorer determines whether the attribute *building* was accurately reported with respect to the corresponding image (the other 104 attributes were also checked, but we will not follow those for the purposes of this example.) Suppose that the scorer indicates that building was described accurately in only one response. The initial evaluation score for the attribute *building* for this image at PT 80 ms is therefore 1/3 or 0.33. Suppose also that the maximum accuracy score achieved in describing this image occurred at PT 500 ms, where two thirds of the responses accurately reported a building. This maximum score of 0.67 would be used to normalize all scores so that the evaluation score PT 80 ms is now 0.5 and the score at 500 ms is 1.0. This normalization allows each image to be its own baseline; therefore, differences in the quality of the image (i.e., simple vs. cluttered, see Figure 7) will not affect scores. Finally, all normalized *building* scores at PT 80 ms—one for each image—are averaged to obtain the final evaluation score at this PT for this particular scorer.

This process of normalization per image and then averaging over all images is done for each PT. Again, the resulting values are per scorer. Thus, in Figure 6, the yellow, blue, green, cyan, and magenta lines each represent the normalized evaluation scores (averaged over images) for one scorer. These curves are then averaged over all the scorers. The resulting means are plotted in the red line in Figure 6, with error bars representing standard error of the mean.

PT = 107 ms

This is outdoors. A black, furry dog is running/walking towards the right of the picture. His tail is in the air and his mouth is open. Either he had a ball in his mouth or he was chasing after a ball. (Subject EC)

PT = 500 ms

I saw a black dog carrying a gray frisbee in the center of the photograph. The dog was walking near the ocean, with waves lapping up on the shore. It seemed to be a gray day out. (Subject JB)

Inside a house, like a living room, with chairs and sofas and tables, no ppl. (Subject HS)

A room full of musical instruments. A piano in the foreground, a harp behind that, a guitar hanging on the wall (to the right). It looked like there was also a window behind the harp, and perhaps a bookcase on the left. (Subject RW)

Figure 7. Subject description samples. In the first row, the scene is relatively easy. Subjects are nearly as good at perceiving the details of the scene at PT 107 ms compared to PT 500 ms. In the second row, the scene is more cluttered and complex.

In addition, there is a black line resting at the bottom of the plot. It consists of scores given by our scorers when the responses/descriptions are randomly matched to the images. This serves as a control in the response evaluation process. As this evaluation process is subjective, scorer bias in judging accuracy of responses could be a potential confound; that is, a scorer might be inclined to generally interpret vague or nebulous responses as "probably correct," giving "the benefit of the doubt" even for inaccurate descriptions. To probe for this bias, we presented each scorer with 220 responses that were paired with an incorrect image (e.g., not the image the subject was viewing when making the response). The scorer had to indicate whether the response accurately described the image with which it was presented, the same task as for the real response–image pairings. Because these are incorrect pairings, responses associated with longer PTs will not contain a more accurate description of any attribute (in this case, *building*) of the image with which it is presented to the scorer. Therefore, assuming no scorer bias, the line should remain low and flat, as observed in Figure 6. The control curves from all scorers were averaged.

Weibull cumulative density functions are also fitted to the evaluation scores for each attribute to further confirm trends indicated by the scores across PTs (see Appendix C for details).

The second kind of analysis is performed to contrast subjects' perception with reality, that is, to determine if systematic discrepancies exist between the stimulus and the perception of the stimulus. For this objective, each image is examined separately, and the normalized evaluation scores for that image are compared with the "ground-truth" classification of that image (Observation II results from this kind of analysis). As an example, we take the attributes *indoor* and *outdoor*. The process of arriving at a normalized evaluation score for each image at a given PT has already been explicated; these scores are then averaged over all scorers, reflecting essentially the percentage of responses indicating that the attribute was perceived in this image. Ground-truth is determined in the following way: for each image, we take all responses of all subjects at PT 500 ms. If most of the subjects accurately described the image as "outdoor," then the ground-truth label for the image is outdoor. The same is true for the "indoor" images. For each PT, a scatter plot is generated (e.g., Figure 9a). On the *x*-axis, we plot the percentage of responses describing the image as outdoor, and the *y*-axis reflects the percentage of responses describing the image as indoor. Each dot represents an image—red dots correspond to ground-truth outdoor images, green dots to ground-truth indoor images. In this way, we can observe how subjects perceive ground-truth indoor and outdoor images and how this perception changes as a function of PT (more detailed explanation follows in Observation II).

Our third form of analysis investigated the correlation between subjects' perception of various kind of attributes. In particular, we were interested in subjects' perception of scene context versus their perception of objects within the scene and whether separate and independent mechanisms operated for these two kinds of perception (Observation IV employs this kind of correlation analysis.) To do this, we created a scatter plot for each PT (e.g., Figure 17a). Each dot on the scatter plot represents one image. One of the attributes, for example, *scene*, is taken as a benchmark. The red dots represent the images with the top 20% of evaluation scores for *scene*, at the baseline condition (PT 500 ms). The green dots are the images with the lowest 20% of evaluation scores for *scene* at the baseline condition. The black dots represent the remaining images. These images' evaluation scores for the *scene* attribute are plotted according to the *x*-axis; their *object* attribute scores are plotted against the *y*-axis. On each scatter plot, we also show the correlation coefficient (between *scene* and *object* scores) computed across all images. This can be done for any pair of attributes.

We now detail the observations that followed from these various analyses.

Results and Observations

Observation I: The "content" of a single fixation

How much of a scene can be initially perceived within the first glance?

Bar and Ullman (1996) and Friedman (1979) proposed that early scene recognition involves the identification of at least one "obligatory" object. In this "priming model," the obligatory object serves as a contextual pivotal point for the recognition of other parts of the scene (Henderson & Hollingworth, 1999). There is also evidence that objects could be independently recognized without facilitation by global scene context (Henderson & Hollingworth, 1999). Biederman's findings however implied that some kind of global context of the scene is registered in the early stages of scene and object recognition (Biederman, 1972). Given the discrepancy between all these models, it is unclear whether the first glance of a scene comprises a mere list of objects, relations of objects, and/or more global information such as background textures and/or layout of space (Wolfe, 1998a).

From subjects' reports of scenes in a single fixation, we try to extract as much information as possible to shed light on this question. While the average fixation length during scene viewing can be as high as 339 ms (Rayner, 1984), numerous previous studies have used PTs between 100 and 200 ms to investigate the effect of single fixation (Biederman, 1982; Boyce et al., 1989; Potter, 1976). Here, we follow the tradition and use 107 ms as an estimate of the length of the first fixation of a scene. Five hundred milliseconds is chosen as a baseline PT for viewing a

scene. It is commonly accepted that this amount of time is sufficient for perceiving a natural scene and most of its contents (e.g., Biederman et al., 1983; Potter, 1976; Thorpe et al., 1996). It is also worthwhile to point out that the 500-ms baseline value is, in a way, too rigorous a criterion. As opposed to the 107-ms viewing time, subjects can make a few saccades within 500 ms. The ability to make eye movements affords them a disproportionate advantage to access visual information from the scene beyond just a longer PT. Our subsequent findings, therefore, are more likely to be a lower limit, and not an upper limit, of the perceived contents. Figure 7 shows two different example scenes and sample descriptions at the two PTs. In the first row, the scene is grasped with relative ease. Subjects are nearly as good at perceiving the details of the scene at PT 107 ms compared to the baseline viewing condition. In the second row, the scene is much more cluttered and complex. We see that the extra PT for PT 500 ms helps greatly in perceiving the details of the scene.

Several attributes were examined, from five branches of the analysis tree and at various levels of abstraction, from superordinate to subordinate. The evaluation scores for each of these attributes were averaged over all images and all scorers. The scores for PT 107 ms and for PT 500 ms were compared; a pair of bars representing the scores at these two PTs is plotted for each attribute of interest.

In Figure 8, we summarize general trends noted through analyzing subject data. In Figures 8a and 8b, we show these comparisons for *objects*. In the superordinate category of *animate objects* (Figure 8a), many attributes—particularly those related to *people*—are equivalently perceived within a single fixation as compared to the baseline viewing condition. Three attributes differ weakly in a one-way ANOVA: *animal*, $F(1,8) = 7.70$, $p = .024$, *mammal*, $F(1,8) = 6.16$, $p = .04$, and *gender*/*age*, $F(1,8) = 9.73$, $p = .01$, and two others strongly differ: *bird*, $t(8) = 73.32$, $p < .001$, and *dogs*/*cats*, $F(1,8) = 33.98$, $p < .001$ (one-way ANOVA). Whereas several detailed attributes of people, such as ethnicity, appearance, and body figures, are perceived with adroitness, recognition of nonhuman animals does not appear to enjoy the same ease. Entry-level animals such as dogs, cats, and birds are more reliably discriminated with longer PTs, with dogs and cats being particularly poorly recognized at 107 ms. These propensities speak to a large body of literature claiming an advantage for visual processing of faces and humans (Farah, 1995; Farah, Wilson, Drain, & Tanaka, 1998; Ro, Russell, & Lavie, 2001; Downing, Jiang, Shuman, & Kanwisher, 2001).

Figure 8b displays the trends for the inanimate objects contained in the image data set. Several attributes pertaining to inanimate object categories are perceived within a single fixation, namely, the superordinate category *inanimate natural objects*, plus more basic-level objects such as *rocks*, *plants*, *mountain*/*hills*, *grass*, *sand*, and *snow*, $4.24e{-}4 < F(1,8) < 4.02$, $p > .05$ (one-way ANOVA). In the realm of manmade objects, the findings are less clear. Superordinate levels, such as *manmade inanimate object*, *furniture*, and *structures* (roads, bridges, railroad tracks), and the basic-level attribute *car* are more accurately reported at 500 ms than at 107 ms ($p < .01$), except for car, which is weakly significant, $F(1,8) = 6.10$, $p = .04$. Other superordinate- and entry-level objects, including *vehicle*, *building*, *chair*, and *desk or table*, exhibit equal accuracy at both PTs ($p > .05$). The lack of an unequivocal advantage for recognition of basic-level categories versus superordinate categories connotes a discrepancy from Rosch's (1978) study on object categories. We observe that one of the main differences between our setup and that of Rosch is the clutter and fullness of our scenes. In her study, objects are presented in isolation, segmented from background. In our setup, objects are viewed under more natural conditions, with clutter and occlusion.

Figure 8c displays comparisons for the scene environments portrayed in our data set. At PT 107 ms, subjects easily name the following superordinate-level categories: *outdoor*, *indoor*, *natural outdoor*, and *manmade outdoor*. In addition, scenes such as *office*/*classroom*, *field*/*park*, *urban streets*, *household rooms* (dining rooms, bedrooms, living rooms), and *restaurant* are recognized within a single fixation, $0.20 < F(1,8) < 5.23$, $p > .05$ (one-way ANOVA). Only shop/store and water scenes require longer presentations, $9.93 < F(1,8) < 50.40$, $p < .02$, except for sky, which is weakly significant, $F(1,8) = 6.73$, $p = .03$ (one-way ANOVA). Compared to objects then, scene context is more uniformly described by our subjects in a single fixation. Our results suggest that semantic understanding of scene environments can be grasped rapidly and accurately after a brief glance, with a hierarchical structure consistent with Tversky and Hemenway (1983).

We have seen that both objects and global scene environments can be processed given a single fixation. These attributes, however, are explicitly denoted by properties of a still image, where the physical features defining an object or the quintessential components of an environment can be readily rendered. Can a more cognitive appraisal of the transpiring scenario be inferred with the same ease? In Figure 8d, we look at attributes related to human activities and social events. Given our data set, only five types of activities are included: sport/game, social interaction, eating/dining, stage performance, and instrument playing. Of the five activities, sport/game, social interactions, and, possibly, stage performance can be reported after a single glance, $0.25 < F(1,8) < 1.54$, $p > .05$ (one-way ANOVA). Only one image each involved humans either eating or playing instruments; thus, these event-related attributes were not statistically meaningful and excluded from our analysis.

