A perceptual paradigm and psychophysical evidence for hierarchy in scene gist processing

Ilan Kadar
Computer Science Department and the Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Ohad Ben-Shahar
Computer Science Department and the Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Beer-Sheva, Israel

What is the order of processing in scene gist recognition? Following the seminal studies by Rosch (1978) and Tversky and Hemmenway (1983) it has been assumed that basic-level categorization is privileged over the superordinate level because the former maximizes both within-category similarity and between-category variance. However, recent research has begun to challenge this view (Oliva & Torralba, 2001; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007; Loschky & Larson, 2010). Here we study these directions more fundamentally by investigating the perceptual relations between scene categories in a way that allows us to identify the order of processing of scene categories across taxonomic levels. We introduce the *category discrimination paradigm* where we briefly present two real scene stimuli simultaneously and ask human observers whether they belong to the same basic-level category or not (i.e., same/different task). As we show, analysis of the obtained data reveals a hierarchical perceptual structure between different scene categories and a corresponding hierarchical structure at the perceptual processing level. In particular, we show a new type of evidence to suggest that the decision whether the scene is manmade or natural is made first, and only then more complicated decisions are taken (such as whether a manmade scene is indoor or outdoor) among a smaller set of viable candidate categories. We argue that this hierarchical structure improves performance and efficiency in both biological and artificial gist recognition systems.

Keywords: scene perception, scene gist processing, gist recognition, scene categorization, hierarchical processing, basic-level, superordinate-level, manmade, natural, indoor, outdoor

Citation: Kadar, I., & Ben-Shahar, O. (2012). A perceptual paradigm and psychophysical evidence for hierarchy in scene gist processing. *Journal of Vision*, *12*(13):16, 1–17, http://www.journalofvision.org/content/12/13/16, doi:10.1167/12.13.16.

Introduction

Consider the four images in Figure 1. Most observers with normal vision would easily match their respective scene classes. More importantly, this can be done even if the presentation time is extremely short. Indeed, the phenomenon of perceiving and categorizing real-world scenes at a glance is a common experience for most of us. Whether we quickly switch television channels, browse video files, or flip rapidly photos in our photo album, our visual system can quickly and effectively recognize visual scenes. This remarkable ability is frequently called *scene gist recognition*, where gist refers to the meaningful information that an observer can extract at a glance, and is often synonymous with its basic-level category, for example *Coast* or *Forest* (Oliva, 2005; Loschky & Larson, 2008).

But what characterizes visual processing underlying scene gist recognition? Since the pioneering perceptual studies by researchers such as Potter and Biederman (Potter & Levi, 1969; Biederman, 1972; Biederman, Glass, & Stacy, 1973; Potter, 1975, 1976), which demonstrated that human observers can recognize the gist of a scene in a fraction of a second, much research has been devoted to understanding the visual process underlying this visual capacity (Oliva & Schyns, 1994; Schyns & Oliva, 1994; Thorpe, Fize, & Marlot, 1996; Oliva & Schyns, 2000; Oliva & Torralba, 2001; Fei-Fei, VanRullen, Koch, & Perona, 2002; Walker & Malik, 2002; BaconMace, Mace, Fabre-Thorpe, & Thorpe, 2005; Rousselet, Joubert, & Fabre-Thorpe, 2005; Fei-Fei, Koch, Iyer, & Perona, 2007; Joubert et al., 2007; Loschky, Sethi, & Simons, 2007; Loschky & Larson, 2008; Peelen, Fei-Fei, & Kastner, 2009; Loschky & Larson, 2010). Although substantial progress has been made, the bulk of this perceptual process remains an open question, both behaviorally and computationally.

One fundamental aspect of scene gist recognition is the order of processing of scene categories across taxonomic levels. Following seminal studies by Rosch (1978) and Tversky and Hemenway (1983), it has been

doi: 10.1167/12.13.16 Received February 1, 2012; published December 18, 2012

Figure 1. Selected scene images from several categories (Oliva & Torralba, 2001), highway, coast, forest, and tall-building.

assumed that distinction of scenes' basic-level (e.g., *Mountain*, *Coast*) is privileged over the superordinate-level (e.g., indoor vs. outdoor) because it maximizes both within-category similarity and between-category variance. In other words, it has been argued that scene images are recognized first at the basic-level, with additional processing (and hence additional resources and time) required to identify them as members of their superordinate category. However, later research has begun to challenge this view. For example, Oliva and Torralba (2001) asked human observers to recursively split 81 scenes presented for unlimited time into two subgroups. They found that the natural/manmade superordinate distinction was the most commonly cited reason given for the first split. Later, Fei-Fei et al. (2007) examined what exactly human subjects perceive and understand when they glance at the world. Their human subjects were asked to view 90 natural scenes for presentation times varying between 27 and 500 ms and then to describe in written free text what they have observed in as much detail as possible. In their results the authors reported that "in general, superordinate-level scene categories seem to require the same amount of information in recognition as the basic-level scenes." Indeed, these results do not brace the findings by Oliva and Torralba (2001) but neither do they support the classical view that basic-level distinction is privileged over superordinate-level distinction.

While the studies mentioned above may question the classical view, in the context of gist recognition they are confounded in at least two ways. While gist recognition is immediate and facilitated by extremely short processing times, Oliva and Torralba (2001) employed long stimulation which could trigger additional higher-level mechanisms. Similarly, while the free text experimental approach (Fei-Fei et al., 2007) can indeed

provide interesting insights, it is less preferable psychophysically due to the implicit nature and the subjective aspects involved in the interpretation of observers' responses and the possibility that higher inference mechanisms are involved in the generation of the descriptions. Alternatively, Joubert et al. (2007) used a go/no-go rapid visual categorization task in which human observers had to respond as fast as possible when they observed either manmade scenes or natural scenes that were presented for 26 ms. By comparing results to an earlier study (Rousselet et al., 2005), the authors found that the natural/manmade task is performed faster than the basic-level task. Later, Loschky and Larson (2010) compared subject performance in both natural/manmade distinction and basic-level distinction as a function of processing time by varying the target-to-mask stimulus onset asynchrony (SOA). In their experiment, subjects were asked to view a briefly presented scene images followed by a mask and then to report if the scene stimulus matched the cue word at the end of each trial. The cue label could be either a basic-level category or the superordinate-levels natural or manmade and the results indicated that subjects are more sensitive to the latter than the former. Furthermore, performance advantage in natural/manmade categorization over the basic-level distinction was greater at earlier processing times (SOAs $<$ 50 ms) compared to longer or unlimited processing time. When restricted to short presentation times, these last results clearly suggest that the natural/manmade distinction is made before basic-level distinctions and question the validity of the default assumption of the basic-level primacy in scene gist recognition.

Interestingly, at the same time when basic-level primacy was challenged in the context of scene categorization, similar results have been shown in the general object categorization literature. For example, event-related potential (ERP) measurements have shown quicker categorizations at the superordinate-level than basic-level object categorizations (Large, Kiss, & McMullen, 2004). Rogers and Patterson (2007) reported that patients with *semantic dementia* lose object categories at the basic-level before the superordinate-level and that the accuracy for superordinate-level tasks is greater than basic-level tasks when subjects are forced to respond rapidly. Later, Mace, Joubert, Nespoulous, and Fabre-Thorpe (2009) used a rapid visual go/no-go categorization task to compare human processing speed when categorizing objects at the superordinate-level (animal/nonanimal) and at the basic-level (bird/nonbird or dog/nondog). The authors found an early temporal window during which the accuracy of subjects increases very fast for superordinate-level responses whereas those at the basic-level have not been initiated yet.

Despite the accumulating evidence in the scene categorization literature and the additional support from the object recognition community, several issues are left unresolved. The first issue concerns existence of other superordinate dichotomies (such as distinction between indoor and outdoor scenes) that may participate in the process. To quote Loschky and Larson (2010), "it will be also more interesting to determine whether the indoor/outdoor distinction is even more primitive than the natural/manmade distinction." The second issue relates to the generalization of this two-stage process to multiple levels, i.e., to a complete hierarchy of processing during gist recognition. Indeed, one may hypothesize that the preference of the natural/manmade distinction over the basic-level distinction is indicative of more levels of distinction that operate in some particular order.

The current study was designed to address both of these last issues. For that, we employ a forced-choice discrimination (same/different) task to study gist recognition in a novel psychophysical way. We introduce the *category discrimination paradigm* where we briefly present two real scene stimuli simultaneously and ask human observers whether they belong to the same basic-level category or not (i.e., same/different task). While this seems a relatively challenging task, evidence for parallel processing in high level categorization of natural images has already been reported, showing that humans are as fast in dual scene presentations as they are for single scene presentations (Rousselet, Fabre-Thorpe, & Thorpe, 2002). As we explain later, the results of such a proposed experiment could give us important insights regarding the perceptual distance between different scene categories over different levels of processing time. Moreover, proper analysis of the obtained data could indicate that both the perceptual distance and the processing of scene categories follow a particular hierarchical structure. Indeed, as our results suggest, the decision whether the scene is manmade or natural is made first and only then is followed by more complicated decisions (such as whether a manmade scene is indoor or outdoor). We argue that with fewer candidate categories left viable at lower levels in the hierarchy, this mechanism facilitates both faster and more accurate categorization.

