Scene categorization in the presence of a distractor

Jiří Lukavský

Institute of Psychology, Czech Academy of Sciences, Prague, Czech Republic

Humans display a very good understanding of the content in briefly presented photographs. To achieve this understanding, humans rely on information from both high-acuity central vision and peripheral vision. Previous studies have investigated the relative contribution of central and peripheral vision. However, the role of attention in this task remains unclear. In this study, we presented composite images with a scene in the center and another scene in the periphery. The two channels conveyed different information, and the participants were asked to focus on one channel while ignoring the other. In two experiments, we showed that (a) people are better at recognizing the central part, (b) the conflicting signal in the ignored part hinders performance, and (c) this effect is true for both parts (focusing on the central or peripheral part). We conclude that scene recognition is based on both central and peripheral information, even when participants are instructed to focus only on one part of the image and ignore the other. In contrast to the zoom-out hypothesis, we propose that the gist recognition process should be interpreted in terms of the evidence accumulation model in which information from the to-be-ignored parts is also included.

Introduction

By glancing at a complex real-world scene, humans can extract a substantial amount of semantic and perceptual information. This ability has been investigated in relation to a large variety of tasks. People can categorize a whole scene (such as an office or a forest; e.g., Rousselet, Joubert, & Fabre-Thorpe, 2005). People can also recognize a corresponding superordinate category (e.g., natural/man-made; Loschky & Larson, 2010) or estimate a global geometrical property (e.g., mean depth or navigability; Greene & Oliva, 2009). People can identify an image based on an earlier verbal description (e.g., "a girl sitting on a bed"; Potter, 1976) and can also determine whether an image contains an animal (e.g., Bacon-Macé, Macé, Fabre-Thorpe, & Thorpe, 2005). The rapidly acquired information in these tasks is referred to as a scene gist.

Scene gist recognition appears effortless and has led to a debate regarding whether attention is required for the task. If minimal or no attention is necessary, scene gist recognition could be similar to preattentive processes. Arguments supporting this view stem from dual task experiments and parallel processing. People have been asked to detect animals or cars in peripherally presented scenes while being focused on a central task (Fei-Fei, VanRullen, Koch, & Perona, 2002; Fei-Fei, VanRullen, Koch, & Perona, 2005). This observation is not limited to the periphery, as negative priming studies have shown that people can better discriminate among scenes than among letters (Otsuka & Kawaguchi, 2007). Furthermore, when people are asked to detect animals across several scenes presented simultaneously, their performance is consistent with a parallel processing model (Rousselet, Fabre-Thorpe, & Thorpe, 2002; Rousselet, Thorpe, & Fabre-Thorpe, 2004).

Arguments supporting the importance of attention in scene gist processing are based on dual tasks involving sustained attention and critical analyses of processing in the animal detection tasks used in previous experiments. Cohen, Alvarez, and Nakayama (2011) showed that dual task experiments involving sustained attention (based on the Multiple Object Tracking paradigm or rapid serial visual presentation) affect the ability to detect unexpected scenes. When attention is consumed in a different task and a surprise scene is presented, people have difficulty detecting its gist, and they are often not aware of its presence (Mack & Clarke, 2012). The conclusions based on the animal detection tasks used in previous experiments may be complicated by the actual level of processing required for a successful response. The task requires few attentional resources, but processing is limited and may be based on the partial processing of unbound features (Evans & Treisman, 2005)—that is, people frequently fail to identify or locate the animals and can be misled by the presence of humans. Gist-related tasks, such as detecting an animal, may require attention when the scene is complex and contains several primary objects (Walker, Stafford, & Davis, 2008).

Citation: Lukavský, J. (2019). Scene categorization in the presence of a distractor. *Journal of Vision, 19*(2):6, 1–11, https://doi.org/10.1167/19.2.6.

https://doi.org/10.1167/19.2.6 Received May 15, 2018; published February 8, 2019 ISSN 1534-7362

The role of attention becomes important when people need to focus on some information and suppress other information. In some tasks, we process multiple information channels that may convey independent and potentially conflicting content. In Stroop's experiment (Stroop, 1935), people were presented with color names printed in conflicting colors (e.g., "red" printed in blue) or color squares, and they were asked to name the ink colors. In this experiment, people read the word even when it was not relevant to the task, and the ink/word incongruence impaired their performance. The picture–word interference paradigm used in language and object recognition research has a similar effect (Starreveld & Heij, 2017). People are presented with a word and a picture of an object and are asked to either read the word or name the object. Incongruent words slow the responses, but incongruent pictures have little effect on the reading time (Rosinski, Golinkoff, & Kukish, 1975). The picture–word interference paradigm highlights the automaticity of reading. However, if more intensive processing is required for the verbal material, the opposite effect can be observed in which the depicted objects interfere with the verbal task. Greene and Fei-Fei (2014) asked people to semantically categorize words (instead of merely reading) accompanied by congruent/conflicting photographs of objects and scenes. These authors found that categorization was impaired under conflicting conditions and argued that visual categorization is an automatic process.

Information interference is not limited to verbal tasks. Navon (1977) presented people with large letters consisting of smaller letters (e.g., a large H consisting of small S's). The participants were tasked with responding to either the global shape (large letter) or the local shape (small letter). Navon's results highlighted the precedence of global processing—that is, people were faster under the global condition and slower under the local condition when the content in the channels conflicted. Other examples include experiments featuring scenes with consistent/inconsistent foreground figures. In inconsistent photographs, the processing of both the foreground figure and background are impaired (Davenport & Potter, 2004). Thus, the results do not show an asymmetrical pattern (a precedence for either the foreground or background), which has been reported in studies similar to Stroop's. Davenport and Potter (2004) concluded that the foreground and background are processed concurrently and interactively. In this experiment, the information conveyed by the central area of an image differed from that conveyed by the peripheral area, and we investigated how people integrated this conflicting information.