In summary, within this brief period, humans seem to be able to recognize objects at a superordinate category level as well as at a variety of basic category levels. Furthermore, a single fixation seems sufficient for recognition of most common scenes and activities, many of

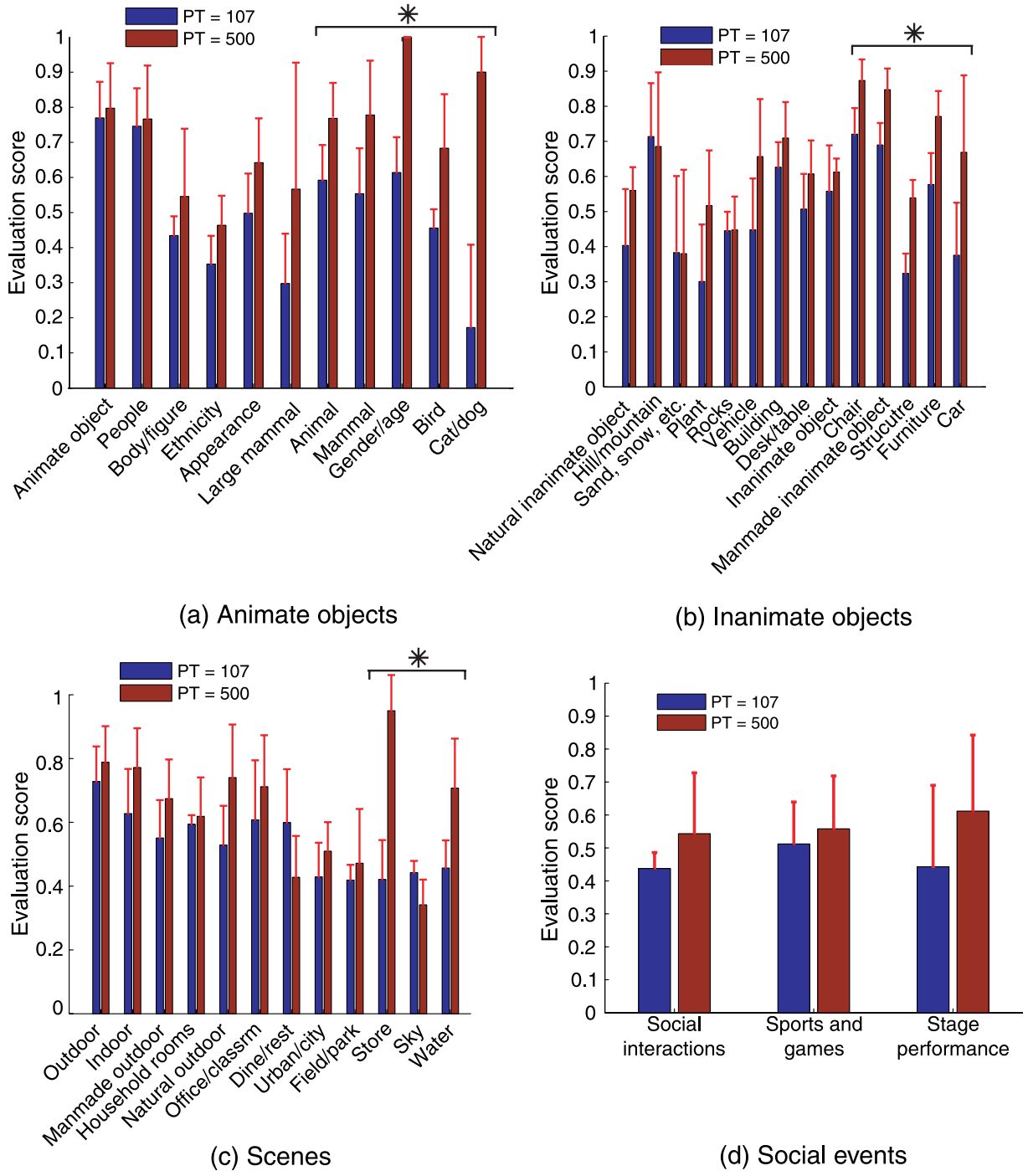

Figure 8. Fixation results for animate objects (a), inanimate objects (b), scenes (c), and social events and human activities (d).

them coinciding with the basic-level scene categories suggested by Tversky and Hemenway (1983).

Observation II: Outdoor and indoor categorization

In recent years, several computer vision studies have suggested efficient algorithms for categorizing scenes, exploiting both global and local image information (Fei-Fei & Perona, 2005; Oliva & Torralba, 2001; Szummer & Picard, 1998; Vailaya, Figueiredo, Jain, & Zhang, 2001; Vogel & Schiele, 2004). Although these methods shed light on how coarse classification of scenes can be achieved in a feed forward fashion after supervised learning, little is known in the human vision literature about the actual cues and mechanisms that allow categorization of different scene classes. In their work on scene taxonomy,

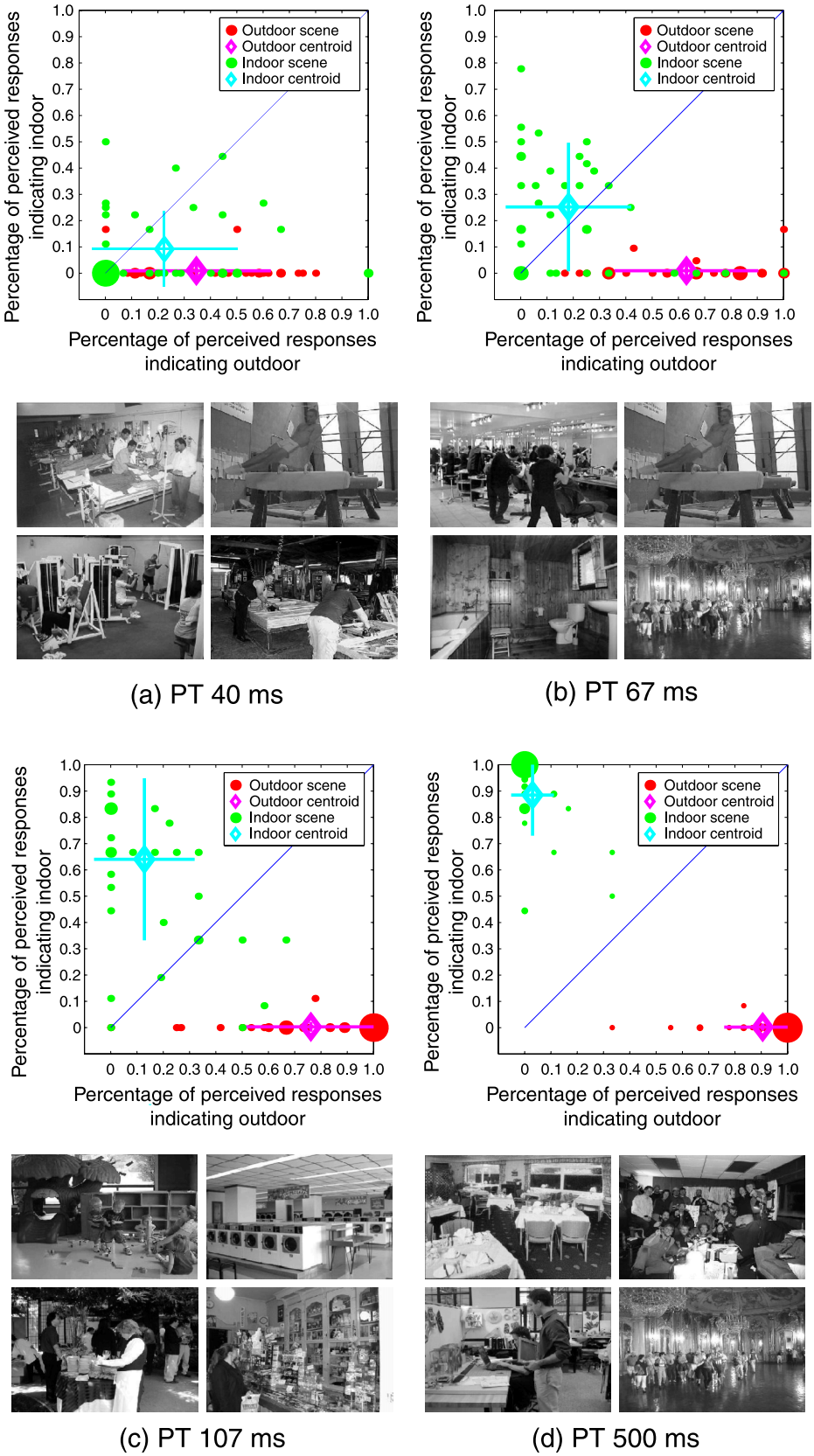

Figure 9. Categorization results of indoor and outdoor scenes. Each sub-figure illustrates the result in a specified PT. The top panel of each sub-figure is a scatter plot of the categorization results. Each dot represents an image in the database: red for ground-truth outdoor and green for ground-truth indoor. A diamond shape with error bars indicates the average performance. The bottom panel shows the four indoor images that were most often confused as outdoor scenes given this PT.

Tversky and Hemenway (1983) examined in particular people's understanding of the disparate components of indoor and outdoor scenes. Their methods, however, treated indoor and outdoor environments symmetrically, presuming no obvious preference or bias.

We examined how the outdoor and indoor images in our data set were classified by our subjects and how this classification changed as a function of PT. For each image, we are able to ascertain the percentage of subjects that labeled the image as indoor or as outdoor at a particular PT time. Figure 9 shows how the images are perceived at different times.

The recall performances for indoor versus outdoor scenes are shown in Figure 9. We sampled the responses as a function of stimulus PTs: 40, 67, 107, and 500 ms. At short PTs, few subjects mentioned the indoor/outdoor category, whereas, at 500 ms, virtually all did. At the baseline PT of 500 ms (Figure 9d), most of the red dots are located on the x-axis, as subjects correctly identified the outdoor images as outdoor. Similarly, most of the green dots are located on the y-axis. In Figures 9a–9d, we observed a very clear trend of an early bias for outdoor images. At PT 40 ms, if subjects chose to make the indoor/outdoor dichotomous distinction in their responses, they tended to identify asymmetrically indoor images as outdoor (one-tailed t test between the x-axis values of the indoor images in Figure 9a and the null hypothesis value 0, $p \ll .001$), despite the fact that there is a similar number of indoor and outdoor images in the data set. This preference for outdoor labeling continues even at PT 107 ms (Figure 9c, one-tailed t test, $p \ll .001$). In Figures 9a–9d, we also present the four indoor images that were most frequently misclassified as outdoor at the corresponding PT. Several of them are consistent over a range of PTs. By considering these images, it is possible that predominantly vertical structures give rise to the

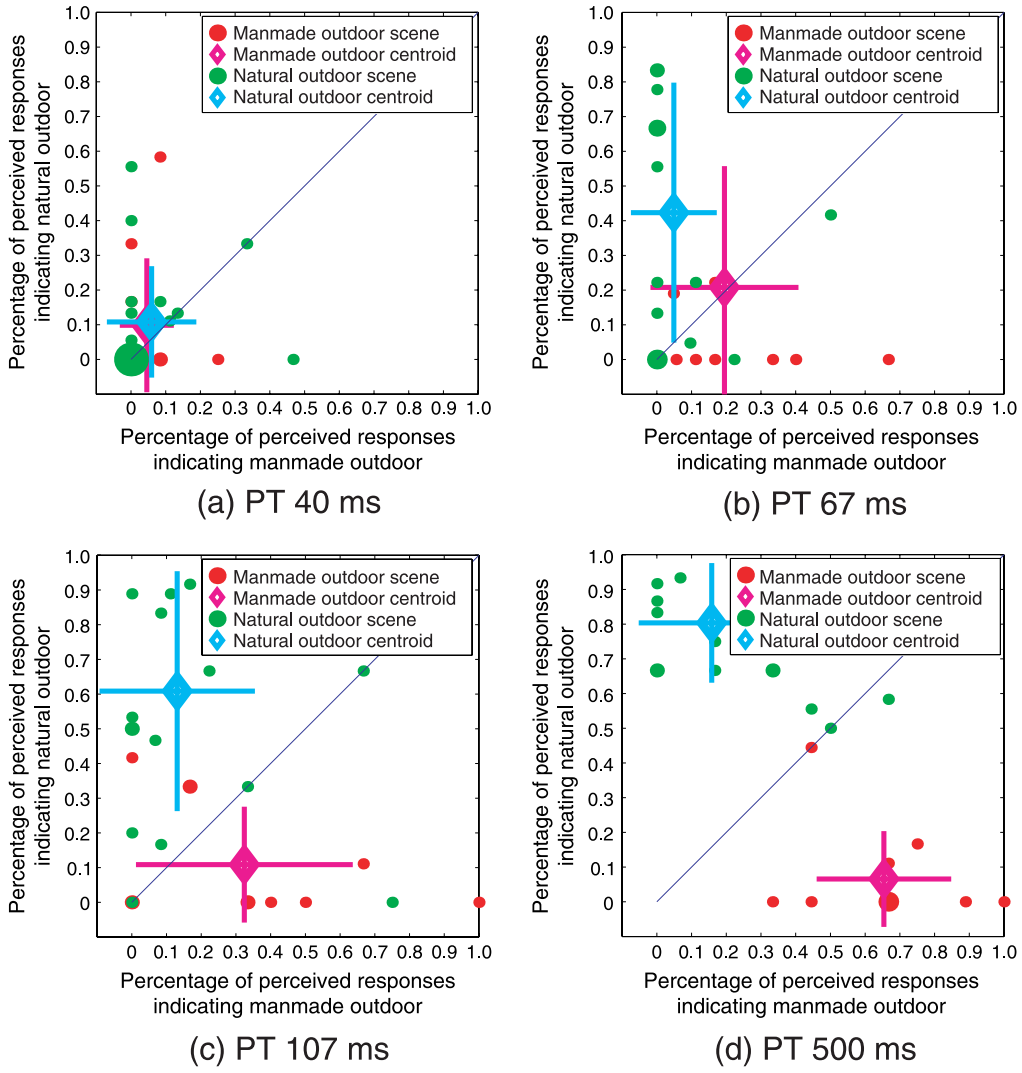

Figure 10. Categorization results of manmade outdoor and natural outdoor scenes. Each dot represents an image in the database: red dots are manmade outdoor scenes and green are natural outdoor scenes. A diamond shape with error bars is also plotted for each class of images (manmade outdoor and natural outdoor) to indicate the average percentage.

outdoor percept more easily when there is less than 107 ms for viewing the image. In Figure 11c, we summarize the change of indoor and outdoor classification over PT in one plot. Each diamond represents the average performance score at one PT.

