Two scenes or not two scenes: The effects of stimulus repetition and view-similarity on scene categorization from brief displays

Martin J. Goldzieher 1 · Sally Andrews 1 · Irina M. Harris 1

Published online: 5 August 2016
© Psychonomic Society, Inc. 2016

Abstract Previous research suggests that understanding the gist of a scene relies on global structural cues that enable rapid scene categorization. This study used a repetition blindness (RB) paradigm to interrogate the nature of the scene representations used in such rapid categorization. When stimuli are repeated in a rapid serial visual presentation (RSVP) sequence (~10 items/sec), the second occurrence of the repeated item frequently goes unnoticed, a phenomenon that is attributed to a failure to consolidate two conscious episodes (tokens) for a repeatedly activated type. We tested whether RB occurs for different exemplars of the same scene category, which share conceptual and broad structural properties, as well as for identical and mirror-reflected repetitions of the same scene, which additionally share the same local visual details. Across 2 experiments, identical and mirror-image scenes consistently produced a repetition facilitation, rather than RB. There was no convincing evidence of either RB or repetition facilitation for different members of a scene category. These findings indicate that in the first 100–150 ms of processing scenes are represented in terms of local visual features, rather than more abstract category-general features, and that, unlike other kinds of stimuli (words or objects), scenes are not susceptible to token individuation failure.

Keywords Scene perception · Categorization · Repetition blindness

A visual scene is a very complex stimulus that confronts the viewer with rich detail to process and an array of complex behaviors to choose from. Imagine, for instance, that you were blindfolded, taken to a familiar location, and the blindfold was then removed revealing a view of a city street. You would be able to rapidly comprehend a number of aspects of the scene within the first second. For instance, you would not only be able to identify what kind of environment you are in (i.e., that this is a street) as well as its general structure and spatial layout (e.g., whether you were looking along the street or across it), but also the presence, identities, and the positions of objects (e.g., cars, other pedestrians, buildings) relative to yourself and each other. You may also be able to readily discern something about the activities currently taking place, as well as contextual associations (e.g., this isn’t just any streetscape, but a particular view of George Street in downtown Sydney).

This highlights the different ways in which a “scene” can be represented: as a member of a general category (e.g., a street or a beach), as a specific familiar place (e.g., George Street or Bondi Beach), or as an even more specific view of a place, with landmarks in specific positions that can provide egocentric information about your own placement in this environment. The utility of these different representations depends to some extent on task requirements. Categorizing a scene as a member of a particular category provides information about the types of objects that one might expect to encounter and the kinds of actions that could be performed in that environment. In contrast, recognizing a scene as a familiar place allows you to orient yourself in a greater environment. Encoding the specific viewpoint from which a scene is viewed can also provide important information about how to navigate around the visual environment and avoid potential obstacles. The overarching question asked in this study is which of these kinds of scene representations have primacy in the very early stages of viewing a scene (the first 100–150 ms)?

In a series of pioneering studies, Potter (1975, 1976) demonstrated that people can process quite abstract conceptual
information about a scene remarkably quickly and efficiently. For example, observers could identify a precued target scene (e.g., a picnic) from a rapid serial visual presentation (RSVP) stream of images presented at rates of about 8/s with relatively little difficulty, despite the fact that the target cue did not provide any specific information about what the target looked like. Potter argued that this conceptual knowledge (or gist) can be rapidly activated (within 100–200 ms), although it must subsequently be consolidated into short-term memory in order to be reportable (Potter, 1999). Similarly, Biederman (1972) and Biederman, Mezzanotte, & Rabinowitz (1982) showed that objects can be detected more efficiently when presented within a coherent scene (as opposed to a scrambled scene), leading him to propose that a scene schema can be activated in as little as 150 ms. Many other behavioral and neurophysiological investigations have since confirmed the speed with which scene gist information is extracted (Fei-Fei, Iyer, Koch, & Perona, 2007; Thorpe, Fize, & Marlot, 1996; VanRullen & Thorpe, 2001).

A number of contemporary theories propose that scene gist recognition relies on analysis of global scene structure, where-by coarse global features such as the major contours and textures of the image rapidly activate high-level conceptual knowledge without the need to represent more local features, such as diagnostic objects, until later in processing (e.g., Bar, 2004; Hochstein & Ahissar, 2002; Oliva & Torralba, 2001; Schyns & Oliva, 1994). Reverse hierarchy theory (Ahissar & Hochstein, 2004; Hochstein & Ahissar, 2002) proposes two distinct phases of scene processing: vision at a glance, which involves rapid, feed-forward activation of high-level cortical areas based on coarse information, followed by vision with scrutiny, which relies on feedback connections to low-level cortical areas and functions to fill in the fine details of a scene. Similarly, according to the spatial envelope model (Oliva & Torralba, 2001; Torralba & Oliva, 2003), scenes are initially processed in terms of global structural cues that determine general spatial envelope properties, such as openness, expansion, ruggedness, and navigability that describe a scene’s spatial extent or general shape. Computational modelling indicates that scenes belonging to the same category have highly similar spatial envelopes. This model successfully emulates the performance of human observers, by making similar kinds of errors (Greene & Oliva, 2009a), requiring less time to accurately classify images according to their spatial envelope properties than their basic-level category (Greene & Oliva, 2009b), and showing similar adaptation effects to global scene properties (Greene & Oliva, 2010). Collectively, this evidence implies that high-level conceptual information about scene category can be extracted very rapidly and that this information relies primarily on coarse global features of a scene rather than local details.

