The contents of the search template for category-level search in natural scenes

Reshanne R. Reeder
Center for Mind/Brain Sciences, University of Trento, Rovereto, Italy

Marius V. Peelen
Center for Mind/Brain Sciences, University of Trento, Rovereto, Italy

Visual search involves the matching of visual input to a "search template," an internal representation of task-relevant information. The present study investigated the contents of the search template during visual search for object categories in natural scenes, for which low-level features do not reliably distinguish targets from nontargets. Subjects were cued to detect people or cars in diverse photographs of real-world scenes. On a subset of trials, the cue was followed by task-irrelevant stimuli instead of scenes, directly followed by a dot that subjects were instructed to detect. We hypothesized that stimuli that matched the active search template would capture attention, resulting in faster detection of the dot when presented at the location of a template-matching stimulus. Results revealed that silhouettes of cars and people captured attention irrespective of their orientation (0°, 90°, or 180°). Interestingly, strong capture was observed for silhouettes of category-diagnostic object parts, such as the wheel of a car. Finally, attentional capture was also observed for silhouettes presented at locations that were irrelevant to the search task. Together, these results indicate that search for familiar object categories in real-world scenes is mediated by spatially global search templates that consist of view-invariant shape representations of category-diagnostic object parts.

Introduction

Real-world visual search involves the selection of target objects among complex and diverse nontargets. In daily life, this selection often operates at the category level (e.g., looking out for cars when crossing a road or for pedestrians when driving). Considering the infinite number of ways in which objects can appear in the real world, humans are remarkably good at selecting behaviorally relevant object categories in natural scenes (Li, VanRullen, Koch, & Perona, 2002; Thorpe, Fize, & Marlot, 1996).

It has been hypothesized that the attentional selection of a target object among distractors is achieved through the matching of incoming visual input to a top-down attentional set that guides visual search to items containing task-relevant features (Duncan & Humphreys, 1989; Wolfe, Cave, & Franzel, 1989). There is currently great interest in how visual features activated in the attentional set, or "search template," guide search. Studies of eye movements have found that searchers may mistakenly fixate distractors that share visual features with targets, which increases search times (Castelhano & Heaven, 2010; Castelhano, Pollatsek, & Cave, 2008; Pomplun, 2006). The more visual information provided by a cue, the better the search performance (Hwang, Higgins, & Pomplun, 2009; Schmidt & Zelinsky, 2009; Wolfe, Horowitz, Kenner, Hyle, & Vasan, 2004); therefore, an image cue that matches the target exactly is most effective in guiding search (providing the most visual information relevant to the target), while an image representing the target category or a feature-and-word cue (e.g., "blue car") are both more effective than a word cue alone (e.g., "car"; Castelhano & Heaven, 2010; Malcolm & Henderson, 2009, 2010; Schmidt & Zelinsky, 2009; Vickery, King, & Jiang, 2005). At the same time, a search template that approximates rather than perfectly replicates an image allows for a certain amount of flexibility between the visual features of the cue and the target (Bravo & Farid, 2009; Vickery et al., 2005). Some visual features may be preferentially represented in the search template. A previous study of feature search found that searchers are faster to detect targets based on color than orientation (Hannus, van den Berg, Bekkering, Roerdink, & Cornelissen, 2006), suggesting that some features are naturally weighted higher than others in the visual/attention system, thus biasing the search template.

Citation: Reeder, R. R., & Peelen, M. V. (2013). The contents of the search template for category-level search in natural scenes. *Journal of Vision*, *13*(3):13, 1–13, http://www.journalofvision.org/content/13/3/13, doi:10.1167/13.3.13

doi: 10.1167/13.3.13 Received December 17, 2012; published June 10, 2013 ISSN 1534-7362 © 2013 ARVO

The contents of the search template may also be determined by task demands (Bravo & Farid, 2012; Foulsham & Underwood, 2007; Underwood, Foulsham, van Loon, & Underwood, 2005; Vickery et al., 2005; Wolfe et al., 2004). For example, the search template may represent different visual information for targets at different category levels. Searchers can rely on specific visual features of a single exemplar to find a target in individual-level search, but in the case of category-level search, activated visual features in the template must generalize across exemplars (Bravo & Farid, 2012; Yang & Zelinsky, 2009). Category-level search in real-world scenes adds even more complexity to the equation. An effective search template must consist of features that optimally distinguish targets from nontargets, but features of targets in real-world scenes may vary in perspective, size, and location, among other aspects; furthermore, nontargets in these scenes typically share many low-level features with targets. Currently little is known about the contents of the search template for category-level visual search in real-world scenes.

What might the template consist of for search tasks in which certain physical properties (size, perspective) and low-level features (color, orientation) are unlikely to be informative of the presence of a target category? One possibility is a holistic representation of the target category, one that contains information about both the shape and typical configuration of object parts. Previous research has proposed that such a prototype template may underlie face and body detection (Lewis & Ellis, 2003; Stein, Sterzer, & Peelen, 2012), which would explain the drop in detection performance when faces and bodies are inverted. Alternatively, the template may consist of view-invariant category-diagnostic shape features of intermediate complexity, such as a car's wheel or a person's leg (Evans & Treisman, 2005; Treisman, 2006). Activating various diagnostic features of an object category in parallel (e.g., arms, legs, torso) may be a more effective preparatory strategy than activating a holistic layout of features when the specific target exemplar is unknown prior to detection. In agreement with this idea, a computational study (Ullman, Vidal-Naquet, & Sali, 2002) showed that features of "intermediate complexity," such as a portion of a face that reveals the mouth and nose area, are optimally suited for classifying novel images, presumably because these features are consistent across variable exemplars (also see Yang & Zelinsky, 2009). Complementing these findings, Delorme, Richard, and Fabre-Thorpe (2010) found that the detection of animals in natural scenes is significantly impaired in the absence of diagnostic features (such as limbs), further supporting the idea that such feature information is critical to rapid and accurate object detection.

