Visual categorization is automatic and obligatory: Evidence from Stroop-like paradigm

Michelle R. Greene Department of Computer Science, Stanford University, Stanford, CA

Li Fei-Fei Department of Computer Science, Stanford University, Stanford, CA

Human observers categorize visual stimuli with remarkable efficiency—a result that has led to the suggestion that object and scene categorization may be automatic processes. We tested this hypothesis by presenting observers with a modified Stroop paradigm in which object or scene words were presented over images of objects or scenes. Terms were either congruent or incongruent with the images. Observers classified the words as being object or scene terms while ignoring images. Classifying a word on an incongruent image came at a cost for both objects and scenes. Furthermore, automatic processing was observed for entry-level scene categories, but not superordinate-level categories, suggesting that not all rapid categorizations are automatic. Taken together, we have demonstrated that entry-level visual categorization is an automatic and obligatory process.

Introduction

Visual categorization is a fundamental cognitive process that allows for efficient action in the world. However, it is not yet known whether categorization proceeds automatically from visual input, or is directed by attentional processes. Automatic processes are those well-learned processes that demand little attention. They can often be computed in parallel, and are obligatory in nature, that is, difficult to ignore, alter, or suppress (Shiffrin & Schneider, 1977). The status of visual categorization remains controversial. Both object (Grill-Spector & Kanwisher, 2005) and scene recognition (Fei-Fei, Iyer, Koch, & Perona, 2007; Greene & Oliva, 2009b; Thorpe, Fize, & Marlot, 1996) occur rapidly and seemingly without effort. The impressive performance of human observers in rapid visual categorization has been taken as evidence for the automaticity of categorization (Grill-Spector & Kanwisher, 2005; Thorpe et al., 1996). However, these results are at odds with theoretical (Fodor, 1983; Pylyshyn, 1999), computational (Riesenhuber & Poggio, 2000) and neurophysiological (Freedman, Riesenhuber, Poggio & Miller, 2001) models that separate categorization from purely visual processes. Furthermore, the attentional requirements for visual categorizations remain controversial (Cohen, Alvarez, & Nakayama, 2011; Evans & Treisman, 2005; Li, VanRullen, Koch, & Perona, 2002). This controversy may be explained in part by the difficulties of separating categorization processes from the processing of visual features that are diagnostic for a class (Delorme, Rousselet, Mace, & Fabre-Thorpe, 2004; Evans & Treisman, 2005; Johnson & Olshaussen, 2003; McCotter, Gosselin, Sowden, & Schyns, 2005). In this work, we examine the extent to which objects and scenes are categorized when the images themselves are task-irrelevant using a Stroop-like paradigm. We will then show that this paradigm can be used to assess the entry-level categories of real-world scenes.

Although a brief glance at a novel visual scene is sufficient to detect a variety of information about the image, it is known that some visual classifications are easier for human observers than others. In particular, classifying an environment as "urban" or "natural" is an easier task for observers than determining that the environment is, say a "beach" or "highway" (Greene & Oliva, 2009b; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007; Loschky & Larson, 2010). However, the reason for this difference is not yet known. Although some have argued that the natural-urban distinction may reflect the true entry-level[1] category of scenes (Loschky & Larson, 2010), urban and natural environments vary in many low-level visual features (Torralba & Oliva, 2003), and early visual processing is sensitive to such differences (Wichmann, Drewes, Rosas, & Gegenfurtner, 2010). We will test the automaticity of both entry-level scene categorization (e.g., "forest" or "street," Tversky &

Citation: Greene, M. R., & Fei-Fei, L. (2014). Visual categorization is automatic and obligatory: Evidence from Stroop-like paradigm. *Journal of Vision*, *14*(1):14, 1–11, http://www.journalofvision.org/content/14/1/14, doi:10.1167/14.1.14.

doi: 10.1167/14.1.14 Received August 2, 2013; published January 16, 2014 ISSN 1534-7362 © 2014 ARVO

Downloaded from jov.arvojournals.org on 08/12/2022

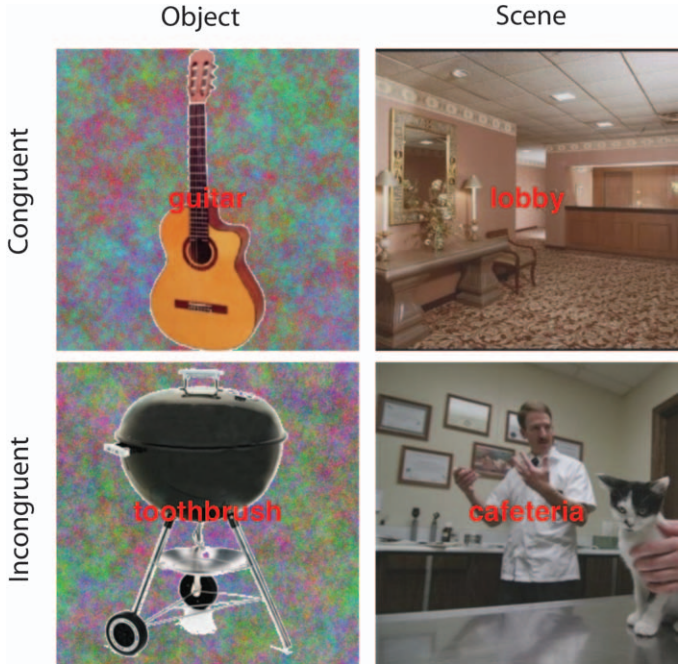

Figure 1. Example stimuli used in Experiment 1. Congruent terms (top) matched the task-irrelevant image while incongruent terms (bottom) came from a different category from the image.

