Does object view influence the scene consistency effect?

Gergo Sastyin · Ryosuke Niimi · Kazuhiko Yokosawa

Published online: 19 December 2014
© The Psychonomic Society, Inc. 2014

Abstract Traditional research on the scene consistency effect only used clearly recognizable object stimuli to show mutually interactive context effects for both the object and background components on scene perception (Davenport & Potter in *Psychological Science*, 15, 559–564, 2004). However, in real environments, objects are viewed from multiple viewpoints, including an accidental, hard-to-recognize one. When the observers named target objects in scenes (Experiments 1a and 1b, object recognition task), we replicated the scene consistency effect (i.e., there was higher accuracy for the objects with consistent backgrounds). However, there was a significant interaction effect between consistency and object viewpoint, which indicated that the scene consistency effect was more important for identifying objects in the accidental view condition than in the canonical view condition. Therefore, the object recognition system may rely more on the scene context when the object is difficult to recognize. In Experiment 2, the observers identified the background (background recognition task) while the scene consistency and object views were manipulated. The results showed that object viewpoint had no effect, while the scene consistency effect was observed. More specifically, the canonical and accidental views both equally provided contextual information for scene perception. These findings suggested that the mechanism for conscious recognition of objects could be dissociated from the mechanism for visual analysis of object images that were part of a scene. The “context” that the object images provided may have been derived from its view-invariant, relatively low-level

visual features (e.g., color), rather than its semantic information.

Keywords Object recognition · Scene perception · Natural image perception · High-level vision

The visual recognition of an object is strongly influenced by the context in which it appears. Biederman (1972) and Palmer (1975) defined context as the environment surrounding an object that includes background information. Together, the object and background form a so-called “scene.” According to studies of context by Biederman (1972), Palmer (1975), and others (Boyce, Pollatsek, & Rayner, 1989; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007), the consistency of a scene, which they defined as the regularity with which an object appears in a particular background, greatly affects object recognition. Figure 1 shows examples of consistent and inconsistent contextual pairings of objects and backgrounds. Palmer (1975) noticed that identification performance improved when object recognition was primed by presenting a consistent (versus an inconsistent) background beforehand. This finding was later validated by studies that used simultaneous presentations of objects with consistent backgrounds (versus inconsistent ones) (Biederman, Mezzanotte, & Rabinowitz, 1982). This phenomenon is known as the “scene consistency effect.”

Recently, Davenport and Potter (2004; see also Davenport, 2007) showed that the context not only provided a scene consistency effect in object recognition, but it also led to a reverse phenomenon where the object provided a facilitating context for the recognition of a consistent background (and other objects). Their conclusion was built upon the use of paradigms involving the usage of colored pictures that were shown to be effective in eliciting scene consistency effects (Oliva & Schyns, 2000). They manipulated the scene

Electronic supplementary material The online version of this article (doi:10.3758/s13414-014-0817-x) contains supplementary material, which is available to authorized users.

G. Sastyin · R. Niimi (✉) · K. Yokosawa
Department of Psychology, Graduate School of Humanities and
Sociology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-0033, Japan
e-mail: niimi@L.u-tokyo.ac.jp

Fig. 1 Example of scene stimulus including all the conditions: scene consistency (consistent/inconsistent) $\times$ object viewpoint (canonical/accidental). In the inconsistent condition, inconsistent background was chosen so that natural/artificial consistency was either maintained (A) or violated (B)

consistency of briefly presented and masked stimuli as a primary independent variable, while the percentage of correct recognition responses served as the main dependent variable. The most important feature of this design however was a task variable; the consistency effect was examined in three different tasks: naming the object in a scene, naming the background in a scene, or naming both object and background in a scene. Since the consistency effect was present in all three tasks, they concluded that the objects and backgrounds were processed simultaneously and that both provided semantic contextual information in an interactive fashion. In other words, the objects and backgrounds functioned as facilitating contexts for each other. In this vein, Joubert et al. (2007) also argued that both background context (i.e., scene gist) and object contexts are processed in parallel and at an equal speed during visual scene perception.

However, it is still not entirely understood how the association between objects and backgrounds actually functions and what information forms the “context.” For instance, it was suggested by Greene and Oliva (2009) that scene categorization (i.e., the background recognition task conducted by Davenport & Potter, 2004), could be described by a computational model giving human-like responses without referring to local cues such as “the location, presence, or identity of particular objects” (Mack & Palmeri, 2010). In addition, Munneke, Brentari, and Peelen (2013) also found that the scene consistency effect could be present regardless of whether spatial attention was directed at the target object. Therefore, this effect was likely using “global scene properties” rather than the local information of the object. These findings indicated that the explicit recognition of every object and high-level object information (e.g., identity, semantic information, viewpoint, etc.) might not be needed for the scene consistency effect to occur.

Another important question we attempted to answer was whether the previous experimental models were capable of being generalized to objects viewed in normal settings. A critical example of such a feature that varies in ordinary settings is the angle from which an object is observed.