To better investigate the contents of the search template for naturalistic visual search, we developed a variant of the contingent attentional capture paradigm (Folk, Leber, & Egeth, 2002; Folk, Remington, & Johnston, 1992). Contingent attentional capture refers to the orienting of attention toward task-irrelevant stimuli that contain task-relevant features. For example, when subjects are instructed to attend to red items in a central rapid serial visual presentation stream, the appearance of an irrelevant red item in the periphery captures attention, as indicated by a decrement in central target identification (Folk et al., 2002). A related study (Downing, 2000) showed that an irrelevant object that matches a target held in working memory captures spatial attention, leading to better discrimination of an immediately subsequent probe presented at the same location as the memory-matching object. Combining these approaches, we used a modified dot-probe paradigm (MacLeod, Mathews, & Tata, 1986) to measure the degree to which particular stimuli capture attention when subjects are prepared to detect real-world object categories.

In the current study, subjects were cued to detect people or cars in diverse photographs of natural scenes. Of interest was a subset of trials (25%) in which task-irrelevant stimuli without scene background appeared instead of scenes. Subjects were instructed to ignore these stimuli and simply indicate the location of a subsequently presented dot that could appear on the left or right of a central fixation cross. We hypothesized that stimuli that matched the active search template would capture attention, resulting in faster detection of the probe when presented at the same location as the template-matching stimulus. In five experiments, we systematically varied the properties of the task-irrelevant stimuli to reveal the contents of the search template during visual search for object categories in real-world scenes.

Methods

Subjects

Sixty-six undergraduate and graduate students from the University of Trento (53 women) participated in the experiments for course credit or payment: 11 in Experiment 1, 17 in Experiment 2, 13 in Experiment 3, 14 in Experiment 4, and 11 in Experiment 5. All subjects had normal or corrected-to-normal visual acuity and were between the ages of 18 and 38 years (mean = 23.1 years). Four subjects participated in more than one experiment separated by at least 1 month. The research protocol of all experiments adhered to the tenets of the Declaration of Helsinki.

Stimuli

All stimuli were presented on a 19-inch Dell 1905 FP monitor with a screen resolution of 1280 × 1024 pixels and 60 Hz refresh frequency (Dell Inc., Round Rock, TX). A fixation cross and letter cues appeared centered on the screen. Letter cues were uppercased with 70-point "strong" Times New Roman font. The fixation cross had dimensions of 31 × 31 pixels subtending 0.92° in height and width, and letters had dimensions of 70 × 70 pixels subtending 2.1° in height and width. Stimuli were presented using A Simple Framework (Schwarzbach, 2011), based on the Psychophysics Toolbox for MATLAB (The MathWorks, Inc., Natick, MA).

Natural scene stimuli

Stimuli presented in the search trials (75% of trials) were 864 color photographs of real-world scenes obtained from the LabelMe online database (Russell, Torralba, Murphy, & Freeman, 2008; see Figure 1a for some examples) and were divided into scenes containing cars (216), people (216), both cars and people (216), or neither cars nor people (216). Two scenes appeared on every trial and no scene was repeated within an experiment.

Scenes were scaled to 548 × 411 pixel resolution, subtending a visual angle of 16.07° × 12.1°. In Experiments 1 to 4, scenes were presented at a visual angle of 8.96° from the center of the screen to the center of the image, to the left and right of fixation. In Experiment 5, scenes were presented at a visual angle of 6.95° from the center of the screen to the center of the image, above and below fixation.

Attentional capture stimuli

Stimuli presented in the dot-probe trials (25% of trials) showed aspects of 192 photographs of cars (96) and people (96) without scene background (see Figure 1b). Most images were obtained from free-access online image sources and were chosen to encompass a variety of viewpoints and features (e.g., a crouching child, a man standing, a truck as seen from the side, a sports car as seen from behind). Heads were removed from all images of people to be consistent with previous imaging studies that used headless bodies as reference stimuli to investigate neural correlates of the search template for the same real-world search task as in the current experiment (Peelen, Fei-Fei, & Kastner, 2009; Peelen & Kastner, 2011; Seidl, Peelen, & Kastner, 2012). None of the stimuli presented in the dot-probe trials were shown in the scenes presented in the search trials.

Upright silhouettes of the person and car photographs were presented in every experiment. Additional transformations of the photographs were presented in Experiments 1 to 4. In Experiment 1, textures were cut from sections of the photographs. For person stimuli, textures consisted mainly of clothing patterns, although skin was used if the original photograph showed large amounts of exposed skin. Car textures were cut from the bodies of cars, which could include streaks, shine, the lines around doors, and the area above the wheel. Color/texture patches were based on the largest surface area occupied by a given color/texture in the original color photograph, with the additional constraint that the patches did not reveal shape features. All textures were adjusted to a constant radius of 150 pixels subtending a visual angle of 4.43°, located 8.52° from the center of the screen. In Experiment 2, silhouettes were rotated 180° to create inverted images. In Experiment 3, silhouettes were rotated clockwise by 90°. In Experiment 4, small diagnostic parts were taken from the upright silhouettes (e.g., an arm, a pair of feet, or the wheel of a car). The size of each part (based on the number of black pixels) never exceeded 25% of the size of the whole silhouette (range: 4.61% to 24.19%, mean = 14.66% for cars and range: 7.68% to 23%, mean = 14.27% for people). The parts were scaled such that they could appear in the same three possible sizes and locations as whole silhouettes during the experiment. In Experiment 5, upright silhouettes remained unchanged. See Figure 1b for examples of each of the stimuli described above.