While we observe this strong bias in favor of outdoor over indoor classification of natural scenes for short display times, we do not see a large difference between manmade outdoor over natural outdoor images (Figure 10). Subjects labeled both natural and manmade outdoor scenes with similar accuracy. Given shorter PTs (<107 ms), manmade outdoor scenes are at times confused with natural outdoor scenes, hence a lower average performance. But overall, the trend is not as pronounced as the bias between indoor and outdoor scenes (Figure 11b).

Figures 11c and 11d summarize average classification results for indoor versus manmade outdoor images and for indoor versus natural outdoor images, respectively. Unlike Figure 11a, there is no indication of a bias in any of these conditions (one-tailed t test between x-axis values of the indoor images and the null hypothesis 0, $p > .05$ for all PTs). This suggests that whereas indoor scenes tend to be confused as outdoor scenes, there is little confusion with manmade or natural outdoor scenes.

From where does this bias arise? Given the limited amount of information available when stimuli are presented very briefly (less than or about a single fixation), did outdoor pictures have an advantage over indoor pictures because subjects could perceive low-level, sensory-related information more clearly? Or was it due to greater ease in identifying objects in the outdoor scenes versus the indoor scenes, as the priming model would predict (Bar & Ullman, 1996; Friedman, 1979)? Figure 12 illustrates the evaluation results in both indoor and outdoor scenes for sensory-level information (Panel a) and object-level information (Panel b), from the shortest PT (27 ms) to the maximum (500 ms). For sensory information perception, we see that the evaluation scores

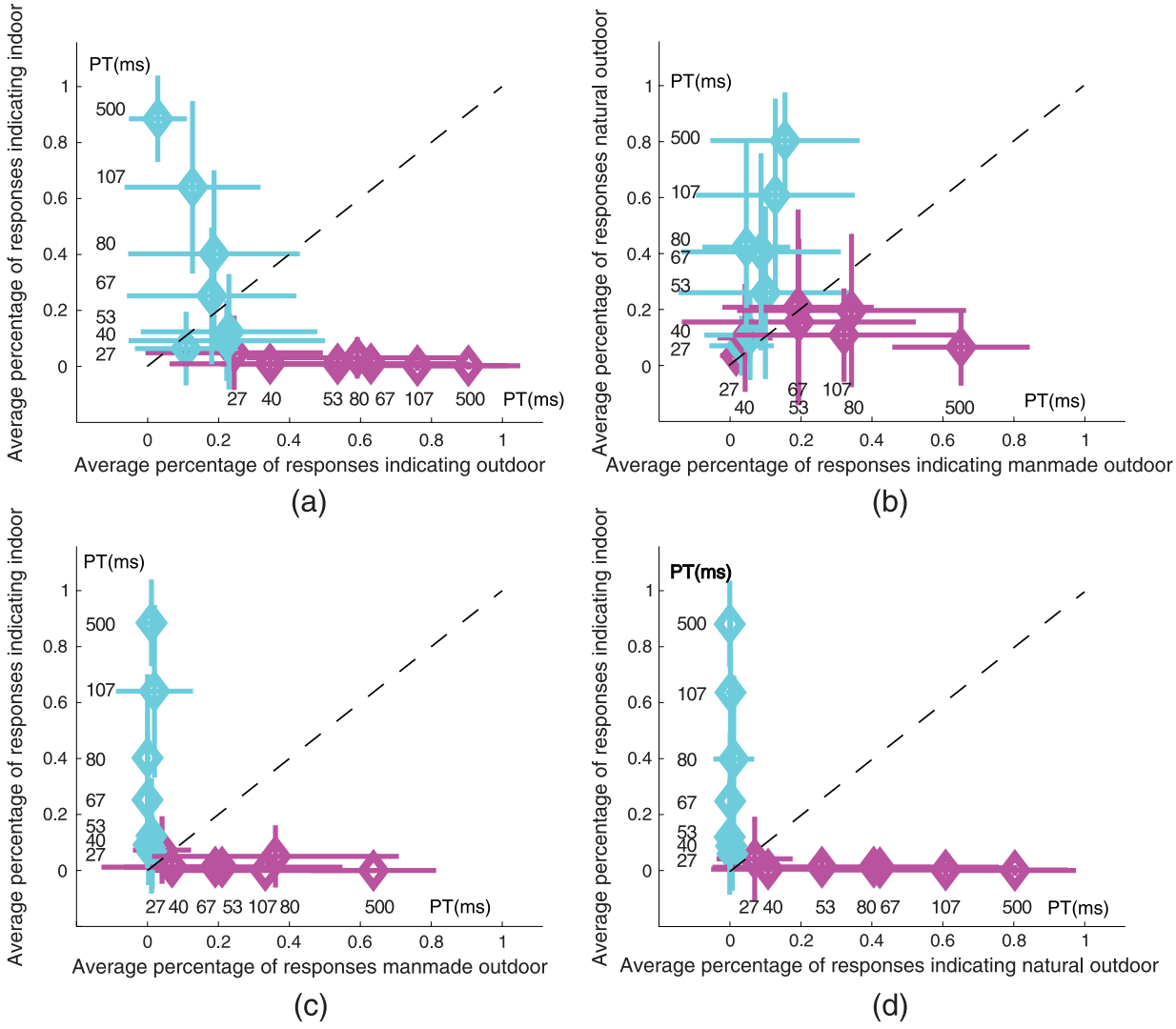

Figure 11. Summary plot of average categorization performances of all seven PTs. (a) Indoor versus outdoor scenes; (b) manmade outdoor versus natural outdoor scenes; (c) indoor versus manmade outdoor scenes; (d) indoor versus natural outdoor scenes.

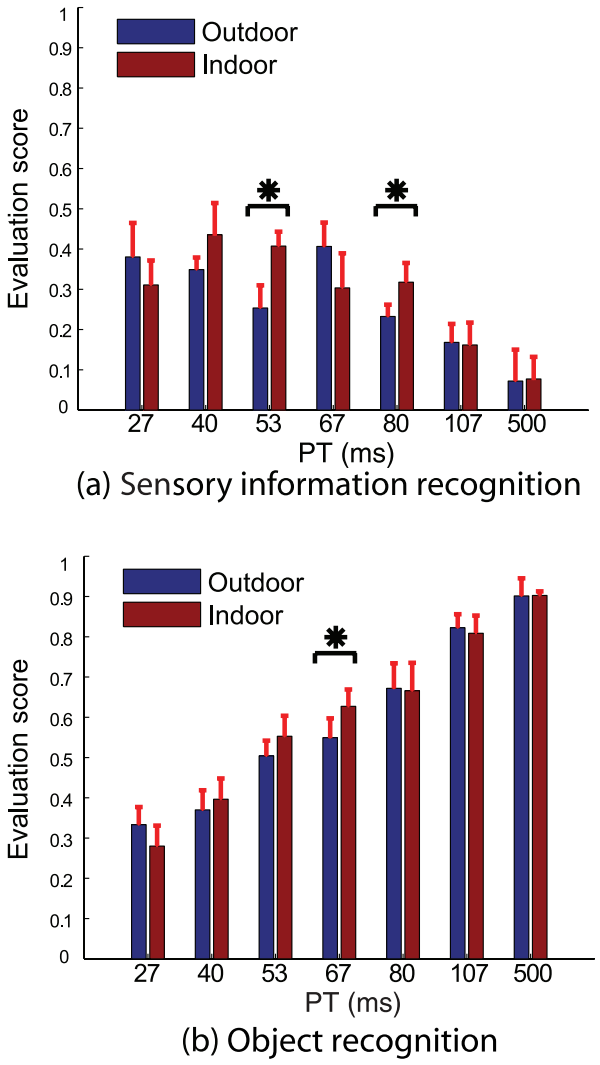

Figure 12. Sensory information and object perception in outdoor and indoor scenes. (a) Sensory information perception performance comparison between indoor and outdoor scenes across all PTs. (b) Overall object recognition performance comparison between indoor and outdoor scenes across all PTs.

for both indoor and outdoor images do not differ significantly at most PTs except for PTs 53 and 67 ms (according to Figure 12). Similarly, little trend is detected with respect to object-level perception (except for PT 67 ms). These results indicate that although there is an obvious preference for discerning outdoor images at short PTs, this bias does not seem to stem from a differential ability to perceive the sensory information or object contents of the different environments.

Lastly, we would like to rule out simple image-level cues such as global frequency or grayscale/intensity value statistics in the explanation of our observed bias. In Appendices A and B, we show that there is little sign of such bias in image-level statistics by using two simple computational models for indoor and outdoor scene categorization.

Observation III: Sensory-level recognition versus object/scene-level recognition

Humans possess a superb ability in categorizing complex natural scenes. Thorpe et al. (1996) have demonstrated that the presence of an animal (or vehicle) in a photograph can be rapidly detected by subjects, and a neurophysiological correlate of this detection is observed in the prefrontal cortex area in as little as 150 ms. Further studies suggest that a low-level, object-independent mechanism precedes the detection or recognition of semantically meaningful scene stimuli (Johnson & Olshausen, 2003; VanRullen & Thorpe, 2001).

Speaking to a similar issue, traditional models of object recognition posit that low-level visual processing precedes higher level object recognition, by which segmentation would occur before recognition (Driver & Baylis, 1996; Nakayama, He, & Shimojo, 1995; Rubin, 1958). Other evidence suggests that semantically meaningful object recognition might in turn influence low-level, object-independent segmentation (Peterson & Gibson, 1993, 1994; Peterson & Kim, 2001). Recently, Grill-Spector and Kanwisher (2005) have found that humans are as accurate at categorizing objects as at detecting their presence and concluded that both processes require a similar amount of information and the same length of neuronal processing time. A key question following these findings is that of the natural evolution of scene perception: What is the time course of object versus more global scene recognition?

The conclusions above are drawn from experiments that rely on a multiple forced-choice paradigm, in which subjects are given a short list of possible answers before viewing the image (e.g., Biederman, Rabinowitz, Glass, & Stacy, 1974). Intuition tells us that different levels of recognition might occur upon processing different levels of information. While coarser or lower frequency information might suffice for the detection of a dog, it is not necessarily adequate to identify the dog as a husky or a German shepherd. We would like to, therefore, scrutinize subjects' descriptions of natural scenes at different PTs to investigate the evolution of different levels of recognition, that is, higher level conceptual information (e.g., object identification, object categorization, scene categorization) versus low-level or "sensory" information (e.g., shape recognition/parsing).