Hemenway, 1983), as well as the superordinate categorization ("natural" and "urban").

The automaticity of a cognitive process can be assessed with a modified Stroop paradigm. Stroop (1935) presented observers with written color names that were either printed with ink matching the color name (e.g., the word "red" written in red ink) or in a different color (the word "red" written in purple ink). Observers were then asked to name the ink color as quickly as possible. Although the meaning of the words was irrelevant to the task, incongruent words resulted in significantly longer reaction times compared to congruent words. Thus, reading (for fluent readers) can be considered an automatic task. Since this influential work, similar logic has been used to study the automaticity of various cognitive processes (see MacLeod, 1991, for a review).

In the domain of object perception, the picture-word interference paradigm (Hentschel, 1973) has been considered a Stroop-like effect. Here, an object name is printed inside a line drawing of an object. As in the original Stroop paradigm, the word can be either congruent or incongruent with the picture. Participants could then be asked either to name the object depicted in the drawing, or to read the word in it. It has been demonstrated that an incongruent word will significantly slow object naming, but that an incongruent object has little effect on reading time (Rosinski, Golinkoff, & Kulish, 1975). However, given the automaticity of reading, we cannot infer that object categorization is not automatic, nor can we make inferences on the nature of scene categorization.

Here, we present a paradigm in which observers view images of objects and scenes with congruent or incongruent nouns superimposed, as in the picture-world interference paradigm. However, we ask participants to semantically categorize words instead of merely reading them. We will show that incongruent images of both scenes and objects interfere with this semantic word categorization task, suggesting that visual categorizations are both automatic and obligatory. Furthermore, we show that this automaticity is limited to entry-level categories for scenes, suggesting that although some superordinate distinctions can be made very easily by human observers, scenes are not automatically processed into "natural" and "urban" environments.

Experiment 1

Methods

Participants

Twelve participants (five female), ages 18–30 with normal or corrected-to-normal vision participated in this experiment. All participants were native English speakers. They provided informed consent and were compensated for their time.

Materials

One hundred unique object categories and 100 unique scene categories were chosen for this experiment. Object categories were taken from (Konkle, Brady, Alvarez, & Oliva, 2010), and scene categories were chosen from the SUN database (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010). As the participants' task was to categorize the category names as being object names or scene names, care was taken to ensure that all category names were unambiguous (for example, "bridge" could describe an object or a location). Four image exemplars were chosen for each of the 200 categories. To equate for stimulus complexity and size, objects were presented on a colored 1/f noise background. The 1/f noise was created in each of the three color channels independently. Category names were written in on top of images in Helvetica font and 40-point size (1.06° of visual angle in height). Words were presented in red as preliminary work showed this to be the most visible for all images. Participants were seated 54 cm from a 21-inch CRT monitor (Sony Trinitron, Tokyo, Japan). Images were presented in 15.8° by 15.8° in size. Matlab with Psychophysics toolbox extensions (Brainard, 1997; Pelli, 1997) was used to present this experiment. Example stimuli are shown in Figure 1.

Downloaded from jov.arvojournals.org on 08/12/2022

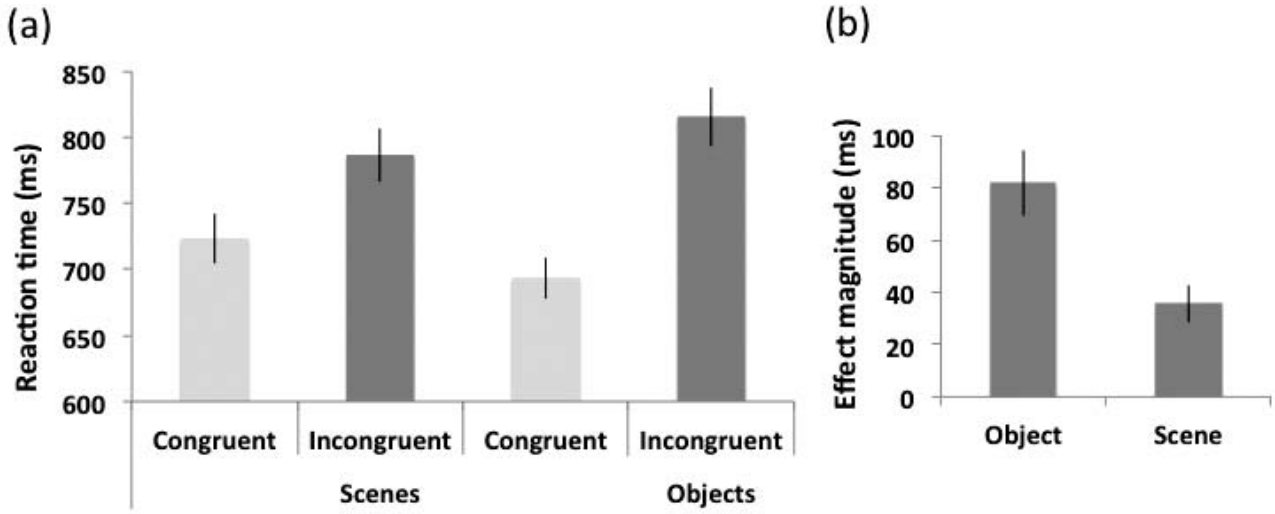

Figure 2. (a) Mean reaction times to categorizing object (right) and scene (left) words on congruent (light gray) and incongruent (dark gray) images. (b) Interference magnitude (congruent RT – incongruent RT) for objects and scenes.