A somewhat different picture emerges from research into how scenes are represented in the brain. Neuroimaging studies have identified a number of brain regions that are dedicated to processing visual scenes in preference to other types of stimuli, such as objects or faces (see Epstein, 2014, for a recent review). These include the parahippocampal place area (PPA) in the posterior parahippocampal gyrus (Epstein & Kanwisher, 1998), the retrosplenial complex (RSC) in the medial parietal lobe (Epstein & Higgins, 2007) and an area near the transverse occipital sulcus, sometimes called the occipital place area (OPA; Dilks, Julian, Paunov, & Kanwisher, 2013). The nature of the representations coded in these areas has been investigated primarily using the fMRI adaptation technique. Typically, this method reveals a reduction in the fMRI signal to repeated stimuli that can be used to index the representational similarity of two stimuli. If such a reduction is observed in a particular brain area, it is inferred that it codes a stimulus attribute that is shared by the stimuli, whereas if there is no adaptation one can infer that that particular area treats the repeated stimuli as representationally distinct. Across a number of studies, it was found that activity in the PPA showed adaptation when the repeated scenes were identical, but not when they were different views of the same scene (Epstein, Graham, & Downing, 2003; Epstein, Higgins, & Thompson-Schill, 2005; S. Park & Chun, 2009; S. Park, Chun, & Johnson, 2010). This implies that the representations in this region are highly viewpoint-specific, although the adaptation does become more view-invariant over multiple exposures to different views of the same scene (Epstein et al., 2005).

There is some evidence that the representations in the PPA are tolerant to one type of viewpoint manipulation, namely mirror-image reflection, whereas the scene responsive area in the transverse occipital sulcus treats mirror-images of scenes as representationally distinct (Dilks, Julian, Kubilis, Speleke, & Kanwisher, 2011). The latter area is considered to be an “earlier” station along the visual processing hierarchy than the PPA, which is consistent with the idea that information becomes less view-specific as it moves along the visual processing hierarchy. However, no adaptation effects were found in these areas for different scenes belonging to the same category (Epstein & Morgan, 2012), suggesting that the PPA and related scene-selective regions represent specific places, and perhaps even specific views of these places, rather than broader categories.

Thus, the findings from these adaptation studies, at least, suggest that these scene-processing areas do not play a major role in the rapid extraction of the type of scene gist information implicated by the behavioral results reviewed above. These conflicting findings prompted Epstein (2014) to conclude that there are two stages of scene processing: (1) analysis of visual features unique to the image (i.e., the view-specific features of a scene) followed by (2) placing this view
within a larger framework, both in terms of what is likely to be outside the current view and conceptual scene information. This would imply that abstract conceptual information about scenes is accessed later than more specific perceptual features of a scene.

This study

In this study, we exploited the phenomenon of repetition blindness to probe the nature of the representations involved in scene categorization by measuring the behavioral effects of different types of representational similarity between two scenes.

Repetition blindness (RB) is the failure to report a second instance of a repeated target stimulus in a RSVP stream. RB is not simply due to difficulty in detecting multiple targets at these rapid presentation rates, because when two targets are unrelated they can both be identified quite successfully. The original explanation proposed for this effect was that it reflects the repeated activation of a “type” that cannot be resolved into separate “tokens,” or visual episodes, due to the temporal constraints of the RSVP procedure (Kanwisher, 1987). Thus, RB allows researchers to interrogate the properties of a type, in much the same way that fMRI adaptation probes the nature of the neural representation of the stimuli. The logic is that if two items induce RB, they must both activate a common type representation.

Repetition blindness has been demonstrated for a variety of stimuli, including letters, words, objects, and even novel 2-D and 3-D figures (Arnell & Jolicoeur, 1997; Bavelier & Potter, 1992; Coltheart, Mundy, & Coltheart, 2005; Harris & Dux, 2005a, 2005b; Hayward, Zhou, Man, & Harris, 2010; Kanwisher, 1987; Kanwisher & Potter, 1990; Kanwisher, Yin, & Wojciulik, 1999). It occurs for identical repetitions, but also for repetitions that are visually distinct, such as letters presented in different cases (Corballis & Armstrong, 2007; Kanwisher, 1987), orthographically similar words (Bavelier, Prasada, & Segui, 1994; Kanwisher & Potter, 1990) and objects presented in different orientations (Harris & Dux, 2005a, 2005b; Hayward et al., 2010). Repetition blindness has also been demonstrated for items presented in different formats, such as the word and picture versions of the same concept (Bavelier, 1994), and between pictures of categorically and conceptually related objects (Kanwisher et al., 1999), though not between semantically related words (Kanwisher & Potter, 1990). This pattern of results is consistent with Bavelier’s (1994) proposal that RB can occur between visually dissimilar items that share a representational code—be it visual, phonological, or semantic—as long as this code is activated prior to tokenization.