Experiment 1

Methods

Subjects

Seventy-nine motivated students from Ben-Gurion University (31 females, 48 males, mean age = 25.66) were paid to participate in the study. All had normal or

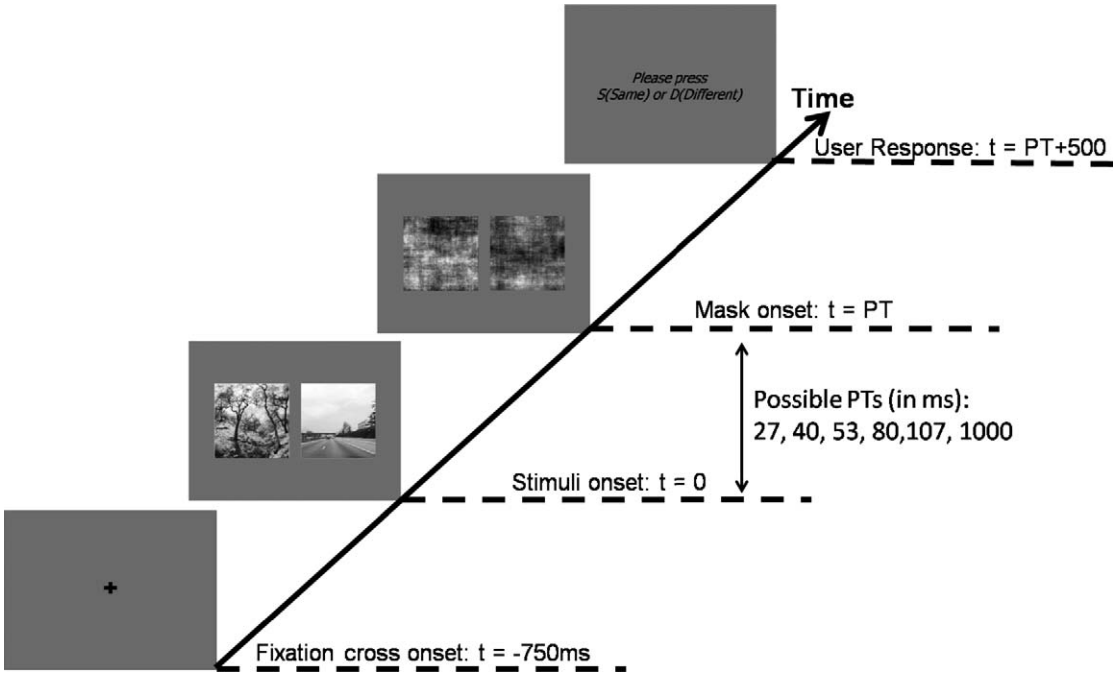

Figure 2. Experimental trials in our experiment began with a fixation point followed by a brief presentation of two images for PT ms and 1/*f* mask patterns for 500 ms. Subjects were then prompted to respond whether the two images belong to the same scene category or not.

corrected-to-normal vision and all were naive about the purpose of the experiment.

Apparatus

Subjects were seated in a dark room especially designed for psychophysical experiments. The seat was approximately 100 cm from the monitor (22-in Iiyama Vision Master Pro 510 CRT monitor with 75 Hz refresh rate). Experimental software was programmed using Matlab and the Psychophysics toolbox (Brainard, 1997) and executed on a Power Mac G5 Macintosh computer.

Stimuli

The underlying pool of scenes used for the experiment consisted of 2,298 images from eight categories borrowed from two published datasets (Oliva & Torralba, 2001; Fei-Fei & Perona, 2005): *Coast* (360 images), *Forest* (328 images), *Mountain* (374 images), *Highway* (260 images), *Tall-buildings* (356 images), *Street* (292 images), *Kitchen* (151 images), and *Bedroom* (177 images). Eighty-four random images were preselected from each category to be used in each experimental session. They were reduced to monochrome and adapted in size to 256 × 256 pixels (10 × 10 cm on screen and 5.72° × 5.72° in visual angle). The size of stimulus including both images and the gap between them was 20.63 × 10 cm on screen and 11.78° × 5.72° in visual angle. The selection of categories was strongly influenced by various earlier studies (Tversky & Hemenway, 1983; Fei-Fei et al., 2007; Loschky & Larson, 2010) and consisted a range of categories as wide as eight classes permits (e.g., natural scenes, manmade scenes, indoor scenes, and outdoor scenes).

Procedure

Figure 2 depicts the sequence of events in an experimental trial. Each trial began with a fixation cross followed by the simultaneous presentation of two images from our dataset for one of six different presentation times (PTs): 27, 40, 53, 80, 107, 1000 ms (all durations are multiplies of a 75 Hz refresh cycle of the computer monitor). PTs were chosen to span a wide range, from very short up to a duration sufficient to allow elaborate perceptual description (Potter & Levi, 1969; Biederman, 1981; Fei-Fei et al., 2007). The longest PT was introduced as a control for how well scene categories were defined semantically (since at 1000 ms, categorization errors are unlikely to be attributed to perceptual confusion).

After a presentation for the selected PT, the two images were then masked for 500 ms by a pair of masks, each selected at random from a pool of eight random patterns having 1/*f* amplitude spectrum (Loschky et al., 2007). The trial concluded with a response cue which remained on screen until the subject's response. Participants pressed "same" if they judged the two images to match in category or "different" if not. They were encouraged to respond

	Bedroom	Kitchen	Street	Tall-building	Highway	Coast	Forest	Mountain
Bedroom	**0.79**	0.64	0.77	0.80	0.80	0.81	0.85	0.84
Kitchen	-	**0.76**	0.72	0.76	0.76	0.82	0.84	0.83
Street	-	-	**0.80**	0.76	0.69	0.82	0.84	0.85
Tall-building	-	-	-	**0.85**	0.84	0.84	0.84	0.86
Highway	-	-	-	-	**0.75**	0.72	0.80	0.79
Coast	-	-	-	-	-	**0.76**	0.77	0.74
Forest	-	-	-	-	-	-	**0.82**	0.72
Mountain	-	-	-	-	-	-	-	**0.81**

Table 1. The Perceptual Data matrix obtained by measuring subjects' average sensitivity between all pairs of scene categories in our dataset. The bolded results along the diagonal represent subjects' average sensitivity in scenarios where the two presented images come from the same category and hence represent a measure of perceptual similarity within each category. All other entries represent the perceptual distance between their corresponding categories for the purpose of gist recognition.

according to their first impression and as quickly and accurately as possible. By design, 50% of the trials constituted a pair of images from the same basic-level category while the other 50% used images from different basic-level categories. Chance level performance was therefore 50%. Each of the eight basic-level categories were presented equally often in both same and different trials to ensure that response for any given category is not biased. PTs were distributed randomly across trials and were counter-balanced across all values and for all of eight basic-level categories.

Before beginning the experiment, participants completed a category learning procedure where they viewed nine images from each of the eight participating categories so that they could get acquainted with the scene category labels. Subjects then completed 12 practice trials so they could become familiar with the experimental task and procedure, after which they started the 336 trial experiment. No learning or practice image was reused in the experiment and no image was presented more than once. Experiments were self-paced and participants were allowed to take breaks any time. In practice, the 336 trial experiment took 20 min to complete on average.

Result and observations

Perceptual distances and evidence for hierarchical perceptual structure

How challenging is it to discriminate *Coast* from *Forest* scenes in a glance? How about *Mountain* versus *Bedroom* scenes? Despite increasing interest in gist recognition research, the answers to such questions and more generally, to the perceptual relationships between different scene categories, remain largely unclear. As we discuss next, the results of our experiment begin to address these issues.

Analyzing subjects' responses first in the control trials with PT = 1000 ms reveals that discrimination under long stimulation yielded near-perfect performance of 96%, indicating that our scene categories are well-defined semantically. Eliminating this possible confound, we then analyzed subjects' responses using the nonparametric signal detection measure A' of sensitivity (Grier, 1971) to exclude the possibility that our results are confounded by certain biases in subjects' responses. In particular, we explored the perceptual distance between all pairs of scene categories in our dataset by measuring subjects' sensitivity for each pair over all trials and PTs (except the control trials with PT = 1000 ms). Table 1, which here we term the Perceptual Data (PD) matrix, shows the obtained sensitivity averaged over all PTs (except 1000 ms) between all pairs of scene categories in our dataset. It is evident that subjects did not discriminate between categories equally accurately (or easily). For example, subjects were able to discriminate with much higher sensitivity *Bedroom* from *Forest* (0.85), *Coast* from *Tall-buildings* (0.84), *Forest* from *Kitchen* (0.84), and *Mountain* from *Street* (0.85). However, sensitivity dropped considerably when discriminating *Bedroom* from *Kitchen* (0.64), *Forest* from *Mountain* (0.72), *Highway* from *Street* (0.69), *Street* from *Tall-buildings* (0.76), and *Highway* from *Coast* (0.72).

