Available online at www.sciencedirect.com
ScienceDirect
Vision Research 47 (2007) 3286–3297
Vision Research
www.elsevier.com/locate/visres

Processing scene context: Fast categorization and object interference

Olivier R. Joubert [a,b], Guillaume A. Rousselet [c], Denis Fize [a,b], Michèle Fabre-Thorpe [a,b,*]

[a] Université, Toulouse 3, CerCo, UPS, France
[b] CNRS, UMR 5549, Faculté de Médecine de Rangueil, 31062 Toulouse cedex 9, France
[c] Centre for Cognitive Neuroimaging (CCNi), Department of Psychology, University of Glasgow, UK

Received 28 November 2006; received in revised form 26 June 2007

Abstract

The extent to which object identification is influenced by the background of the scene is still controversial. On the one hand, the global context of a scene might be considered as an ultimate representation, suggesting that object processing is performed almost systematically before scene context analysis. Alternatively, the gist of a scene could be extracted sufficiently early to be able to influence object categorization. It is thus essential to assess the processing time of scene context. In the present study, we used a go/no-go rapid visual categorization task in which subjects had to respond as fast as possible when they saw a ''man-made environment'', or a ''natural environment'', that was flashed for only 26 ms. ''Man-made'' and ''natural'' scenes were categorized with very high accuracy (both around 96%) and very short reaction times (median RT both around 390 ms). Compared with previous results from our group, these data demonstrate that global context categorization is remarkably fast: (1) it is as fast as object categorization [Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience, 13*(2), 171–180]; (2) it is faster than contextual categorization at more detailed levels such as sea, mountain, indoor or urban contexts [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]. Further analysis showed that the efficiency of contextual categorization was impaired by the presence of a salient object in the scene especially when the object was incongruent with the context. Processing of natural scenes might thus involve in parallel the extraction of the global gist of the scene and the concurrent object processing leading to categorization. These data also suggest early interactions between scene and object representations compatible with contextual influences on object categorization in a parallel network.

Keywords: Natural scenes; Fast categorization; Context categorization; Object–context interactions; Parallel processing; Congruency; Object saliency

1. Introduction

Previous studies from our group have demonstrated the very high accuracy and fast speed of the visual system in categorizing different kinds of objects like animals, humans, means of transport or food items. Images flashed for about 20 ms are typically categorized by human observers with high accuracy (94% correct or more), median reaction times around 400 ms, and shortest response latencies around 250 ms (Delorme, Richard, & Fabre-Thorpe, 2000; Fabre-Thorpe et al., 2001; Fabre-Thorpe, Richard, & Thorpe, 1998; Rousselet, Macé, & Fabre-Thorpe, 2003; Thorpe, Fize, & Marlot, 1996; VanRullen & Thorpe, 2001). These short reaction times provide an upper estimate of processing time, as they include the time necessary not only for image processing, but also decisional and motor mechanisms (Bacon-Macé, Macé, Fabre-Thorpe, & Thorpe, 2005; VanRullen & Thorpe, 2001). Despite this limitation, experiments on object categorization in natural scenes have been instrumental in providing temporal constraints on object processing speed.

But typically, these experiments have ignored the relationship between target objects and other elements in the scene. Indeed, in pictures of natural scenes, objects are never isolated; they are seen on a background, surrounded by other objects and various contextual elements.

* Corresponding author.
E-mail addresses: olivier.joubert@cerco.ups-tlse.fr (O.R. Joubert), michele.fabre-Thorpe@cerco.ups-tlse.fr (M. Fabre-Thorpe).

0042-6989/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2007.09.013

Therefore, it is important to determine to what extent scene context might influence object recognition. Information relative to the context of a scene, like semantic consistency (Biederman, Mezzanotte, & Rabinowitz, 1982; Boyce & Pollatsek, 1992; Ganis & Kutas, 2003; Palmer, 1975) or repeated spatial configuration (Chun, 2000), could interact with object information by either facilitating or impairing object visual search and object processing. Although there is strong evidence that the processing of objects is influenced by contextual information, it is still unclear whether context might facilitates object recognition *per se* or might instead facilitate later stages of processing, for instance a decision making stage (Ganis & Kutas, 2003; Henderson, 1992; Henderson & Hollingworth, 1999; Hollingworth & Henderson, 1998). However, in this debate, we lack information about the speed of processing of contextual information, a crucial element needed to determine how early context information might be able to influence object recognition. Mechanisms by which scenes are recognized are still poorly understood, in part because of their complexity. Scenes not only contain objects, but also several non-movable elements with fixed spatial locations such as floor, walls, ceiling, sky, fields, trees, etc. which contribute to the 'gist' of the scene. Different layouts of such fixed elements might rely on different global image features such as spatial envelope properties (openness, naturalness, expansion, symmetry, Oliva & Torralba, 2001, 2006). The fast extraction of such spatial structure of a scene would allow an estimation of the meaning of the scene. Beside this "scene-centered approach", other theories describe scene recognition as the result of the successful identification of some objects in the scene (Friedman, 1979), or the evaluation of spatial links between objects (De Graef, Christiaens & d'Ydewalle, 1990). According to these hypotheses, objects would be systematically processed before scenes (see also Biederman, 1987; Riesenhuber & Poggio, 2000).