Both central and peripheral vision contribute to scene gist recognition. Central vision (foveal and parafoveal) provides a high-resolution signal suitable for object recognition. Information about the presence of a specific or diagnostic object can be indicative of a scene belonging to a certain category (Oliva & Schyns, 1997). Although the resolution in the periphery is lower, low-frequency signals are processed faster and can convey useful information for gist recognition (Schyns & Oliva, 1994).

Larson and Loschky (2009) investigated the relative contribution of both parts. These authors presented grayscale photographs either centrally (window condition) or without the center—that is, by covering the central part (scotoma condition), and tested the effect of different window/scotoma sizes on scene gist recognition. The authors found that the contribution of the peripheral signal was larger than expected based on cortical magnification models (Florack, 2007; Van Essen, Newsome, & Maunsell, 1984). Thus, the central window must cover a larger area of the visual field (i.e., a larger corresponding area of V1) than the respective scotoma scene to yield an equal performance (central circle with a 7.4° radius vs. annulus with an outer radius of 13.55°).

In their subsequent experiments, the authors manipulated the presentation times to reveal how the relative contributions of the central and peripheral areas change over time (Larson, Freeman, Ringer, & Loschky, 2014). The authors found that during the first 100 ms, there is an advantage in favor of central processing, and subsequently, the relative contribution of peripheral vision increases. The authors proposed the zoom-out hypothesis to interpret the results as follows: Attention starts allocated centrally at the fixation and expands outward over time. The zoom-out hypothesis combines the sequential attention model of eye movements (Henderson, 1992) and the zoom-lens model of attention (Eriksen & St. James, 1986). According to the sequential attention model, attention is focused at the foveated object at the beginning of each fixation and moves farther only when the foveated object is processed. The zoom-lens model states that the attention span can vary from a small, narrow, and focused area to a large, broadly distributed, and diffuse area.

In this study, we investigate how people balance the signals from their central and peripheral vision when the signals are independent and potentially conflicting. In our experiments, the participants were presented with a scene in the central window and another scene in the periphery. The participants were asked to attend to one part and ignore the other. The attentional focus did not change within a block to discount the necessity of fast attentional shifts. When people expect a peripheral signal, the early advantage of central vision is eliminated (Larson et al., 2014). Building on previous research investigating peripheral/central visual processing and the attentional processing of conflicting signals, we tested several hypotheses.

- Hypothesis 1: Parallel and isolated processing. One

possibility is that central and peripheral signals are processed in parallel and in isolation. If the information in both channels is conflicting, no impact on performance is expected because people can focus solely on the required channel. In addition to the anatomical separation during the early stages of visual processing, this selection mechanism may be based on the ability of humans to distribute attention over several noncontiguous locations (Franconeri, Alvarez, & Enns, 2007; Kramer & Hahn, 1995). Alternatively, the selection may be based on a strategic decision—that is, whether to pay attention to the local features and diagnostic objects or ensemble characteristics, which may favor central or peripheral vision, respectively.

- Hypothesis 2: Automaticity and precedence. Alternatively, one channel may have priority in processing and affect the outcome of the second channel. This pattern is observed in several Stroop-like paradigms. Thus, we predicted that the prioritized channel would not be affected by the conflicting stimuli, but the processing of the second channel would be impaired. In particular, the zoom-out hypothesis (Larson et al., 2014) supports the prioritization of central processing, especially with short presentation times. Alternatively, the concept of coarse-to-fine recognition (e.g., Navon, 1977; Schyns & Oliva, 1994) assumes that the low-frequency features available in the peripheral vision are processed earlier than the detailed information (available from central vision).
- Hypothesis 3: Concurrency and interactivity. Finally, the processing of central and peripheral information may be concurrent and interactive. When presented with conflicting stimuli, performance may be impaired in both channels. This pattern is observed in experiments featuring inconsistent figure/background compositions (Davenport, 2007; Davenport & Potter, 2004).

To distinguish between the hypotheses, we fitted the data with a generalized mixed model and tested whether there were significant decreases in performance caused by inconsistent distractors. We tested these differences separately for each focused area (central/peripheral).

Experiment 1

Method

Participants

Twenty-four university students (16 women) participated in Experiment 1 (age range = 19–32 years; M = 22.5 years). Three additional participants were originally tested, but their data were not used because they had difficulty maintaining central fixation (see Procedure and Design). The participants signed an informed consent form and received course credit for their participation.

Stimuli

We presented the participants with composite scenes in which the central circle depicted a scene (radius = 5.54°) and the surrounding annulus (radius = 11.08°) depicted a different scene. The radius values were based on the study of Larson and colleagues (2014) and were later validated in a control study.[1] The stimuli were presented on a midgray background (128 RGB pixel luminance). To avoid abrupt transitions, we covered the boundary between the circles and the boundary between the annulus and the background with a smooth midgray gradient. The transparency of the gradient was modulated with a raised cosine function (the gradient band was 1.8° wide from 0% cover to 0% cover). The same background color was used on blank screens within the trial to minimize the interscreen luminance contrast (Freeman, Loschky, & Hansen, 2015).

We used 200 images selected from 10 scene categories (man-made: house, market, pool, street, and train/station; natural: beach, desert, forest, mountain, and river). Example composite stimuli are shown in Figure 1. Each original image was used twice (central and peripheral part) to create 200 composite scenes. No photograph was repeated within the same attention condition. For each attention condition (central/peripheral), we balanced the combination of superordinate categories in the central and peripheral areas (2 × 2: central: man-made/natural; peripheral: man-made/natural). Thus, the central and peripheral scenes were consistent in their superordinate category in half the trials and differed in the other half of the trials.