Categories that are substantially less accurate to tell apart in the brief presentation times used in our experiments must share enough perceptual properties to make the discrimination process more difficult (again, within these short PTs). Therefore, the sensitivity for the different pairs of categories can be interpreted as the perceptual distance between these pairs of categories: Lower sensitivity means that they are perceptually closer while higher sensitivity implies they are perceptually distant. Hence, Table 1 depicts the perceptual distance between all pairs of categories for the purpose of gist recognition (e.g., with 0.84 compared to 0.69, the *Forest* and *Kitchen* are far less perceptually related than the *Highway* and *Street* categories, respectively). Importantly, the results along the diagonal of Table 1 represent subjects' average accuracy in trials where the expected response was

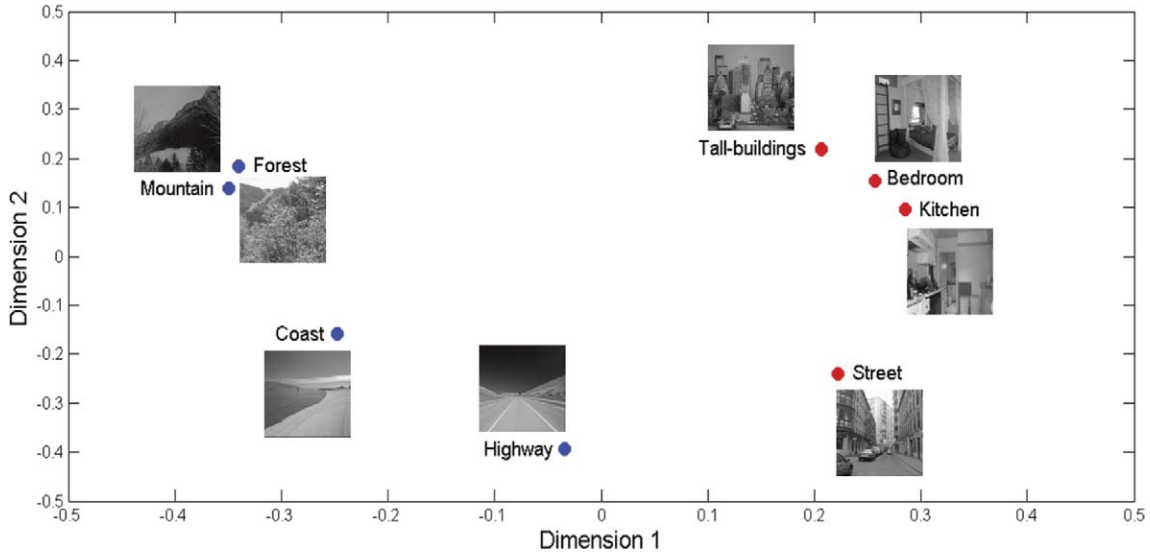

Figure 3. Applying MDS and k-means analyses on the perceptual distances obtained in our experiment and visualizing the results as points in a 2D perceptual space reveal a configuration which clusters natural scenes separately from manmade scenes, with *Highways* being an anomalous borderline case.

"same" as opposed to the "different" response expected in all other entries. Hence, to represent a measure of distance, one should consider the complement of PD(i, i), and since it applies within category, it represents the variance, or the perceptual diversity, within each category. For example, as can be appreciated from the table, humans are more accurate at saying that two *Tall-buildings* images are from the same category ($A' = 0.85$) than they are saying that two coast images are from the same category ($A' = 0.76$). Therefore, with $1 - \mathrm{PD}(i, i) = 0.24$, the coast category is far more perceptually diverse than the *Tall-buildings* category, whose $1 - \mathrm{PD}(i, i) = 0.15$.

With these results in mind, our next step is to make a further analysis of the obtained perceptual distances in order to obtain insights regarding the structure of the perceptual space composed of the different scene categories. Fortunately, analysis of this sort can be done computationally using *multidimensional scaling* (MDS)—a technique from statistical inference and data visualization to embed a set of objects in Euclidean space while preserving their distance as much as possible (Torgerson, 1952). This technique fits very well our goal because the data it requires is typically a measure of the dissimilarity between the objects under investigation (in our case, the scene categories). The output is a spatial organization in which similar objects are placed nearby while dissimilar objects remain apart in the embedding space, all while distorting the input distances as little as possible. (Note that some distortion may be unavoidable if the data comes from a non-Euclidean space.)

As can be appreciated from the results of the MDS analysis in two dimensional space (Figure 3), the different classes that participated in our experiment appear to split into two main groups, grossly described as natural (left group) and manmade (right group) scenes. This intuitive division is also obtained formally once we apply the *k-means* clustering method to cluster the scenes into two clusters. Indeed, the result (coded in color in Figure 3) shows a division along the natural/manmade classification. Clearly, these results are consistent with recent work (Joubert et al., 2007; Loschky & Larson, 2010) and provide additional perceptual support to the observation that human observers tend to prefer the natural/manmade distinction when segregating scene images into two groups (Oliva & Torralba, 2001). Still, one piece of these results is perhaps a bit surprising and warrants a second look. The clustering algorithm groups the highway category into the natural cluster, indicating that the highway category is perceptually more related to the natural categories rather than to the manmade categories. While this appears to be inconsistent with previous work (Oliva & Torralba, 2001; Loschky & Larson, 2010), or with the fact that *Highways* are not natural objects, inspecting many *Highway* images raises some serious doubts whether this type of scene should be considered manmade (see Figure 4). Although more analyses may be needed to make conclusive statements, it appears as if the *Highway* category indeed shares more perceptual properties with natural scene categories than with manmade scene categories.

With the natural/manmade division established as the first level of decision, it is now natural to examine the perceptual structure of each of the two subclasses separately. This can be done by repeating the same MDS analysis followed by clustering but this time to

Figure 4. Selected scene images from the highway category (see subsection Stimuli). Although more analyses may be needed to make conclusive statements, it appears as if the *Highway* category indeed shares more perceptual properties with natural scene categories than with manmade scene categories.

each of the natural/manmade categories in isolation. Figure 5 illustrates the result of this operation and provides several new insights. Interestingly, manmade scenes seem to divide naturally between indoor and outdoor scenes, and natural scenes seem to separate *Coasts* and *Highways* from *Forests* and *Mountains*. While this latter distinction may be related to the openness perceptual property of a scene (Oliva & Torralba, 2001), further research with more natural scene categories is needed to better understand this part of the hierarchy (see also Experiment 2).

With the results and analysis discussed thus far, several insights are obtained. In particular, under brief presentation humans are able to discriminate much more easily manmade scenes from natural scenes rather than between different manmade scenes or natural scenes. This can be observed intuitively from the visualization of the obtained perceptual space (Figure 3) and formally from the output of the unsupervised clustering algorithm whose result indeed seems to group the scenes into natural and manmade clusters. If we break this division further, a second level in a hypothetical hierarchy is revealed. Indeed, human observers are able to discriminate indoor manmade scenes from outdoor manmade scenes more easily than the discrimination of scene types inside each of these classes. Evidence for similar behavior is obtained for natural scenes, although the small numbers of natural scenes in our experiment requires further verification of this structure using an extended experiment (see Experiment 2). Such results might predict that it would be much easier to discriminate *Mountain* from *Bedroom*