Viewpoints of objects in everyday environments are rarely static because observers are not stationary. In addition, even with a stationary observer, if the object is animated or moved by other external forces, it could shift the angular orientation relative to the observer. Examples of different viewpoints of an object are shown in Fig. 1. A special viewpoint, known as the “accidental view,” has been shown to yield relatively poor object recognition performance (Biederman, 1987; Marr, 1982). There are multiple factors that could lead to poor performance: a low familiarity with a particular viewpoint, a foreshortening of the object’s elongation axis that occludes a significant portion of the structural information as well as impairing depth perception, or a foreshortening of the symmetry plane that occludes parts of the object that can no longer be inferred from the available information (Humphrey & Jolicoeur, 1993; Mitsumatsu & Yokosawa, 2002; Newell & Findlay, 1997; Niimi & Yokosawa, 2009). Another common viewpoint category is known as the “canonical view” (Fig. 1), where object properties usually tend to facilitate object recognition (Palmer, Rosch, & Chase, 1981). The effect a viewpoint change can have on object recognition is known as the “viewpoint dependency effect.” Considering the fact that the viewpoint could drastically alter the content of visual information that is available, and therefore influence the outcome of the recognition tasks, it was crucial to examine the scene consistency effect while manipulating the object viewpoint. This sort of investigation would provide a more explicit answer to the general question regarding which components of the scene affect its perception and how this process occurs.

Van Eylen, De Graef, Steyaert, Wagemans, and Noens (2013) showed that it is worth examining the effects of scene consistency and viewpoint separately, because they were able to find significant scene consistency and viewpoint main effects, as well as a significant interaction between these effects for both individuals with autistic spectrum disorder (ASD) and typically developing (TD) individuals. Their experimental paradigm, however, was very different from that of the present study, and hence differed with the paradigm used

by Davenport and Potter (2004). These differences included the usage of black-and-white line drawings, the repetition of the same object-background pair within one session, the peripheral location of the target object, the duration of the extremely short presentation time (10 ms), a smaller sample size (20 in the ASD group and 20 in the TD group), and the task itself, which consisted of counting the non-object entities on each image while the participant's eye movements were being tracked. Their stimuli also were more complex than those used in the current study, because they always included both a consistent and inconsistent object in all their images. Due to these differences in the methodology and purpose of the studies, and more importantly, because they did not include any background recognition tasks, we conducted the present study.

The present study assesses the generality of the scene consistency effect with a special focus on the validity of the high-level object information's (e.g., viewpoint) influence on background recognition. We used the same experimental design as Davenport and Potter (2004), but with a broader range of object viewpoints. Thus, comparing performances between the canonical and accidental viewpoints allowed us to examine the origin of the scene consistency effect in more detail. By using this experimental procedure, we were able to observe the fundamental differences between the levels of information used by the different recognition systems (i.e., the object and background recognition systems).

We hypothesized that the object recognition task would replicate earlier results from both scene consistency and view dependency effect studies. Therefore, we predicted that objects would be more accurately identified when paired with consistent backgrounds compared with inconsistent ones and that objects would be more accurately identified from canonical than from accidental viewpoints. We also expected a significant interaction between these effects because accuracy for object recognition in the accidental view condition should benefit more from scene consistency than the background provides. On the other hand, objects in canonical view should be easy enough to recognize and minimize the need to rely on clues from the backgrounds, whereas accidental view objects should maximize this need (Bar, 2004; Van Eylen et al., 2013). There were two object recognition experiments (Experiments 1a and 1b) in this study with different presentation times, where the second one was a complementary experiment for Experiment 1a.

A background recognition task (Experiment 2) also was conducted. We hypothesized that the consistency effect would occur, because it was observed in previous studies. However, this difference between the consistent and inconsistent conditions may be a simple by-product of scene categorization (Mack & Palmeri, 2010; Munneke et al., 2013) that leads to a nonsignificant viewpoint effect, or a result of combining early object processing and scene statistics (see Bar, 2004). On

the other hand, an alternative hypothesis could argue that the viewpoint effect holds true in this task as well, which would suggest that the background recognition is influenced by the ease of object recognition.

In addition, based on the results of Davenport and Potter (2004), it was hypothesized that the accuracy for object recognition tasks would be significantly higher than for background recognition task. It is still not perfectly clear why this difference exists, but one could postulate that object and background recognitions rely mostly on separate sets of source information in each scene. For example, object recognition may require detailed and local information processing, whereas background recognition may only need rough and holistic information, which leads to differences in both their processing speed and efficiency (see Greene & Oliva, 2009; Landau & Jackendoff, 1993). Therefore, three different experiments were conducted to test some of the aforementioned hypotheses.

General method

Participants

Sixty fluent Japanese speakers (37 males and 23 females, average age = 22.9 years, standard deviation [*SD*] = 2.55) participated in the three experiments and two preliminary experiments (combined). Participants were either students or were in some way affiliated with the university. All reported normal or corrected-to-normal vision.

Factorial design

All the experiments employed the same design: a 2×2 within-participant design with two levels of scene consistency (consistent and inconsistent) crossed with two levels of viewpoint (canonical and accidental). The main dependent variable was the proportion of correct recognition answers.

Materials

The study used 240 colored images as stimuli (224 in the experimental trials and 16 in the practice trials; see Fig. 1 for examples). Sixty of them were used in the consistent/canonical condition, 60 others in the consistent/accidental condition, 60 in the inconsistent/canonical condition, and 60 in the inconsistent/accidental condition. All images were created by placing object images onto backgrounds. The object and background images were obtained from commercially available photo collections and the Internet. The backgrounds often included smaller (non-target) objects, but they were irrelevant to the task. The target objects were placed prominently in the center of the backgrounds.