Stimuli could appear in three possible sizes (100 × 100, 180 × 180, or 200 × 200 pixels, or 2.95° × 2.95°, 5.31° × 5.31°, or 5.90° × 5.90° of visual angle, respectively) and at three different screen locations along the X-axis, subtending 6.46°, 7.99°, or 10.75° of visual angle. Size and location values were chosen randomly on each trial and independently for the left and right stimulus. On each prime trial, a single aspect of a car and person appeared to the left and right of fixation. Aspects of cars and people appeared on the left and right an equal number of times. Each image was repeated once per experiment under two different transformations (e.g., 192 upright silhouettes and 192 rotated silhouettes). Transformations were randomly intermixed within the experiment but were not mixed within a trial, that is, an upright silhouette of a person always appeared with an upright silhouette of a car, and never with a rotated silhouette of a car.

Procedure

All subjects completed one practice block followed by nine blocks of 64 trials each. Each block was made up of two tasks as illustrated in Figure 2. The search task made up 75% of trials in a block to ensure that subjects actively prepared to detect the cued object category. Trials were randomized so subjects did not

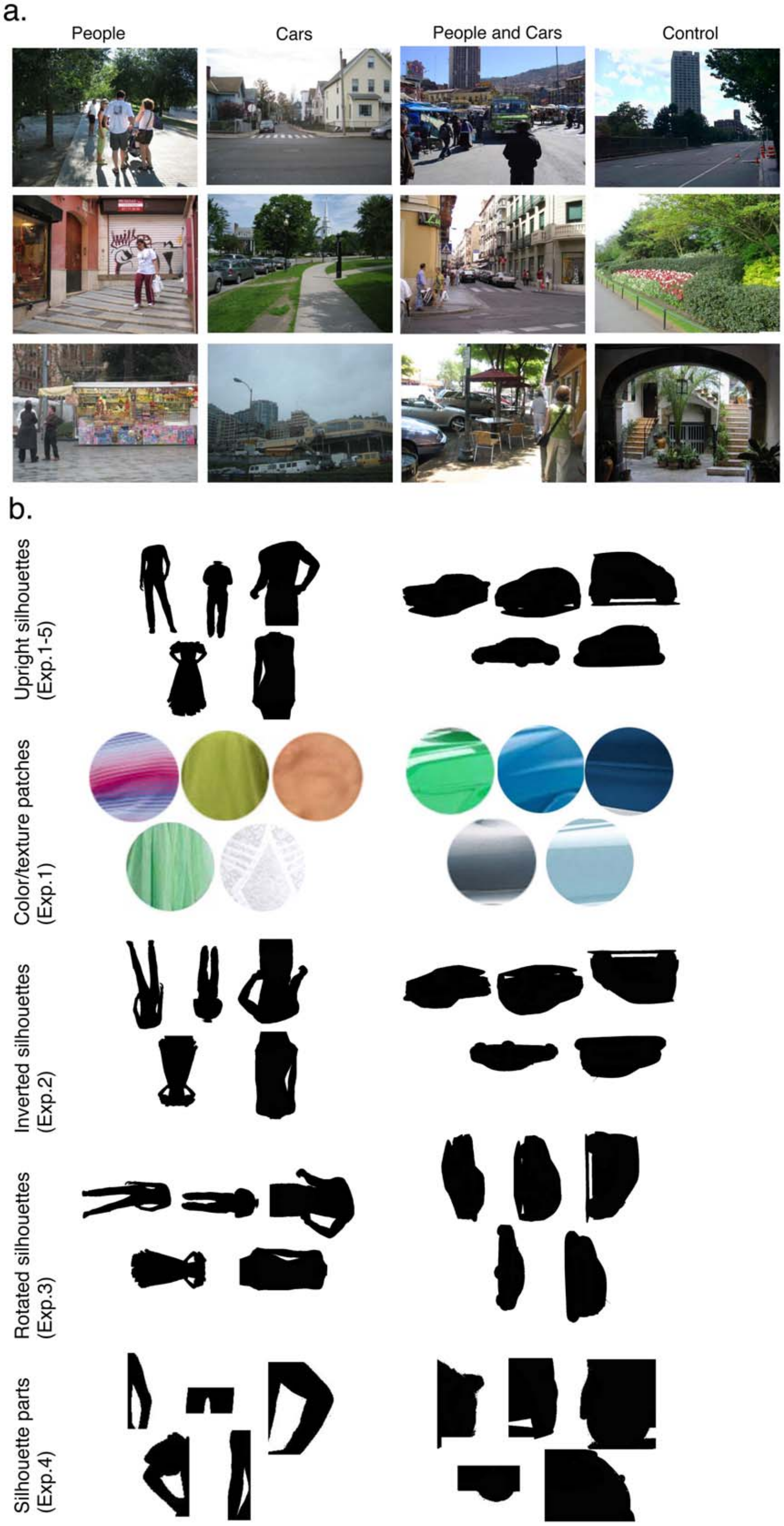
a.
People
Cars
People and Cars
Control
b.
Upright silhouettes (Exp.1-5)
Color/texture patches (Exp.1)
Inverted silhouettes (Exp.2)
Rotated silhouettes (Exp.3)
Silhouette parts (Exp.4)

Figure 1. Examples of stimuli used in (a) the search task and (b) the prime task.