These examples show that RB has successfully been used as a tool to investigate how both words and objects are represented at early stages of processing. Here we extend this approach to scenes. To our knowledge, there is only one previous study that has investigated RB using scene stimuli. This was an fMRI study conducted by Naughtin, Tamber-Rosenu, and Dux (2013), which looked at the neural correlates of token individuation and used scene stimuli because they are associated with a reliable pattern of brain activation (in the PPA). Naughtin et al. (2013) asked participants to report whether a three-target RSVP sequence contained a repeated scene, no repetition, or was a “catch” trial on which only two scenes were presented. They found that accuracy in detecting the occurrence of a repeated scene was low. Although this suggests that RB can occur for scenes, their experiment was not designed to investigate how scenes are represented, and indeed their task did not require any explicit categorization of the scenes. The experiments reported in the present study therefore represent the first systematic investigation of RB for scene stimuli in the context of a scene categorization task.

In two experiments, we presented observers with short RSVP streams that contained three scenes, with the first and last being the critical items (denoted as C1 and C2), and we varied the level of similarity between them. These items could be (1) completely identical (same scene from the same viewpoint); (2) mirror images of each other (same scene, containing all the same local features, but seen from a different viewpoint); (3) different members of the same basic-level category (e.g., two different beaches); or (4) unrelated scenes, from different basic-level categories (as the control condition). These different conditions tap into different levels of scene representations (specific view vs. a particular place vs. scene category) as well as differing in the amount of shared local features, with identical images sharing both local detail and spatial layout; mirror images sharing the same local detail, but in a different spatial configuration; and category repetitions sharing broad conceptual and spatial envelope features, but differing in terms of local details.

If scenes are initially represented in terms of global structure and/or more abstract conceptual information (gist, for short), then RB should occur between members of the same category which share these features, even when they are otherwise visually dissimilar in terms of their local detail. In contrast, if scenes are represented in terms of specific local details—that is, a representation of a particular place—then RB should be restricted to items that share these details (i.e., identical repetitions and possibly mirror-image repetitions). Conversely, if scenes are represented as very specific views of a place—as might be useful for navigating around a scene—then RB should be restricted only to identical repetitions, but not occur for mirror images. Given the evidence...
reviewed above suggesting the existence of more or less view-specific scene representations at different levels of the processing hierarchy, we also compared different presentation times to determine whether the time available to encode the stimuli interacted with the pattern of RB, which would suggest that different types of scene representations (category vs. place vs. view) are activated as visual information flows through the system.

**Experiment 1**

This experiment manipulated both the level of similarity between the scenes of interest and the presentation rate of the stimuli. Trials were grouped into two blocks. The *category block* comprised equal numbers of trials with identical, same-category, and different-category (nonrepeat) critical items; this block tests whether scene gist common to members of the same category induces RB and allows a direct comparison with the RB induced by identical repetitions. The other *mirror block* contained trials with identical, mirror-image and different-category (nonrepeat) critical items; this block tests how RB is modulated by changing the spatial configuration of otherwise identical local features and allows comparison of this with the RB induced by identical repetitions. We separated trials into these distinct blocks, rather than intermixing all types of trials, to reduce ambiguity about the level of processing required for the nonidentical repetitions. Blocking the trials in this manner also limited the proportion of critical items that shared local details, as occurred for both identical and mirror images, to avoid a bias to use such features as the basis for category judgements. Each participant completed both blocks, with presentation time held constant within a testing session.

Based on the parameters used in previous RB studies employing word and picture stimuli, as well as pilot testing, we used a presentation rate of 117 ms/item as the briefest presentation time, and compared this with three longer durations (129, 141, and 153 ms) to observe how presentation duration modulates repetition effects. Presentation rate was manipulated between subjects: four groups of participants completed the experiment, each at a different presentation duration.

Given previous evidence of early extraction of scene gist, RB was expected to occur for identical scenes at the very least. If scene recognition relies on global image features or a rapid semantic activation, this should result in RB for both mirror images and category repeats as well, because both of these types of repetitions share higher order gist information, and this RB should be evident even with the shortest stimulus durations. Conversely, if scene recognition relies primarily on view-specific representations of place, then we would expect RB to only occur for identical scene images and perhaps show some viewpoint generalization for mirror images with longer exposure durations, as information flows from lower levels of the visual system to areas that can represent scenes in a more view-invariant manner, such as the PPA. In this case, no RB would be expected for different exemplars of the same scene category, as these do not activate a common type representation.