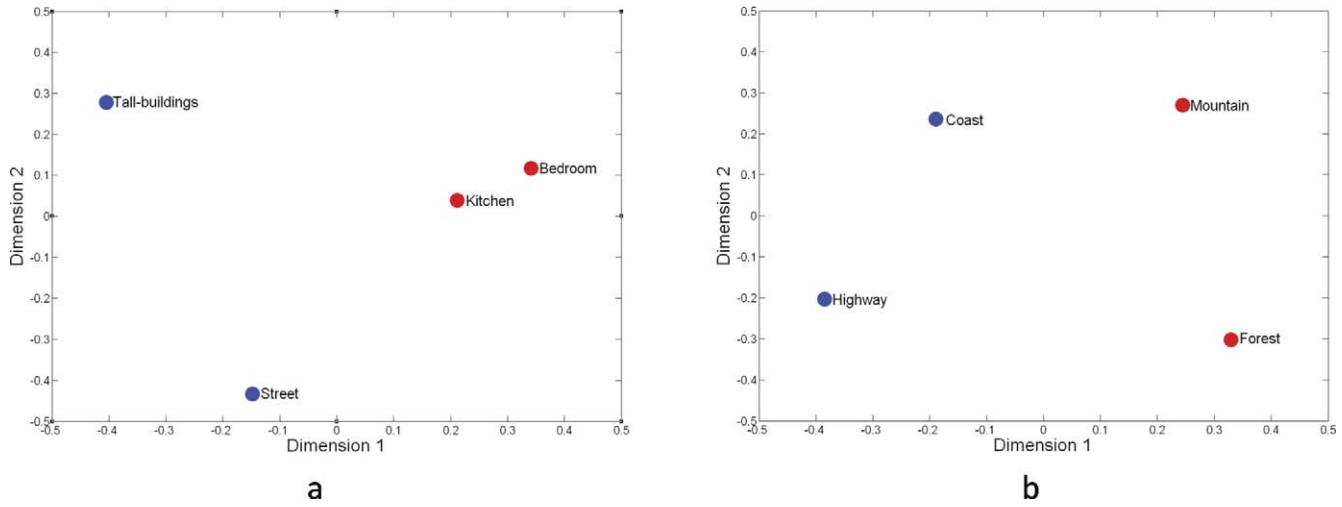

Figure 5. MDS + clustering analysis of the second level of the scene categorization hierarchy. (a) Results on the manmade scene categories can be naturally interpreted as division between indoor and outdoor manmade scenes. (b) Results on the natural scene categories separate *Coasts* and *Highways* from *Forests* and *Mountains*, possibly by their degree of their perceptual openness (Oliva & Torralba, 2001).

than it would be to discriminate *Forest* from *Coast*, *Street* from *Highway*, *Kitchen* from *Bedroom*, and so forth.

Temporal dynamics and evidence for hierarchical processing

Recall that Table 1 presents subjects' sensitivity averaged over all presentation times. Next we examine what additional insights can be obtained by breaking down these averages by presentation times. Indeed, in conjunction with the use visual masking, the wide range of stimulus onset asynchrony (SOA) values used in our experiment facilitates the analysis of our results in terms of their processing time. This is based on the assumption that variation of SOAs can be used to vary processing time, which is supported by over 100 years of visual masking research (for a review, see Breitmeyer & Ogmen, 2006), including recent studies using single cell recording in macaques (Kovacs, Vogels, & Orban, 1995; Rolls, Tovee, & Panzeri, 1999) and brain imaging in humans (BaconMace et al., 2005; Rieger, Braun, Bulthoff, & Gegenfurtner, 2005).

More specifically, we examined subjects' responses in all trials, broken down by all levels of the hierarchical structure (i.e., manmade vs. natural, etc.) and plotted as a function of stimulus PT. Again, to exclude the possibility that such analysis (and its results) is confounded by certain biases in subjects' responses, we analyzed these responses using the nonparametric signal detection measure A' of sensitivity (Grier, 1971). The results, presented in Figure 6 and in Table 2, indicate several observations. First, subjects' sensitivity at the shortest presentation times was significantly better in discriminating manmade from natural scenes $A'(27\text{ ms}) = 0.66$ and $A'(40\text{ ms}) = 0.77$, compared to discriminating manmade-indoor from manmade-outdoor scenes, $A'(27\text{ ms}) = 0.55$, $p < 0.0005$ and $A'(40\text{ ms}) = 0.66$, $p \leq 0.0001$, two-sample t test, or compared to discriminating natural-open from natural-closed scenes, $A'(27\text{ ms}) = 0.58$, $p < 0.0095$ and $A'(40\text{ ms}) = 0.67$, $p < 0.0005$, two-sample t test. We argue that this is a strong evidence that the distinction between manmade versus natural scenes is processed prior to these finer distinctions.

Second, the advantage of discriminating natural from manmade scenes over discriminating natural-open from natural-closed scenes drops significantly and becomes statistically insignificant once presentation time is increased to PT = 53 ms, with sensitivity level of 0.85 in the manmade versus natural trials, compared to 0.81 ($p > 0.073$, two-sample t test) in the natural open versus natural closed scenes. In a similar way, the advantage of discriminating natural from manmade scenes over discriminating manmade-indoor from manmade-outdoor scenes drops significantly and becomes statistically insignificant once presentation time is increased to PT = 80 ms, with sensitivity level of 0.9 in the manmade versus natural trials, compared to 0.87 ($p > 0.084$, two-sample t test) in the manmade-indoor versus manmade-outdoor scenes. In other words, the discrimination between manmade versus natural scene categories becomes equally effective to the discrimination of natural open versus natural closed scenes once presentation time is increased to PT = 53 ms and equally effective to the discrimination of manmade-indoor versus manmade-outdoor scenes once presentation time is increased to PT = 80 ms.

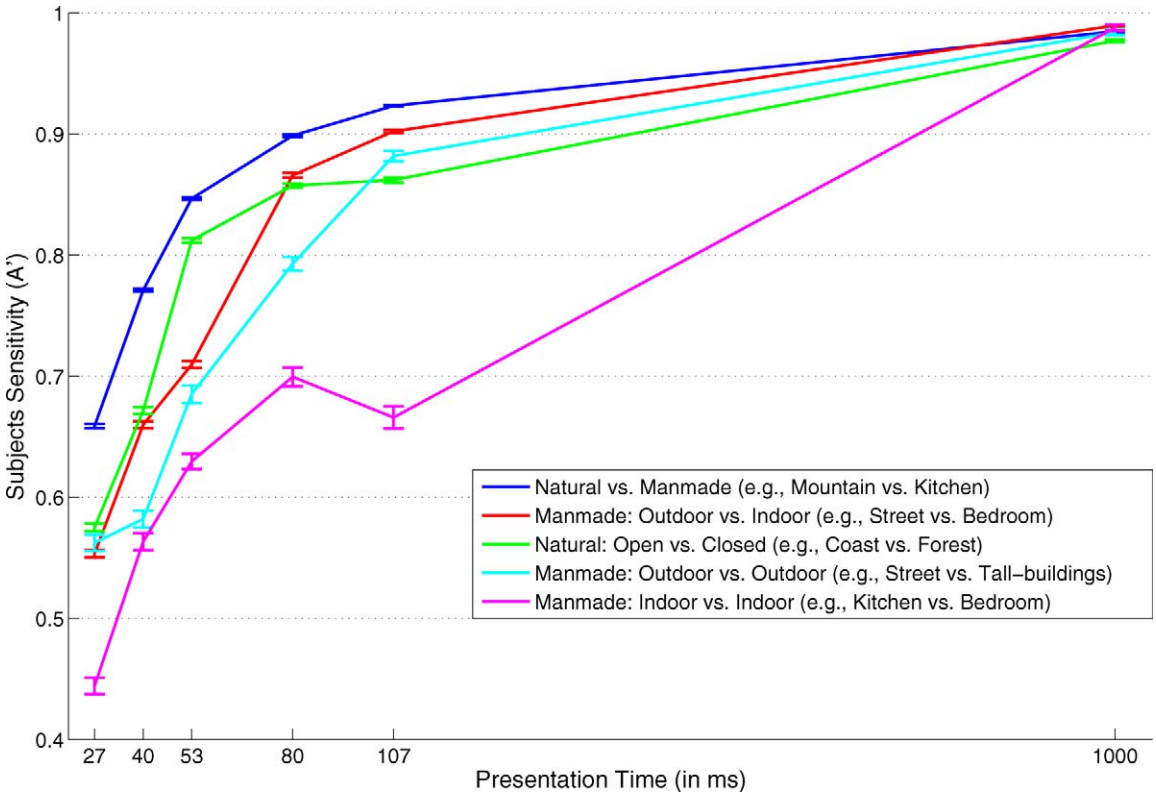

Figure 6. Hierarchical processing is revealed in the analysis of all trials with subjects' sensitivity presented as a function of stimulus PT. Note how the advantage of each level of the hierarchical structure compared to its subsequent levels is decreased once presentation time increases. See text and Table 2 for more details and quantitative data.

While discriminating scene categories at the first level of the hierarchical structure loses its advantage when PTs increase to 53 ms and 80 ms, the second-level in the hierarchy (e.g., manmade-indoor vs. manmade-outdoor) still exhibits a significant advantage over the third-level division at these very same PTs. With a sensitivity level of 0.87 at PT = 80 ms in manmade-indoor versus manmade-outdoor trials, it is significantly better than discrimination within each of these classes ($A' = 0.7$, $p \leq 0.0001$ between different manmade-indoor scenes and $A' = 0.79$, $p \leq 0.05$ between different manmade-outdoor scenes, two-sample t test). This suggests that the distinction between manmade-indoor and manmade-outdoor is made prior to distinctions within these two subclasses, a behavior that becomes a repeating pattern over the levels of the hierarchy.