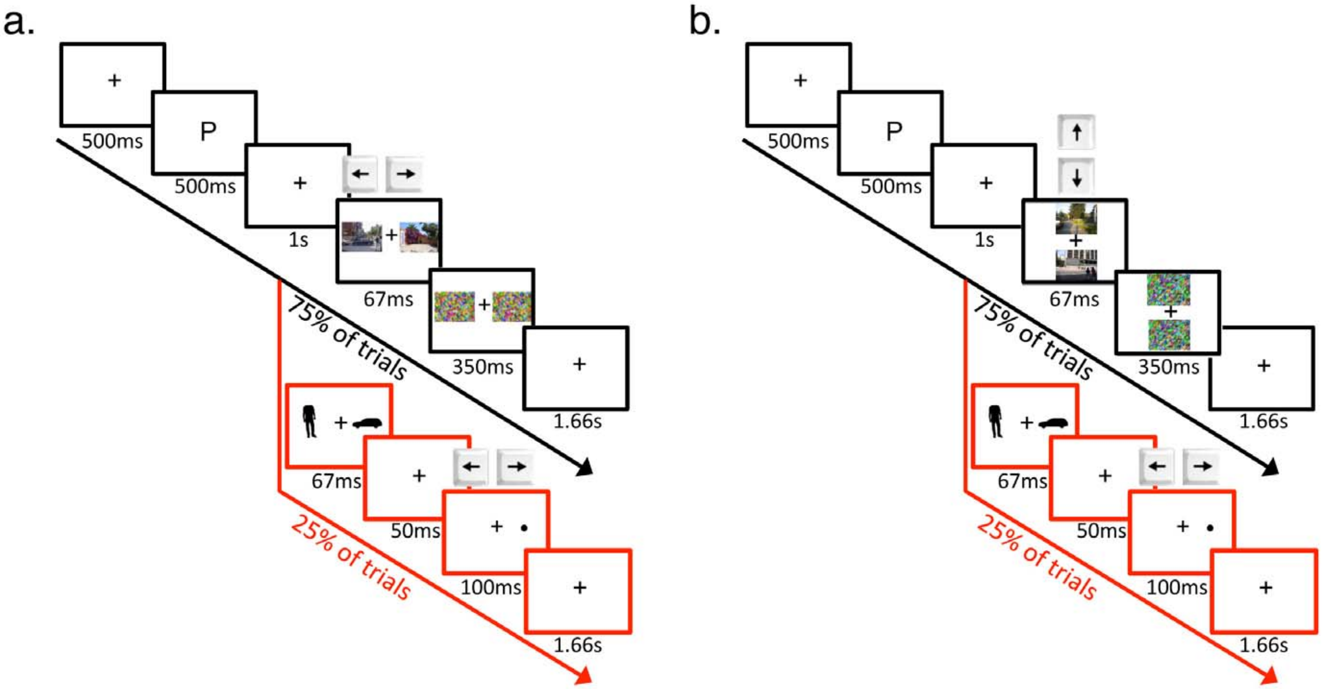

Figure 2. Schematic outline of the experimental paradigm for (a) Experiments 1 to 4 and (b) Experiment 5. In the prime task (25% of trials), subjects were required to respond whether the dot probe appeared on the left or right of fixation. In the search task (75% of trials), subjects were required to respond whether the cued object appeared in the left or right scene (Experiments 1 to 4) or in the top or bottom scene (Experiment 5).

know whether they would perform the search task or the prime task on any given trial.

For both tasks, a trial began with the presentation of a fixation cross for 500 ms, followed by a single letter for 500 ms: "P" for "persona" or "M" for "macchina" (the Italian words for person and car, respectively). After the letter, another fixation cross appeared for 1 s. On search trials, subjects would then see two photographs of real-world scenes for 67 ms, followed by a 350-ms mask. On prime trials, subjects would instead see the primes representing a person and a car for 67 ms, then a fixation cross for 50 ms, followed by a dot probe that would appear for 100 ms, 8.52° from the center of the screen on the left or right. The trial sequence for both tasks ended with a 1.66-s fixation.

For the search task in Experiments 1 to 4, subjects were instructed to respond whether a cued object category (person or car) appeared in the scene on the left or right using the left and right arrow keys, respectively (Figure 2a). In Experiment 5, subjects were instructed to respond whether the cued object category appeared in the scene above or below fixation using the up and down arrow keys, respectively (Figure 2b). The cued object category always appeared in one of the two scenes. The two scenes that appeared could either be one containing cars and the other containing people, or one containing both cars and people and the other containing no cars or people. This structure allowed us to present cars and people on every trial without making the presence of one category informative of the absence of the other category. Each of the four scene types appeared in each possible location an equal number of times (left and right for Experiments 1 to 4, or up and down for Experiment 5).

For the prime task, subjects were instructed to respond using the arrow keys whether the dot probe appeared to the left or right of fixation. Subjects were instructed to ignore the prime images (i.e., the various transformations of silhouettes or color/texture patches) that appeared prior to the probe. Prime images did not predict the probe's location, and the probe appeared on the left and right an equal number of times.