**Method**

**Participants**

Ninety-six undergraduate students at the University of Sydney participated for course credit. They were randomly allocated to four groups, which determined the presentation rate of all images presented in the experiment (117, 129, 141, and 153 ms). Four participants were excluded from the 117-ms group for having zero correct responses in more than two conditions. This left 20 participants in the 117-ms version (13 female, mean age 19 years), and 24 in each of the 129-ms (14 female, mean age 20 years), 141-ms (20 female, mean age 19 years), and 153-ms (20 female, mean age 20 years) versions. The sample size was decided on the basis of previous RB experiments in our lab, which found that 20–24 subjects usually yield robust effects.

**Stimuli**

Stimuli were 256 × 256 pixel grayscale images of scenes drawn from the set developed by Konkle, Brady, Alvarez, and Oliva (2010), obtained from the Oliva Lab website (http://cvcl.mit.edu/MM/sceneCategories.html). This is a large database of 4,672 scene images, from 160 categories. Four scenes were selected from 15 categories to be used in this experiment (60 images total). The categories used were: airport, bathroom, beach, cemetery, classroom, field, forest, garden, mountain, petrol station, port, restaurant, street, supermarket, and waterfall (see Appendix). Mirror-image versions of this stimulus set were created by reflecting each original image around its y-axis. A number of masks were generated from these images by taking small fragments from all of these scenes and using them to compose a mosaic such that the resulting image had no coherent structure or meaning.
Design

The experiment comprised two blocks, category block and mirror block, which followed the same design, differing only in one of three experimental conditions. The order of the blocks was counterbalanced between participants. The three kinds of experimental trials in each block were defined by the relationship between the first (C1) and third (C2) scene in the sequence. Both blocks contained identical-repeat and nonrepeat trials (15 trials each). On identical-repeat trials, C1 and C2 were exactly the same images, whereas on nonrepeat trials, they were scenes from different categories (e.g., beach and bathroom). The third condition differed between the blocks. In the category block, these were category-repeat trials, where C1 and C2 were different exemplars of the same category. In the mirror block, these were mirror-repeat trials, on which C2 was the mirror image of C1. In all experimental conditions the intervening item in the sequence was always a scene from a different category (e.g., beach and bathroom). The third condition differed between the blocks. In the category block, these were category-repeat trials, where C1 and C2 were different exemplars of the same category. In the mirror block, these were mirror-repeat trials, on which C2 was the mirror image of C1. In all experimental conditions the intervening item in the sequence was always a scene from a different category to both C1 and C2. In addition to these 45 experimental trials, each block contained six “catch” trials, on which either C1 or C2 was replaced with a mask. See Fig. 1 for an illustration of the experimental conditions in both Experiments 1 and 2. Data from the two blocks were analyzed separately to allow direct comparison of the nonidentical-repeat condition against the identical-repeat and nonrepeat conditions.

Procedure

The experiment was run using Presentation software (Neurobehavioral Systems, www.neurobs.com) on a Dell T3500 computer. An 85 Hz Dell Trinitron Multiscan G420 monitor (1,280 × 1,024 resolution) was used for testing. Viewing distance was not controlled, but was approximately 50 cm. At this distance the scenes subtended approximately 7° of visual angle. Participants were first familiarized with the different images and categories that were to be used in the experiment. They viewed all images on the computer screen, in the position they would appear in the experiment. Scenes were displayed one at a time with the category label displayed below. Participants could view each image for as long as they liked, pressing the space bar to move to the next one. Both before and during this familiarization period, they had access to a printout of the same images. All four images from a category were printed on a single page, with the category label at the bottom of the page.

Participants then completed 15 practice trials comprising examples of all the different conditions, for which feedback was given to illustrate the experimental instructions before moving on to the main experiment. They initiated each trial by pressing the space bar. A fixation cross was displayed for 1 s, followed by a rapid sequence of nine images. The initial and final three images were masks, with two or three scene images displayed in between. The items in the RSVP sequence were presented at one of four durations (117, 129, 141, or 153 ms/item), which was held constant for each individual participant.

Participants were instructed to view the sequence, and then verbally report the categories of any images they had seen and to mention if any category had occurred twice. They were specifically instructed that either the same image repeated, or two different members of the same basic-level category counted as a repetition. Responses were checked by the experimenter but no feedback was given in the experiment proper.

Results

Category block

Trials were included in the analysis only if at least one of the critical items (C1 or C2) was correctly reported. Across the four incremental exposure duration conditions this resulted in the exclusion of an average of: 2.25, 1.54, 1.50, and 1.28 identical trials (out of 15); 2.29, 1.46, 1.25, and 1.12 category-repeat trials; and 2.29, 1.42, 1.13, and 0.6 nonrepeat trials. Thus, participants were slightly more likely to miss both critical items at the shortest exposure duration, but there were no differences between the different repetition conditions.