In the Method section, we gave a detailed account of how subjects viewed and recorded their responses to each of the natural scene images in our database. Figure 13 shows three of the images and some of their free-recall responses at four different PTs. When the PT is short (e.g., PT = 27 or 40 ms), shape- and low-level sensory-feature-related (such as "dark," "light," and "rectangular") terminology predominates in the free-recall responses. As the display time increases, subjects more often identify objects as well as scene categories. More conceptual and semantic terms such as "people," "room," and "chair"



PT 27 ms	There was a range of dark sploches in the middle of the picture, running from most of the way on the left side, to all the way on the right side. This was surrounded primarily by a white or light gray color. (Subject: KM)	Couldn't see much; it was mostly dark w/ some square things, maybe furniture. (Subject: AM)	Looked like something black in the center with four straight lines coming out of it against a white background. (Subject: AM)
PT 40 ms	I saw a very brght object, shaped in a pyramidal shape. There was somethin black in the front, but I couldn't tell what it was. (Subject: JB)	This looked like an indoor shot. Saw what looked like a large framed object (a painting?) on a white background (i.e., the wall). (Subject: RW)	The first thing I could recognize was a dark sploch in the middle. It may have been rectangular-shaped, with a curved top…but, that's just a guess. (Subject: KM)
PT 67 ms	Possibly outdoors. maybe a few ducks, or geese. Water in the background. (Subject: JL)	I saw the interior of a room in a house. There was a picture to the right, that was black, and possibly a table in the center. It seemed like a formal dining room. (Subject: JB)	A person, I think, sitting down or crouching. Facing the left side of the picture We see their profile mostly. They were at a table or were some object was in front of them (to their left side in the picture). (Subject: EC)
PT 500 ms	It was definately on a coast byt hte ocean with a large [r]ock in the forground and atleast three bird sitting on the rock. (Subject: CC)	Some fancy 1800s living room with ornate single seaters and some portraits on the wall. (Subject: WC)	This looks like a father or somebod helping a little boy. The man had something in his hands, like a LCD screen or laptop. they looked like they were standing in a cubicle. (Subject: WC)

Figure 13. Samples of subjects' free-recall responses to images at different PTs.

appear with increasing frequency. We quantify the above observation by comparing the evaluation scores of the shape/sensory-related attribute, as a function of PT, with scores of more semantically meaningful attributes.

Figure 14 summarizes our results. The *y*-axis of each panel is the evaluation score of each selected attribute(s). For comparison, we plot the sensory information response in all three panels of Figure 14. The general trend in sensory information accuracy indicates that its score decreases, relative to other attributes, as the PT increases; subjects cease to report shape- or sensory-related information when they are able instead to ascribe higher level descriptions to the image. In contrast, evaluation scores for attributes such as object names and scene types rise as the PT lengthens. The accuracy and frequency with which these attributes are reported increase as more information becomes available.

In Figure 14a, we compare the responses of low-level visual/sensory information to the high-level information related to object and scene superordinate categorizations.

Sensory information dominates over object information—object, inanimate object, and animate object curves until PT 53 ms, $2.21 < F(1,8) < 36.86$, $p < .05$ (one-way ANOVA; also confirmed by Weibull fit, see Appendix C for details). Scene information is more heterogeneous: The outdoor scene attribute becomes indistinguishable to that for sensory-level information at PT 53 ms, $F(1,8) = 0.003$, $p = .96$ (one-way ANOVA; confirmed by Weibull fit, Appendix C), whereas the indoor scene curves overtake the sensory curve slightly before 80 ms, $F(1,8) = 36.86$, $p = .03$ (one-way ANOVA; confirmed by Weibull fit, Appendix C). Once again, we find an obvious advantage for accurate report of outdoor scenes over indoor scenes, confirming Observation II.

Figures 14b and 14c allow us to inspect more closely scene and object perception at finer levels of detail. While outdoor recognition begins at about 53 ms, all other levels of scene recognition transpire at approximately 67–80 ms (confirmed by Weibull fit, Appendix C). In an analogous assessment, Figure 14c displays evaluation scores as a function of PT for object information. Somewhere between 40 and 67 ms PT, various levels of object perception (except for some indoor furniture categories) become more pronounced than sensory-level information (at PT 67 ms, animate and inanimate objects are both significantly more reported than sensory information), $5.34 < F(1,8) < 7.30$, $p < .05$ (one-way ANOVA; confirmed by Weibull fit, Appendix C). This switch in the predominant information reported transpires with shorter PTs as compared to reports of scene-related attributes.

While our results cannot attest directly for the time course of information processing while viewing an image, our evidence suggests that, on average, less information is needed to access some level of nonsemantic, sensory-related

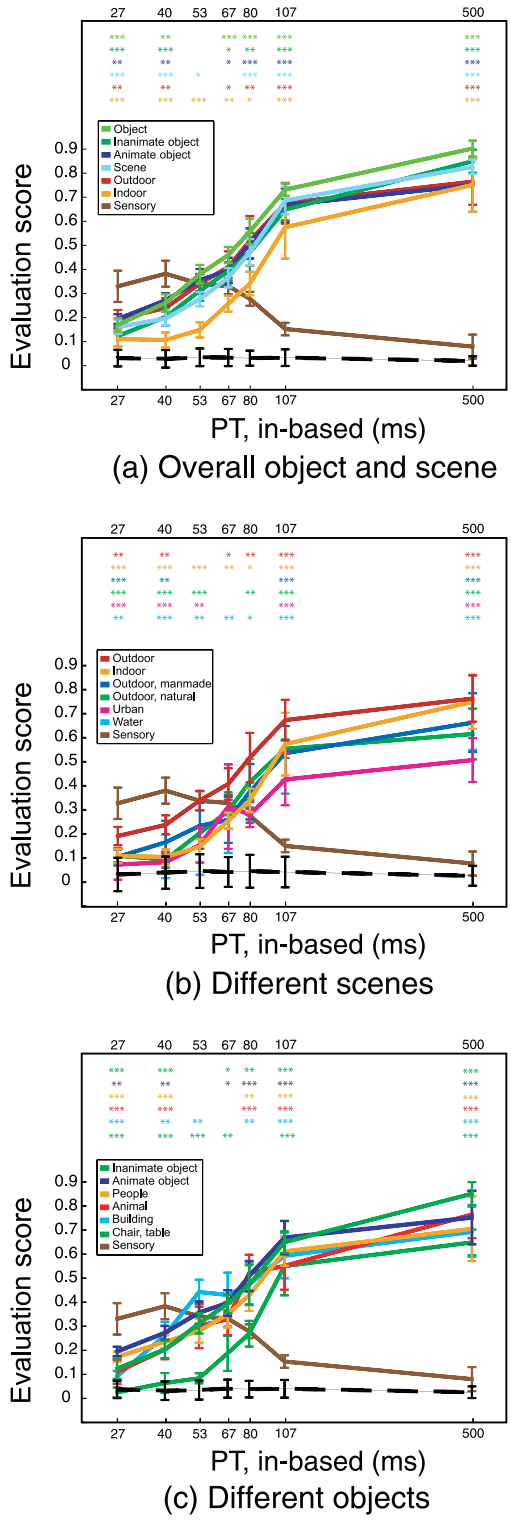

Figure 14. Perceptual performances of different attributes across all seven PTs. The perceptual performance is based on evaluation scores detailed in the Method section. The sensory-related perception is plotted as a benchmark in all three panels. Perceptual performances for (a) overall object and scene attributes, (b) scene-level attributes, and (c) object-level attributes are shown. The black line (at the bottom) of each panel corresponds to the random control responses.

information in a scene compared to semantically meaningful, object- or scene-related information. This result differs from what Grill-Spector and Kanwisher (2005) reported in their study. One major difference in our experimental design is that their subjects are forced to make a multiple choice, whereas our subjects are instructed to write down whatever they recall. In addition, in their database, scenes that contain objects have very different statistics compared to the scenes that do not contain objects, namely, randomized pixels. An earlier study by Bruner and Potter (1964) has already suggested that sensory attributes are more likely to be reported when the viewed scenes are blurred (as against nonblurred). Moreover, studies have suggested that some reliable structural information of a scene may be quickly extracted based on coarse spatial scale information (Oliva & Schyns, 2000). Consistent with these findings, our data seem to also show that coarse spatial information about shape segmentation can be perceived with less presentation of the image.

Observation IV: Hierarchies of objects and scenes

It has been shown that some level of categorization of objects is most natural for identifying the object as well as for discriminating it from others. Rosch developed this category hierarchy for object recognition and identification; Tversky and Hemenway (1983) suggested a similar taxonomy for natural environments. We were therefore interested to see if any correlation existed between our subjects' reports of scene and object recognition, as a function of PT, and the findings in the studies of Rosch (1978) and Tversky and Hemenway (1983). We follow the same method described in the Observation III section and the Experimental Stage II section to track perceptual content of subjects' responses over time.

First, we explored the relationship between levels of the *animate object* hierarchy. We show in Figure 15a three levels of animate objects: the superordinate levels *animate objects*, *animal*, and *mammal*. At PT 27 ms, there exists an advantage for more accurate and frequent report of animate objects versus the other three categories ($p < .05$; confirmed by Weibull fit, Appendix C). This advantage decreases by PT 40 ms, although it still retains statistical significance with respect to *animal* and *large mammal*: *animal*, $F(1,8) = 9.99$, $p = .01$, *mammal*, $F(1,8) = 1.25$, $p = .30$, *large mammal*, $F(1,8) = 6.55$, $p = .03$ (one-way ANOVA; confirmed by Weibull fit, Appendix C). In short, given a very limited amount of information, subjects tend to form a vague percept of an animate object, but little beyond that.

A comparable advantage is found for manmade inanimate objects. Figure 15b shows that while the evolution of structure and road/bridge are very similar, subjects tend to accurately report an overall impression of a manmade

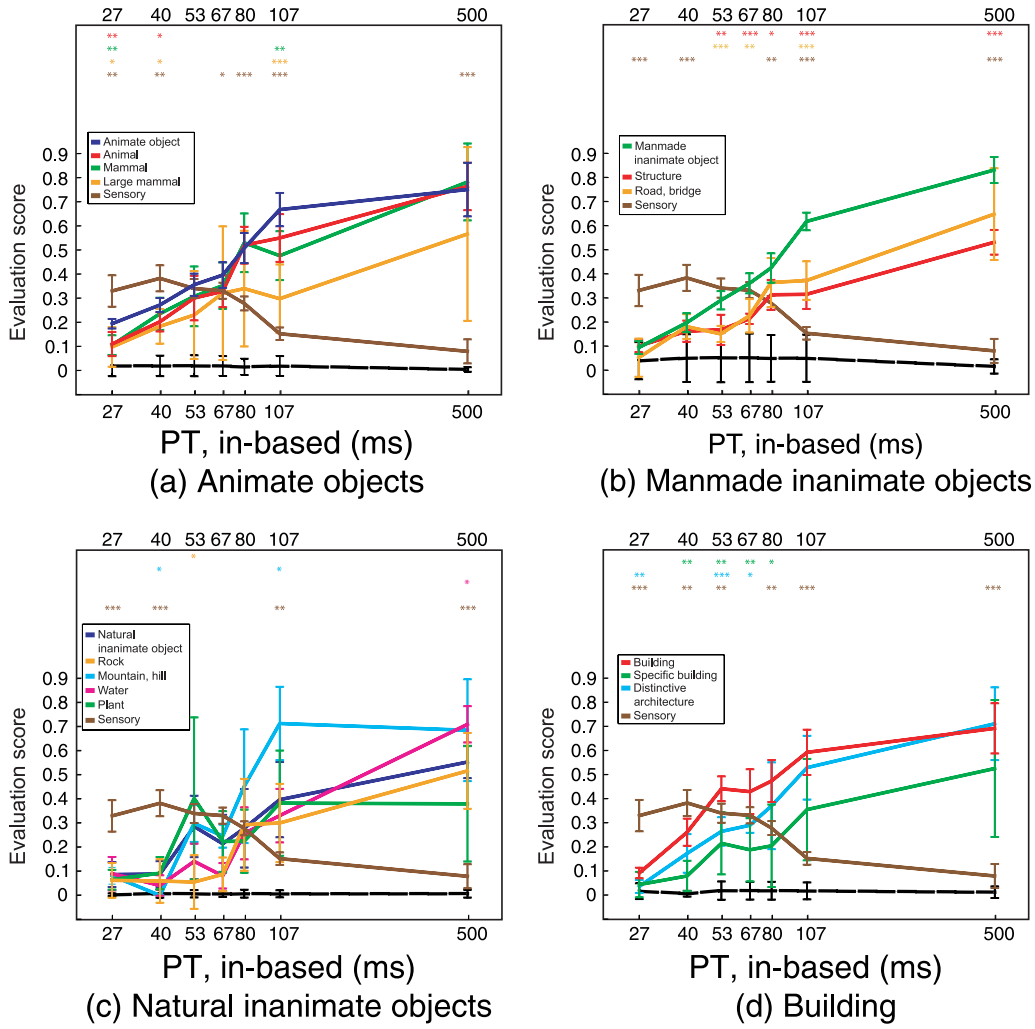

Figure 15. Perceptual performances of different object attributes across all seven PTs. The perceptual performance is based on evaluation scores detailed in the Method section. The shape-segmentation-related perception is plotted as a benchmark in all three panels. (a) Animate-object-related attributes; (b) manmade-inanimate-object-related attributes; (c) natural-inanimate-object-related attributes; (d) building and subordinate building categories.

inanimate object rather than provide a more detailed level of categorization. At short PTs (27 and 40 ms), recognition of all levels of this hierarchy is poor. With longer PTs (from PT 53 ms onward), recognition improves, preferentially for the most superordinate level of "manmade inanimate object" (significantly greater than structure and road/bridge for PTs 53–500 ms, $p < .05$, except vs. road/bridge at 80 ms, $p = .30$, and at 500 ms, $p = .08$; confirmed by Weibull fit, Appendix C). The trend is replicated in the hierarchy of structure recognition (Figure 15d). In this plot, we observe that there is very clear gradation in terms of perception accuracy among buildings, distinctive architectural styles (e.g., Gothic building, triangular roof), and specific buildings (e.g., Capitol Hill, Golden Gate). As with Figure 15b, accuracy is poor for all levels at PT 27 ms. From 40 to 80 ms, "building" evaluation scores are significantly greater than those for the finest level of descriptive resolution "specific building" ($p < .05$; confirmed by Weibull fit, Appendix C); for the earlier part of the same interval (53 and 67 ms), building perception is also superior to the intermediate level attribute of "distinctive architectural features" ($p < .05$; confirmed by Weibull fit, Appendix C). Less overall trend is seen in natural inanimate objects, largely due to the high noise level of the plot (Figure 15c).

