Automatic basic-level object and scene categorization

Michelle R. Greene and Li Fei-Fei

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

Categorization is the act of generalizing over multiple instances of an object—abstracting over some object differences while highlighting others. Human observers categorize both objects and scenes with remarkable efficiency (Thorpe, Fize, & Marlot, 1996), so fast that it has been claimed to be an automatic process (Grill-Spector & Kanwisher, 2005). However, the degree of conceptual penetrability of perception is deeply debated (Pylyshyn, 1999). Evidence for automatic categorization has primarily come through examining the time course of detection versus recognition. Whereas some have found that object detection and categorization follow indistinguishable perceptual time courses (Grill-Spector & Kanwisher, 2005), categorization tasks can be made slower without affecting detection ability (Bowers & Jones, 2008; Mack, Gauthier, Sadr, & Palmeri, 2008), suggesting that categorization is not automatic.

A classic method for testing automatic processing is the Stroop task. Originally presented (MacLeod, 1991; Stroop, 1935), a Stroop experiment presents observers with colour nouns printed in either the congruent colour (the word “red” typed in red ink) or an incongruent colour (the word “red” written in blue ink). Observers are asked to read the ink colour of each word. Although the words are irrelevant to the task, incongruent words result in significantly longer reaction times as reading is automatic and obligatory. We tested the automaticity of basic-level scene and object categorization using a Stroop-like paradigm.

Please address all correspondence to Michelle R. Greene, Department of Computer Science, Room 241, Stanford University, 353 Serra Mall, Stanford, CA 94305, USA. E-mail: mrgreene@stanford.edu

http://dx.doi.org/10.1080/13506285.2012.726470
METHOD

We presented 12 naive observers with scene and object category names superimposed on images of scenes and objects. A total of 100 object categories and 100 scene categories were used in this experiment. Object images were shown over coloured 1/f noise to equate the visual complexity and viewing angle of the object and scene images. Observers classified the words as being a type of object or a type of scene, while ignoring the image, and were instructed to respond as quickly and accurately as possible. Images remained on screen until response. Observers completed 200 trials with object images and names, and 200 trials of scene images and names. Half of the trials were congruent (e.g., the word “chair” on a picture of a chair, or the word “kitchen” on the picture of a kitchen, see Figure 1) and half of the trials were incongruent (e.g., the word “lamp” on the picture of a microscope, or the word “office” on a picture of a gym). If basic-level categories are automatically processed, then we predicted that it would take observers longer to categorize a word superimposed over an incongruent image compared to a congruent image.

RESULTS AND DISCUSSION

Participants performed the word classification task with high accuracy (mean: 95% correct). Overall, there was no main effect of stimulus type on reaction time (730 ms for objects, 735 ms for scenes), $F(1, 11) < 1$. We found

![Figure 1](image-url)
a significant main effect of stimulus congruency, $F(1, 11) = 52.4, p < .001$, with congruent trials performed faster than incongruent trials (708 ms vs. 767 ms). Furthermore, there was a significant interaction between congruence and stimulus type, $F(1, 11) = 16.5, p < .01$. Exploring further, objects had a significantly larger Stroop-like effect (defined as incongruent RT – congruent RT) than scenes (81 ms vs. 36 ms), $t(22) = 3.2, p < .05$. Together, these results show that basic-level categorization is an automatic process, but that this effect is more pronounced for objects than scenes.

Why did objects exhibit a larger Stroop-like effect than scenes? One possibility is that scenes, unlike objects, can be members of more than one category. For example, a scene depicting a lake in the mountains can be a full member of both lake and mountain categories. Berlin and Kay (1969) demonstrated higher subject agreement for category names when participants had fewer names to choose between. Does reducing the number of category names increase the Stroop-like effect for scenes? Four additional participants completed a follow-up experiment that was identical to the main experiment, except that the number of categories was reduced from 100 to eight for both objects and scenes. However, the results of this experiment were very similar to the main finding: A 79 ms Stroop-like effect for objects and a 36 ms Stroop-like effect for scenes.

Perhaps there is less automatic processing of scenes at these putative basic-level categories because scenes have a different entry level. To test, we ran a follow-up experiment on scenes alone. Forest and street scenes were shown with one of 16 words superimposed: Eight nouns and eight adjectives. Among the nouns were the terms “forest” and “street” and among the adjectives were the terms “urban” and “natural”. The other terms were matched to these in terms of word length and frequency. Five participants viewed these stimuli, classifying the words as adjectives or nouns. If scenes are automatically classified at the superordinate (natural vs. urban) level, we would expect a larger Stroop-like effect than that found at the basic-level (forest vs. street). As before, we found a small but reliable basic-level Stroop-like effect (28 ms). However, congruence at the superordinate level yielded no significant Stroop-like effect, $t(4) < 1$, suggesting that the categorization of scenes as natural or manmade, although rapid, is not automatic.

A satisfying explanation of the smaller Stroop-like effect for scenes remains elusive. Further work will examine whether a scene may be automatically classed as a specific instance rather than category (e.g., my kitchen, rather than any kitchen). However, the possibility remains open that scenes are rapidly processed for the actions taken within them, rather than for their semantic content per se.
Altogether, the current results demonstrate that visual objects and scenes are automatically categorized at the basic-level, even when doing so hinders task performance. This study might recall the well-known picture–word interference effect (Lupker, 1979), where incongruent words delay the naming of a picture. The current study demonstrates a complementary effect: That the automatic semantic processing of a picture interferes with the semantic classification of words. Although pictures do not interfere with simply reading words (Roninski, Golinkoff, & Kukish, 1975), our results suggest that semantic processing of words can be disrupted by the semantic content of image. These results speak to the integrated nature of semantic and perceptual representations through demonstrating an automatic and obligatory semantic categorization of images.

REFERENCES