In total, there were 60 object identities (30 natural and 30 artificial objects), and each was presented with both canonical and accidental views to give a total of 120 object images. Two background images were then matched to each object identity; one of them was a consistent background and the other was an inconsistent background. In total, there were 60 unique background images. They were chosen to serve as the consistent backgrounds of the 60 object identities (e.g., an office background for a chair object). Every background also was used in the inconsistent condition with a different object (e.g., an office background was used for the inconsistent condition of sea lion object). To make the inconsistent stimuli, we made 30 pairs of object identities (e.g., chair and sea lion) and then swapped their backgrounds. The 30 pairs comprised 15 within-category pairs (e.g., natural object and natural object) and 15 cross-category pairs (natural and artificial objects). This counterbalancing was done to ensure that the observed consistency effect (if any) was derived from the object-background consistency and not the consistency of the natural or artificial environments.

In the consistent conditions, the objects in the scenes were carefully positioned based on their size, location, and color, so that their presence in the scene would not look strange (i.e., their size, location, perspective, and hue were always adjusted to the background as well as possible). For example, the size of a picture of a dog was adjusted based on the size of its doghouse, the location of the dog was adjusted so that it would conform to the laws of gravity, and the hue of the sea turtle was adjusted so that it matched the hue of the aquarium. Therefore, in this study, “consistency” does not merely mean semantic associations but also includes the preceding visual factors. The representativeness of images (for example, how typical a particular image of a camel was in the camel category), consistency, and validity of the object viewpoints (whether the canonical and accidental view images actually corresponded to each viewpoint criteria) were evaluated by all the authors before the study began, and only images that had unanimous consensus were used. Both the consistency and object viewpoint were evaluated relatively rather than absolutely so that the consistent scenes were always more consistent than the inconsistent scenes, and the canonical views had more canonical characteristics on average than the accidental views. All the object-background pairs can be viewed in the [Supplementary Material](#) section.

Each participant was presented 60 different stimulus images (15 images per condition) from the whole set of the stimuli (240) to avoid repeated presentations of object and background identities. Namely, we divided the 240 stimuli into 4 subsets of 60 images (4 for the practice trials and 56 for the main trials), and each participant was randomly assigned only one of the subsets. Each subset included 30 consistent and 30 inconsistent stimuli (2 for the practice trials and 28 for experimental trials), 30 canonical and 30 accidental view

condition stimuli (2 for practice trials and 28 for experimental trials), and 30 natural and 30 artificial objects (2 for practice trials and 28 for experimental trials). For the inconsistent scene trials, the number of within-category and cross-category swapped image trials were also equal (15 in total; 1 for practice trials and 14 for the main trials). Each participant only saw one of these subsets, where the images were presented in randomized order. All four subsets were used equally often (as there were 20 participants within each experiment, any one of these subsets was used 5 times in 1 experiment). The practice trials were presented before the experimental trials so the participants could familiarize themselves with the procedure.

The mask images were generated from six scenes (not used as scene stimuli in the experiments) by cutting each into 20×20 pixel rectangles and then rearranging them in an arbitrary fashion. They appeared in random order in each trial.

Apparatus

All the scenes were 800×480 pixel JPEG files (the horizontal-vertical ratio was the same as that used in Davenport & Potter, 2004). They were presented from an observation distance of 80 cm and formed a 32×19.4 cm (i.e., $22^\circ \times 13^\circ$ visual angle) rectangle on the display. We used a 21-inch CRT display (Mitsubishi Electric Corporation, RDF22PII) with $1,024 \times 768$ resolution and a 75 Hz refresh rate. The experimental program was written in Matlab, using the Psychophysics Toolbox. Each scene was shown on the uniformly dark background. In addition, as participants provided handwritten answers during the experiment, the experiment booth was lit by indirect, modest illumination so that they could see what they were writing.

Experiment 1a: object recognition task with long presentation time

In Experiment 1a, it was examined whether the object viewpoint modified the scene consistency effect. Prior to this experiment, a preliminary experiment (Preliminary Experiment 1) that included both the object and background recognition tasks was conducted with four participants. It was designed to examine whether accuracy remained constant at a certain level (i.e., no obvious floor or ceiling effect was observed).

It was hypothesized, based on earlier studies, that both scene consistency and object viewpoint effects would be observed, as well as that the impact of the consistency effect would be greater in the accidental view condition than in the canonical view condition.

Method

Participants

Twenty fluent speakers of Japanese (10 males and 10 females, average age = 22.8 years, SD = 2.61) served as participants. The data of two participants from Preliminary Experiment 1, who did the object recognition task first, were included in this analysis. All participants reported normal or corrected-to-normal vision.

Design and procedure

Experiment 1a consisted of four practice trials and 56 experimental trials. All trials began with an instruction in Japanese to “Please press the space key,” which was presented in white letters in the middle of the black screen. This display changed into a white fixation cross for 300 ms after the participants pressed the space key. This was followed by 200 ms of a blank black screen, and then by the scene stimulus for exactly 93.3 ms. After the scene stimulus disappeared, a 300-ms mask appeared, and then the display returned to the first screen with the same message and the trial number in the top-right corner. At this point, the participants wrote the name of the recognized target object on an answer sheet next to the corresponding trial number. However, if a participant recognized the object but could not remember its exact name, he or she was allowed to briefly describe the object’s features or functions. The participants were instructed to write “I don’t know” if they could not recognize the target object. They were allowed to answer in any Japanese writing system (either kanji or kana characters). All participants were told before the start of the sessions that the target objects would appear near the center of the scenes and that the objects and backgrounds would either match (i.e., consistent condition) or not match (i.e., inconsistent condition). There was a break in the middle of each session (after 28 trials), which allowed the participants to restart the experiment any time they wanted to after taking a rest.