Considered together, these results based on PT suggest that in accordance with the perceptual distance between categories and the overall structure of the

Hierarchical structure level	27 ms	40 ms	53 ms	80 ms
Manmade vs. natural	0.66	0.77	0.85	0.90
Natural open vs. natural closed	**0.58** (0.0094)	**0.67** (0.0003)	0.81 (0.073)	
Manmade indoor vs. manmade outdoor	**0.55** (0.0004)	**0.66** (0.0001)		0.87 (0.084)
Manmade indoor vs. manmade indoor				**0.70** (0.0001)
Manmade outdoor vs. manmade outdoor				**0.79** (0.05)

Table 2. Explicit sensitivity values and statistical significance compared to previous level in the hierarchy (in parentheses) for selected data points from Figure 6. Boldface highlights statistically significant entries. Note how subjects' sensitivity at the shortest PTs (27 and 40 ms) was significantly worse in discriminating natural-open from natural-closed scenes or manmade-indoor from manmade-outdoor scenes compared to discrimination at the top hierarchical division of manmade versus natural scenes. The advantage of the latter over natural-open versus natural-closed scenes diminishes and becomes statistically insignificant once presentation time is increased to PT = 53 ms. Similarly, the same advantage over manmade-indoor versus manmade-outdoor scenes diminishes at PT = 80 ms. Still, at this last PT, discrimination at the third-level hierarchical division (i.e., discrimination between different manmade-indoor scenes or between different manmade-outdoor scenes) remains statistically significantly worse.

corresponding perceptual space, the visual process underlying scene gist recognition is hierarchical as well. Indeed, it seems as if decisions related to the top part of our hierarchy tree (i.e., whether the scene is manmade or natural) are made first while more complicated decisions (such as whether a manmade scene is indoor or outdoor) are taken later. This strategy is ecologically viable since with the pruning of complete classes of scene categories at earlier stages, more complicated decisions need to cope with fewer candidate categories and hence can be made both faster and more accurately. A proof of this concept requires computational modeling of the categorization process and is left here as future work (see the Discussion and conclusions section).

Experiment 2

The results of Experiment 1 reveal a hierarchical perceptual structure between different scene categories and a corresponding hierarchical structure at the perceptual processing level to suggest that scene gist recognition is hierarchical in nature. Importantly, these results are established without assuming the existence of a particular hierarchical structure ahead of time, as had been done in previous studies (Loschky & Larson, 2010), but rather they emerge directly from subjects' responses. This, however, raises the issue of the range of scene categories involved, since one cannot avoid noticing that our experimental evaluation was limited to eight scene categories. Indeed, the scale of Experiment 1 is on par with previous work in the perceptual and psychophysical literature (e.g., eight classes in Oliva & Torralba, 2001; Loschky & Larson, 2008, 2010 or 10 classes in Loschky et al., 2007). Still, the use of more categories can reveal a hierarchical structure that is less dependent on the specific choice of basic-level categories. This consideration indicates that a second experiment with a wider range of categories is needed to exclude the possibility that the results are confounded by a particular selection of scene categories.

Unfortunately, it is very difficult to obtain reliable perceptual data for many scene categories while using the controlled lab procedure described in the Methods section of Experiment 1. There are two main reasons for this. First, since subjects need to remember the scene categories that participate in the experiment in order to facilitate their "same" or "different" decision, they fail to do so when the number of categories exceeds some memory threshold. Second, the number of trials in such an experiment grows quadratically with the number of participating categories and exceeds the capacity of normal human subjects for more than 10 classes.

As a result of the above, it is clear that the data collection procedure must be amended to facilitate the acquisition of perceptual relationships between many categories, with the ambitious goal possibly one that targets the SUN database of 908 categories (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010). One way to achieve this is by repeating numerous times the procedure from Experiment 1, each time for a small number of categories selected randomly from the large database. Pooling data from a large number of participants may then overcome both limitations and provide the knowledge base to construct the perceptual relations between all categories in the database. We note that relying on the collective effort of a large population of users had already proven successful (Von Ahn & Dabbish, 2004; Torralba, Russell, & Yuen, 2010) and toward the goal above we have developed an online experimental system that harnesses the power of the web to facilitate such data collection.

The online experimental system was developed as a Silverlight application that allows users to participate in our experiment from any place with an internet connection (Kadar & Ben-Shahar, 2011). This application executes an experiment similar to the one described in Experiment 1, but in a form of a game that motivates users to participate. From a participant's perspective, the goal is to get his or her name into the top-ten online ranking by maximizing one's score for guessing whether two briefly presented scenes belong to the same category or not. We believe that with an online experimental system like that that the perceptual relations between any number of scene categories (and in particular, those in the SUN database, Xiao et al., 2010) could be established in reasonable time. Experiment 2 takes a first step in this direction by exploring the relations between all scene categories in the Scene-15 dataset (Oliva & Torralba, 2001; Fei-Fei & Perona, 2005; Lazebnik, Schmid, & Ponce, 2006), which was the largest available dataset of scene categories until recently and is now included in the SUN database.

Methods

Subjects

The web-based experiment was shared through social and professional networks with students, friends, and colleagues. In total, 389 subjects from 33 countries (according to Google Analytics) volunteered to participate in the web-based experiment.

Stimuli

The experiment was performed with the Scene-15 dataset which was compiled by several researchers (Oliva & Torralba, 2001; Fei-Fei & Perona, 2005;

Lazebnik et al., 2006). It consists of 4,485 images spread over 15 categories: *Bedroom*, *Living room*, *Kitchen*, *Office*, *Store*, *Street*, *Tall-buildings*, *Suburb*, *Inside city*, *Industrial*, *Highway*, *Coast*, *Open-country*, *Mountain*, and *Forest*. This dataset expands on that used in Experiment 1 with seven additional categories, three that are considered manmade-indoor categories (*Living room*, *Office*, and *Store*), three manmade-outdoor categories (*Suburb*, *Inside city*, and *Industrial*) and one natural category (*Open-country*). Sixty-nine scene images from each basic-level category were randomly selected, reduced to monochrome, and adapted in size to 256 × 256 pixels. The same set of masks from Experiment 1 was used in this experiment.

Procedure

In the beginning of each experiment participants were shown the instructions while the system randomly selected four different categories out of the total 15. Participants then needed to complete a category learning procedure using 24 images (six from each category) so that they could get acquainted with the scene category labels. Then they ran five practice trials so they could become familiar with the experimental procedure and task. The experiment itself followed all these steps and consists of 50 trials of the form discussed in the Methods section of Experiment 1. Including category learning and practice phases, the entire experiment lasted around 5 min for each subject.

Unlike in the lab, using a web application results in the inability to control precisely various experimental parameters, the most critical of which are presentation times. Very short presentation times are excluded because of the inability to ensure small relative error in their value when executed on unknown computer platform and display device. Hence, we currently limit PTs to 50, 100, and 200 ms, using the latter also as catch trials to validate subject's awareness. (High error rates in this PT would indicate unreliable subject.) Except as noted, the sequence of events in an experimental trial are identical to those of Experiment 1.

Results and observations

We first analyze subjects' responses in the catch trials with PT = 200 ms to validate subject's awareness. Note that catch trials in Experiment 2 used stricter PT than Experiment 1 (i.e., 200 ms vs. 1000 ms) to compensate for the inability to control various aspects of the experiment due to its online and remote nature. To exclude unreliable subjects, we set a threshold of 0.75 on average discrimination accuracy (i.e., at the midpoint between chance level and perfect discrimination) in PT = 200 ms trials. Table 3 shows the sensitivity A' between all pairs of scene categories in the Scene-15 dataset obtained by averaging the results from the 293 reliable subjects over all PTs.

Figure 7 illustrates the result of the MDS and k-means analyses on the perceptual distances obtained in the web-based experiment. Indeed, the results show a similar configuration to Experiment 1, which cluster natural scenes separately from manmade scenes, with *Highways* being an anomalous borderline case marginally grouped with manmade categories.[1] Figure 8 further illustrates the result of the MDS and k-means analyses of the natural/manmade categories in isolation. Again, similar to Experiment 1, manmade scenes seem to divide between indoor and outdoor scenes while natural scenes seems to divide between open and closed scenes. Considered together, this provides a strong support to the results obtained in Experiment 1 but this time with twice as many categories to exclude the possibility that the original conclusions are confounded by a particular selection of basic-level categories.

Interestingly, not unlike *Highways* in Experiment 1, one piece of the results of Experiment 2 is perhaps a bit surprising and warrants a closer look. The clustering algorithm groups the *Store* category with (what would naturally be described as) manmade-outdoor classes, indicating that the *Store* category is perceptually more related to manmade-outdoor than to manmade-indoor categories. While this appears to be inconsistent with the fact that most of the *Store* scenes in the web-based experiment are taken indoors, inspecting many of them next to manmade scenes from various categories raises some serious doubts as to their perceptual relation to indoor scenes rather than to outdoor scenes (see Figure 9). Further research with more manmade categories may be needed to better understand and explain this observation.