The proportion of trials on which both C1 and C2 were correctly reported was calculated for each repeat type at each presentation time and averaged across participants (see Fig. 2a). These data were entered into a 3 × 4 mixed ANOVA with repetition type (identical, category and nonrepeat) as the repeated-measure, and presentation time (117 ms, 129 ms, 141 ms, 153 ms) as the between-subjects factor. The main effects of repetition, F(2, 176) = 50.308, p < .001, ηp² = 0.364, and presentation time, F(3, 88) = 4.043, p = .01 ηp² = 0.121, were both significant, as was the interaction between these two factors, F(6, 176) = 3.41, p = .003, ηp² = 0.104.

A series of planned pairwise comparisons (Bonferroni corrected) was conducted to compare the different repetition conditions at each presentation time. Accuracy was higher for
identical trials relative to nonrepeats for the 117-ms ($p < .001$) and 153-ms ($p = .001$) presentation times. This difference was not significant for images presented for 129 ms and 141 ms ($p > .05$). Thus, rather than RB, we observed repetition facilitation for identical-repeats at the shortest and longest presentation durations.

In contrast, performance on category-repeat trials was significantly lower than on nonrepeat trials for presentation durations of 129 ms ($p = .022$) and 141 ms ($p = .015$), suggestive of RB. Category repeats did not differ from nonrepeats when scenes were presented for 117 ms or 153 ms ($p > .05$). Accuracy was also significantly lower on category-repeat than identical-repeat trials for all presentation times (all $p s < .001$).

Mirror block

As in the category block, only trials in which at least one of the critical items was reported were included in the analysis. Across the four incremental exposure duration conditions this resulted in the exclusion of an average of 1.88, 1.08, 1.21, and 1.2 identical trials (out of 15); 2.38, 1.33, 1.25, and 1.08 mirror-repeat trials; and 0.67, 0.88, 1.08, and 0.8 nonrepeat trials. Although the numbers are generally small, there were slightly more trials excluded in the mirror-repeat condition, especially at the short presentation duration, than in the other conditions.

The data for this block were analyzed with the same $3 \times 4$ mixed ANOVA with repetition type as the within-subjects factor, and presentation duration as the between-subjects factor. These data are presented in Fig. 2b. There were significant effects of both repetition type, $F(2, 176) = 22.421, p < .001, \eta_p^2 = 0.203$, and presentation duration, $F(3, 88) = 4.519, p = .005, \eta_p^2 = 0.133$, and no significant interaction, $F(6, 176) = 1.635, p = .14, \eta_p^2 = 0.053$. Bonferroni-adjusted pairwise comparisons showed that, when collapsed across exposure duration, accuracy in both the identical-repeat and the mirror-repeat conditions was significantly higher than in

Fig. 1 Illustration of the typical trial structure and experimental conditions used in both Experiments 1 and 2. A. The four possible trial types used across Experiments 1 and 2, as defined by the relationship between the two critical items. Identical-repeat trials, where C1 and C2 were the same image; mirror-repeat trials, where C2 was the mirror image of C1; category-repeat trials, where they are different exemplars of the same category; and nonrepeat trials where they are drawn from different categories. B. The structure of the RSVP sequence in both experiments. A sequence consisted of nine images, with scene images presented in between three initial and three final masks. In Experiment 1, the scene images are the critical items plus an additional filler scene of a different category to both C1 and C2. In Experiment 2, only C1 and C2 were presented, with an additional mask displayed in between. In Experiment 1, presentation durations was varied across participants (117 ms, 129 ms, 141 ms, or 153 ms/item). In Experiment 2, the presentation rate was held constant at 129 ms/item.
the nonrepeat condition (ps < .005). Accuracy was also significant higher in the identical-repeat than in the mirror-repeat condition (p < .001).

Discussion

The results of Experiment 1 revealed a consistent advantage for identifying both critical items when they were identical repetitions relative to when they were scenes from different categories. This was most clearly seen at the shortest (117 ms) and longest (153 ms) presentation rates. A repetition advantage was also observed for mirror-image repetition trials across all presentation rates. In contrast, accuracy for reporting category repeats was significantly lower than accuracy on nonrepeat trials at the two intermediate presentation rates (129 and 141 ms/item).

These findings run counter to our prediction that identical scenes should induce RB, although they suggest that RB may be present when a scene category is repeated, as we predicted would be the case if scene gist is activated rapidly, but this information fails to be individuated into distinct perceptual episodes. However, if scene gist underlies the putative RB effect for category repeats, a similar repetition effect would be expected for identical and mirror-image critical items, as these also share conceptual information and global structure, as well as specific local details. Instead, a repetition advantage was observed for identical and mirror-image items. A possible explanation for this facilitation is that it is mediated by overlap in local scene details, which is maximal in the identical repeats, and also present in the mirror repeats, which share the same local features, albeit in a different spatial configuration. Thus, the pattern of results seen here may arise from a combination of RB mediated by scene gist and facilitation mediated by specific image features. This explanation would be consistent with a model proposed by Hayward et al. (2010) in the context of RB for rotated objects. That study found RB for objects that were repeated in different orientations, but a reduction in the size of RB (sometimes even crossing into facilitation) when the objects were repeated in the identical orientation. This led the authors to suggest that image-based similarity between the critical items counteracts the RB effect that is based on viewpoint-independent features of the object.