Design

Participants completed 800 experimental trials. Two hundred of these trials (100 for objects and 100 for scenes) were congruent (category name matched the picture, such as the word "gym" over an image of a gym, see Figure 1) and 200 trials were incongruent (e.g., the word "kitchen" over an image of a gym). Incongruent images were created by randomly selecting a different category name for an image. To decouple the correct response from the type of image presented (object or scene), the 400 remaining trials consisted of object category names written on scene images and scene category names presented on objects (400 of each). These trials were excluded from analysis. Images were not repeated.

Procedure

Each participant completed 10 practice trials before completing experimental trials. Images and category names used in the practice trials were not used in the main experiment. Each experimental trial commenced with a fixation point for 300 ms. In each trial, participants would view an image-word pair, and were instructed to classify the word as being an object name or a scene name as quickly and accurately as possible. The image remained on the screen until response. Participants were given performance feedback: Incorrect responses were followed by a 200 ms tone. Reaction times less than 200 ms and greater than 2 s were discarded from analysis (1.4% of data and no more than 5.8% of trials from any one participant).

Results

Participants performed the word classification task with high accuracy (mean: 95% correct, range: 88%–100%). Overall, there was no significant difference in reaction time to classify object words (M = 721 ms) versus scene words (M = 733 ms, $F(1, 11) < 1$). We found a significant main effect of stimulus congruence, $F(1, 11) = 52.4$, $p < 0.001$) with congruent trials performed faster than incongruent trials (708 ms vs. 767 ms). Furthermore, there was a significant interaction between congruence and stimulus type, $F(1, 11) = 16.5$, $p < 0.01$) such that objects had a significantly larger interference effect (defined as incongruent RT − congruent RT) than scenes (81 ms vs. 36 ms, $t(22) = 3.2$, $p < 0.05$), see Figure 2.

Why did we observe more interference for objects than for scenes? One possibility is that category boundaries are fuzzier for scenes than objects (Boutell, Luo, Shen, & Brown, 2004). If some of the images used could belong in more than one scene category, perhaps some of the incongruent trials had some degree of congruence, driving down the amount of semantic interference. For example, both "baggage claim" and "airport terminal" were separate categories in this experiment. However, these categories might be more conceptually related than say, "baggage claim" and "bedroom." Thus, the word "airport terminal" on an image of a baggage claim might be treated as congruent by observers. To test this hypothesis, we examined the semantic similarity between the word-scene pairs for all incongruent scene trials used in the experiment using latent semantic analysis (LSA, Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990). LSA uses a large

Downloaded from jov.arvojournals.org on 08/12/2022

text corpus to aggregate contexts in which words are used. The more similar the contexts two words share, the more similar the words are to one another. Here, word similarities can range from 1 (perfectly interchangeable) to −1 (completely unrelated). If the fuzzy nature of scene categories makes highly conceptually related scene pairs, then we would expect higher similarity scores for scenes compared to objects. Indeed, this was what was found: Incongruent object pairs had a mean LSA similarity of 0.10, while incongruent scene pairs were 0.18, $t(1836) = 13.7$, $p < 0.001$. Does the higher conceptual similarity between scene categories drive the smaller interference effect? To test, we omitted any incongruent scene trials that were out of the 95% confidence interval for the LSA similarity range for objects. The resulting interference effect was 42 ms, well within the original 95% confidence interval for the scene interference effect (20–51 ms). Thus, although semantic similarity between incongruent words and images may affect the interference magnitude, the larger semantic similarity for incongruent scene pairs does not explain the smaller interference effect for scenes.

Perhaps the scene categories were on average more obscure than the object categories. Although it has been estimated that there are up to 30,000 basic-level categories of objects (Biederman, 1987), the most comprehensive scene database (SUN, Xiao et al., 2010) contains only 899 categories. Thus, the 100 scene categories used in this experiment may represent a larger proportion of all types of environments in the world than do the 100 object categories. If observers were not familiar with some environments or scene category labels, perhaps these would not be automatically categorized, therefore driving down the magnitude of the interference effect. We tested both the frequency of the written category labels as well as the relative prevalence of the scene images themselves. Term frequency was assessed using Google's ngrams (Michel et al., 2011), a word-frequency analysis of 5 million books. For all object and scene category labels, term frequency was defined as the peak frequency of that term between 1990 and 2008. The date range was chosen to approximate the date when our participants, on average, began reading. Unlike LSA, this technique does not disambiguate between senses of a word, so it can overestimate the frequency of some terms. For example, "fan" can be a device for creating a breeze (the sense used in this experiment), or a person with a strong interest in something. As both senses make up the frequency, this method is biased towards overestimating word frequency of some terms. Overall, word frequencies did not significantly differ between objects and scenes ($M = 0.0009$ vs. $M = 0.0005$, two-sample t test: $t(188) = 1.6$, $p = 0.11$). Furthermore, word frequency had low correlation with interference magnitude for both objects ($r = -0.12$) and scenes ($r = 0.04$). Thus, word frequency does not seem to be responsible for the smaller interference effect for scenes.