The masks were synthetic textures generated with Portilla and Simoncelli's (2000) algorithm. The masks were created separately for the central and peripheral parts and were identical in size, and their boundaries were also smoothed with a gradient.

All stimuli (targets and masks) were equalized in terms of their mean luminance and root mean square (RMS) contrast (Loschky et al., 2007). The stimuli and data from both experiments can be found at https://osf.io/849wm/.

Procedure and design

The participants were asked to attend to either the central or peripheral part and categorize the scenes into corresponding superordinate categories (man-made/

Figure 1. Examples of composite stimuli. The stimuli contained one scene in the central part and a different scene in the peripheral part. The scenes were selected from two superordinate categories (man-made or natural scenes). The resulting composite image featured either scenes from the same superordinate categories (congruent condition, top row) or different superordinate categories (incongruent condition, bottom row). Man-made scene categories: house, market, pool, street, and train/station. Natural scenes: beach, desert, forest, mountain, and river.

natural scenes). The central fixation was controlled by an eye tracker (EyeLinkII, SR Research, Ltd., Ontario, Canada). The experiment was presented on a 22-in. monitor (resolution 1680 × 1050, refresh rate 60 Hz), which the participants viewed from a distance of 60 cm. Their head movements were restricted using a chin rest. The experiment was presented using a MATLAB (MathWorks, Natick, MA) script with Psychophysics and Eyelink Toolbox extensions (Brainard, 1997; Cornelissen, Peters, & Palmer, 2002; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997).

Each trial started with a drift correction (see Figure 2). Then, the fixation point was presented for 1000 ms at the location of the drift correction target. The composite scene was presented for 33 ms, followed by a blank screen (for 167 ms), and then a corresponding mask (presented for 33 ms, with target-to-mask stimulus onset asynchrony [SOA] = 200 ms). Finally, the screen turned blank again until the participant responded. The responses were conveyed via the arrow keys on a keyboard. During the course of a trial, black text at the lower edge of the screen reminded the participants of the response assignment (left = man-made; right = natural) and the part (i.e., central or peripheral) on which they should focus.

The timing followed the design of experiment 1 by Larson et al. (2014). Due to the differences in the refresh rate (60 vs. 85 Hz), we chose slightly longer presentation times (33 vs. 24 ms). Larson et al. (2014) did not find difference between scotoma/window conditions for SOAs within the 94–376 ms range. The SOA of 200 ms was selected to be close to the 188-ms condition used in their experiment.

The experiment was divided into two training blocks and four experimental blocks. Each training block included 20 trials, and the stimuli were featured only at the peripheral or central part (the other part was filled with a midgray color). Each experimental block included 50 trials. The central/peripheral focus did not vary within the blocks and alternated between the blocks. Half of the participants started with the central focus.

When the fixation point was presented, we checked whether the gaze was within a circular area around the fixation point (radius = 1°). If the gaze moved beyond this limit, the trial was interrupted (no stimulus was

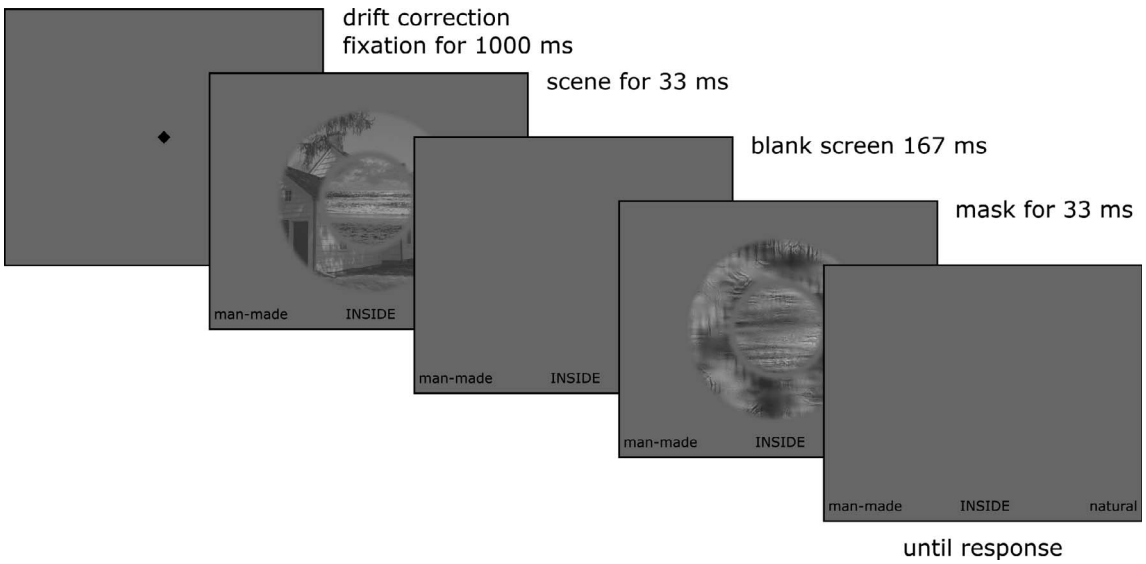

Figure 2. The course of a trial in Experiment 1. The response labels were shown in all phases of the trial and were located further near the screen edge.