However, an alternative explanation for this pattern of results is that it stems from the difficulty of the task. The overall level of accuracy in this experiment was quite low, even for nonrepeat trials, which suggests that participants found it challenging to process three scenes under these conditions. Morris and Still (2008) have suggested that, when item identification is difficult and subjects are aware of the presence of repetitions, they may make guesses based on partial information (e.g., local image elements) gleaned from C1. Such a strategy could result in guessing the presence of repeated items, resulting in a repetition advantage, while at the same time impairing performance on category-repeat trials, where subjects may have low levels of confidence that they had seen a “repetition,” given the lower perceptual overlap of the partial information gleaned from the two scenes. It is therefore important to replicate these results in a less difficult context. This was achieved in Experiment 2 by reducing the number of scenes presented to two instead of three.

Experiment 2

In this experiment, we reduced the working memory load of the task used in Experiment 1 by presenting only two scenes in
the sequence (C1 and C2) separated by a mask. Experiments using a similar design have observed RB for single letters (Chun, 1997; J. Park & Kanwisher, 1994). Only one presentation time was used in this experiment. The 129-ms condition was selected because it was the presentation time for which both the repetition advantage for identical scenes and the apparent RB effect for category repeats were clearly present in the previous experiment. With the exception of these changes, this experiment closely followed the design of Experiment 1. One block contained identical and category repetitions, as well as nonrepeat trials (category block), with the other containing identical, mirror-repeat and nonrepeat trials (mirror block). If the pattern of results in Experiment 1 is replicated here, then we can be more confident that the effects are due to the characteristics of the stimuli rather than an artefact of task difficulty.

Method

Participants

Twenty-five first year psychology students (15 female, mean age 20.5 years) from the University of Sydney participated in this experiment for course credit. None had taken part in Experiment 1.

Procedure

The experiment was identical to Experiment 1 except that a single group was tested at a presentation time of 129 ms/item, and the trials consisted of an RSVP sequence containing only two scene images (C1 and C2) separated by an intervening mask (as well as the three initial masks and three masks following the scenes).

Results

Category block

As in Experiment 1, only trials on which at least one of the critical items was reported were analyzed. On average, 0.44 (out of 15) identical trials, 0.32 category-repeat trials and 0.16 nonrepeat trials were excluded. This low rate of exclusions confirms that this experiment was generally easier than Experiment 1.

The proportion of trials on which both C1 and C2 were correctly reported was calculated for each repetition condition and averaged across participants. These data are presented in Fig. 3a and reveal at a glance that performance is generally higher in this experiment. A within-subjects ANOVA and Bonferroni corrected pairwise comparisons were used to test for differences between the three conditions (identical, category repeats, and nonrepeats). There was an overall effect of repetition type, \( F(2, 48) = 4.986, p = .011, \eta_p^2 = 0.172 \). Accuracy was higher on identical-repeat trials than on category-repeat trials \((p = .001)\), but not reliably higher than nonrepeat trials \((p > .05)\), despite these differences being of similar magnitude. Accuracy was very similar on non-repeat and category-repeat trials in this experiment \((p > .05)\).

Mirror block

Trials were only included in the analysis if at least one of the critical items was reported. As in the category block, only a very small fraction of trials were excluded: on average, 0.52
(out of 15) identical trials, 0.44 mirror-repeat trials and 0.48 nonrepeat trials, which represents lower exclusion rates than in the corresponding block in Experiment 1.

The proportion of trials on which both C1 and C2 were correctly reported was averaged across participants (see Fig. 3b), and analyzed in a within-subjects ANOVA, which showed significant differences in performance between identical, mirror-image, and nonrepeat trials, $F(2,48) = 7.682, p = .001, \eta^2 = 0.242$. Pairwise comparisons (Bonferroni corrected) revealed that accuracy was significantly higher for both the identical and mirror-repeat conditions compared to nonrepeat trials ($p = .021$, and $p = .016$, respectively). Identical-repeat performance did not differ from mirror-repeat performance ($p > .05$).

**Discussion**

Reducing the task demand to reporting two scenes rather than three resulted in higher overall accuracy in this experiment compared to Experiment 1. Under these conditions, a repetition advantage for identical and mirror-repeat images was again obtained, but the RB effect observed for category-repeat trials in Experiment 1 was not replicated. On category-repeat trials, participants were just as likely to detect a shared category as they were to detect two different categories on nonrepeat trials. Therefore, this experiment does not provide any evidence that RB occurs on the basis of shared conceptual and/or global structural scene information. The robust repetition advantage for identical and mirror-image scenes reinforces the initial conclusion from Experiment 1 that specific scene details shared between the two targets mediate the priming effect, and we can be confident that this is not an artefact of task difficulty.