Image frequency was assessed using the number of images returned using a Google Image search. Thus, insofar as the frequency of images on the internet can be a proxy for the frequency of the images that we observe,[2] examining the number of images in each category can provide information about the frequency of scene environments. However, we found that queries for object and scene categories yielded a similar number of terms ($M = 803$ million for objects vs $M = 993$ for scenes, two-sample t test: $t(197) < 1$), suggesting that both object and scene categories have similar frequencies on the internet. Furthermore, there was little correlation between object or scene frequency in Google and the magnitude of the interference effect ($r = -0.05$ and $r = -0.06$ respectively). Therefore, the difference in magnitude of the interference effect for objects and scenes cannot be explained by the relative frequencies of categories in the world.

Discussion

The results of this experiment demonstrate that human observers automatically categorize meaningful visual stimuli, even when doing so is task irrelevant and harmful to performance. Thus, visual categorization seems to be an automatic and obligatory process. Unexpectedly, we found that this effect was more pronounced for objects than for scenes. Although increased semantic similarity between incongruent words and pictures reduced the interference effect, it does not completely explain the reduced magnitude of the scene interference effect. Furthermore, the effect could not be explained by differences in familiarity with the object and scene categories.

These results immediately beg the question of why the scene interference effect is so much smaller than the object interference effect. A remaining possibility is that the larger size of the scene images distributed the meaning of these images over a larger space, making them easier to ignore. Although scene images show a strong bias towards presenting diagnostic information in the center (Tatler, 2007), the possibility remains that the spatially distributed nature of scenes reduced the magnitude of the interference effect. This hypothesis is tested in Experiment 2.

Experiment 2

To test whether the spatial extent of the stimuli was driving the interference effect, we ran a new experi-

Downloaded from jov.arvojournals.org on 08/12/2022

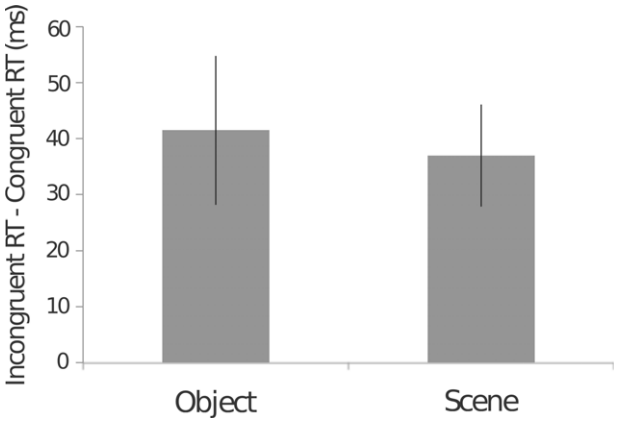

Figure 3. Interference effect (incongruent RT – congruent RT) for object and scene words in Experiment 2.

ment, identical to Experiment 1, except that scene size was reduced to the median size for objects.

Methods

Participants

Ten participants (eight female), age 18–30 with normal or corrected-to-normal vision participated in this experiment. They provided informed consent and were compensated for their time. All participants were native English speakers, and none had participated in Experiment 1.

Materials

The materials for Experiment 2 were identical to those of Experiment 1 with the exception that scene images were reduced to 62% of their original size, equating to the median size for objects (9.5° of visual angle). The overall stimulus size remained the same as the scenes, like the objects, were presented on top of colored 1/f noise.

Design and procedure

Experiment 2 used the same design and procedure as Experiment 1. As in Experiment 1, reaction times less than 200 ms or greater than 2 s were discarded from further analysis (4% of trials overall, no more than 15% from any participant).

Results and discussion

Replicating the results of Experiment 1, task performance was high (mean: 95% correct, range, 91%–99%). Again, there was no significant difference in reaction time to categorize object words versus scene words ($M = 771$ ms and $M = 799$ ms, respectively, $F(1, 9) = 3.08$, $p = 0.11$). Furthermore, we found a significant main effect of word-picture congruence, with words categorized significantly faster on a congruent background than an incongruent background ($M = 773$ and $M = 809$, $F(1, 9) = 28.3$, $p < 0.001$). There was no significant interaction between stimulus (object versus scene words) and stimulus congruency, $F(1, 9) < 1$), demonstrating that the magnitude of the interference effect was similar for objects ($M = 41$ ms) and scenes ($M = 37$ ms), see Figure 3.

In addition to providing a replication of the original result, Experiment 2 demonstrates that when object and scene images are small and centralized over the text being categorized, that the degree of interference created by an incongruent picture is similar for objects and scenes. This implies that the smaller interference observed for scenes in Experiment 1 can be explained by the larger spatial extent of the scene stimuli. As the meaning of the scenes was distributed over a larger spatial area, it is possible that observers could more effectively filter out scene meaning, resulting in less interference. As spatial attention was focused on the word, some of this attention was placed on the scene immediately beneath the word. When the size of the scenes was reduced, more attention was focused on more of the scene, resulting in greater interference. Although photographs tend to be composed in a way that places the most informative regions in the center (Tatler, 2007), scene meaning is largely global and often cannot be solely predicted from a local region (Greene & Oliva, 2009a). Furthermore, scene meaning is largely invariant to changes in size and human observers can accurately recognize scenes as small as 32 × 32 pixels (Torralba, 2009).