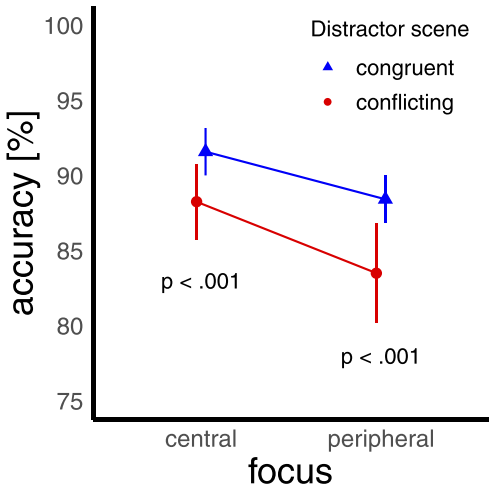

Figure 3. Results of Experiment 1. Accuracy as a function of the focused area (central/peripheral) and category consistency (both parts with same/different superordinate category). The *p* values refer to pairwise contrasts between congruent and conflicting conditions in the full model.

shown) and was scheduled to be performed later during the block. If more than 20% of the trials had to be rescheduled, the block was terminated, and the participant finished the remaining blocks, but his/her data were discarded (three participants in Experiment 1).

Results

The average accuracy was 87.9% ($SD = 8.2$, range = 69.5%–99.5%). The results are shown in Figure 3. We analyzed the performance data using generalized mixed models (logistic regression; see Supplementary File S1). First, we specified the null model with no fixed effect and two random intercept effects (participant and attended image). The dependent variable was the accuracy of the particular responses (correct/incorrect). We gradually added more fixed factors and checked for statistically significant improvements in the model with likelihood ratio tests. We report Bayesian information criterion (BIC) values; lower values are preferable in the model selection process. We added the fixed factors in the following order: superordinate category of the attended scene (man-made/natural; to discount the differences between the categories), focused area (central/peripheral; to check for a foveal benefit), category conflict (conflicting/congruent scenes; to check whether the unattended signal affected the response), and area × conflict interaction (to check whether the conflict effect was similar in both areas). We report the odds ratios (ORs) and 95% CIs in the final model.

The participants showed greater accuracy in making decisions about the natural scenes, $\chi^2(1) = 23.59$, $p < 0.001$, BIC = 3159.4, null model BIC = 3174.5. Assessing the scenes shown in the periphery was more difficult, $\chi^2(1) = 22.64$, $p < 0.001$, BIC = 3145.2. Stimuli comprising two conflicting scenes were more difficult to analyze, $\chi^2(1) = 25.15$, $p < 0.001$, BIC = 3128.5. However, no significant interaction was observed between the presence of a conflict and the attended area, suggesting that the cost of the conflict does not differ between the fovea and periphery, $\chi^2(1) = 0.10$, $p = 0.747$, BIC = 3136.9. Since adding the interaction term did not improve the model, the final model featured the following: superordinate category of the attended scene (man-made/natural), focused area (central/peripheral), and the presence of a category conflict.

To summarize the final model, the natural scenes were easier to recognize (OR = 2.22, 95% CI [1.62; 3.06]). Performance was higher in response to stimuli presented in the central area (OR = 1.60, 95% CI [1.31; 1.95]). When a scene from a conflicting category was presented in the unattended area, the accuracy decreased (congruent OR = 1.66, 95% CI [1.36; 2.03]).

In the model selection process, adding the Area × Conflict interaction term did not improve the model. The nonsignificant interaction is not an argument for one of the hypotheses we formulated in the introduction. It only means that the difference between congruent and conflicting stimuli is likely not different for central and peripheral conditions. Pairwise contrast in the final model showed a significant decrease in conflicting stimuli across both the central and peripheral conditions (log OR = 0.504, $z = 5.122$, $p < 0.001$). When the interaction term was added, the decrease was also significant in both conditions (central: log OR = 0.541, $z = 3.637$, $p < 0.001$, peripheral: log OR = 0.475, $z = 3.606$, $p < 0.001$).

Discussion

In Experiment 1, we presented two different scenes simultaneously and asked people to attend to one part defined by location and eccentricity (central/peripheral). We found that people were generally better at recognizing central stimuli. Importantly, when the stimuli belonged to opposing categories and potentially required different responses, we observed decreased performance for both attended areas.

Regarding our hypotheses, our observations oppose the hypothesis of parallel and isolated processing. This hypothesis predicted that no decrease in performance would occur under conflicting conditions because people are able to isolate the information channels. The observed pattern is consistent with the hypothesis of concurrent and interactive processing, which suggests

that people simultaneously process both parts (with a small advantage for the central part). While the results also oppose the hypothesis of automaticity and precedence, it is possible that at 200 ms SOA, the signals from central and peripheral areas are already processed, and we cannot determine the precedence. Because the time course of scene gist processing differs between the central and peripheral areas (Larson et al., 2014), we further manipulated the presentation times and tested the impact on the processing of conflicting signals.

Experiment 2

In Experiment 2, we used a similar procedure and the same stimuli used in Experiment 1. The main difference was that we manipulated the timing (SOA). In each block, the participants focused on either the central or the peripheral part, but the presentation time varied (SOA = 33, 200, or 400 ms). The middle value (SOA = 200 ms) was identical to the timing used in Experiment 1. The other two durations corresponded approximately to the lower and upper limits used by Larson et al. (2014). They selected their minimum SOA (24 ms) as the shortest duration, which yielded above-chance performance in their pilot testing. Their longest SOA (376 ms) was set to be longer than the duration of an average fixation on scene images (330 ms) and they showed it yielded identical results to the no-mask condition.

Method

Participants

Twenty-four university students (18 women) participated in Experiment 2 (age range = 18–40 years; M = 22.4 years). Four additional participants were originally tested, but their data were not used because they had difficulty maintaining central fixation (see Procedure). The participants signed an informed consent form and received course credit for their participation.

Stimuli, procedure, and design

In Experiment 2, we used the same stimuli, but we manipulated the presentation times. In one-third of the trials, we used the same SOA used in Experiment 1 (SOA = 200 ms). In another third of the trials, the mask immediately followed the scene (SOA = 33 ms). In the remaining trials, the mask was delayed (SOA = 400 ms). We presented 48 trials within each experimental block (with 16 trials of each SOA in random order). The training blocks included 20 trials (as in Experiment 1) with six to seven trials of each SOA condition.