Analysis

For the prime task, we analyzed accuracy and reaction time (RT) for consistent and inconsistent trials. Consistent trials were those in which the cued prime (e.g., the person prime following the "P" cue) appeared on the same side of fixation as the dot probe. Inconsistent trials were those in which the cued prime appeared on the opposite side of fixation. Only correct trials were included in the RT analysis. Subjects were excluded from analysis if their mean prime task accuracy fell 2.5 standard deviations below the group

Experiment	Reaction time (ms)	Accuracy (% correct)
1	613 ± 96	80.4 ± 5.6
2	606 ± 124	82.3 ± 5.4
3	664 ± 127	83.6 ± 6
4	655 ± 123	80.4 ± 8.5
5	737 ± 76	76.4 ± 5

Table 1. Mean RT and accuracy, with standard deviation, in the search task for Experiments 1 to 5.

mean for the experiment. Three subjects were excluded based on this criterion (one each from Experiments 1, 2, and 4).

For the search task, we analyzed accuracy and RT. Only correct trials were included in the RT analysis. Results of the search task are reported in Table 1. The Results section reports the results of the prime task only.

Results

Experiment 1: Upright silhouettes versus color/texture patches

Experiment 1 was conducted to test whether the search template for category-level search consists of object shape and/or surface features (texture and color; see Figure 1b). In the prime trials, subjects had to ignore the prime images (half upright silhouettes and half color/texture patches) and respond whether a dot probe appeared on the left or right.

Reaction time

Figure 3 illustrates mean RT for consistent and inconsistent trials in the upright silhouette and color/texture conditions. A repeated-measures ANOVA with prime type (silhouette, color/texture) and consistency (consistent, inconsistent) as factors revealed a significant interaction, $F(1, 9) = 7.18$, $p = 0.025$, $\eta_p^2 = 0.444$, reflecting a larger consistency effect for the upright silhouette condition than for the color/texture condition. Paired-samples *t*-tests revealed a significant consistency effect for the upright silhouette condition, $t(9) = 3.82$, $p = 0.004$, $d = 0.44$, but not for the color/texture condition, $t(9) = 1.05$, $p = 0.32$, $d = 0.06$. These results indicate that subjects' attention was captured by consistent prime images in the upright silhouette condition only.

Accuracy

A repeated-measures ANOVA with prime type (silhouette, color/texture) and consistency (consistent, inconsistent) as factors did not reveal a significant interaction, $F(1, 9) = 1.55$, $p = 0.244$, $\eta_p^2 = 0.147$. There was a main effect of prime type: subjects performed the prime task with higher accuracy in the color/texture condition (98.6%) than in the upright silhouette condition (96.5%), $F(1, 9) = 12.27$, $p = 0.007$, $\eta_p^2 = 0.577$. A main effect of consistency did not reach significance, $F(1, 9) = 3.61$, $p = 0.09$, $\eta_p^2 = 0.286$.

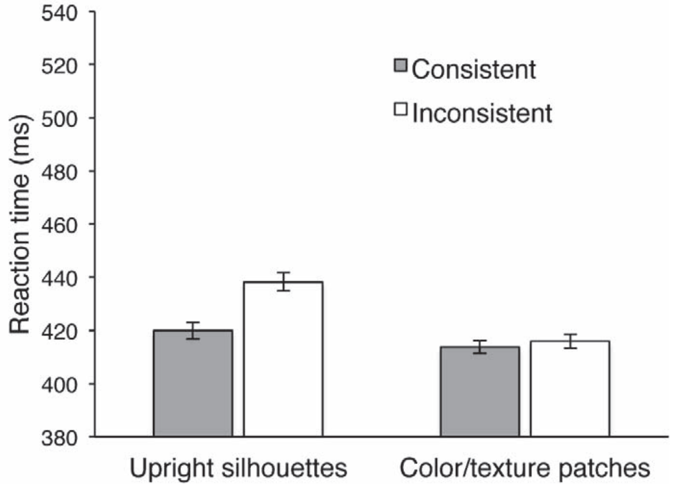

Figure 3. RT for consistent and inconsistent trials in the upright silhouette and color/texture conditions of Experiment 1. Error bars represent the standard error of the mean after adjusting for between-subjects variance (Loftus & Masson, 1994).

Recognition of color/texture patches

To ensure that the color/texture patches we used could be recognized as belonging to either cars or people, a subset of subjects ($n = 8$) performed a color/texture discrimination task after the main experiment. Subjects were shown 96 pairs of color/texture patches that appeared in the prime task, and were required to respond on each trial whether the patch associated with a car ($n = 4$) or person ($n = 4$) appeared on the left or right, using the left and right arrow keys. The position and size of the patches were identical to those used in the prime task. Color/texture patches remained on screen until subjects made a decision, which triggered the onset of the next pair of images. Accuracy on this task ranged from 85.4% to 94.8% with a mean of 90.9%, which was significantly higher than chance (50%), $t(7) = 30.37$, $p = 0.0001$, $d = 1.92$.

Discussion

Results from Experiment 1 indicate that the current paradigm can successfully reveal the attentional capture effect of task-irrelevant stimuli when those stimuli contain features that match the search template.

Subjects experienced a significantly stronger capture effect by upright silhouettes compared to color/texture patches, as indicated by significant RT differences between consistent and inconsistent trials in the silhouette condition only. These results suggest that the search template may be dominated by object shape information rather than color and texture, although it cannot be ruled out that color and texture may be activated in the template when manipulated in other ways.