So far, we have demonstrated that both scenes and objects are automatically categorized by human observers at the basic-level (Rosch, Mervis, Gray, Johnson, Boyes-Braem, 1976; Tversky & Hemenway, 1983). Although objects have a clear hierarchical category structure, less is known about the category structure of scenes. Human observers can categorize a scene as "natural" or "urban" quicker and easier than they can categorize it at a putative entry-level (such as "kitchen" or "beach"). Does this mean that the urban/natural distinction reflects the true entry-level category for scenes? On the other hand, the ease of distinguishing between urban and natural scenes could reflect the low-level statistical differences between these environments. In Experiment 3, we test this directly by examining interference patterns for both putative entry-level categories ("street" and "forest") as well as the superordinate natural/urban distinction.

Downloaded from jov.arvojournals.org on 08/12/2022

Experiment 3

The image in the top left corner of Figure 4a can be validly categorized as a "natural environment," a "forest," or a "deciduous forest in autumn" with increasing levels of specificity. As with objects, human observers tend to prefer naming scenes at the middle level of specificity (i.e., "forest") as this level maximizes within category similarity, and between-category distinctiveness (Rosch et al., 1976; Tversky & Hemenway, 1983).

However, not all of the basic-level advantages observed for objects are observed for scenes. For example, although human observers are faster to classify an object at the basic-level (e.g., "chair") than at the superordinate level (e.g., "furniture"), the opposite appears to be true for scenes: Participants are actually faster to categorize a scene as an "urban" or "natural" environment than as, for example, a "highway" or "beach" (Fei-Fei et al., 2007; Greene & Oliva, 2009b; Joubert et al., 2007; Loschky & Larson, 2010).

These differences could arise due to a number of factors: Some have argued that easier tasks may be processed earlier by the visual system (Crouzet, Joubert, Thorpe, & Fabre-Thorpe, 2012), while others (Loschky & Larson, 2010) have argued that easier classifications reflect entry-level category status, while still others have suggested that low-level image differences could be driving the effect (Wichmann et al., 2010). These possibilities are difficult to disentangle because categorization performance depends both on the information requirements of the task as well as they availability of visual features that are relevant for the task (Schyns, 1998), and we have limited knowledge of both. However, our modified Stroop paradigm allows us to directly test the hypothesis that the natural/urban distinction could reflect an entry-level category for scenes, as interference reflects automatic processing.

Methods

Participants

Twenty native English-speaking participants (14 female, ages 18–30) with normal or corrected-to-normal vision participated in this experiment. They provided informed consent and were compensated for their time. None of these participants took part in Experiments 1 or 2.

Materials

The images consisted of 200 images: 100 images of forests and 100 images of streets, all taken from the SUN database (Xiao et al., 2010). The words consisted of 10 adjectives and 10 nouns, matched for word length and lexical frequency (from Project Guttenberg). Of these, "natural" and "urban" were the experimentally critical adjectives while "forest" and "street" were the critical nouns.

Design and procedure

As Experiment 3 consisted of scenes only, the semantic categorization task was changed to categorizing words as adjectives or nouns. Participants completed 200 experimental trials: 50 trials with adjectives on top of forest images, 50 trials with nouns on forests, 50 with adjectives on street scenes, and 50 with nouns on street scenes. Each noun or adjective was viewed 10 times throughout the experiment. As only four words were experimentally critical, 40 of the 200 trials were experimental (see Figure 4a for examples of each condition), while the remainder served as a baseline for assessing whether any congruence effect was due to interference in incongruent conditions or facilitation of congruent conditions. Images were not repeated.

The experiment commenced with 10 practice trials. In each experimental trial, participants viewed a word-image pair, and then indicated as rapidly and accurately as possible whether the word was an adjective or a noun. As before, they were told to ignore the picture and focus only on the words. Stimuli remained on the screen until response, and performance feedback (a 200-ms tone following incorrect responses) was given. Reaction times less than 200 ms and greater than 2 s were discarded from analysis (2.1% of data, no more than 4.8% from any given participant). One participant was omitted from analysis for having more than 10% of trials flagged, and having high error rates (~25%).

Results and discussion

Overall, noun/adjective word classification was 90% correct (range: 83%–97%). Although word-image congruence did not have a significant overall effect on reaction time, $F(1, 18) = 1.8$, $p = 0.19$, there was a significant interaction between congruence and categorization level (basic versus superordinate, $F(1, 18) = 4.5$, $p < 0.05$), such that a significant interference effect was observed for scenes at the basic level, 97 ms, $t(18) = 2.4$, $p < 0.05$, but not the superordinate level, $t(18) = 1.1$, $p = 0.28$), see Figure 4b. Additionally, there was a significant main effect of word type (noun vs. adjective), with nouns categorized significantly faster than adjectives, $M = 760$ for nouns vs. $M = 809$ for adjectives, and $F(1, 18) = 4.7$, $p < 0.05$.