Our results on object hierarchies and the change of perceptual accuracy over increasing PTs are not necessarily in conflict with the findings of Rosch (1978). In her study, the goal is to determine the level of categorical representation that is most "informative" and useful to identify and distinguish an object. An unspoken assumption is that this categorization is achieved given full amount of perceptual information. In our setup, however, subjects do not have unlimited access to the images. Under this setting, coarser level object categorization is in

general more accurate than finer level ones. As more information becomes available (i.e., longer PT), this difference becomes smaller.

We adopted a similar strategy in examining the evolution of scene-related perceptions, as represented in Figure 16. Figure 16a shows, as a function of PTs, the accuracy scores of "indoor scenes" and three different "basic-level" indoor environments: "household rooms" (e.g., living room, bedroom), "office/classroom," and "dining/restaurant" (Tversky & Hemenway, 1983). Unlike the hierarchical perception of objects, different levels of indoor scenes do not exhibit clear discrepancies in recognition frequency and accuracy at any PT ($p > .05$; confirmed by Weibull fit, Appendix C). The accuracy scores for store show a minor but significant deviation from the indoor curve at lesser PTs (e.g., at 27 and 53 ms, $p < .05$). However, only three images in our data set correspond to store environments, and it is difficult to generalize from such a small sample.

Figure 16b shows the evaluation results for different levels of natural outdoor scenes (natural outdoor scene, field, beach, and water). The coarsest level of the hierarchy, "outdoor scene," has a clear advantage over all other levels from the shortest PT until about 500 ms ($p < .05$, except at 80 ms: outdoor natural, $p = .11$, water, $p = .14$; confirmed by Weibull fit, Appendix C). However, at more detailed levels of the hierarchy, the situation is analogous to the indoor scenario. Once subjects have classified an image as a natural outdoor scene, they are capable of further identifying its basic-level category. There is no statistical difference among the evaluation scores for natural outdoor and many of its subordinate categories such as field, mountains, and water (an exception is the entry-level scene "beach," which is significantly lower at all PTs until 107 ms, $p < .05$).

A commensurate hierarchical trend is observed in manmade outdoor scenes (Figure 16c). The perceptual

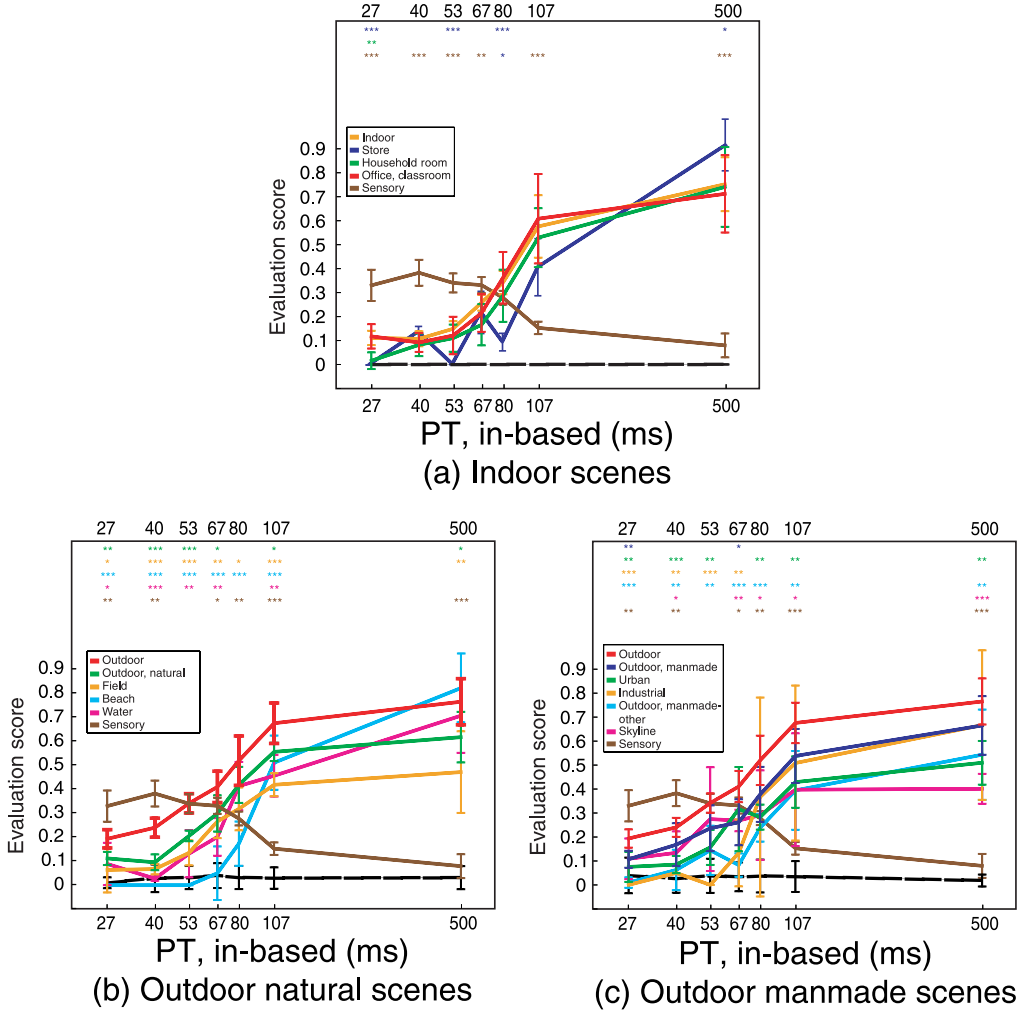

Figure 16. Perceptual performances of different scene attributes across all seven PTs. The perceptual performance is based on evaluation scores detailed in the Method section. The shape-segmentation-related perception is plotted as a benchmark in all three panels. (a) Indoor scenes, (b) outdoor natural scenes, (c) outdoor manmade scenes.

accuracy scores of manmade outdoor scene, urban centers, skylines, industrial environments, and other manmade outdoor environments are essentially indistinguishable. A few instances of significant but small differences were noted between manmade outdoor and industrial and between manmade outdoor and "other manmade" scenes, for example, for other manmade, $F(1,8) = 27.41$, $p < .001$ (one-way ANOVA; confirmed by Weibull fit, Appendix C). These categories comprised images of construction sites, parking lots, and swimming pools; such scenes have not been mapped out in terms of their taxonomy and could conceivably be specific subordinate rather than basic-level categories. This may in part account for these findings.

Tversky and Hemenway (1983) have suggested a taxonomy of scenes similar to that of objects. Their study follows a similar line of arguments as Rosch (1978). Our observations, however, imply that scene perception differs from object perception. While object recognition reveals some hierarchical structure, only the overall categorization of outdoor environment seems to need less information than recognition of other scene types. In general, superordinate-level scene categories (e.g., indoor, manmade outdoor, natural outdoor) seem to require the same amount of information in recognition as the basic-level scenes (e.g., field, beach, skyline, urban centers).

Observation V: Object and scene perception: Are they correlated?

Intuitively, much of the meaning of a scene is defined by the objects that comprise it. Biederman (1972) has shown that recognition of objects is impaired when embedded in jumbled scenes rather than coherent scenes. On the other hand, recent computational work has suggested that global features such as the spatial frequencies of the images are often sufficient for categorizing different environments without explicit recognition of the objects (Torralba & Oliva, 2003). Thus, are the objects in the scene perceived first? Or is the scene context grasped independently and perhaps prior to recognizing the objects? How are the two perceptions related? Such questions have been open for debate for more than two decades (De Graef, Christiaens, & d'Ydewalle, 1990; Germeys & d'Ydewalle, 2001; Hollingworth & Henderson, 1999).

If scene and object perception follow from unrelated and disparate mechanisms as the functional isolation model states, little correlation between the two should be observed regardless of the PT. Conversely, if they share computational resources or facilitate each other in some way, we expect a correlation between the perception of objects and scenes. Furthermore, if there is a correlation between object and scene, we would like to know how this correlation is affected by the amount of available information—in other words, how different levels of object categorization relate to overall scene perception.

We show the relationship between object-level information and scene-level information in Figure 17. Each of the eight panels in Figure 17 is a scatter plot of the evaluation scores for these two attributes. The Experimental Stage II section describes in detail how these plots were obtained. Each dot on the scatter plot represents one image. If more than one image falls on the same coordinate, the size of the dot increases linearly with the number of images. Figures 17a–17d use the *scene* attribute as a benchmark. The red dots represent the images with the top 20% of evaluation scores for *scene*, at the baseline condition (PT 500 ms). The green dots are the images with the lowest 20% of evaluation scores for *scene* at the baseline condition. The black dots represent the remaining images. From 40 to 107 ms, there is a weak correlation between the scene attribute and the object attribute, $\rho(40\text{ ms}) = 0.38$, $\rho(80\text{ ms}) = 0.26$, $\rho(107\text{ ms}) = 0.29$, suggesting that subjects will perceive objects a little more accurately when they perceive scenes more accurately. At PT 500 ms, this correlation becomes nearly 0. However, both scene and object scores cluster near the upper right corner of the plot, indicating very high accuracy of perception for both these attributes. Similar to Figures 17a–17d, Figures 17e–17h show the relationship between scene and object recognition using the *object* attribute as a benchmark. In this case, the red dots are images that have the top 20% of evaluation scores for *object* under the baseline condition, and the green dots are those images with the lowest 20% of evaluation scores. Because correlation does not reflect causality, we should obtain the same correlation score whether the object or the scene attribute is used as a benchmark. Our data in Figures 17e–17h show the same correlation scores as each of their counterpart plots in Figures 17a–17d.

We can further explore the different relationships between scene perception and various-level object attributes at different PTs (Figure 18). The *x*-axis is the log scale of PT times, ranging from 40 to 500 ms. Most of the object attributes receive very low evaluation scores at 27 ms, hence the omission. The *y*-axis is the correlation score between a given attribute (e.g., inanimate object) and overall scene perception.