Results and discussion

The average accuracy was 81.3% (SD = 6.9, range = 65.1%–91.1%). Overall, Experiment 2 was more difficult than Experiment 1, $t(44.75) = 3.02$, $p = 0.004$, Cohen's $d = 0.87$. When we calculated the performance selectively for the SOA 200-ms condition, which was used in Experiment 1, we found that the performance in both experiments was comparable, $t(44.81) = 0.92$, $p = 0.362$. The results are shown in Figure 4.

Similar to Experiment 1, we used a generalized linear model but also included the effect of SOA. We added the fixed factors in the following order: superordinate category of the attended scene (man-made/natural; to discount the differences between the categories), SOA (three levels 33, 200, or 400 ms; to determine the effect of the limited presentation time), focused area (central/peripheral; to check for a foveal benefit), category conflict (conflicting/congruent scenes; to check whether the unattended signal affects the response), Area × Conflict interaction (to check whether the conflict effect is similar in both areas), and SOA × Area × Conflict interaction. We report the ORs and 95% CIs in the final model.

In contrast to Experiment 1, we found no significant difference between the man-made and natural scenes, $\chi^2(1) = 0.35$, $p = 0.552$, BIC = 4337.0, null model BIC = 4328.9. Consequently, we omitted this term in the following models. We found that the presentation time affected the performance, $\chi^2(2) = 172.5$, $p < 0.001$, BIC = 4173.3. Assessing the scenes shown at the periphery was more difficult, $\chi^2(1) = 85.99$, $p < 0.001$, BIC = 4095.8. Stimuli comprising two conflicting scenes were more difficult to analyze, $\chi^2(1) = 53.74$, $p < 0.001$, BIC = 4050.5. Adding the interaction terms did not improve the model (Area × Conflict: $\chi^2[1] < 0.01$, $p = 0.992$; BIC = 4058.9, SOA × Area × Conflict: $\chi^2[1] = 11.49$, $p = 0.119$, BIC = 4098.0). As shown in Figure 4, an interaction appeared to be present under the SOA = 33-ms condition. To test this interaction, we analyzed the data for SOA = 33 ms separately, but again, adding the interaction term to the focused area and conflict predictors did not improve the model ($\chi^2[1] = 1.45$, $p = 0.228$, BIC = 1820.5 vs. BIC = 1814.6 with no interaction, null model BIC = 1863.3). The final model featured the following: SOA, focused area (central/peripheral), and presence of category conflict.

To summarize the final model, we found a significant effect of presentation time. Relative to SOA = 200 ms (also used in Experiment 1), the shorter presentation time (SOA = 33 ms) led to a decrease in performance (OR = 0.35, 95% CI [0.29, 0.43]), while the longer

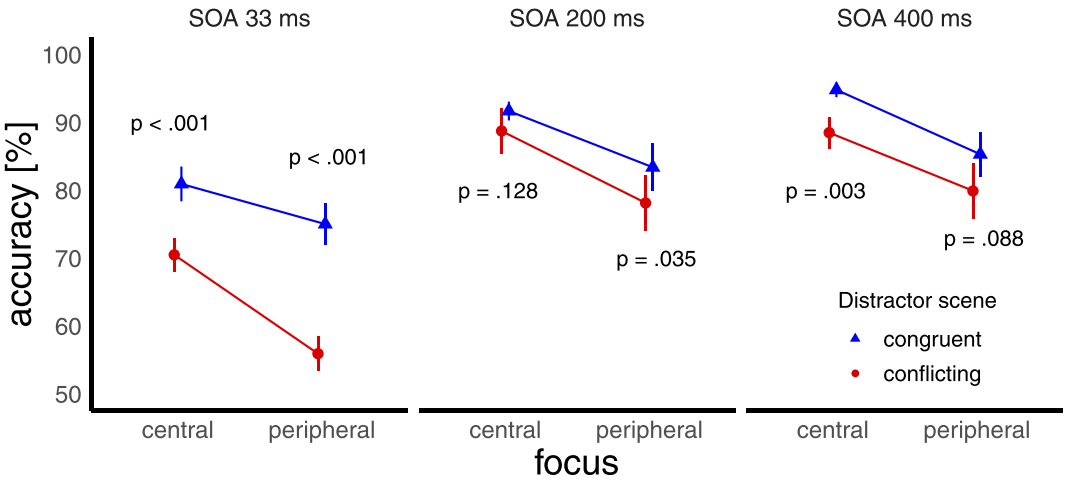

Figure 4. Results of Experiment 2. Accuracy as a function of timing (stimulus onset asynchrony 33, 200, or 400 ms), focused area (central or peripheral) and category consistency (both parts with same or different superordinate category). The *p* values refer to pairwise contrasts between congruent and conflicting conditions in the full model. The final model led to equivalent and significant contrast across all conditions ($z = 7.33$, $p < 0.001$).

presentation time did not provide a significant benefit (OR = 1.18, 95% CI [0.95, 1.47]). Performance was higher in response to the stimuli presented in the central area (OR = 2.17, 95% CI [1.84; 2.58]). When a scene from a conflicting category was presented in the unattended area, the accuracy decreased (congruent OR = 1.86, 95% CI [1.58; 2.21]).

Similar to Experiment 1, the final model suggested that we should treat the congruent-conflicting difference as equivalent across all conditions (area and SOA). Calculated in this way, the conflicting distractors caused a significant decrease in performance (log OR = 0.622, $z = 7.33$, $p < 0.001$). When the interaction terms were added, the decrease was also significant in both conditions for SOA = 33 ms, but at longer presentation times we could see few nonsignificant differences (see Table 1 for detailed results).