Downloaded from jov.arvojournals.org on 08/12/2022

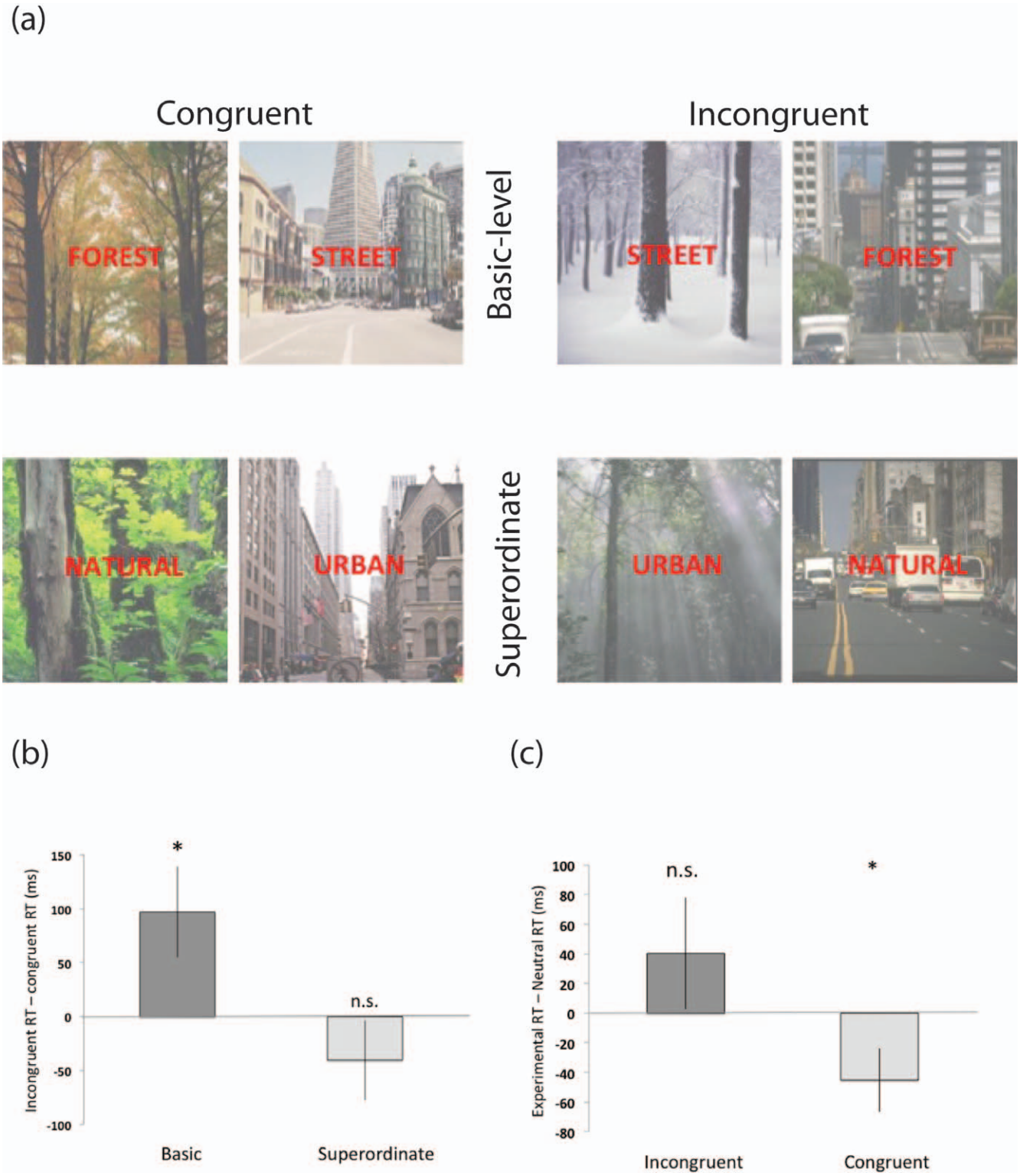

Figure 4. (a) Example stimuli in each of the experimental conditions of Experiment 3. Participants were instructed to categorize words as adjectives or nouns while ignoring the scene. (b) Interference, measured as the difference (in ms) between congruent and incongruent RTs. Basic-level congruence showed significant interference while superordinate-level congruent showed none. (c) Difference between experimental RTs and neutral RTs. Congruent scenes showed significant facilitation of RT.

Downloaded from jov.arvojournals.org on 08/12/2022

The experimental design allowed us to ask whether the Stroop-like effect was due primarily to facilitation of congruent trials, or interference from incongruent trials. For each subject, we subtracted reaction times to each of the experimental conditions at the basic-level (congruent and incongruent) from the reaction times to neutral, non-experimental trials. Longer RTs for incongruent trials compared to neutral can be taken as evidence of interference, while shorter RTs for congruent trials relative to neutral suggest facilitation. We found a significant facilitation effect for congruent trials, $t(18) = 2.1$, $p < 0.05$, see Figure 4c, as well as a trend for interference for incongruent trials, $t(18) = 1.1$, $p = 0.29$, that seems driven by four participants who showed large (> 100 ms) interference effects. Therefore, a congruent scene on top of the word "forest" or "street" will facilitate word categorization.

The results of Experiment 3 demonstrate that the automaticity of visual scene categorization is limited to basic-level categorizations. Although human observers can very rapidly classify a scene as a natural or urban environment (Greene & Oliva, 2009b; Joubert et al., 2007; Loschky & Larson, 2010), it does not appear to be the case that observers automatically classify scenes in this way. Instead, our results demonstrate that upon viewing a novel scene, we automatically generate only the basic-level class label for this scene.