One cannot avoid noticing the omission of a temporal analysis of the Experiment 2 data (of the sort shown in Figure 6), a fact that may seem surprising given the main thesis of this paper. However, there are several reasons for this. First, owing to the fact that Experiment 2 is executed remotely via the web on unknown computer platforms and display devices, it runs the risk of high relative error in short SOAs (e.g., the program may request to present the stimulus for 27 ms but in practice the presentation could be very different for any one of many reasons, from the speed of the unknown computer through the implementation level of the unknown web client to the refresh rate of the unknown display. This relative error drops significantly for higher SOAs). The sensitivity of the experiment to higher SOAs indeed limited our implementation to much fewer SOAs (50, 100, and 200 ms only) and once missing the shortest SOAs, the

	Bedroom	*Living room*	*Kitchen*	*Office*	*Store*	*Street*	*Tall-building*	*Suburb*	*Inside city*	*Industrial*	*Highway*	*Coast*	*Open-country*	*Mountain*	*Forest*
Bedroom	**0.83**	0.60	0.71	0.77	0.85	0.88	0.89	0.80	0.86	0.83	0.88	0.90	0.86	0.88	0.91
Living room	-	**0.83**	0.72	0.72	0.82	0.85	0.88	0.81	0.79	0.84	0.86	0.90	0.88	0.90	0.89
Kitchen	-	-	**0.84**	0.71	0.79	0.85	0.88	0.85	0.74	0.82	0.88	0.89	0.92	0.92	0.91
Office	-	-	-	**0.83**	0.73	0.82	0.86	0.86	0.79	0.84	0.88	0.86	0.92	0.89	0.90
Store	-	-	-	-	**0.83**	0.81	0.87	0.84	0.74	0.74	0.87	0.89	0.91	0.89	0.92
Street	-	-	-	-	-	**0.83**	0.85	0.80	0.73	0.72	0.71	0.87	0.87	0.88	0.87
Tall-building	-	-	-	-	-	-	**0.88**	0.88	0.78	0.74	0.85	0.93	0.91	0.91	0.93
Suburb	-	-	-	-	-	-	-	**0.86**	0.76	0.79	0.82	0.88	0.89	0.92	0.92
Inside city	-	-	-	-	-	-	-	-	**0.83**	0.77	0.89	0.88	0.88	0.91	0.91
Industrial	-	-	-	-	-	-	-	-	-	**0.78**	0.80	0.82	0.86	0.88	0.89
Highway	-	-	-	-	-	-	-	-	-	-	**0.84**	0.82	0.81	0.90	0.90
Coast	-	-	-	-	-	-	-	-	-	-	-	**0.86**	0.66	0.77	0.80
Open-country	-	-	-	-	-	-	-	-	-	-	-	-	**0.85**	0.83	0.81
Mountain	-	-	-	-	-	-	-	-	-	-	-	-	-	**0.89**	0.80
Forest	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**0.90**

Table 3. The Perceptual Data matrix obtained by measuring reliable subjects' average sensitivity between all pairs of scene categories in the Scene-15 dataset. As before, the bolded results along the diagonal represent a measure of perceptual similarity within each category. All other entries represent the perceptual distance between their corresponding categories for the purpose of gist recognition.

motivation for temporal analysis is essentially lost. Second, recall that the only way to obtain mass data with human subjects in larger group of categories was to limit each subject to four categories selected randomly from the larger pool of classes and then to accumulate and compile a joint dataset. As a result, the sensitivity over all levels of distinctions for each subject simply cannot be evaluated (e.g., it is not inconceivable that one subject would be tested on four natural scene categories only, while another would be tested on four manmade indoor categories only).

Discussion and conclusions

The current study investigates the perceptual relations between scene categories to explore the perceptual structure and the order of processing in scene gist recognition across taxonomic levels. For that, we introduce the scene category discrimination paradigm based on the common discrimination procedure, from which we obtain novel perceptual data and insights about gist recognition. In particular, we analyze

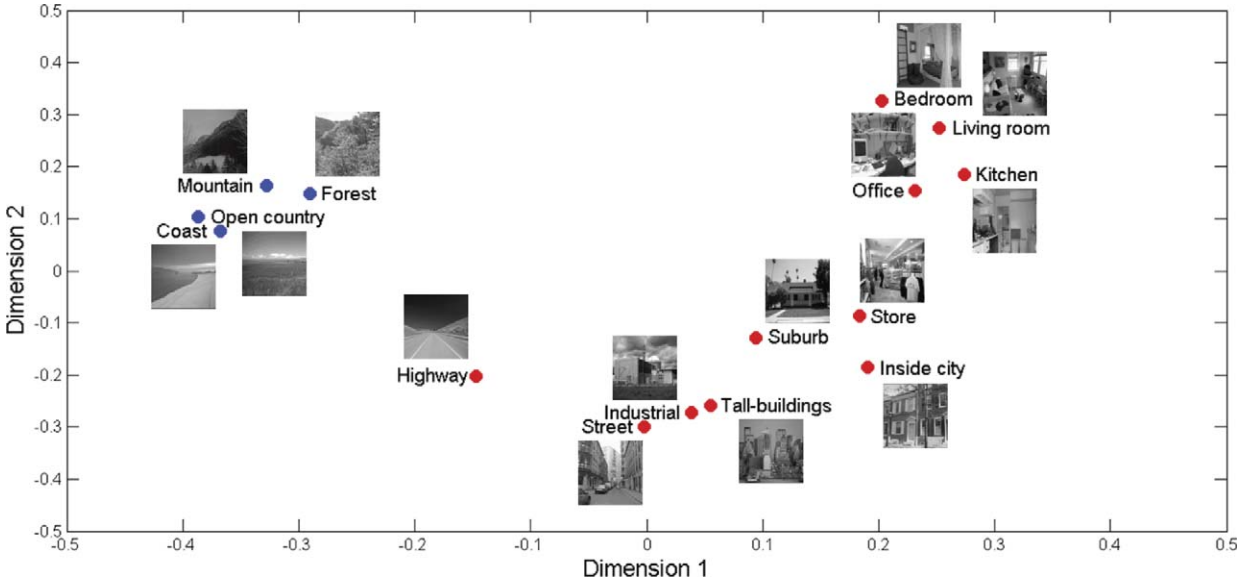

Figure 7. Applying MDS and k-means analyses on the perceptual distances obtained in the web-based experiment and visualizing the results as points in a 2D perceptual spaces reveal a similar configuration to Experiment 1 which clusters natural scenes separately from manmade scenes, with *Highways* being an anomalous borderline case.

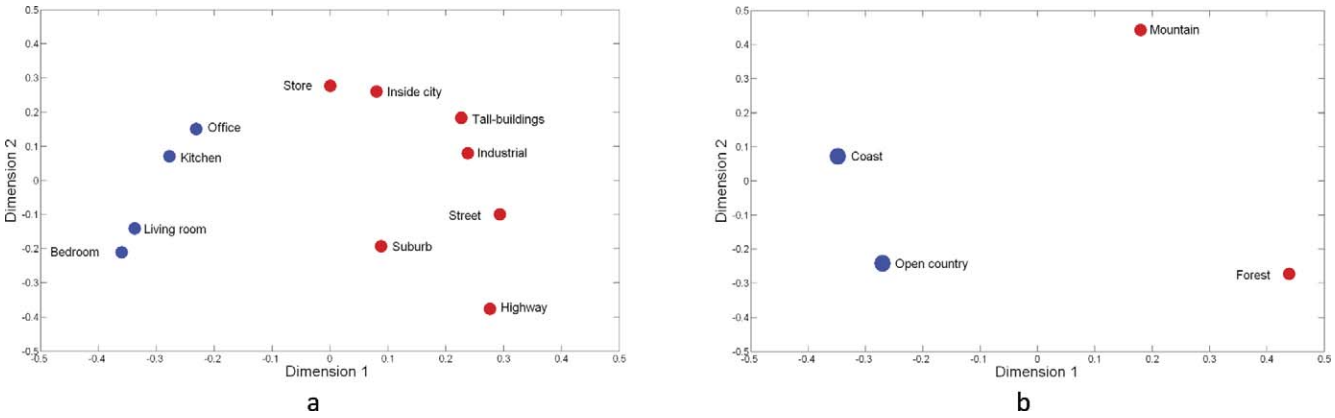

Figure 8. MDS + clustering analysis of the second level of the gist recognition hierarchy. (a) Results on the manmade scene categories can be naturally interpreted as a division between indoor and outdoor manmade scenes with *Store* being an anomalous borderline case. (b) Results on the natural scene categories can be interpreted as a division between open and closed scenes.

observers' responses to extract perceptual distance between scene categories, from which a complete hierarchical perceptual structure is revealed. Analysis as a function of stimulus presentation time and SOA further reveals a similar hierarchical structure at the perceptual processing level.