**General Discussion**

This study used an RB paradigm to investigate the way scenes are represented at the early stages of visual processing. Surprisingly, and contrary to our predictions, the repetition effects manifested primarily as a repetition advantage for identical and mirror-image repetitions of the same scenes, rather than as a repetition deficit (i.e., RB). Nevertheless, regardless of direction, the presence of repetition effects across two different experiments indicates that the relevant scene representations were accessed during our task. At the same time, we found no convincing evidence of either facilitation or RB for repetitions of a scene category; rather, different exemplars of the same basic-level scene category appear to be treated the same as scenes of different categories. This pattern of results provides strong evidence that in the first 100–150 ms, scenes are primarily represented in terms of local details and specific views, rather than in terms of more general conceptual categories or global structural properties (i.e., gist). These findings have important implications both for scene processing and for our understanding of RB, which we discuss below.

**Implications for scene processing**

One of the main goals of this study was to test the claim that abstract scene properties, such as conceptual gist, and global structural features, such as the spatial envelope of a scene, are represented at very early stages of processing. These claims are an explicit part of many models of scene gist perception (Bar, 2004; Hochstein & Ahissar, 2002; Oliva & Torralba, 2001; Potter, 1975, 1976). If this were true, then we should have observed similar patterns of results for category-repeat trials as for identical or mirror-image repetitions, given that all of these types of repetitions share these properties. At the very least, one might expect a graded effect in the same direction as identical and mirror-image repetitions, given that different exemplars of the same category share fewer local details. This is not what we found.

Instead, identical repetitions and mirror-image repetitions both yielded a repetition advantage compared to nonrepeated items, although the advantage was stronger for identical than for mirror-image repeats in Experiment 1. In Experiment 1 there was also a suggestion that the advantage for identical repetitions was most pronounced with stimulus durations of 117 ms and 153 ms. These results suggest that local details exert an influence on very early scene identification, and again at a later time point in processing. The later effect may reflect a feedback sweep of activation of local image details during later stages of vision with scrutiny, as suggested by the reverse hierarchy theory (Hochstein & Ahissar, 2002). Alternatively, the early and late influence of local details may arise from distinct mechanisms underlying different stages of scene processing (Epstein, 2014). According to the latter proposal, the first stage of scene processing involves visual analysis of a scene, whereas the second stage involves placing the initial view within a larger framework, both in terms of the spatial context outside the immediate view and conceptual scene information.

In contrast to critical items that shared local scene details, detection of category repeats was largely indistinguishable
from that of unrelated scenes. In Experiment 1, there was a hint of RB between different exemplars of the same category that was only apparent at certain exposure durations (129 and 141 ms/item), but it was not replicated in Experiment 2. Thus, it is likely that this finding was an artefact of the task difficulty in Experiment 1 rather than a true RB effect. We therefore conclude that early representations of scenes rely on local details present in the scene, which implies that scenes are coded primarily as unique places at this early stage of processing, rather than as broader categories.

These findings are consistent with those of the fMRI adaptation studies reviewed in the Introduction, which show a decrease in signal strength in the PPA in response to repetitions of the same scene, but not in response to different exemplars of the same category (Epstein & Morgan, 2012). Thus, the scene representations tapped by our RSVP task seem to have the same characteristics as those coded in the PPA. These representations might become activated primarily to prepare the observer for navigation or action within a scene, but our results indicate that they also influence performance in a scene categorization task. It is interesting to note that, although fMRI studies have not revealed adaptation effects for different exemplars of the same scene category, information about the category nonetheless seems to be coded in distributed patterns of activity, as revealed by multivoxel pattern analyses (Walther, Caddigan, Fei-Fei, & Beck, 2009). This may be able to support recognition of scene gist when monitoring incoming visual information for particular cued categories (e.g., Potter, 1975, 1976; Thorpe et al., 1996), but does not appear to influence the behavioral and neural response to scenes in a rapidly updating context, such as these experiments or the fMRI adaptation studies.

Implications for repetition blindness

The most surprising finding of this study was that our task did not elicit RB for scene stimuli under the same experimental conditions that typically yield robust RB for other types of stimuli, such as letters, words, and pictures of objects (Bavelier, 1994; Harris & Dux, 2005a; Hayward et al., 2010; Kanwisher, 1987; Kanwisher et al., 1999). A repetition advantage (i.e., priming) was found instead. Priming and repetition blindness are related phenomena. Both involve the repeated presentation of stimuli that are either identical or similar on some critical dimension. Repetition causes stronger type activation, which leads to priming in most circumstances. Under certain conditions, mostly characterized by insufficient time to consolidate both stimuli, RB occurs instead. Previous research indicates that whether priming or RB occurs depends on whether C1 is attended and tokenized—tokenizing C1 leads to RB, whereas ignoring C1 typically results in priming (Kanwisher, 1987)—as well as how easily C1 and C2 can be distinguished from each other and from other items in the RSVP stream, in terms of low-level perceptual features (Chun, 1997; Dux & Coltheart, 2008). Our finding of repetition priming rather than RB for scene stimuli might reflect the fact that scene processing relies heavily on perceptual features present in the image (as discussed above), which may help to individuate the critical items from the surrounding items in the stream.