In Experiment 1, the final and full models yielded similar patterns of results. In Experiment 2, the apparent Area × Conflict interaction at 33 ms SOA, as well as two pairwise contrasts, were not significant, which might be caused by the lower power of the design. We reduced the number of trials in each condition to keep the overall number of trials similar in both experiments and fit the experiments into a single experimental session. Despite this limitation, we argue that the results allow us to draw conclusions about the postulated hypotheses. First, the model selection procedure showed that adding the interaction terms did not improve the model fit. We report the results of the full model to improve transparency regarding our results. Second, the critical pairwise contrasts in Experiment 2 were for the 33-ms SOA condition because we wanted to test whether the distractors affected performance for short presentation times. Both contrasts were significant.

General discussion

In a series of two experiments, we presented people with composite images that featured a photograph in the central area (up to 5.54°) and a different photograph in the periphery (5.54°–11.08°). Previous experiments with isolated photographs have found balanced performance with an advantage for the central photograph during brief presentation times (Larson et al., 2014). In the current task, people needed to focus on one part of the image and ignore the other.

In designing our experiments, we tried to follow the design in Larson et al. (2014). We used the same critical radius values, and we chose stimuli from similar scene

SOA	Focus area	log OR	*z*	*p*
Final model				
For all conditions		0.622	7.330	<0.001
Full model				
33	Central	0.688	3.646	<0.001
33	Peripheral	0.945	5.590	<0.001
200	Central	0.396	1.524	0.128
200	Peripheral	0.411	2.107	0.035
400	Central	0.836	2.944	0.003
400	Peripheral	0.346	1.705	0.088

Table 1. Pairwise contrasts between congruent and conflicting conditions. Note. The final model refers to the best fitting model. The full model also includes nonsignificant interactions (SOA × Area × Conflict). SOA = stimulus onset asynchrony; log OR = logarithm of the odds ratio.

categories and we processed them in a similar way (equalizing luminance and contrasts, using the same algorithm for generating masks). We also used similar presentation intervals. However, differences exist. Most notably, our participants were asked to identify the superordinate category (man-made vs. natural), while the participants in Larson et al.'s study were asked about the basic level (e.g., "Was this a forest?"). While both tasks are common in scene gist experiments, the latter is more difficult and may require more attention (Evans & Treisman, 2005; Loschky & Larson, 2010). In our task, we needed the stimuli to be congruent or incongruent. If we defined congruent as coming from the same basic level category, 50% of the stimuli would share the same category and would look too congruent relative to the rest. In addition to the category level difference, the two studies differed in their research questions, experimental conditions, sample size, stimuli images, and likely in the setup details. Consequently, our experiments should not be regarded as replications of Larson et al. (2014). However, despite differences, we consider the experiments to be sufficiently comparable to involve similar underlying processes.

In both experiments, people were better at recognizing the central stimulus. In the current design, we did not directly compare performance between isolated photographs and composites. However, based on our control study, the isolated stimuli lead to comparable performance (within ±2% equivalence bounds). Note that in the control study the performance for peripheral stimuli was actually higher by 0.9% compared to the central stimuli. While the better performance for central stimuli is interesting, our primary interest was to test how performance changed with congruent/conflicting distractors.

The results showed that the content of the distractor image affected the scene recognition in the target area. Importantly, this effect was present with both central and peripheral distractors. This observation contradicts the predictions of the zoom-out hypothesis (Larson et al., 2014), which suggests that people should focus on the central area and gradually expand their focus to peripheral areas. In this study, we observed the detrimental effect of peripheral distractors even with very short presentation times (10.4% decrease with a 33-ms SOA).

Alternatively, we can model the performance as an evidence accumulation or random walk model (Gold & Shadlen, 2007). In a random walk model, a single variable can represent the current state of evidence for the man-made scenes (positive values) and natural scenes (negative values). Over time, each piece of evidence accumulates until a decision boundary is reached, at which point we can declare the stimulus category with sufficient confidence. We propose that people gradually gather evidence from the whole image during gist recognition. Particular object features or ensemble characteristics are considered partial evidence for the classification into superordinate categories. The evidence from the attended regions in the composite images is weighted more (or the evidence from to-be-ignored parts is weighted less). Nevertheless, information from the distractor areas contributes to the decision. If conflicting, the information increases uncertainty, and more time is required to reach the decision boundary. Given that the presentation time is limited, the participants often cannot reach a decision and are required to guess.

Based on the results of this study and Larson et al. (2014), we speculate that two main factors affect the weights of evidence in the model. First, the content in the attended area is weighted more; otherwise, we could not discriminate under the conflicting condition. Second, the content in the central area is weighted more, which could explain the overall effect of the central focus. The magnitudes of both effects are comparable; that is, these effects cancel out when people are asked to attend to a single peripheral stimulus (Experiment 3 in Larson et al., 2014).

The observed pattern of distractor effects on recognition is analogous to observations in experiments with inconsistent figure/background composites (Davenport & Potter, 2004); the recognition of both the figure and background is affected by an inconsistent distractor. However, the current experiment differed in two aspects. First, people were presented with composites of two scenes rather than a figure on a potentially inconsistent background. Neuroimaging studies have shown that many different areas contribute to the processing of inconsistent scenes (Gronau, Neta, & Bar, 2008; Mudrik, Lamy, & Deouell, 2010; Rémy, Vayssière, Pins, Boucart, & Fabre-Thorpe, 2014). Changes in activity have been reported in the lateral occipital complex, parahippocampal place area, and prefrontal cortex. In particular, Rémy and colleagues (2014) showed that activity in the right anterior parahippocampal cortex is correlated with an increase in reaction time in inconsistent scenes.