Consistent with several previous studies (Oliva & Torralba, 2001; Joubert et al., 2007; Loschky & Larson, 2010), our work provides a new and solid type of evidence that the natural/manmade distinction is made before basic-level categorization, a result that conflicts with the default assumption of the basic-level primacy in scene gist recognition (Rosch, 1978; Tversky & Hemenway, 1983). Moreover, the current results are also consistent with the coarse-to-fine processing order (Schyns & Oliva, 1994), which first derive a coarse description of the scene that may be sufficient for superordinate distinction before processing more detailed information to infer basic-level distinctions.

Unlike previous work, however, our results are obtained without imposing or assuming any prior structure on the decision process or the comparison of specific alternatives. (For example, Loschky & Larson, 2010, essentially assumed prior division of the stimuli to manmade and natural categories while comparing performance in two predefined alternatives—superordinate vs. basic level categorization.) Furthermore, our study provides a research methodology and practical results to argue that the precedence of natural/manmade distinction is only one (possibly the first) step in a hierarchy of decisions, where each level deals with a finer subdivision of its parent classes.

A possible critique of the current study is that it may encourage a different type of processing at different

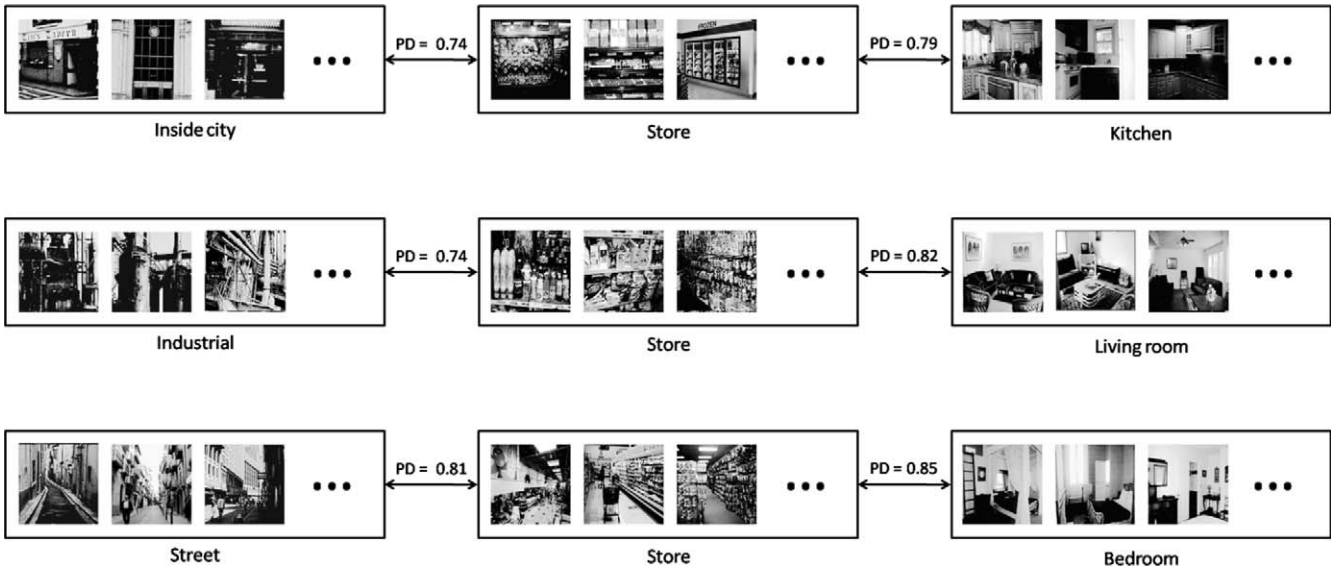

Figure 9. Selected images from the *Store* category and various manmade categories suggest that the *Store* category indeed shares more perceptual properties with manmade-outdoor scene categories rather than with manmade-indoor scene categories.

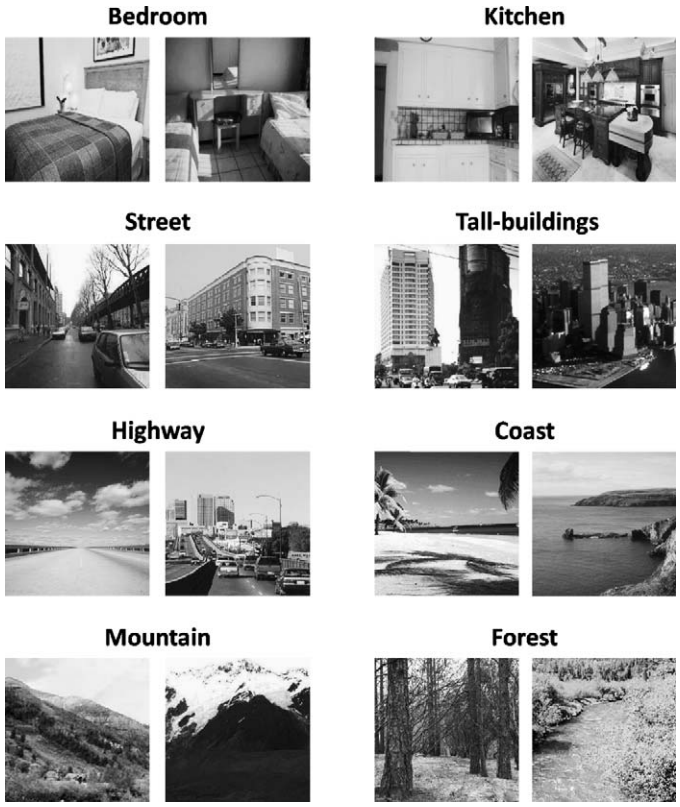

Figure 10. Looking at individual images from different scene categories suggests that a pair of images from the same category may not necessarily be similar. Despite their lack of similarity, subjects responded "same" for these demonstrated pairs at the shortest SOA of 27 ms. This, and a bias analysis of subjects' responses (see text), suggests that the possibility that subjects switch from category judgment to similarity judgment at shortest SOAs in unlikely.

SOAs. In particular, in the absence of a full and clear representation of both stimuli at very short SOAs, subjects may unconsciously be making an image similarity judgment rather than a category judgment. However, there are at least two pieces of evidence that weaken this possibility. First, at the base of this concern is the assumption that category members are perceptually similar. This is an intuitive assumption, but is it true? Naturally, if it was strictly true, there was little reason to differentiate between category judgment and similarity judgment at all (i.e., regardless of SOA). But looking at individual instances from different categories suggests differently. As demonstrated in Figure 10, pairs of images from the same category may not necessarily be similar at all. Second, despite their lack of similarity, subjects responded "same" for these demonstrated pairs at the shortest SOA of 27 ms (which is inconsistent with the discussed switch from category judgment to similarity judgment at shortest SOAs). Finally, the observation demonstrated in Figure 10 (i.e., that members of a particular category may not look similar) suggests that two images from different categories will be judged "not similar" virtually always, while two images from the same category may be judged "not similar" at least some of the time. In other words, if this possibility was fully valid, making judgments by similarity rather than category membership would produce a significant bias to respond "different" at the shortest SOA in our experiment. However, an analysis of subjects' responses at the shortest SOA excludes this possibility. Indeed, subjects did not have any special tendency to respond "different" at the shortest SOA, with probability 0.506 ($p > 0.72$, one-sample t test) to respond "different" at SOA = 27 ms.

While considering the interpretation of data at various SOAs, it is important to note that the current findings are based on the assumption that sensitivity differences at different SOAs reflect different time courses of information integration and use (Breitmeyer & Ogmen, 2006). However, other interpretations may be possible as well. For example, it may be that the precedence of the superordinate distinction over basic-level distinctions follows a more efficient extraction of those features that are relevant for superordinate distinction compared to those features relevant for basic-level distinctions (whose extraction requires additional processing time). Further studies measuring the time course of brain processes may be required to fully address this issue.