Scoring

The scoring process for identifying correct answers was undertaken by four scorers, one of the authors, and three individuals who did not participate in any experiment reported in this paper and did not know the purpose of this study. The scorers judged whether each answer accurately represented the corresponding target object. During the scoring process, the scorers could look up both the object names chosen in advance by the experimenters (see Table 1 or [Supplementary Material](#) section for the complete list) and the actual pictures used in the experiments. However, any answer that clearly identified the target object (e.g., “a construction vehicle to

Table 1 Example of objects and background pairs used as scene stimulus

Natural	Artificial
Duck–Lake	Bench–Park
Camel–Desert	Satellite–Outer Space
Monkey–Jungle	Cannon–Castle
Lizard–Rock	Mug–Cafe

scoop and dig” for a power shovel) was considered to be correct because this study did not focus on the lexical knowledge of the participants. On the other hand, vague answers (if instead of the correct answer “monkey,” a participant answered “mammal”) were considered to be incorrect. The “I don’t know” answers were treated as incorrect responses. In addition, the following answers were also considered correct: taxonomically different organisms that were hard to distinguish at a glance (e.g., “seal” instead of “sea lion”); and answers that were more detailed than the experimenters’ name labels (e.g., “baboon” instead of “monkey”). Any answers that contained typographical errors or incomplete answers, but were clearly indicative of the correct answer were also considered correct.

These criteria were given to each of the scorers before the scoring process, and then each of them independently completed the scoring procedure. Finally, answers that were considered correct by at least three of the scorers were included in the statistical analyses to calculate the percentage of correct answers. These last two steps were included to enhance the objectivity of the procedure.

Results

A two-factor repeated measures ANOVA (consistency $\times$ viewpoint) on the accuracy data showed that the proportion of correct responses was significantly higher for canonical views ($M = .82$) than for accidental views ($M = .70$) ($F(1,19) = 32.59, p < .001, \eta_p^2 = .632$), and for consistent scenes ($M = .80$) compared with inconsistent scenes ($M = .73$) ($F(1,19) = 11.69, p = .003, \eta_p^2 = .381$). However, there was no significant interaction between the consistency and viewpoint variables ($F(1,19) = 2.75, p = .113, \eta_p^2 = .127$; Fig. 2).

Discussion

These results suggested that there was better object recognition for consistent than for inconsistent scenes. In this respect, the results were in agreement with findings from previous studies. Regardless of the object viewpoint, even with a brief presentation time (93.3 ms), a strong context effect from the background facilitated object perception in the consistent condition. Of particular interest was the finding that the new

Fig. 2 Mean accuracy of Experiment 1a (object recognition task). The error bars show standard error of mean

variable of viewpoint also significantly influenced performance in this object recognition task, with better performance observed with the canonical view than with the accidental view. However, the interaction effect that was expected between these two factors was not observed in this experiment. This outcome suggested that these two factors might operate separately and that there may be no advantage to using background information when observing accidental view objects (as opposed to canonical view objects) in this particular setting.

Experiment 1b: object recognition task with short presentation time

In Experiment 1a, we did not observe the interaction effect of viewpoint and consistency, which was expected to be found if the recognition of the objects with an accidental view (more difficult than canonical view condition) relied more on the background context. Experiment 1b was conducted to examine the influence of the task's difficulty. The exposure duration of the scene stimuli was shortened to make the task more difficult.

In a preliminary experiment (Preliminary Experiment 2; only an object recognition task as opposed to Preliminary Experiment 1) it was found that even at 53.3 ms, which was a much shorter presentation time than in Experiments 1a, it was possible for the participants to adequately recognize objects. Hence the 53.3 ms duration was used in Experiment 1b.

The research hypothesis for the second experiment postulated that if we were able to lower the accuracy due to higher difficulty, an interaction between scene consistency and viewpoint effects might arise so that accidental view objects would receive more benefit from consistent backgrounds than canonical view objects, as suggested by previous studies (see Bar, 2004; Van Eylen et al., 2013).

Method

The stimuli, apparatus, and experimental design (including factors and levels) were all the same as in Experiment 1a.

Participants

Twenty fluent Japanese speakers (13 males and 7 females, average age = 23.7 years, $SD = 2.32$) participated in this experiment, none of which had participated in Experiment 1a. All had normal or corrected-to-normal vision.

Design, procedure, and scoring

The experimental procedure was identical to the one used in Experiment 1a, with one exception: the stimulus presentation time was reduced to 53.3 ms. The scoring procedure was the same as used in Experiment 1a.