The variety of involved regions corresponds to the variety of contributing processes (e.g., semantic processing, visual contextual processing, object-related processing, and attention and memory-related processes; Gronau et al., 2008). In the current experiment, the importance of object recognition may differ. Some central stimuli may feature objects that facilitate gist recognition (windows, cars, or trees), but in many photographs, the central part consisted of a texture (rocky slope, water surface, or sand). Thus, the conclusion that the central parts are primarily recognized by object recognition in the current experiments is imprecise.

Second, the current experiments differed in the definition of consistency. The congruent condition

featured scenes with identical superordinate categories (e.g., man-made). Strictly speaking, the composite image was always semantically and geometrically inconsistent (pool and street). In a minority of trials (10%), the composite image consisted of images from the same basic-level category (e.g., house/house). Even in these cases, the composite image was defined by eccentricity with no attempt to introduce a meaningful layout in the resulting scene. Therefore, some conflict was always present in the stimuli, but the congruent stimuli required the same response for both parts.

The relative contribution of signal enhancement and distractor suppression when people are asked to focus on a part of a composite scene remains an open question. Selective attention is traditionally associated with signal enhancement, but the suppression of salient distractors is an important factor in visual search tasks (Gaspar & McDonald, 2014; Gaspelin & Luck, 2018). Our data cannot distinguish between these accounts.

In conclusion, the present study suggests that scene recognition is based on both central and peripheral information, even when the participants are instructed to focus only on one part of the image and ignore the other part. Contrary to the zoom-out hypothesis, we propose that this process should be interpreted in terms of the evidence accumulation model, which posits that information from the to-be-ignored parts is also included.

Keywords: scene perception, gist, peripheral vision, attention, evidence accumulation

Acknowledgments

The work was supported by Czech Science Foundation grant (GA16-07983S) and RVO68081740; this study is part of the research program of the Czech Academy of Sciences Strategy AV21. The author would like to thank Andrea Dally for conducting the pilot experiments and Filip Děchtěrenko for his helpful comments on earlier versions of the manuscript.

Commercial relationships: none.
Corresponding author: Jiří Lukavský.
Email: jirilukavsky@gmail.com.
Address: Institute of Psychology, Czech Academy of Sciences, Prague, Czech Republic.

Footnote

[1] To validate the assumption of the critical radius in our stimuli and setup, we ran a control study ($N = 15$) in which we presented stimuli from Experiment 1 in isolation (with no distractor scene) and with stimulus onset asynchrony (SOA) of 400 ms. The performance was high, but in most cases below the maximum of 100% (central scenes: M accuracy = 96.5%, $SD = 1.9$, range = 92%–100%; peripheral scenes: M accuracy = 97.3%, $SD = 1.9$, range = 94%–100%). The difference of 0.9% (95% CI [–0.09; 1.82]) was not significant, $t(14) = -1.94$, $p = 0.072$. The Bayes factor was not decisive for the null hypothesis ($BF_{10} = 1.16$), but the test of equivalence (Lakens, 2017) showed that both conditions could be considered equivalent within the bounds of ±2%, $t(14) = 2.541$, $p = 0.012$.

References

Bacon-Macé, N., Macé, M. J.-M., Fabre-Thorpe, M., & Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision Research*, *45*(11), 1459–1469, https://doi.org/10.1016/j.visres.2005.01.004.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436, https://doi.org/10.1163/156856897X00357.

Cohen, M. A., Alvarez, G. A., & Nakayama, K. (2011). Natural-scene perception requires attention. *Psychological Science*, *22*(9), 1165–1172, https://doi.org/10.1177/0956797611419168.

Cornelissen, F. W., Peters, E. M., & Palmer, J. (2002). The Eyelink Toolbox: Eye tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 613–617, https://doi.org/10.3758/BF03195489.

Davenport, J. L. (2007). Consistency effects between objects in scenes. *Memory & Cognition*, *35*(3), 393–401, https://doi.org/10.3758/BF03193280.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15*(8), 559–564, https://doi.org/10.1111/j.0956-7976.2004.00719.x.

Eriksen, C. W., & St. James, J. D. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics*, *40*(4), 225–240, https://doi.org/10.3758/BF03211502.

Evans, K. K., & Treisman, A. (2005). Perception of objects in natural scenes: Is it really attention free? *Journal of Experimental Psychology: Human Perception and Performance*, *31*(6), 1476–1492, https://doi.org/10.1037/0096-1523.31.6.1476.

Fei-Fei, L., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the*

National Academy of Sciences, *99*(14), 9596–9601, https://doi.org/10.1073/pnas.092277599.

Fei-Fei, L., VanRullen, R., Koch, C., & Perona, P. (2005). Why does natural scene categorization require little attention? Exploring attentional requirements for natural and synthetic stimuli. *Visual Cognition*, *12*(6), 893–924, https://doi.org/10.1080/13506280444000571.

Florack, L. (2007). Modeling foveal vision. In *Scale space and variational methods in computer vision* (pp. 919–928). Berlin, Germany: Springer. https://doi.org/10.1007/978-3-540-72823-8_79

Franconeri, S. L., Alvarez, G. A., & Enns, J. T. (2007). How many locations can be selected at once? *Journal of Experimental Psychology: Human Perception and Performance*, *33*(5), 1003–1012, https://doi.org/10.1037/0096-1523.33.5.1003.

Freeman, T. E., Loschky, L. C., & Hansen, B. C. (2015). Scene masking is affected by trial blank-screen luminance. *Signal Processing: Image Communication*, *39*, 319–327, https://doi.org/10.1016/j.image.2015.04.004.

Gaspar, J. M., & McDonald, J. J. (2014). Suppression of salient objects prevents distraction in visual search. *Journal of Neuroscience*, *34*(16), 5658–5666, https://doi.org/10.1523/JNEUROSCI.4161-13.2014.