A strong argument against these theories is the demonstration that the gist of a scene can be accessed rapidly and accurately even when an image is displayed too briefly to allow an exhaustive processing of the objects in the scene (Biederman, 1972; Biederman et al., 1982; Oliva & Schyns, 1997, 2000; Potter, 1975; Rousselet et al., 2005). The fast processing of briefly presented natural scenes might be explained by the existence of scene specific features that might be used to categorize a scene independently of the objects it contains. To perform scene categorization tasks, subjects could rely on low-level features such as patches of diagnostic colours (Goffaux, Jacques, Mouraux, Oliva, Schyns, & Rossion, 2005; Oliva & Schyns, 2000; Schyns & Oliva, 1994). Alternatively, the spatial structure of the scene might be sufficient on its own to identify scene contexts (Henderson & Hollingworth, 1999; Oliva & Schyns, 2000; Sanocki & Epstein, 1997). Indeed, scene context can still be extracted from filtered scenes containing only low spatial frequencies at which objects cannot be categorized (Schyns & Oliva 1994). Moreover, modelling work suggests that scene classification could rely on specific visual filters that would capture the 'layout of the scene', (Oliva & Torralba, 2001, 2006; Torralba & Oliva, 2003). Such global image signature could be used to determine the general meaning of the scene, or 'gist'. This framework is consistent with the idea that a high-level categorization process does not necessarily depend on high-level representations if representations of lower levels are sufficient to categorize a stimulus in a given task (Schyns, 1998; Ullman, Vidal-Naquet, & Sali, 2002).

Overall, the literature suggests that fast processing of scene context relies to a large extent on visual information that is independent from that used to perform object categorization. However, whether scenes can be categorized as fast or even faster than objects is still a much debated question. Recently, by using a go/no-go paradigm in a 'gist' categorization task, we showed that subjects could discriminate "sea", "mountain", "indoor" and "street" scenes with a very good accuracy (>90%) and short median reaction times (RT) (400–460 ms) (Rousselet et al., 2005). Although such reaction times are relatively fast, object categorization can be faster, with median RT around 400 ms for animal targets (Delorme et al., 2000; Delorme, Rousselet, Macé, & Fabre-Thorpe, 2004; Fabre-Thorpe et al., 1998; Fize, Fabre-Thorpe, Richard, Doyon, & Thorpe, 2005; Rousselet et al., 2003; Thorpe et al., 1996; VanRullen & Thorpe, 2001). However, the reaction time distributions for scenes and objects categorization showed a considerable overlap, arguing against the idea of a systematic processing speed advantage for objects over scenes and leaving open the possibility of large interactions between the two systems in a parallel network.

In the present study, we used broader categories such as natural contexts and man-made contexts. Human subjects might be faster at categorizing scene context at a more general level than the 4 categories (mountain, sea, indoor, and street) used in our previous experiment, allowing more time for interaction between object and context processing. To test this hypothesis, we used the same fast visual categorization task but subjects were asked to categorize the briefly flashed photographs as either "natural" or "man-made" environments. Indeed, compared to the "sea/mountain/indoor/street" experiment, subjects were faster at completing the task. Moreover, when scenes required long processing times to be categorized, a post-hoc analysis revealed a strong interference due to the presence of salient objects.

2. Methods

2.1. Participants

Twelve volunteers (8 men and 4 women, mean age 31, range 23–39, 3 of them left handed) gave their informed written consent. All of them had normal or corrected to normal vision.

2.2. Stimuli

We used photographs of natural scenes from a large commercial CD-ROM library (Corel Stock Photo Libraries). Images (either horizontal or vertical) were in 24-bits jpeg format (16 millions colours), with a size of 768×512 pixels sustaining approximatively a visual angle of $16° \times 11°$. The 1440 images were selected in order to represent equally two categories,