Our experimental design allowed for us to test one proposed explanation for the facility of natural/urban scene classifications: that "natural" and "urban" represent the entry-level categories for scenes (Loschky & Larson, 2010). If this were the case, then we would expect an interference effect for the words "natural" and "urban" on incongruent scenes. We did not observe any interference at the superordinate level, suggesting that the natural/urban distinction is not the entry-level category for scenes. As natural and urban environments differ in a number of low-level visual features (Torralba & Oliva, 2003), it is likely that observers capitalize on these differences in performing rapid categorization.

General discussion

It is well established that we can rapidly recognize visual information (Fei-Fei et al., 2007; Greene & Oliva, 2009b; Grill-Spector & Kanwisher, 2005; Potter, 1976; Thorpe et al., 1996). Although we can extract a wide variety of information in a brief glance, we often approximate understanding in terms of entry-level categorization performance, partially because these category labels are especially informative (Rosch et al., 1976). It has not yet been established whether categorization is a necessary step in visual processing, or whether rapid categorization is made possible through the extraction of one or more diagnostic features of a target category for preferential processing. This work demonstrates that human observers automatically categorize object and scene stimuli, even when doing so is irrelevant and harmful to performance at the task at hand. Thus, object and scene categorization may be viewed as automatic and obligatory processes.

Categorization is the act of abstracting over certain stimulus features to create equivalence classes of objects so that we may act similarly to similar objects. Although some categories exist to facilitate moment-to-moment goals (Barsalou, 1983), the results of these experiments suggest that entry-level categorization is not limited by the observer's task. On the one hand, the remarkable results of rapid categorization experiments has led some to assume automatic categorization (Grill-Spector & Kanwisher, 2005; Thorpe et al., 1996), while on the other hand, those in the theoretical and modeling communities have posited that categorization occurs in a separate stage, after visual processing (Fodor, 1983; Pylyshyn, 1999; Riesenhuber & Poggio, 2000). The current work fills this gap by providing evidence for visual categorizations being intrinsic to visual processing, rather than a separate, possibly later stage.

Our results further emphasize the distinction between classifications that are easy for observers to do versus those that are automatic. Although a number of studies have found that classifying a scene as a "natural" or "urban" environment is easier than classifying it as, say a "mountain" or "forest" (Fei-Fei et al., 2007; Greene & Oliva, 2009b, Joubert et al, 2007; Loschky & Larson, 2010), the results of Experiment 3 demonstrated that scenes are not automatically classified into natural and urban categories. This interference technique can be used to determine more about the conceptual structure of scene categories. Theoretical insights into scene understanding are currently limited by a lack of knowledge about this structure, as basic scientific inquiry requires a common characterization of what is being studied. Although entry-level categories have been identified for some types of environments (Tversky & Hemenway, 1983), scenes represent continuous spaces and ongoing activities, making their classification challenging.

Although these experiments provide evidence for the requisite nature of object and scene categorization, they do not say whether categorization processes are a part of early detection and recognition mechanisms (Grill-Spector & Kanwisher, 2005), a necessary step in feed-forward visual processing (Serre, Oliva, & Poggio, 2007), or a result of rapid feedback (Kveraga, Ghuman, & Bar, 2007). Grill-Spector and Kanwisher (2005) demonstrated that the time courses of object detection

Downloaded from jov.arvojournals.org on 08/12/2022

and basic-level categorization were similar, and that these processes were correlated on a trial-by-trial basis, suggesting that basic-level categorization and object detection could be part of the same mechanism. However, object detection and categorization can be decoupled when the task is made more difficult, such as when stimuli are inverted or noisy (Mack, Gauthier, Sadr, & Palmeri, 2008), or when the task requires finer-grained category distinctions (Mack & Palmeri, 2010). Another possibility is that rapid categorization is made possible through top-down signals that make category predictions on the basis of early visual input (Kveraga, Ghuman, & Bar, 2007).

Our results are consistent with recent work showing that the category status of an object can alter visual detection ability. Lupyan, Thompson-Schill, and Swingley (2010) had participants perform a same-different stimulus judgment on letter stimuli, and found that when letters were different, but shared the same category (e.g., B and b), that participants were slower to reject them as being different. Thus, category membership can alter very simple task-irrelevant visual discriminations. Similarly, observers presented with a seemingly random stream of pictures can pick out patterns of repeating scene categories (e.g., images of classrooms follow images of fields; Brady & Oliva, 2008), suggesting that scene category information is automatically and implicitly processed from these image streams.

More broadly, these results show that our perceptual system is deeply connected with conceptual stored knowledge. This view is congruent with recent neuroimaging work that has demonstrated that semantic and visual information can co-exist within brain regions. For example, visual scene categories can be decoded from patterns of neural activity in the frontal gyrus (Broadman areas 44–45, Walther, Caddigan, Fei-Fei, & Beck, 2009), and written object names can be predicted from activity in occipitotemporal areas (Kherif, Josse, & Price, 2011). Altogether, these results suggest the possibility that shared visual and semantic representations can serve as top-down signals to facilitate rapid visual recognition.

In summary, we have demonstrated that visual objects and scenes are automatically categorized into entry-level categories, even when doing so is harmful to task performance. We have demonstrated that not all rapid categorizations are automatic, and that scenes seem to be automatically categorized only at the entry-level. These results provide evidence for strong interactions between visual and semantic systems, and show the interconnection between perceptual representations and stored conceptual knowledge.