Compared to objects, the inanimate object attribute possesses a much stronger correlation with scene perception (average correlation score between 40 and 107 ms is .55 for inanimate object and .30 for overall object, $p < 10e-3$). This relatively stronger correlation between scene and inanimate object perception continues as we break it down to manmade inanimate objects and natural inanimate objects. They each have an average correlation score of .39 ($p \leq .01$) and .32 ($p \leq .04$), respectively (for PT 40 to 107 ms). In Figure 18, we also show two manmade objects, *vehicle* and *building*. Interestingly, whereas *building* is very similar to *manmade inanimate object* in terms of correlation between its recognition accuracy with scene perception (average correlation score of .31 for PT 40 to 107 ms, $p \leq .02$,

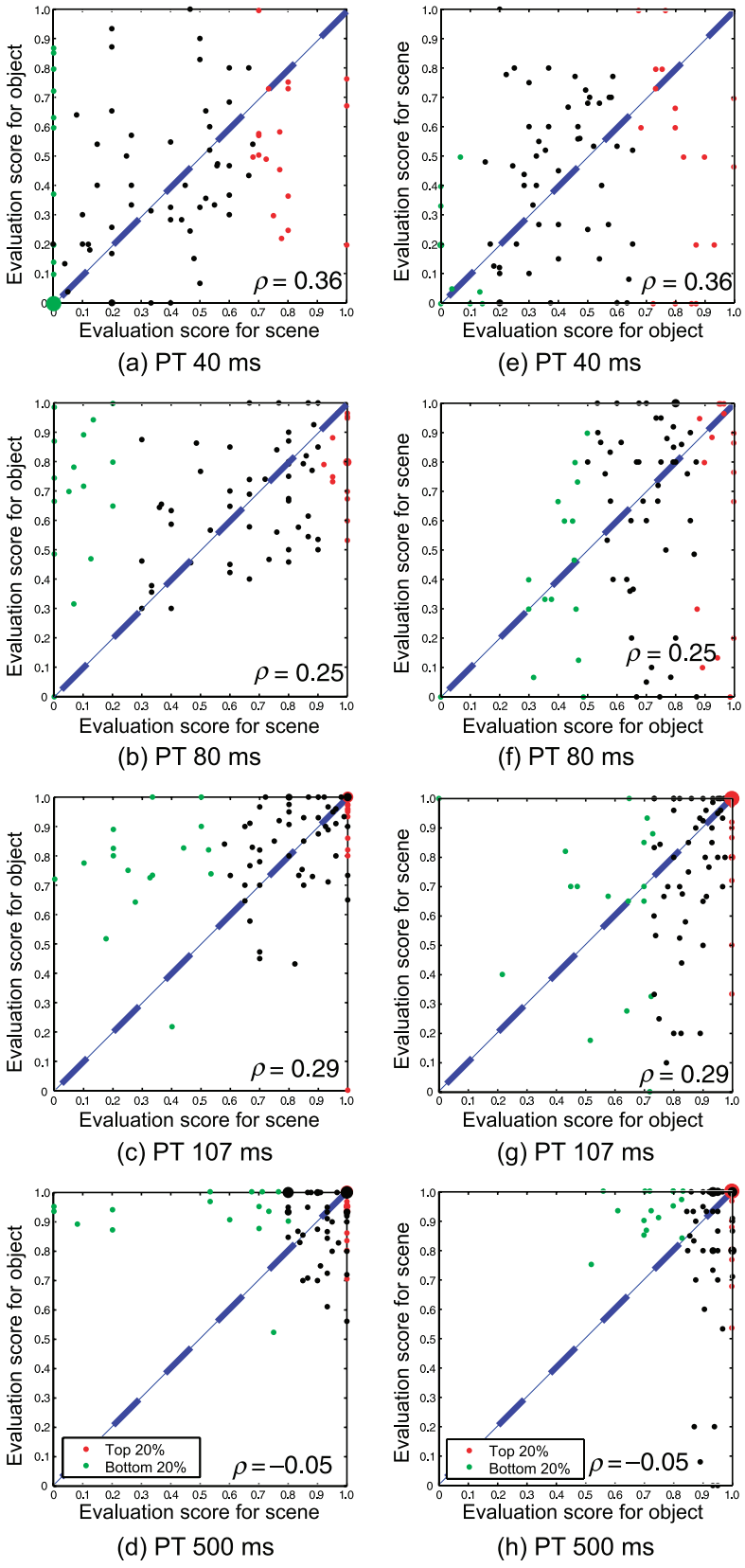

Figure 17. Object recognition performance versus scene recognition performance at various PTs. Performance is based on evaluation scores. See the Results and Observations section for detailed explanations. Note that the relatively close to zero correlation (ρ) for PT 500 ms reflects a mathematical property that when the evaluation scores are close to perfect, little correlation is possible due to the lack of variance.

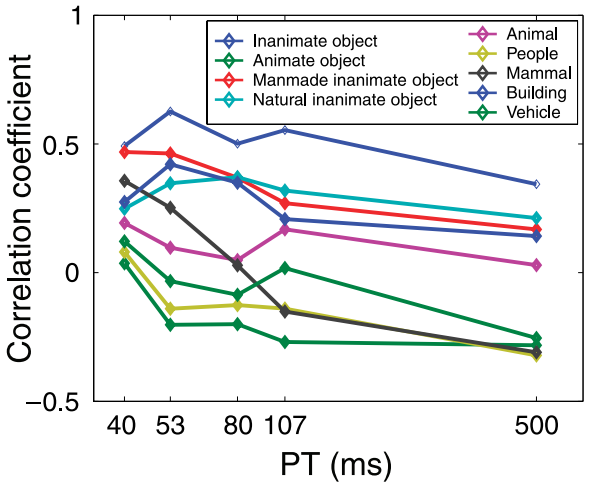

Figure 18. Overall correlation coefficients for scene versus objects and breakdowns.

except for PT 107 ms, $p = .09$), the *vehicle* attribute seems to have a near-zero correlation with the scene (average correlation score of .01 for PT 40 to 107 ms, $.40 \leq p \leq .92$).

Curiously, the predominantly strong correlation between inanimate object perception and scene perception does not hold for those attributes involving animate objects. At the coarsest level, *animate object* recognition has an average correlation score of −.15 with scene perception (for PT 40 to 107 ms, $.02 \leq p \leq .77$). At various levels of animate object recognition, the correlations with scene perception oscillate between no correlation (e.g., people, an average correlation of −.08 for PT 40 to 107 ms, $.25 \leq p \leq .51$) and a very weak correlation (e.g., animal and mammal, both with an average correlation of .12 for PT 40 to 107 ms, $.12 \leq p \leq .92$).

Conclusion

We have shown a novel method to study scene perception. We collected free-recall responses from subjects who were instructed to view 90 different real-world scenes under different PTs. An independent group of subjects then evaluated the free-recall responses. From this approach, we have gleaned several trends, many of which are consistent with those cited in known literature and with others that suggest interesting lines of questioning to pursue in future studies. In this vein, we hope that our design might broaden the scope of scene perception research.

The gist of gist

Information contained in the gist of a real-world scene seems to enjoy a tremendous privilege in visual processing. Temporally, this privilege is reflected through the ultrarapid speed with which the brain categorizes natural scenes (Thorpe et al., 1996). Spatially, this complex scene categorization is not affected when spatial attention is deployed elsewhere (Li et al., 2002; Fei-Fei, VanRullen, Koch, & Perona, 2005). Yet the central question remains as to what actually constitutes this scene gist. We would like to suggest that the term "gist" is used to denote the perceived contents of a scene given a certain amount of viewing time. A sensible and intuitive proposal for a discrete viewing time is single fixation, as many studies have shown that much can be seen within a single glance of a scene (Biederman, 1972; Boyce et al., 1989; Grill-Spector & Kanwisher, 2005; Li et al., 2002; Thorpe et al., 1996; VanRullen & Koch, 2003). These experiments, however, are all conducted with some form of forced multiple choices. Our observations suggest that a rich collection of perceptual attributes is represented and rises to conscious memory within a single fixation. In Observation I, we have collected a list of scene attributes perceived by subjects. Beyond a list of objects and scene environments (Wolfe, 1998b), more cognitive appraisals of the event—such as social interaction and sports events—can be recognized effortlessly. It would be highly interesting for future studies to investigate into the neural correlates that are responsible for such superb ability of real-world scene perception.

Shapes, objects, and scenes

A key issue in perception is the neuronal time course a given perceptual task follows, in other words, the stages through which a stimulus is processed to manifest as semantically meaningful concepts.

The ventral visual pathway, linking the primary visual cortex through the inferior temporal cortex to the prefrontal cortex, is generally known as the "what" visual pathway, as it is responsible for object recognition through integrating features (Kosslyn, Flynn, Amsterdam, & Wang, 1990; Mishkin, Ungerleider, & Macko, 1983; Ungerleider & Mishkin, 1982; Van Essen, 1985). Given the hierarchical structure of the visual system, many have proposed a model in which elementary features of objects are first processed and then bound together for object recognition (Treisman & Gelade, 1980; Wolfe, 1998b). An ongoing debate in this picture is whether shape segmentation is a necessary intermediate step between low-level feature processing and high-level object recognition (Driver & Baylis, 1996; Nakayama et al., 1995; Rubin, 1958). Recently, Grill-Spector and Kanwisher (2005) have found that categorization of superordinate- to basic-level objects (e.g., vehicle, musical instrument, bird, car, dog) is as accurate and fast as the mere detection of the object. Their conclusion is based on an experiment in which subjects are asked to either choose one of the possible object categories or respond simply if an object is

detected. Comparing their nonobject distractors, it is obvious that the low-level image statistics of the distractors (mostly pixel noise) are drastically different from the images that contain objects (all containing a central blob). Given this expectation, subjects are likely to heighten their search for a centrally located blob when detecting objects. In our experiments, subjects viewed freely a naturally cluttered real-world scene. Because our scenes are highly variable, they cannot expect a centrally located blob when looking at an image. In Observation III, we found that shape-related information has a slight advantage over semantically meaningful information of a scene. Our data set shows that compared to higher level semantically meaningful recognition, lower level shape recognition seems to need less information. This temporal constraint implicates a lower feature-level processing in facilitation of the initial stages of complex scene recognition.

Another major question regards object recognition in cluttered scenes. Several psychological models have been proposed to suggest different mechanisms of scene and object perception (Bar & Ullman, 1996; Biederman, 1972, 1982; Friedman, 1979; Hollingworth & Henderson, 1999; Mandler & Parker, 1976; Palmer, 1975). Supported by studies of scene consistency and object detection, the *perceptual schema model* proposes that expectations derived from knowledge about the composition of a scene type interact with the perceptual analysis of objects in the scene (Biederman, 1982; Boyce et al., 1989; Metzger & Antes, 1983; Palmer, 1975). This view suggests that scene context information can be processed and accessed early enough to influence recognition of objects contained in scene, even inhibiting recognition of inconsistent ones (Biederman et al., 1983).

The *priming model*, on the other hand, proposes that the locus of the contextual effect is at the stage when a structural description of an object is matched against long-term memory representations (Bar & Ullman, 1996; Friedman, 1979). Regardless of the mechanism, both the priming model and the perceptual schema model claim that scene context facilitates consistent objects more so than inconsistent ones. These theories predict that we should observe a correlation of object identification performance with scene context categorization performance.

In contrast, a third theory called the *functional isolation model* proposes that object identification is isolated from expectations derived from scene knowledge (Hollingworth & Henderson, 1999). It predicts that experiments examining the perceptual analysis of objects should find no systematic relation between object and scene recognition performance (Hollingworth & Henderson, 1999).

In this article, we do not attempt to resolve the debate between these models directly. Instead, we look at the correlation between subjects' perception of different levels of object categorization with scenes and find a weak but significant correlation (Figure 17) at and up to PTs of 107 ms. This correlation might suggest several possibilities: (i) object and scene perceptions might share at least some resources in processing and/or (ii) object (or scene) perception facilitates processing of scene (or object) perception.

In general, the question of the processing stages of cluttered scenes is still largely unsolved. Our experiments add evidence that there might exist a mutual facilitation between overall scene recognition and object recognition. In addition, both low-level shape- and sensory-related processing seem to require less information and, possibly, less time compare to more high-level, semantically meaningful categorizations of objects and scenes. Traditionally, scene comprehension tends to be viewed in a serial fashion—in the order of sensory information, object features, objects, and the overall scene. Many new studies have now suggested that contrary to this view, high-level perception of natural scenes might be a highly efficient and parallel process (Grill-Spector & Kanwisher, 2005; Li et al., 2002; Rousselet, Fabre-Thorpe, & Thorpe, 2002; Thorpe et al., 1996). It would be interesting to examine an alternative hypothesis in which most of the recognition stages occur in parallel and constantly feed back information to each other to enhance the overall recognition of various components of the scene. In this possible scenario, early sensory information extraction stages still precedes most of the semantic recognition stages. But as soon as there is any information for any possible level(s) of recognition, our brain takes advantage of this.

Two puzzling asymmetries?

In Observation II, we observe a strong preference for outdoor scenes over indoor scenes when visual information is scarce. Subjects seem to assume by default that an ambiguous image is likely to be outdoor than indoor. This effect diminishes as the PT lengthens. At 500 ms, outdoor and indoor scene categorizations become nearly perfect. Our results further show that the bias only appears at the most superordinate level. When indoor scenes are compared with manmade or natural outdoor scenes, the bias disappears. Furthermore, neither segmentation nor object recognition seems influenced by this bias between these two categories of scenes. Hence, what is it that causes this bias? Recent computational models have shown that, using global and local cues such as edge and color information, it is possible to separate most outdoor and indoor scenes (Fei-Fei & Perona, 2005; Szummer & Picard, 1998; Torralba & Oliva, 2003; Vailaya et al., 2001). This strongly suggests that any feature that enables this discrimination is either missing or inaccessible when information is scarce. More studies should be performed to pinpoint exactly what it is. This might be a very useful entry point for someone who is investigating the features needed for rapid scene categorization.