But what are the benefits that such hierarchical processing may provide? We argue that this strategy is beneficial not only in decision accuracy but in processing time as well, since the deeper one goes in the hierarchy, the fewer candidate categories remain viable. While the latter hypothesis requires further research, it is likely to hold for both biological and artificial visual systems. In fact, we argue that lack of proper knowledge about such a hierarchical perceptual structure may be the reason why traditional artificial visual systems categorize natural scene images in a linear (one against all) fashion rather than hierarchically (Vogel & Schiele, 2004; Fei-Fei & Perona, 2005; Bosch, Zisserman, & Munoz, 2006; Lazebnik et al., 2006; Xiao et al., 2010). To our best knowledge, the only exception to this prevailing approach are the two related models proposed by Oliva, Torralba, Guerin-Dugue, and Herault (1999) and Oliva and Torralba, 2001, where a predetermined, two stage hierarchical process was utilized (i.e., superordinate-level distinction followed by basic-level categorization). Unfortunately, the hierarchical nature of these two models was never followed up, nor was it tested rigorously. Indeed, the hierarchy in the first model (Oliva et al., 1999) was selected mostly intuitively following informal observations while the latter (Oliva & Torralba, 2001) was derived from the most cited criteria on which subjects

segregated a group of 81 images to two groups after unlimited presentation time. Here, however, we provide experimental (perceptual) evidence to support and validate this hierarchical approach, to suggest its extension to multiple levels, and to imply that it is readily extended to other artificial visual systems for scene classification and gist recognition.

A main take-home message in our paper is that advancing our understanding of scene gist processing must involve additional insights about the hierarchical structure of the corresponding perceptual space. While one could suggest that this may be done quite conveniently using existing hierarchical semantic structures such as Wordnet (Miller, 1995), we argue that semantic relations between categories do not necessarily agree with their perceptual relationship. For example, concepts such as "snowy mountains" and "skiing activity" are far from each other semantically (with semantic distance $SD = 0.88$ based on a popular distance measure used in Wordnet, Fergus, Bernal, Weiss, & Torralba, 2010) while perceptually they are very much related. This observation is particularly salient in the context of visual scenes. Indeed, even our modest experimental setups reveal that the *Highway* category is perceptually closer to *Coast* than to *Kitchen*, and that the *Store* category is perceptually closer to *Street* than to *Bedroom*, although semantically the opposite holds, $SD(Highway, Coast) = 0.57 > SD(Highway, Kitchen) = 0.38$; $SD(Store, Street) = 0.5 > SD(Store, Bedroom) = 0.4$. Arguing that the hierarchical structure involved in human gist recognition should be based on perceptual criteria that could be inferred or determined directly from human vision, this research is aimed to make an important step in this direction, in part by suggesting a research methodology, analysis methods, and an online experimental system for obtaining perceptual relations for large collections of scene categories. We call upon the community to help collect these data, and we hope it will facilitate the construction of the complete hierarchical structure involved in the perception and processing of the gist of a scene.

Acknowledgments

This work was funded in part by the European Commission in the Seventh Framework Programme (CROPS GA no. 246252), the Frankel fund, the Paul Ivanier center for Robotics Research, and the Zlotowski Center for Neuroscience at Ben-Gurion University. Some information included in this paper was presented in the 2011 annual meeting of the Vision Sciences Society (VSS). The authors thank the editor and three anonymous reviewers for their helpful comments.

Commercial relationships: none.
Corresponding author: Ohad Ben-Shahar.
Email: ben-shahar@cs.bgu.ac.il.
Address: Computer Science Department and the Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Beer-Sheva, Israel.

Footnote

[1] To understand how borderline *Highways* are by increasing the reliability threshold to 0.80 and repeating the analysis, we obtain a similar perceptual space configuration with *Highways* remaining a borderline case which now groups with natural categories.

References

BaconMace, N., Mace, M., Fabre-Thorpe, M., & Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision Research, 45*, 1459–1469.

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

Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kubovy & J. Pomerantz (Eds.), *Perceptual organization* (pp. 213–254). Hillsdale, NJ: Lawrence Erlbaum.

Biederman, I., Glass, A., & Stacy, E. J. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology, 97*(1), 22–27.

Bosch, A., Zisserman, A., & Munoz, X. (2006). Scene classification via pLSA. *Proceedings of the European Conference on Computer Vision, 3954*, 517–530.

Brainard, D. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Breitmeyer, B. G., & Ogmen, H. (2006). *Visual masking: Time slices through conscious and unconscious vision*. New York: Oxford University Press.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we see when we glance at a scene? *Journal of Vision, 7*(1):10, 1–29, http://www.journalofvision.org/content/7/1/10, doi:10.1167/7.1.10. [PubMed] [Article]

Fei-Fei, L., & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2,* 524–531.

Fei-Fei, L., 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 USA, 99*, 9596–9601.

Fergus, R., Bernal, H., Weiss, Y., & Torralba, A. (2010). Semantic label sharing for learning with many categories. *Proceedings of the European Conference on Computer Vision, 6311*, 762–775.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75*, 424–429.

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research, 47*(26), 3286–3297.

Kadar, I., & Ben-Shahar, O. (2011). *An online experimental system*. Retrieved May, 2011, http://www.cs.bgu.ac.il/~vision/pmc.

Kovacs, G., Vogels, R., & Orban, G. A. (1995). Cortical correlate of pattern backward-masking. *Proceedings of the National Academy of Sciences of the USA, 92*(12), 5587–5591.

Large, M., Kiss, I., & McMullen, P. (2004). Electrophysiological correlates of object categorization: Back to basics. *Cognitive Brain Research, 20*(3), 415–426.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2,* 2169–2178.

Loschky, L., & Larson, A. (2008). Localized information is necessary for scene categorization, including the natural/man-made distinction. *Journal of Vision, 8*(1):4, 1–9, http://www.journalofvision.org/content/8/1/4, doi:10.1167/8.1.4. [PubMed] [Article]

Loschky, L., & Larson, A. (2010). The natural/man-made distinction is made before basic-level distinctions in scene gist processing. *Visual Cognition, 18*, 513–536.

Loschky, L., Sethi, A., & Simons, D. (2007). The importance of information localization in scene gist recognition. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 1431–1450.

Mace, M. J. M., Joubert, O. R., Nespoulous, J. L., & Fabre-Thorpe, M. (2009). The time-course of visual categorizations: You spot the animal faster than the bird. *PLoS ONE, 4*(6), e5927.

Miller, G. (1995). Wordnet: A lexical database for English. *Communications of the ACM, 38*, 39–41.

Oliva, A. (2005). Gist of the scene. In L. Itti, G. Rees, & J. Tsotsos. (Eds.), *Neurobiology of attention* (pp. 251–256). San Diego, CA: Elsevier.

Oliva, A., & Schyns, P. (1994). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology, 34*, 72–107.

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

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*(3), 145–175.

Oliva, A., Torralba, A., Guerin-Dugue, A., & Herault, J. (1999). Global semantic classification of scenes using power spectrum templates. *Proceedings of the Challenge of Image Retrieval, Newcastle. Electronic Workshop Computer Series,* Springer-Verlag.

Peelen, M. V., Fei-Fei, L., & Kastner, S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature, 460*, 94–97.

Potter, M. (1975). Meaning in visual search. *Science, 187*, 965–966.

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

Potter, M., & Levi, E. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology, 81*, 10–15.

Rieger, J. W., Braun, C., Bulthoff, H. H., & Gegenfurtner, K. R. (2005). The dynamics of visual pattern masking in natural scene processing: A magnetoencephalography study. *Journal of Vision, 5*(3):10, 275–286, http://www.journalofvision.org/content/5/3/10, doi:10.1167/5.3.10. [PubMed] [Article]

Rogers, T. T., & Patterson, K. (2007). Object categorization: Reversals and explanations of the basic-level advantage. *Journal of Experimental Psychology, 136*, 451–469.

Rolls, E., Tovee, M. J., & Panzeri, S. (1999). The neurophysiology of backward visual masking: Information analysis. *Journal of Cognitive Neuroscience, 11*(3), 300–311.

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*(7), 629–630.

Rousselet, G. A., Joubert, O. R., & Fabre-Thorpe, M.

(2005). How long to get to the "gist" of real-world natural scenes. *Visual Cognition, 12*(6), 852–877.

Schyns, P., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time and spatial scale dependent scene recognition. *Psychological Science, 5*, 195–200.

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

Torgerson, W. S. (1952). Multidimensional scaling: Theory and method. *Psychometrika, 17*(6), 401–419.

Torralba, A., Russell, B. C., & Yuen, J. (2010). Labelme: Online image annotation and applications. *Proceedings of the IEEE, 98,* 1467–1484.

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

Vogel, J., & Schiele, B. (2004). Semantic typicality measure for natural scene categorization. *Pattern Recognition Symposium* DAGM 2004, Tubingen, Germany.

Von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* Vienna, Austria, pp. 319–326.

Walker, L., & Malik, J. (2002). When is scene recognition just texture recognition? *Vision Research, 44*, 2301–2311.

Xiao, J., Hays, J., Ehinger, K., Oliva, A., & Torralba, A. (2010). Sun database: Large scale scene recognition from abbey to zoo. *Proceedings of the 23rd IEEE conference on Computer Vision and Pattern Recognition*, pp. 3485–3492. IEEE Computer Society.