Thus far, it is unclear whether the shape representations active in the template are restricted to canonical, upright orientations (as used in Experiment 1), or whether they are view invariant such that unexpected or unnatural orientations of objects (e.g., inverted images) may still capture attention. The generalization of object shape to unexpected orientations would suggest that the search template is composed of combinations of local features (e.g., arms, legs) that are not canonically grounded. We compared prime task performance between upright and inverted silhouettes in Experiment 2 to explore this possibility.

Experiment 2: Upright versus inverted (180°) silhouettes

The task in Experiment 2 was the same as in Experiment 1. Half of prime trials showed upright silhouettes and the other half showed silhouettes rotated 180° (Figure 1b).

Reaction time

Figure 4 depicts mean RT for consistent and inconsistent trials in the upright and inverted silhouette conditions. A repeated-measures ANOVA with prime type (upright, inverted) and consistency (consistent, inconsistent) as factors did not reveal a significant interaction, $F(1, 15) = 2.47$, $p = 0.137$, $\eta_p^2 = 0.141$, indicating a comparable consistency effect for upright and inverted silhouette conditions. There was a main effect of consistency, $F(1, 15) = 16.1$, $p = 0.001$, $\eta_p^2 = 0.518$, with faster responses on consistent trials compared to inconsistent trials. There was no main effect of prime type, $F(1, 15) = 1.41$, $p = 0.254$, $\eta_p^2 = 0.086$. These results suggest that attention was captured by consistent prime images regardless of image orientation (upright or inverted).

Accuracy

A repeated-measures ANOVA with prime type (upright, inverted) and consistency (consistent, inconsistent) as factors revealed a significant interaction, $F(1, 15) = 9.5$, $p = 0.008$, $\eta_p^2 = 0.388$, reflecting a larger consistency effect in the upright silhouette condition (98.8% for consistent trials vs. 89.6 % for inconsistent trials) than the inverted silhouette condition (98.3% for consistent trials vs. 92.9% for inconsistent trials). Paired-samples *t*-tests revealed a significant consistency effect for both the upright silhouette condition, $t(15) = 3.9$, $p = 0.001$, $d = 1$, and the inverted silhouette condition, $t(15) = 2.85$, $p = 0.012$, $d = 0.73$.

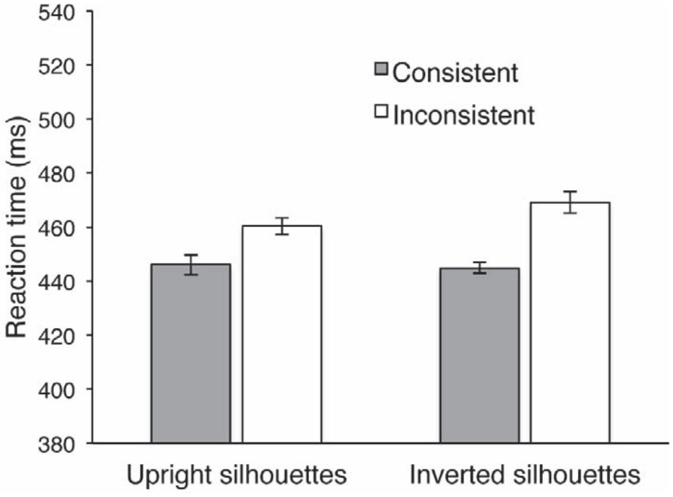

Figure 4. RT for consistent and inconsistent trials in the upright and inverted silhouette conditions of Experiment 2. Error bars represent the standard error of the mean after adjusting for between-subjects variance.

Discussion

Results from Experiment 2 suggest that object shape is part of the active search template regardless of canonical orientation. This may indicate that the template for natural search is, to an extent, composed of view- and orientation-invariant shape features. Alternatively, it is possible that the search template consists mainly of simple orientation features: cars typically have many horizontally oriented features while people typically have many vertically oriented features. These category-related orientation features were largely maintained in the inverted silhouette condition of Experiment 2. Thus, in Experiment 3, we presented 90°-rotated silhouettes on half of prime trials so that people appeared along a horizontal plane and cars appeared along a vertical plane.

Experiment 3: Upright versus rotated (90°) silhouettes

The task in Experiment 3 was the same as in Experiments 1 and 2. Half of prime trials showed upright silhouettes and the other half showed silhouettes rotated clockwise by 90° (Figure 1b).

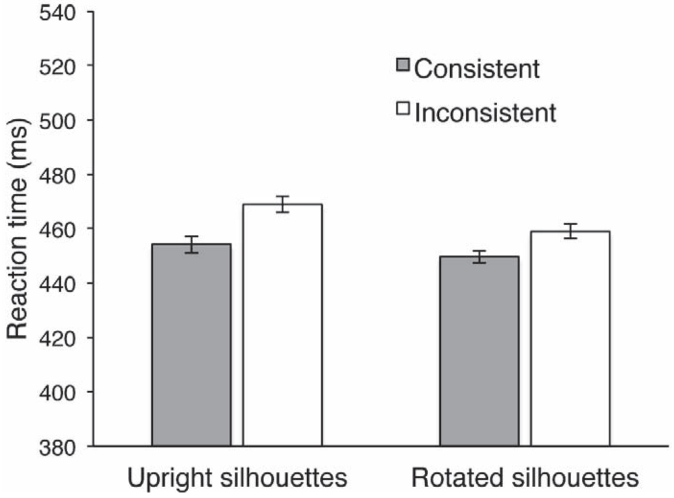

Figure 5. RT for consistent and inconsistent trials in the upright and rotated silhouette conditions of Experiment 3. Error bars represent the standard error of the mean after adjusting for between-subjects variance.