Another curious asymmetry we observe in Observation V is the stronger correlation between inanimate object recognition and overall scene context versus that between animate object recognition and overall scene context. One

possible explanation of this phenomenon is the effect of familiarity. It has been long known that there might be special neuronal resources designated for human parts such as faces and bodies (Farah, 1995; Farah et al., 1998; Kanwisher, 2001; Kanwisher, McDermott, & Chun, 1997; Ro et al., 2001). We have also found recently that familiarity might modulate the level of attentional requirement in object recognition tasks (Fei-Fei, Fergus, & Perona, 2004). If there is indeed an innate preference for animate objects such as animals and humans, there might also be efficient computational mechanisms for the visual system to process this information rapidly and accurately. Compared to other object categorization, it might, hence, be less dependent on possible mutual facilitation mechanisms with scene gist perception. Interestingly, vehicle is among the least correlated object categories with scenes. Given our modern lifestyle, subjects are, in general, very familiar with various kinds of vehicles in the pictures in our database. Another highly speculative hypothesis would be that there is less mutual facilitation between the recognition of mobile objects (such as animals, people, and vehicles) and scenes. If prior knowledge of these objects informs us that they are likely to move from scene to scene, there might be less expectation for recognizing them in any particular scene. Admittedly, much still needs to be done to fully understand this unexpected asymmetry between inanimate and animate objects. As this is largely speculation, more experiments need to be done to address these hypotheses and to account for this asymmetry.

Appendix A: Control Experiment 1 for Observation II

We wished to know whether a bias in subject performance could be accounted for by simple, low-level global cues. Indeed, many studies have explored the usage

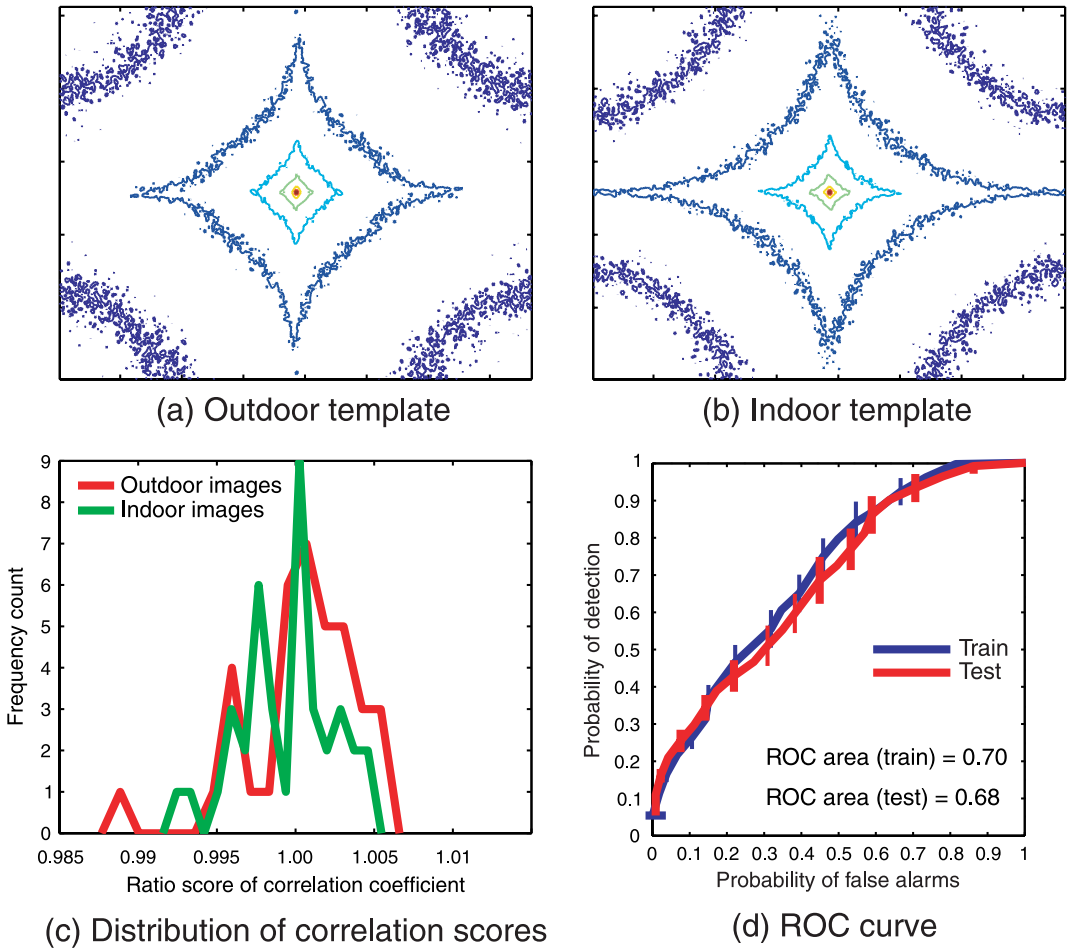

Figure A1. Power spectral analysis. (a) A sample outdoor template, which averaged the power spectra of all outdoor images in the data set (excluding the image itself if it is outdoor). (b) A sample indoor template, which averaged the power spectra of all indoor images in the data set (excluding the image itself if it is indoor). (c) Distribution of the ratio score for outdoor and indoor images. The ratio score of correlation coefficients is obtained from the outdoor correlation coefficient and indoor correlation coefficient for each image. Panel d shows two ROC curves (training and testing) of the classification results based on the correlation ratios. A weak classification result of 68.0% is achieved for separating indoor images from outdoor ones based on the average power spectra in the testing case.

of global cues for categorizing natural scenes, and computer vision algorithms have demonstrated relative success in utilizing such cues to accurately achieve a variety of classifications (Oliva & Torralba, 2001; Szummer & Picard, 1998; Vailaya et al., 2001). Following the same line of reasoning, we carried out two control analyses of the global statistics of the scenes in our data set.

In the first control experiment, we assessed whether indoor and outdoor scenes in our database could be separated by simple frequency information (Oliva & Torralba, 2001). Both the indoor and outdoor images were randomly divided into halves—a "training set" and a "test set." Two power spectrum templates were then created: (i) an outdoor template, which averaged the power spectra of all outdoor images in the outdoor training set, and (ii) an indoor template, which averaged the power spectra of all indoor images in the indoor training set. Figures A1(a) and A1(b) show two example outdoor and indoor templates for randomly drawn training sets. For the images in the test sets, a two-dimensional correlation was performed between the power spectrum of each image and the outdoor template and between the power spectrum of each image and the indoor template. We then obtained a ratio of correlation coefficients (outdoor correlation coefficient:indoor correlation coefficient) for each image in the test sets. This correlation analysis was repeated, with training and test sets reversed; that is, the images previously in the training sets formed the new test sets, and the images formerly used in the test sets were used to generate the templates. Ratios of correlation coefficients were obtained for images of the new test sets. In this way, correlations were performed on every image in the data set, with templates formed from a disjoint set of images. This procedure was reiterated 10 times, with a random segregation of images into either the training sets or the test sets each time.

Figure A1(c) shows the distribution of this ratio score for all of the outdoor and indoor images. We use this ratio score of the images to perform indoor versus outdoor classification. Figure A1(d) is a receiver operating characteristic (ROC) curve of the result. A weak classification result of 68.0% is achieved for separating indoor images from outdoor ones based on the average power spectra (chance classification by an ROC analysis is

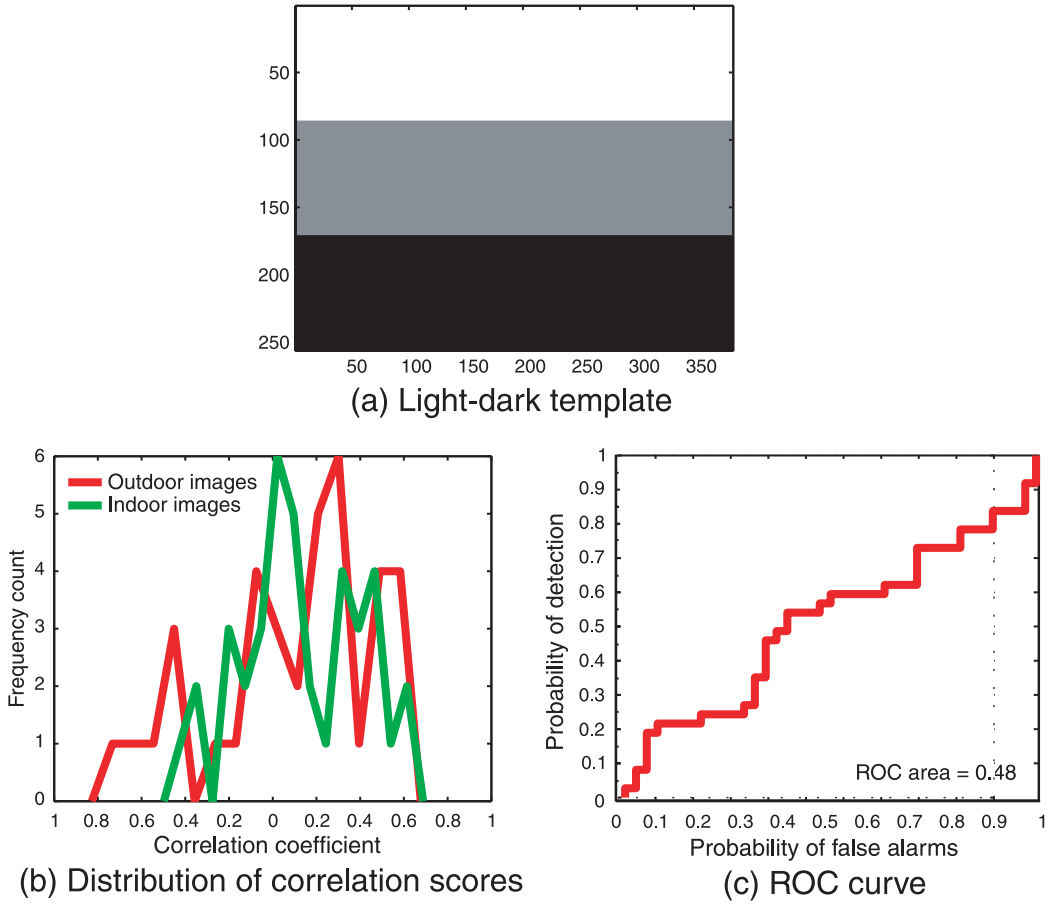

Figure B1. Light top dark bottom correlation analysis. (a) Two horizontal layers constituted this template, the top consisting of high-intensity pixels, and the bottom of low-intensity pixels. (b) A two-dimensional correlation was performed between each image in the dataset and the template. The correlation coefficient for each image was used for classification. (c) shows the classification results in ROC curve. Only a 47.5% performance is achieved by using the template method.

considered to be 50%). Compared to the average performance of human observers at PT 500 ms (90.5% in Figure 9), this result indicates that little information could be used to classify indoor and outdoor scenes based on low-level power spectral information.

It is worth noticing that our results are not entirely consonant with recent computational models that suggest that global frequency cues are adequate for the separation of natural scene categories (Oliva & Torralba, 2001). We would like to point out that the success of these models is demonstrated on a set of relatively typical and canonical natural scenes, where beaches are nearly all uncluttered and expansive and where streets are nearly all photographed from a similar angle (Oliva & Torralba, 2001). Our data set of indoor and outdoor scenes is significantly more challenging. In particular, all indoor and outdoor scenes are cluttered and taken from a variety of angles. It is, thus, not surprising that the indoor and outdoor images in this data set are less separable based on global frequency information.