Results

As in Experiment 1a, a two-factor repeated measures ANOVA (consistency $\times$ viewpoint) was conducted on the accuracy data. The results showed that accuracy was significantly higher in the canonical view condition ($M = .84$) than in the accidental view condition ($M = .70$) ($F(1,19) = 20.38$, $p < .001$, $\eta_p^2 = .518$). Accuracy also was significantly higher when participants viewed consistent scenes ($M = .79$) as opposed to inconsistent scenes ($M = .74$) ($F(1,19) = 6.27$, $p = .022$, $\eta_p^2 = .248$). The interaction between consistency and viewpoint also was significant ($F(1,19) = 5.18$, $p = .035$, $\eta_p^2 = .214$; Fig. 3), even though it had not been significant in Experiment 1a. We followed up on this interaction by conducting a paired-samples t -test (two-tailed). The consistency variable yielded a significant effect in the accidental view condition ($t(19) = 3.18$, $p = .005$), but not in the canonical view condition ($t(19) = 0.12$, $p = .904$). The viewpoint variable yielded significant effects for both the consistent

Fig. 3 Mean accuracy of Experiment 1b (object recognition task with shorter stimulus duration). The error bars show standard error of mean

condition ($t(19) = 2.44, p = .025$) and the inconsistent condition ($t(19) = 4.49, p < .001$).

The results from Experiments 1a and 1b were compared by conducting a three-factor ANOVA (viewpoint and consistency were within-participant factors, and experiment was a between-participant factor). It revealed that the three-way interaction was not significant ($F(1,38) < 1$). Besides the significant main effects of viewpoint ($F(1,38) = 48.76, p < .001, \eta_p^2 = .562$) and consistency ($F(1,38) = 17.66, p < .001, \eta_p^2 = .317$), there was an overall significant interaction between viewpoint and consistency ($F(1,38) = 7.84, p = .008, \eta_p^2 = .171$). Therefore, we again followed up on the interaction with paired-samples t -tests (two-tailed). The consistency variable had a significant effect on the accuracy in the accidental view condition ($t(39) = 4.89, p < .001$), but not in the canonical view condition ($t(39) = 0.76, p = .452$). The viewpoint variable had a significant effect on accuracy in both the consistent condition ($t(39) = 3.89, p < .001$) and the inconsistent condition ($t(39) = 6.55, p < .001$). Given these results, it was concluded that the viewpoint $\times$ consistency interaction was significant in the object recognition task (Experiments 1a and 1b), though it was not significant in the analysis for Experiment 1a alone.

Discussion

The results from Experiment 1b revealed a significant viewpoint effect as well as an overall scene consistency effect. As expected, there was a significant interaction between consistency and viewpoint. This interaction was driven by the absence of a significant scene consistency effect in the canonical view condition, even though it was clearly present in the accidental view condition. A similar trend was found in Experiment 1a as well, and a significant interaction effect was found when the results from Experiments 1a and 1b were compared. Therefore, we surmised that the object viewpoint did affect the scene consistency effect and that the object recognition system relied more on the background context when the object was in the accidental view condition. These findings partly supported our research hypothesis. It should be noted, however, that there was no statistically significant difference in the overall accuracy in Experiments 1a and 1b. This meant that the shorter stimulus duration in Experiment 1b did not significantly reduce the overall accuracy.

Experiment 2: background recognition task

In Experiment 2, our purpose was to see if the object viewpoint modulates the effects of scene consistency when participants were asked to recognize a scene's background (rather

than a target object). Besides these changes, the procedure, stimuli, and conditions were the same as in Experiment 1a.

The research hypothesis for this experiment was that the scene consistency effect should be observed and that the viewpoint effect should disappear. These patterns were expected because previous studies suggested that high-level object identity information may not be necessary for basic scene categorization (Greene & Oliva, 2009) or in the computation of scene consistency (Munneke et al., 2013). However, because the objects were processed as part of the backgrounds, they would have had an obvious influence on the hue and spatial frequency statistics used for categorization, which meant the scene consistency effect may have arisen independent of which viewpoint was offered. On the other hand, an alternative hypothesis might state that both scene consistency and object viewpoint effects should be present, because even during background recognition tasks the high-level (i.e., semantic, viewpoint etc.) information of objects is processed.

Method

The stimuli, apparatus, and experimental design (including factors and levels) were the same as those used in Experiment 1a.

Participants

Twenty fluent Japanese speakers served as participants in this experiment (14 males and 6 females, average age = 22.3 years, $SD = 2.43$); none of which had participated in any of the previous experiments. However, the data of two participants from Preliminary Experiment 1, who did the background recognition task first, were included. All participants reported normal or corrected-to-normal vision.

Design, procedure, and scoring

The design and procedure were the same as those used in Experiment 1a except that the participants had to report perceived backgrounds (e.g., office, beach) instead of objects. For the scoring criterion, answers were also considered correct if they accounted for a relatively large and relevant part of a whole scene (e.g., "trees" for the "park" background, but not "sky" for the same background).

Results

A two-factor repeated measures ANOVA of the accuracy data revealed significantly higher accuracy for the consistent condition ($M = .64$) compared with the inconsistent condition ($M = .50$) ($F(1,19) = 24.86, p < .001, \eta_p^2 = .567$). However,

neither the viewpoint variable ($F(1,19) < 1$) nor its interaction with consistency were significant ($F(1,19) < 1$; Fig. 4).