Reaction time

Figure 5 depicts mean RT for consistent and inconsistent trials in the upright and rotated silhouette conditions. A repeated-measures ANOVA with prime type (upright, rotated) and consistency (consistent, inconsistent) as factors did not reveal a significant interaction, $F(1, 12) = 0.53$, $p = 0.482$, $\eta_p^2 = 0.042$, indicating a comparable consistency effect for upright and rotated silhouette conditions. There was a main effect of consistency, $F(1, 12) = 16.18$, $p = 0.002$, $\eta_p^2 = 0.574$, with faster responses on consistent trials than inconsistent trials. Additionally, there was a main effect of prime type, $F(1, 12) = 7.31$, $p = 0.019$, $\eta_p^2 = 0.379$, with faster responses in the rotated silhouette condition compared to the upright silhouette condition.

Accuracy

A repeated-measures ANOVA with prime type (upright, rotated) and consistency (consistent, inconsistent) as factors revealed no significant interaction, $F(1, 12) = 0.0$, $p = 0.995$, $\eta_p^2 = 0.00$. There was a main effect of consistency, $F(1, 12) = 8.5$, $p = 0.013$, $\eta_p^2 = 0.415$, with higher accuracy on consistent trials (98.9%) than inconsistent trials (97.2%). There was no main effect of prime type, $F(1, 12) = 0.0$, $p = 0.997$, $\eta_p^2 = 0.00$.

Discussion

Results of Experiment 3 suggest that the search template is made up of object shape information that is not dependent on simple vertical and horizontal discrimination when searching for people and cars, respectively. This is further evidence that the search template consists of view- and orientation-invariant representations of object shape. However, we have not yet determined whether that shape information is based on the whole object or on object parts. The capture of objects regardless of orientation suggests that searchers activate representations of diagnostic object parts independent of their global layout. We conducted Experiment 4 to directly address the possibility of a part-based template.

Experiment 4: Whole silhouettes versus silhouette parts

The task in Experiment 4 was the same as in Experiments 1 to 3. Half of prime trials showed whole upright silhouettes and the other half showed diagnostic parts of those silhouettes (see Figure 1b). Object parts, on average, were made up of about 15% of the whole silhouettes (see Methods).

Reaction time

Figure 6 depicts mean RT for consistent and inconsistent trials in the whole silhouette and silhouette parts conditions. A repeated-measures ANOVA with prime type (whole silhouettes, silhouette parts) and consistency (consistent, inconsistent) as factors did not reveal a significant interaction, $F(1, 12) = 0.012$, $p = 0.915$, $\eta_p^2 = 0.001$, suggesting a comparable consistency effect for whole silhouettes and silhouette parts. There was a main effect of consistency, with significantly faster responses on consistent trials than on inconsistent trials, $F(1, 12) = 6.39$, $p = 0.027$, $\eta_p^2 = 0.347$. There was no main effect of prime type, $F(1, 12) = 0.485$, $p = 0.499$, $\eta_p^2 = 0.039$.

Accuracy

A repeated-measures ANOVA with prime type (whole silhouettes, silhouette parts) and consistency (consistent, inconsistent) as factors revealed a significant interaction, $F(1, 12) = 5.21$, $p = 0.042$, $\eta_p^2 = 0.303$, with a larger consistency effect in the whole silhouette condition (98.5% for consistent trials vs. 93.6% for inconsistent trials) compared to the silhouette parts condition (98.9% for consistent trials vs. 98.3% for inconsistent trials). Paired-samples t-tests revealed a significant consistency effect for the whole silhouette condition, $t(12) = 2.76$, $p = 0.017$, $d = 0.96$, but not for the silhouette parts condition, $t(12) = 0.74$, $p = 0.47$, $d = 0.35$.

Recognition of object parts

In addition to the main experiment, a subset of subjects ($n = 8$) performed a discrimination task after

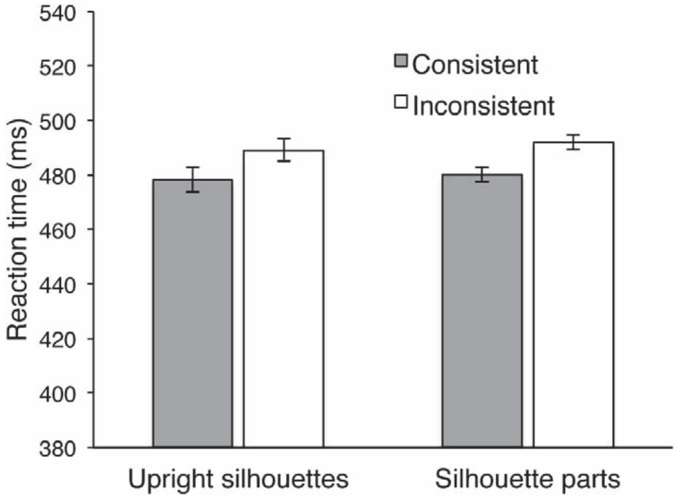

Figure 6. RT for consistent and inconsistent trials in the upright silhouette and silhouette parts conditions of Experiment 4. Error bars represent the standard error of the mean after adjusting for between-subjects variance.

the experiment to ensure that the object parts were recognized as belonging to cars or people. This task was identical to that described for color/texture patches in Experiment 1, except that subjects were shown 96 pairs of object parts instead of color/texture patches. Accuracy on this task ranged from 95.8% to 99% with a mean of 97.1%, which was significantly higher than chance (50%), $t(7) = 109.86$, $p < 0.0001$, $d = 1.94$. These results indicate that the object parts we used in this experiment could be reliably recognized as belonging to people or cars.