Appendix B: Control Experiment 2 for Observation II

Our second control addressed the argument that outdoor scenes tend to have a lighter top partly due to the contrast

Attribute	α	γ	Relevant figures
Object	80.27 ± 1.25	1.53 ± 0.06	Figure 14a
Inanimate object	88.29 ± 1.32	1.63 ± 0.06	Figures 14a and 14c
Animate object	72.01 ± 2.88	1.42 ± 0.14	Figures 14a, 14c, and 15a
People	78.76 ± 3.86	1.48 ± 0.47	Figure 14c
Animal	85.60 ± 4.27	1.55 ± 0.19	Figures 14c and 15a
Mammal	94.20 ± 8.68	1.25 ± 0.23	Figure 15a
Large mammal	106.22 ± 15.21	0.95 ± 0.21	Figure 15a
Natural inanimate object	99.99 ± 10.21	1.44 ± 0.33	Figure 15c
Rock	107.48 ± 11.17	2.31 ± 0.64	Figure 15c
Plant	58.08 ± 11.83	1.59 ± 0.91	Figure 15c
Body of water	137.94 ± 20.24	1.87 ± 0.47	Figure 15c
Mountain, hill	76.86 ± 4.92	3.26 ± 0.90	Figure 15c
Manmade inanimate object	92.57 ± 2.08	1.65 ± 0.09	Figure 15b
Structure	109.59 ± 9.54	1.12 ± 0.16	Figure 15b
Road, bridge	110.72 ± 11.20	1.44 ± 0.27	Figure 15b
Building	66.36 ± 4.09	1.52 ± 0.25	Figures 14c and 15d
Specific building	105.98 ± 9.94	1.60 ± 0.31	Figure 15d
Distinctive architecture	91.69 ± 3.73	1.75 ± 0.18	Figure 15d
Chair	91.91 ± 1.81	3.58 ± 0.29	Figure 14c
Scene	84.90 ± 3.26	1.77 ± 0.18	Figure 14a
Outdoor	73.18 ± 2.51	1.54 ± 0.14	Figures 14a, 14b, 16b, and 16c
Outdoor, manmade	87.20 ± 3.54	1.74 ± 0.18	Figures 14b and 16c
Industrial	91.82 ± 3.89	3.87 ± 0.71	Figure 16c
Skyline	59.28 ± 3.69	1.60 ± 0.28	Figure 16c
Outdoor, mm-other	99.66 ± 5.68	2.75 ± 0.52	Figure 16c
Outdoor, natural	77.25 ± 2.47	2.34 ± 0.25	Figures 14b and 16b
Urban	80.27 ± 4.20	2.03 ± 0.31	Figures 14b and 16c
Water	98.68 ± 6.64	2.33 ± 0.45	Figures 14b and 16b
Field	76.28 ± 2.08	2.51 ± 0.24	Figure 16b
Beach	107.34 ± 0.83	5.30 ± 0.27	Figure 16b
Indoor	95.38 ± 3.74	2.32 ± 0.27	Figures 14a, 14b, and 16a
Household room	100.56 ± 1.93	3.01 ± 0.20	Figure 16a
Office, class	89.24 ± 3.66	3.00 ± 0.45	Figure 16a
Store	133.98 ± 20.56	2.79 ± 1.00	Figure 16a

Table C1. Attributes and parameters from Weibull cdf fitting. Object-related attributes are listed in the upper half of the table; scene-related attributes are listed in the lower half. α and γ are parameters, and their errors are reported. Only those attributes whose evaluation scores are plotted for Observations III and IV are listed here. The figures where they appeared are documented in the last column.

of the sky, whereas there is no such cue in an indoor image. We therefore used a simple "sky" template to explore this possibility (Figure B1(a)). Three horizontal layers constituted this template: the top, which consists of high-intensity pixels; the middle, which consists of median-intensity pixels; and the bottom, which consists of low-intensity pixels. A two-dimensional correlation was performed between each image in the data set and the template. The correlation coefficient for each image was used for classification. Figure B1(b) shows the distributions of the correlation coefficients of all the indoor and outdoor images, whereas Figure B1(c) shows the classification results in the ROC curve. Only a 47.5% performance is achieved by using the template method. This is no better than chance, as compared to a high human observer performance at PT 500 ms (90.5% in Figure 9).

Appendix C: Weibull curve fits for evaluation scores as a function of PT

For Observations III and IV, evaluation scores are presented as a function of PT (in milliseconds; see Figures 14, 15, and 16). These scores represent the degree to which attributes are perceived, which approximates the probability with which they are reported. Thus, the scores for each attribute are also fitted with a cumulative density function, the Weibull cdf:

$$F(t) = 1 - e^{-\left(\frac{t}{\alpha}\right)^{\gamma}}.$$

Here, t is PT and α and γ are the parameters determined by the fitting procedure. In addition to the ANOVAs run on the actual scores, the Weibull fitted curves (and the 95% confidence intervals) are examined to compare the various attributes. Note that in Observation III, we compare these semantic-related attributes to the original curve sensory-level attribute. All details for attributes and their corresponding fitted curves are shown in Table C1.

Acknowledgments

This work was supported by an NSF ERC grant from Caltech. L. F.-F. was supported by Paul and Daisy Soros Fellowship for New Americans as well as an NSF Graduate Fellowship. The authors thank Irv Biederman, Jochen Brown, Shin Shimojo, Dan Simons, Rufin VanRullen and three anonymous reviewers for their helpful comments.

L. F.-F. and A. I. contributed equally in this work.

Commercial relationships: none.
Corresponding author: Li Fei-Fei.
Email: feifeili@cs.princeton.edu.
Address: 35 Olden St. Princeton, NJ 08540, USA.

Footnotes

[1]The quality of the images of the stimuli might influence the performance of subjects. We are, therefore, careful to choose images that are of decent quality and contrast although they are likely to be taken by amateurs.

[2]It is not hard to infer these words or phrases from Figures 1 and 2. A recent rerun of 10 naive subjects suggests a list of the following 31 words or phrases that are very similar to the ones obtained before: park, pool, desert, animals, ocean, yard, beach, suburb or residential area, urban environ or skyline, parking lot, city or streets, mountain, forest, airplane or airport, river, various sport games, office, library, cafe or restaurant, store, lab, bathroom, kitchen, living room, bedroom, store or mall, concert or stage, hospital, museum, party, and gym.

References

Bar, M., & Ullman, S. (1996). Spatial context in recognition. *Perception, 25,* 343–352. [PubMed]

Biederman, I. (1972). Perceiving real-world scenes. *Science, 177,* 77–80. [PubMed]

Biederman, I. (1982). On the semantics of a glance at a scene. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 213–254). Hillsdale, NJ: Erlbaum.

Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W., Jr. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology, 103,* 597–600. [PubMed]

Biederman, I., Teitelbaum, R. C., & Mezzanotte, R. J. (1983). Scene perception: A failure to find a benefit from prior expectancy or familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 411–429. [PubMed]

Boyce, S. J., Pollatsek, A., & Rayner, K. (1989). Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 556–566. [PubMed]

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436. [PubMed]

Bruner, J. C., & Potter, M. C. (1964). Interference in visual recognition. *Science, 114,* 424–425. [PubMed]

De Graef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research, 52,* 317–329. [PubMed]

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science, 293,* 2470–2473. [PubMed]

Driver, J., & Baylis, G. C. (1996). Edge-assignment and figure-ground segmentation in short-term visual matching. *Cognitive Psychology, 31,* 248–306. [PubMed]

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience, 13,* 171–180. [PubMed]

Farah, M. J. (1995). Dissociable systems for visual recognition: A cognitive neuropsychology approach. In S. M. Kosslyn & D. N. Osherson (Eds.), *Visual cognition: An invitation to cognitive science* (pp. 101–119). Cambridge, MA: MIT Press.

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review, 105,* 482–498. [PubMed]

Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *IEEE International workshop on Generative-Model Based Vision, in conjunction with CVPR'05*.

Fei-Fei, L., & Perona, P. (2005). A Bayesian hierarchy model for learning natural scene categories. In *IEEE Computer Vision and Pattern Recognition*. San Diego, USA. VII: 524–531.

Fei-Fei, L., VanRullen, R., Koch, C., & Perona, P. (2005). Why does natural scene recognition require little attention? Exploring attentional requirements for natural and synthetic stimuli. *Visual Cognition, 12,* 893–924.

Friedman, A. (1979). Frame pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General, 108,* 316–355. [PubMed]

Germeys, F., & d'Ydewalle, G. (2001). Revisiting scene primes for object locations. *Quarterly Journal of Experimental Psychology, 54,* 683–693. [PubMed]

Goffaux, V., Jacques, C., Mauraux, A., Oliva, A., Schynsand, P. G., & Rossion, B. (2005). Diagnostic colors contribute to the early stages of scenes categorization: Behavioral and neurophysiological evidence. *Visual Cognition, 12,* 878–892.

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science, 16,* 152–160 [PubMed]

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50,* 243–271. [PubMed]

Hollingworth, A., & Henderson, J. M. (1999). Object identification is isolated from scene semantic constraint: Evidence from object type and token discrimination. *Acta Psychologica, 102,* 319–343. [PubMed]

Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. *Journal of Vision, 3*(7), 499–512, http://journalofvision.org/3/7/4/, doi:10.1167/3.7.4. [PubMed] [Article]

Kanwisher, N. (2001). Faces and places: Of central (and peripheral) interest. *Nature Neuroscience, 4,* 455–456. [PubMed] [Article]

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience, 17,* 4302–4311. [PubMed] [Article]

Kosslyn, S. M., Flynn, R. A., Amsterdam, J. B., & Wang, G. (1990). Components of high-level vision: A cognitive neuroscience analysis and accounts of neurological syndromes. *Cognition, 34,* 203–277. [PubMed]

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences of the United States of America, 99,* 9596–9601. [PubMed] [Article]

Lumet, S. (Director). (1965). *The pawnbroker* [Motion picture]. United States: Republic Pictures.

Mandler, J. M., & Parker, R. E. (1976). Memory for descriptive and spatial information in complex pictures. *Journal of Experimental Psychology, 2,* 38–48. [PubMed]

Metzger, R. L., & Antes, J. R. (1983). The nature of processing early in picture perception. *Psychological Research, 45,* 267–274. [PubMed]

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences, 6,* 414–417.

Nakayama, K., He, Z. J., & Shimojo, S. (1995) Visual surface representation: A critical link between lower-level and higher level vision. In Kosslyn S. M. and Osherson D. N. (Eds.), *An Invitation to Cognitive Science: Visual Cognition* (pp. 1–70). Cambridge, MA: M.I.T. Press.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology, 41,* 176–210. [PubMed]

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Visions, 42*, 145–175.

Palmer, S. E. (1975). Visual perception and world knowledge: notes on a model of sensory–cognitive interaction. In D. A. Norman & D. E. Rumelhart (Eds.), *Explorations in cognition* (pp. 279–307). San Francisco: LNR Res. Group.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442. [PubMed]

Peterson, M. A., & Gibson, B. S. (1993). Shape recognition contributions to figure-ground organization in three-dimensional display. *Cognitive Psychology, 25,* 383–429.

Peterson, M. A., & Gibson, B. S. (1994). Must shape recognition follow figure-ground organization? An assumption in peril. *Psychological Science, 5,* 253–259.

Peterson, M. A., & Kim, J. H. (2001). On what is bound in figures and grounds. *Visual Cognition, 8,* 329–348.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 509–522. [PubMed]

Potter, M. C., Staub, A., Rado, J., & O'Connor, D. H. (2002). Recognition memory for briefly presented pictures: The time course of rapid forgetting. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 1163–1175. [PubMed]

Rayner, K. (1984). Visual selection in reading, picture perception and visual search: A tutorial review. In H. Bouma & D. Bouwhuis (Eds.), *Attention and Performance* (vol. 10, pp. 67–96). Hillsdale, MJ: Erlbaum.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8,* 368–373.

Ro, T., Russell, C., & Lavie, N. (2001) Changing faces: A detection advantage in the flicker paradigm. *Psychological Science, 12,* 94–99. [PubMed]

Rosch, E. (1978). "Principles of categorization". In E., Rosch, B., Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience, 5,* 629–630. [PubMed] [Article]

Rubin, E. (1958). Figure and ground. In D. C. Beardslee & M. Wertheimer (Eds.), *Readings in perception.* (pp. 194–203). Princeton, NJ: Van Nostrand & Company.

Szummer, M., & Picard, R. (1998). Indoor–outdoor image classification. In *IEEE International workshop on content-based access of image and video databases, in conjunction with ICCV '98* (pp. 42–51). Bombay, India.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381,* 520–522. [PubMed]

Torralba, A., & Oliva, A. (2003). Statistics of natural images categories. *Network: Computation in Neural Systems, 14,* 391–412. [PubMed]

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12,* 97–136. [PubMed]

Tversky, B., & Hemenway, K. (1983). Categories of environmental scenes. *Cognitive Psychology, 15,* 121–149.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549–586). Cambridge, MA: MIT Press.

Vailaya, A., Figueiredo, M., Jain, A., & Zhang, H. J. (2001). Image Classification for Content-Based Indexing. *IEEE Transactions on Image Processing, 10*, 117–130.

Van Essen, D. C. (1985) Functional organization of primate visual cortex. In E. G. Jones & A. Peters (Eds.), *Cerebral Cortex* (vol. 3, pp. 259–329). New York: Plenum Press.

VanRullen, R., & Koch, C. (2003). Competition and selection during visual processing of natural scenes and objects. *Journal of Vision, 3*(1), 75–85, http://journalofvision.org/3/1/8/, doi:10.1167/3.1.8. [PubMed] [Article]

VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience, 13,* 454–461. [PubMed]

Vogel, J., & Schiele, B. (2004). A semantic typicality measure for natural scene categorization. In *DAGM '04 Annual Pattern Recognition Symposium,* Tubingen, Germany.

Wolfe, J. M. (1998a). Visual memory: What do you know about what you saw? *Current Biology, 8,* R303–R304. [PubMed] [Article]

Wolfe, J. M. (1998b). Visual Search. In H. Pashler (Ed.), *Attention* (pp. 13–74). Hove, UK: Psychology Press Ltd.