Keywords: basic-level categorization, Stroop, object recognition, scene recognition

Acknowledgments

Commercial relationships: none.
Corresponding author: Michelle R. Greene.
Email: mrgreene09@gmail.com.
Address: Department of Computer Science, Stanford University, Stanford, CA.

Footnotes

[1]We use the term "entry-level" rather than "basic-level" for scene categories as less is known about their conceptual structure. "Entry-level" makes no claims about the hierarchical level of entry, and allows for shifts due to typicality, experience and other factors.

[2]Americans spend over 7 hours a day, on average, looking at television and computer screens outside of work (American Time-Use Study http://www.bls.gov/tus), suggesting that the distribution of images on the internet does indeed make up a large portion of our daily visual experience.

References

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition, 11*(3), 211–227.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*(2), 115–147.

Boutell, M. R., Luo, J., Shen, X., & Brown, C. M. (2004). Learning multi-label scene classification. *Pattern Recognition, 37*(9), 1757–1771.

Brady, T. F., & Oliva, A. (2008). Statistical learning using real-world scenes extracting categorical regularities without conscious intent. *Psychological Science, 19*(7), 678–685.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*(4), 433–436.

Cohen, M. A., Alvarez, G. A., & Nakayama, K. (2011). Natural-scene perception requires attention. *Psychological Science, 22*(9), 1165–1172.

Crouzet, S. M., Joubert, O. R., Thorpe, S. J., & Fabre-Thorpe, M. (2012). Animal detection precedes access to scene category. *PLoS ONE, 7*(12), e51471.

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science, 41*(6), 391–407.

Downloaded from jov.arvojournals.org on 08/12/2022

Delorme, A., Rousselet, G. A., Mace, M., & Fabre-Thorpe, M. (2004). Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Cognitive Brain Research, 19*(2), 103–113.

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

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://www.journalofvision.org/content/7/1/10, doi:10.1167/7.1.10. [PubMed] [Article]

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science, 291*(5502), 312–316.

Greene, M. R., & Oliva, A. (2009a). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology, 58*(2), 137–176.

Greene, M. R., & Oliva, A. (2009b). The briefest of glances: The time course of natural scene understanding. *Psychological Science, 20,* 464–472.

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.

Hentschel, U. (1973). *Two new interference tests compared to the Stroop Color-Word Test*. Lund, Sweden: Lund University.

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

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.

Kherif, F., Josse, G., & Price, C. J. (2011). Automatic top-down processing explains common left occipito-temporal responses to visual words and objects. *Cerebral Cortex, 21*(1), 103–114.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General, 139*(3), 558–578.

Kveraga, K., Ghuman, A. S., & Bar, M. (2007). Top-down predictions in the cognitive brain. *Brain and Cognition, 65*(2), 145–168, doi:10.1016/j.bandc.2007.06.007.

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, USA, 99*(14), 9596–9601.

Loschky, L. C., & Larson, A. M. (2010). The natural/man-made distinction is made before basic-level distinctions in scene gist processing. *Visual Cognition, 18*(4), 513–536.

Lupyan, G., Thompson-Schill, S. L., & Swingley, D. (2010). Conceptual penetration of visual Processing. *Psychological Science, 21*(5), 682–691.

Mack, M., Gauthier, I., Sadr, J., & Palmeri, T. (2008). Object detection and basic-level categorization: Sometimes you know it is there before you know what it is. *Psychonomic Bulletin & Review, 15*(1), 28–35.

Mack, M. L., & Palmeri, T. J. (2010). Decoupling object detection and categorization. *Journal of Experimental Psychology: Human Perception and Performance, 36*(5), 1067–1079.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*(2), 163–203.

McCotter, M., Gosselin, F., Sowden, P., & Schyns, P. G. (2005). The use of visual information in natural scenes. *Visual Cognition, 12*(6), 938–953.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176–182.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology. Human Learning and Memory, 2*(5), 509–522.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *The Behavioral and Brain Sciences, 22*(3), 341–365; discussion 366–423.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience, 3,* 1199–1204.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382–439.

Rosinski, R. R., Golinkoff, R. M., & Kukish, K. S. (1975). Automatic Semantic Processing in a Pic-

Downloaded from jov.arvojournals.org on 08/12/2022

ture-Word Interference Task. *Child Development, 46*(1), 247–253.

Schyns, P. G. (1998). Diagnostic recognition: Task constraints, object information, and their interactions. *Cognition, 67*(1-2), 147–179.

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

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review, 84*(2), 127–190.

Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643–662.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14):4, 1–17, http://www.journalofvision.org/content/7/14/4, doi:10.1167/7.14.4. [PubMed] [Article]

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

Torralba, A. (2009). How Many Pixels Make an Image? *Visual Neuroscience, 26*(01), 123–131.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network, 14*(3), 391–412.

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

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *The Journal of Neuroscience, 29*(34), 10573–10581.

Wichmann, F. A., Drewes, J., Rosas, P., & Gegenfurtner, K. R. (2010). Animal detection in natural scenes: Critical features revisited. *Journal of Vision, 10*(4):6, 1–27, http://www.journalofvision.org/content/10/4/6, doi:10.1167/10.4.6. [PubMed] [Article]

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. *2010 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3485–3492). Presented at the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, San Francisco, CA, USA.

Downloaded from jov.arvojournals.org on 08/12/2022