Gaspelin, N., & Luck, S. J. (2018). The role of inhibition in avoiding distraction by salient stimuli. *Trends in Cognitive Sciences*, *22*(1), 79–92, https://doi.org/10.1016/j.tics.2017.11.001.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*(1), 535–574, https://doi.org/10.1146/annurev.neuro.29.051605.113038.

Greene, M. R., & Fei-Fei, L. (2014). Visual categorization is automatic and obligatory: Evidence from Stroop-like paradigm. *Journal of Vision*, *14*(1):14, 1–11, https://doi.org/10.1167/14.1.14. [PubMed] [Article]

Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, *20*(4), 464–472, https://doi.org/10.1111/j.1467-9280.2009.02316.x.

Gronau, N., Neta, M., & Bar, M. (2008). Integrated contextual representation for objects' identities and their locations. *Journal of Cognitive Neuroscience*, *20*(3), 371–388, https://doi.org/10.1162/jocn.2008.20027.

Henderson, J. M. (1992). Visual attention and eye movement control during reading and picture viewing. In *Eye movements and visual cognition* (pp. 260–283). New York, NY: Springer. https://doi.org/10.1007/978-1-4612-2852-3_15

Kleiner, M., Brainard, D. H., & Pelli, D. G. (2007). What's new in Psychtoolbox-3? *Perception 36 ECVP Abstract Supplement*.

Kramer, A. F., & Hahn, S. (1995). Splitting the beam: Distribution of attention over noncontiguous regions of the visual field. *Psychological Science*, *6*(6), 381–386, https://doi.org/10.1111/j.1467-9280.1995.tb00530.x.

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *1*, 1–8, https://doi.org/10.1177/1948550617697177.

Larson, A. M., Freeman, T. E., Ringer, R. V., & Loschky, L. C. (2014). The spatiotemporal dynamics of scene gist recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 471–487, https://doi.org/10.1037/a0034986.

Larson, A. M., & Loschky, L. C. (2009). The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision*, *9*(10):6, 1–16, https://doi.org/10.1167/9.10.6. [PubMed] [Article]

Loschky, L. C., & Larson, A. M. (2010). The natural/man-made distinction is made before basic-level distinctions in scene gist processing. *Visual Cognition*, *18*(4), 513–536, https://doi.org/10.1080/13506280902937606.

Loschky, L. C., Sethi, A., Simons, D. J., Pydimarri, T. N., Ochs, D., & Corbeille, J. L. (2007). The importance of information localization in scene gist recognition. *Journal of Experimental Psychology: Human Perception & Performance*, *33*(6), 1431–1450, https://doi.org/10.1037/0096-1523.33.6.1431.

Mack, A., & Clarke, J. (2012). Gist perception requires attention. *Visual Cognition*, *20*(3), 300–327, https://doi.org/10.1080/13506285.2012.666578.

Mudrik, L., Lamy, D., & Deouell, L. Y. (2010). ERP evidence for context congruity effects during simultaneous object–scene processing. *Neuropsychologia*, *48*(2), 507–517, https://doi.org/10.1016/j.neuropsychologia.2009.10.011.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*(3), 353–383, https://doi.org/10.1016/0010-0285(77)90012-3.

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, *34*(1), 72–107, https://doi.org/10.1006/cogp.1997.0667.

Otsuka, S., & Kawaguchi, J. (2007). Natural scene categorization with minimal attention: Evidence from negative priming. *Perception & Psychophysics*, *69*(7), 1126–1139, https://doi.org/10.3758/BF03193950.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442, https://doi.org/10.1163/156856897X00366.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–70, https://doi.org/10.1023/A:1026553619983.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(5), 509–522, https://doi.org/10.1037/0278-7393.2.5.509.

Rémy, F., Vayssière, N., Pins, D., Boucart, M., & Fabre-Thorpe, M. (2014). Incongruent object/context relationships in visual scenes: Where are they processed in the brain? *Brain and Cognition*, *84*(1), 34–43, https://doi.org/10.1016/j.bandc.2013.10.008.

Rosinski, R. R., Golinkoff, R. M., & Kukish, K. S. (1975). Automatic semantic processing in a picture-word interference task. *Child Development*, *46*(1), 247–253, https://doi.org/10.2307/1128859.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience*, *5*(7), 629–630, https://doi.org/10.1038/nn866.

Rousselet, G. A., Joubert, O., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition*, *12*(6), 852–877, https://doi.org/10.1080/13506280444000553.

Rousselet, G. A., Thorpe, S. J., & Fabre-Thorpe, M. (2004). Processing of one, two or four natural scenes in humans: The limits of parallelism. *Vision Research*, *44*(9), 877–894, https://doi.org/10.1016/j.visres.2003.11.014.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, *5*(4), 195–200, https://doi.org/10.1111/j.1467-9280.1994.tb00500.x.

Starreveld, P. A., & Heij, W. L. (2017). Picture-word interference is a Stroop effect: A theoretical analysis and new empirical findings. *Psychonomic Bulletin & Review*, *24*(3), 721–733, https://doi.org/10.3758/s13423-016-1167-6.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology: General*, *18*(6), 643–662, https://doi.org/10.1037/h0054651.

Van Essen, D. C., Newsome, W. T., & Maunsell, J. H. R. (1984). The visual field representation in striate cortex of the macaque monkey: Asymmetries, anisotropies, and individual variability. *Vision Research*, *24*(5), 429–448, https://doi.org/10.1016/0042-6989(84)90041-5.

Walker, S., Stafford, P., & Davis, G. (2008). Ultra-rapid categorization requires visual attention: Scenes with multiple foreground objects. *Journal of Vision*, *8*(4):21, 1–12, https://doi.org/10.1167/8.4.21. [PubMed] [Article]