Discussion

The scene consistency effect in the background recognition task was replicated. These data supported the hypothesis of Davenport and Potter (2004), which predicted that backgrounds would exert contextual effects on object recognition and vice versa. Importantly, we found that viewpoint showed no effect on the results. The objects in both the accidental and canonical views equally provided the context, though the former was more difficult to recognize in the object recognition task. This might have occurred because of the inclusion of objects as part of the backgrounds, which means the apparent scene consistency effect could have been a collateral effect caused by the global scene statistics. These findings suggested that the cognitive functions used in the explicit recognition of the object and those used in background recognition tasks must process their target information extremely differently to fulfill their roles. Another interpretation could be that early object processing that excludes high-level information, such as viewpoint, and scene statistics were combined to deal with this task.

Additional analyses

Natural/artificial consistency

We conducted a statistical analysis on the results of inconsistent condition trials by comparing the trials in which natural/artificial consistency was maintained with trials in which the consistency was violated (Fig. 1). The results showed that natural/artificial inconsistent trials had significantly better accuracy than natural/artificial consistent trials in Experiment 1a (consistent mean .64 vs. inconsistent mean .81) ($F(1,19) =$

43.23, $p < .001$, $\eta_p^2 = .70$) and Experiment 1b (consistent mean .69 vs. inconsistent mean .80) ($F(1,19) = 13.55$, $p = .002$, $\eta_p^2 = .42$). On the other hand, natural/artificial consistent trials had significantly better accuracy than natural/artificial inconsistent trials in Experiment 2 (consistent mean .53 versus inconsistent mean .44) ($F(1,19) = 4.90$, $p = .039$, $\eta_p^2 = .21$).

Additionally, we tested these findings for possible viewpoint influence. Viewpoint effect was significant in Experiment 1a ($F(1,19) = 24.74$, $p < .001$, $\eta_p^2 = .57$) and Experiment 1b ($F(1,19) = 20.13$, $p < .001$, $\eta_p^2 = .51$), but not significant in Experiment 2 ($F(1,19) < 1$).

These results can be explained by the same logic we have been using across this paper: in Experiments 1a and 1b (object recognition task) natural/artificial inconsistent trials were more accurate due to the saliency of visual as well as semantic contrast between the object and its background compared with natural/artificial consistent trials, thus allowing a more accurate segmentation and recognition/naming of the target object. As for Experiment 2 (background recognition task), consistency within the natural/artificial aspect presumably helped the integration of object with the background hence reaching a higher performance compared with natural/artificial inconsistent trials. The viewpoint effect, which was present in this analysis in Experiments 1a and 1b, was absent in Experiment 2, for exactly the same reasons this effect appeared or disappeared when all conditions were taken into account (see individual Discussion sections for each experiment above and General discussion below).

Inter-rater reliability

An inter-rater reliability score (e.g., 75% if 3 of the 4 scorers made same judgment) also was computed for all the experiments, yielding the following results: the mean reliability was 97% for Experiment 1a, 95% for Experiment 1b, and 85% for Experiment 2. Because these values were well above 75% (i.e., the minimum agreement a trial needs to be considered correct), we concluded that the consistency of decisions made by the scorers was high enough.

Additionally, all reliability scores were significantly negatively correlated with (Experiment 1a: $r(54) = -.48$, $p < .001$; Experiment 1b: $r(54) = -.49$, $p < .001$; Experiment 2: $r(54) = -.45$, $p < .001$) and predicted by the variety in labels (see below) used by the participants as well as explained its significant proportion of variance (Experiment 1a: (unstandardized) $b = 1.01$, $t(54) = 83.23$, $p < .001$; $R^2 = .23$, $F(54) = 16.52$, $p < .001$; Experiment 1b: (unstandardized) $b = 1.01$, $t(54) = 75.60$, $p < .001$; $R^2 = .24$, $F(54) = 17.35$, $p < .001$; Experiment 2: (unstandardized) $b = 0.95$, $t(54) = 32.28$, $p < .001$; $R^2 = .20$, $F(54) = 13.81$, $p < .001$). Thus, we can assume that the lower reliability score in Experiment 2 was

Fig. 4 Mean accuracy of Experiment 2 (background recognition task). The error bars show standard error of mean

mainly the result of high variability in descriptive labels used by the participants.

Label variety

We counted the number of different labels or descriptions used by the participants to name each object and background, and we found that backgrounds had significantly more variety (507 in total) compared to objects in both Experiments 1a and 1b (229 and 323 respectively). These values were not only significantly negatively correlated with scorer reliability, but also with accuracy data for objects and backgrounds (Experiment 1a: $r(54) = -.66$, $p < .001$; Experiment 1b: $r(54) = -.76$, $p < .001$; Experiment 2: $r(54) = -.72$, $p < .001$).

Effect of task difficulty in background recognition

The overall accuracy of background recognition (Experiment 2) was lower than object recognition (Experiments 1a and 1b). One might assume that the absence of viewpoint effect in Experiment 2 (but present in Experiment 1) would be attributable to the higher task difficulty. To address the issue, we examined the effect of task difficulty in Experiment 2.