Of course, we cannot exclude the possibility that RB might be demonstrated for scene stimuli if the timing parameters were different or if the task demands are changed. However, it is interesting that under very similar temporal parameters and employing a similar task of basic-level naming, RB is routinely obtained for object stimuli (and corresponding words), but not for pictures of scenes. This implies that rapidly changing visual input affects processing of objects and words differently to scenes. Some researchers have speculated that RB reflects a heuristic employed by the visual system to deal with its limited rate of temporal updating (Chun, 1997; Kanwisher, 1987; Naughtin et al., 2013). That is, if two identical stimuli appear in the same spatial location at almost the same time, the system “assumes” that this is the same object, because under most circumstances that is the case—typically, any temporal discontinuity in perceiving the object is due to the observer looking away or blinking, but the object itself has not moved or changed. As a result, the two instances are collapsed into a single object token, and the observer is blind to the repetition. This may not hold when viewing scenes, because we interact with scenes in a very different way. Unlike static objects, scenes are large environments that the observer frequently views while in motion. This means that the visual system is used to dealing with an ever changing view of the scene, as the observer moves through the scene and local details shift in relation to the observer’s viewpoint. Thus, to enable effective navigation, brain systems involved in scene perception would need to be able to maintain and rapidly update viewpoint-specific representations of the environment.

In conclusion, this study found that scenes are represented primarily in terms of local details and specific views at early stages of processing, implying that we encode scenes as particular places first and foremost. The discrepancy between our results and other findings of early extraction of scene gist may be the result of differences in the tasks used: tasks that require rapid updating of the representations of the critical scene items may rely on mechanisms that code visual features of the scenes with great fidelity.

Author Note  This research was supported by ARC grant DP0879206. M. Goldzieher was supported by an Australian Postgraduate Award and I. M. Harris was supported by an ARC Future Fellowship (FT0992123). Correspondence regarding this article should be addressed to mgoldzieher@gmail.com or irina.harris@sydney.edu.au.
Appendix

Example scenes from each of the 15 scene categories used in Experiments 1 and 2. Both mirror-image versions and a different exemplar from the same category are provided to illustrate how similar the critical items would appear on mirror-repeat and category-repeat trials.

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Mirror Image</th>
<th>Same Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airport</td>
<td><img src="image" alt="Airport Original" /> <img src="image" alt="Airport Mirror Image" /> <img src="image" alt="Airport Same Category" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bathroom</td>
<td><img src="image" alt="Bathroom Original" /> <img src="image" alt="Bathroom Mirror Image" /> <img src="image" alt="Bathroom Same Category" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beach</td>
<td><img src="image" alt="Beach Original" /> <img src="image" alt="Beach Mirror Image" /> <img src="image" alt="Beach Same Category" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cemetery</td>
<td><img src="image" alt="Cemetery Original" /> <img src="image" alt="Cemetery Mirror Image" /> <img src="image" alt="Cemetery Same Category" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classroom</td>
<td><img src="image" alt="Classroom Original" /> <img src="image" alt="Classroom Mirror Image" /> <img src="image" alt="Classroom Same Category" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Category</td>
<td>Image 1</td>
<td>Image 2</td>
<td>Image 3</td>
</tr>
<tr>
<td>-----------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>Field</td>
<td><img src="image1" alt="Field Image 1" /></td>
<td><img src="image2" alt="Field Image 2" /></td>
<td><img src="image3" alt="Field Image 3" /></td>
</tr>
<tr>
<td>Forest</td>
<td><img src="image4" alt="Forest Image 1" /></td>
<td><img src="image5" alt="Forest Image 2" /></td>
<td><img src="image6" alt="Forest Image 3" /></td>
</tr>
<tr>
<td>Garden</td>
<td><img src="image7" alt="Garden Image 1" /></td>
<td><img src="image8" alt="Garden Image 2" /></td>
<td><img src="image9" alt="Garden Image 3" /></td>
</tr>
<tr>
<td>Mountain</td>
<td><img src="image10" alt="Mountain Image 1" /></td>
<td><img src="image11" alt="Mountain Image 2" /></td>
<td><img src="image12" alt="Mountain Image 3" /></td>
</tr>
<tr>
<td>Petrol Station</td>
<td><img src="image13" alt="Petrol Station Image 1" /></td>
<td><img src="image14" alt="Petrol Station Image 2" /></td>
<td><img src="image15" alt="Petrol Station Image 3" /></td>
</tr>
<tr>
<td>Port</td>
<td><img src="image16" alt="Port Image 1" /></td>
<td><img src="image17" alt="Port Image 2" /></td>
<td><img src="image18" alt="Port Image 3" /></td>
</tr>
<tr>
<td>Restaurant</td>
<td><img src="image19" alt="Restaurant Image 1" /></td>
<td><img src="image20" alt="Restaurant Image 2" /></td>
<td><img src="image21" alt="Restaurant Image 3" /></td>
</tr>
</tbody>
</table>
References