We sorted the background stimuli into two categories, those that yielded a better performance and those that yielded a worse performance. The overall accuracy for the better performing half was comparable with that of the object recognition task (Experiments 1a and 1b); (consistent/canonical: 83%, consistent/accidental: 80%, inconsistent/canonical: 70%, inconsistent/accidental: 73%). We tested these results with a two-factor repeated measures ANOVA (consistency $\times$ viewpoint) and we once again found a significant consistency effect where consistent scenes ($M = .81$) had more accuracy than inconsistent scenes ($M = .72$) ($F(1,19) = 7.89$, $p = .011$, $\eta_p^2 = .29$), along with a nonsignificant viewpoint effect ($F(1, 19) < 1$) as well as a nonsignificant interaction between these two effects ($F(1,19) < 1$). This pattern of result was identical to that found in the analysis on the whole data of Experiment 2. We further confirmed that the pattern was replicated for the worse-performing half.

In addition, a three-factor ANOVA (the better performing half versus whole original dataset $\times$ consistency $\times$ viewpoint) showed a significant overall difference between the two sets of data (as expected, because we only used the better performing half of the original data here), but there was no interaction, which means that the original pattern did not change significantly. Therefore, we can assume that the conclusions of Experiment 2 were independent of overall task difficulty.

However, these ad-hoc analyses are not completely conclusive. Future examination is warranted to identify the reason

why background naming is often more difficult than object naming.

General discussion

The present study observed a significant scene consistency effect (Davenport & Potter, 2004; Davenport, 2007), both in the object recognition and background recognition tasks. We also found a viewpoint dependency effect in the object recognition task in both Experiments 1a and 1b, which revealed that when a target object is oriented in a canonical view, it is more accurately recognized than when presented in an accidental view. The critical finding was significant interaction between the scene consistency and viewpoint variables. These two factors influenced object recognition accuracy in a dependent fashion. Namely, they supported the research hypothesis that the recognition of canonical view objects relies less on the scene context, which is likely to be derived from low-level visual information (e.g., color statistics). Van Eylen et al. (2013) also found similar results for their “first fixation duration” and “time to first fixation” (i.e., time elapsed between the beginning of the trial and the first fixation on the target object) variables to the extent that there was a significantly larger context effect (known as the scene consistency effect in the current study) in the accidental view condition compared to the canonical view condition. Moreover, for their “first fixation duration” variable, the context effect completely disappeared in the canonical view condition. Munneke et al. (2013) observed scene consistency effect regardless of whether attention was directed at the target object or not, which they attributed to a possibility that computation of scene consistency might mainly use global scene statistics. Therefore, the current lack of the scene consistency effect in the canonical view condition also implied the inactive or at least a less active status of the background recognition system.

In Experiment 2, we conducted a background naming task similar to the one used by Davenport and Potter (2004). The results from the present study showed a significantly higher accuracy for the object-background consistent scenes than for the inconsistent scenes. The results also indicated that the object view had no influence on the effect (independent of overall accuracy, as pointed out in the [Additional analyses](#) section), which suggested that certain information regarding the objects, such as their orientation, may not have been processed. Nevertheless, the presence of the consistent objects enhanced background recognition. In other words, the contexts provided by the objects were derived from view-invariant visual features of the object images. As one possible interpretation, the significant scene consistency effect might have been a by-product of pure scene statistics, such as the consistency in hues or luminance across the images, as if the

objects were included as parts of the backgrounds without any segmentation (Mack & Palmeri, 2010; Munneke et al., 2013). The scene statistics could be considered distinct, or at least at an earlier computational stage compared to the scene semantics. This was speculated by Munneke et al. (2013), where their results showed a possibility that scene consistency (viz. scene semantics) could be predicted purely from general scene statistics. In fact, Oliva and Torralba (2006) have already postulated that the “semantic category of most real-world scenes can be inferred from their spatial layout.” As another interpretation, it is possible that there also was an early, rough processing of the object identities, which could have had significant influence on the consistency effect but was not sufficient to obtain their viewpoints nor their exact identities (knowing the identities of the objects would have been sufficient to diminish the viewpoint dependency effect, but this was not available in the inconsistent scene condition) (see Bar, 2004).

We may explain these results by defining two independent putative systems: the object recognition system, which extracts rich information about an attended object, and the background or scene analysis system, which skims through the entire environment to create a coarse visual description based on global and relatively low-level visual information (e.g., low spatial frequency components, color statistics). The object recognition task requires retrieving detailed object information, including viewpoints for conscious object recognition, and coarse object and background information for calculating scene consistency. In other words, the visual system synthesizes both detailed object information and low-level background information. On the other hand, for the background recognition task, the visual system does not need detailed object information to perform conscious background recognition (Mack & Palmeri, 2010). It does, however, require relying on low-level information of both the background and objects. If the visual system needs to process high-level information from any part of a background, the information has to be separated from the whole scene (e.g., attentional selection). The object-like representation (figure as opposed to ground) becomes available, then the visual system switches to the object recognition system.

These results do not contradict the mutually interactive model of Davenport and Potter (2004). However, they suggest a nonnegligible asymmetry in the level of information processed by the two recognition subsystems; the object recognition system, which is responsible for high-level image processing, and the background recognition system, which is responsible for dealing with low-level image processing. This distinction between the two subsystems is similar to the richness of “language of objects and places” concept proposed by Landau and Jackendoff (1993). They found a significant correlation between the richness of the object descriptions and concreteness of questions that were asked about those

particular objects, as well as between the coarseness of the spatial relation descriptions and the questions asked about them. This is in line with our finding where participants applied a bigger variety of labels to backgrounds compared to objects. The background recognition system may not be sensitive enough to differentiate between even slightly similar places (hence the coarseness of descriptors), whereas the object recognition system may be able to differentiate even between very similar objects (hence the richness of descriptors).

One may assume that the current findings would not be replicated with participants with different cultural backgrounds (the current study was conducted in Japan), because it is well established that there are cultural differences in both scene perception and visual contextual effects (Chua, Boland, & Nisbett, 2005; Ji, Nisbett, & Zhang, 2004; Kitayama, Duffy, Kawamura, & Larsen, 2003; Masuda & Nisbett, 2001). Overall, the scene consistency effect reported in previous studies was replicated, while several new findings, such as the absence of the viewpoint effect in the background recognition task (Experiment 2), and the reduced scene consistency effect for the recognition of objects in canonical views also were observed (Experiments 1a and 1b). For future research, it is recommended that the generality of these findings across cultures should be tested.

References

- Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, 5, 617–629. doi:10.1038/nrn1476
- Biederman, I. (1972). Perceiving real-world scenes. *Science*, 177, 77–80. doi:10.1126/science.177.4043.77
- Biederman, I. (1987). Recognition by components: A theory of human image understanding. *Psychological Review*, 94, 115–147. doi:10.1037/0033-295X.94.2.115
- Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14, 143–177. doi:10.1016/0010-0285(82)90007-x
- Boyce, S. J., Pollatsek, A., & Rayner, K. (1989). Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception and Performance*, 15, 556–566. doi:10.1037/0096-1523.15.3.556
- Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 12629–12633. doi:10.1073/pnas.0506162102
- Davenport, J. L. (2007). Consistency effects between objects in scenes. *Memory & Cognition*, 35, 393–401. doi:10.3758/BF03193280
- Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, 15, 559–564. doi:10.1111/j.09567976.2004.00719.x
- Greene, M., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, 58, 137–176. doi:10.1016/j.cogpsych.2008.06.001

- Humphrey, G. K., & Jolicoeur, P. (1993). An examination of the effects of axis foreshortening, monocular depth cues, and visual field of object identification. *The Quarterly Journal of Experimental Psychology*, 46A, 137–159. doi:[10.1080/14640749308401070](https://doi.org/10.1080/14640749308401070)
- Ji, L., Nisbett, R. E., & Zhang, Z. (2004). Is it culture or is it language? Examination of language effects in cross-cultural research on categorization. *Journal of Personality and Social Psychology*, 87, 57–65.
- Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research*, 47, 3286–3297.
- Kitayama, S., Duffy, S., Kawamura, T., & Larsen, J. T. (2003). Perceiving and object and its context in different cultures: A cultural look at new look. *Psychological Science*, 14, 201–206. doi:[10.1111/1467-9280.02432](https://doi.org/10.1111/1467-9280.02432)
- Landau, B., & Jackendoff, R. (1993). “What” and “where” in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16, 217–265. doi:[10.1017/S0140525X00029733](https://doi.org/10.1017/S0140525X00029733)
- Mack, M. L., & Palmeri, T. J. (2010). Modeling categorization of scenes containing consistent versus inconsistent objects. *Journal of Vision*, 10(3), Article 11. doi:[10.1167/10.3.11](https://doi.org/10.1167/10.3.11)
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W.H. Freeman. doi:[10.1002/neu.480130612](https://doi.org/10.1002/neu.480130612)
- Masuda, T., & Nisbett, R. E. (2001). Attending holistically vs. analytically: Comparing the context sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology*, 81, 922–934.
- Mitsumatsu, H., & Yokosawa, K. (2002). How do the internal details of the object contribute to recognition? *Perception*, 31, 1289–1298. doi:[10.1068/p3421](https://doi.org/10.1068/p3421)
- Munneke, J., Brentari, V., & Peelen, M. V. (2013). The influence of scene context on object recognition is independent of attentional focus. *Frontiers in Psychology*, 4, 552. doi:[10.3389/fpsyg.2013.00552](https://doi.org/10.3389/fpsyg.2013.00552)
- Newell, F. N., & Findlay, J. M. (1997). The effect of depth rotation on object identification. *Perception*, 26, 1231–1257. doi:[10.1068/p261231](https://doi.org/10.1068/p261231)
- Niimi, R., & Yokosawa, K. (2009). Viewpoint dependence in the recognition of non elongated familiar objects: Testing the effects of symmetry, front-back axis, and familiarity. *Perception*, 38, 533–551. doi:[10.1068/p6161](https://doi.org/10.1068/p6161)
- Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41, 176–210. doi:[10.1006/cogp.1999.0728](https://doi.org/10.1006/cogp.1999.0728)
- Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 155, 23–36. doi:[10.1016/S0079-6123\(06\)55002-2](https://doi.org/10.1016/S0079-6123(06)55002-2)
- Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, 3, 519–526. doi:[10.3758/BF03197524](https://doi.org/10.3758/BF03197524)
- Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and Performance IX* (pp. 135–151). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Van Eylen, L., De Graef, P., Steyaert, J., Wagemans, J., & Noens, I. (2013). Children with autism spectrum disorder spontaneously use scene knowledge to modulate visual object processing. *Research in Autism Spectrum Disorders*, 7(7), 913–922. doi:[10.1016/j.rasd.2013.04.005](https://doi.org/10.1016/j.rasd.2013.04.005)