Discussion

Results from Experiment 4 suggest that the search template is composed of a flexible layout of diagnostic object parts (e.g., an arm connected to a torso) rather than global shape (e.g., a person standing in the center of the frame), a finding that complements the results of Experiments 2 and 3.

Experiment 5: Differing locations of scenes and primes

Previous research on contingent attentional capture has found evidence that attending to low-level features can result in the capture of attention by these features even if presented at task-irrelevant locations (Folk et al., 2002). We conducted Experiment 5 to examine whether this was similarly the case for the capture effects observed in our paradigm. Such a result would provide evidence that the search template is spatially global in its representation of category-level object parts.

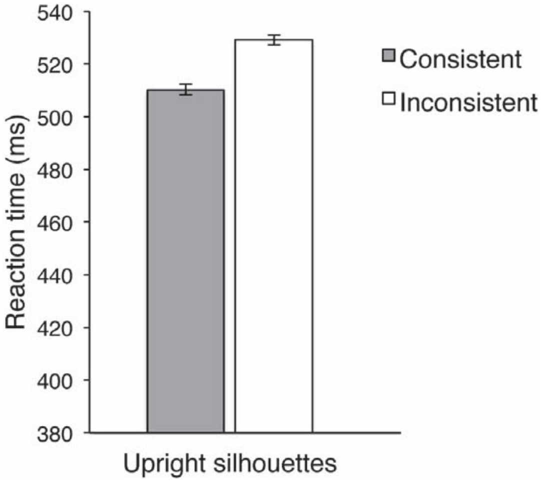

Figure 7. RT for consistent and inconsistent trials in the upright silhouette condition of Experiment 5, in which silhouettes were presented at search-task-irrelevant locations. Error bars represent the standard error of the mean after adjusting for between-subjects variance.

The search task differed from Experiments 1 to 4 in that subjects now had to respond whether a car or person appeared in a scene above or below fixation using the up and down arrow keys. The prime task remained the same as in the previous experiments, in which subjects were required to ignore prime images presented to the left and right of fixation and to respond whether a dot probe appeared on the left or right using the left and right arrow keys. Prime images were all upright silhouettes, with each image repeated once in the experiment. Because only one type of prime was presented, paired samples *t*-tests were conducted for consistent versus inconsistent trials on RT and accuracy results.

Reaction time

Figure 7 depicts mean RT for consistent and inconsistent trials in the prime task. A paired-samples t-test on consistency revealed that subjects responded significantly faster on consistent trials than inconsistent trials, $t(10) = 4.76$, $p = 0.001$, $d = 0.26$.

Accuracy

A paired-samples *t*-test revealed that subjects performed the prime task with similar accuracy on consistent (99.4%) and inconsistent trials (98.6%), $t(10) = 1.58$, $p = 0.145$, $d = 0.62$.

Discussion

The results of Experiment 5 suggest that the template activated for our search task was spatially global, modulating the processing of visual input across the visual field even at locations that were known to be irrelevant for the search task.

Attentional capture for car versus person primes

To investigate whether the reported consistency effects in Experiments 1 to 5 differed for car and person primes, we categorized trials as belonging to "person" or "car" based on the correspondence in the location of the prime category and the subsequent dot probe. For example, trials in which the dot probe followed the car prime were labeled as car trials. Cue consistency was then determined by the category of the preceding cue. For example, an inconsistent car trial would be a trial with a person cue and with the dot probe appearing at the location of the car prime. Using this method, we split capture trials into car and person categories and explored the consistency effects for upright silhouettes combined across Experiment 1 to 5 ($n = 59$; RT was averaged across experiments for subjects who participated in two experiments, $n = 4$). A repeated-measures ANOVA with category (cars, people) and consistency (consistent, inconsistent) as factors revealed, as expected, a highly significant main effect of consistency, $F(1, 58) = 26.4$, $p < 0.001$, $\eta_p^2 = 0.313$. There was also a significant interaction between category and consistency, $F(1, 58) = 6.92$, $p = 0.011$, $\eta_p^2 = 0.107$, with somewhat stronger consistency effects for person than car trials. Importantly, however, the consistency effect was significant for both person, $t(58) = 5.85$, $p < 0.001$, $d = 1.54$, and car, $t(58) = 2.70$, $p = 0.009$, $d = 0.71$, conditions separately, indicating that both categories contributed to the overall consistency effects observed.

General discussion

What do people look for when searching for an object category in a natural scene? We developed a novel attentional capture paradigm to explore this question. On the majority of trials, subjects searched for people or cars in real-world scenes. On a subset of trials, the search cue was followed by task-irrelevant stimuli instead of scenes, directly followed by a dot that subjects were instructed to detect. Attentional capture was defined as the RT difference in detecting a dot presented at the location of the consistent, putatively template-matching stimulus, versus the location of the inconsistent stimulus. In Experiment 1, there was a strong capture effect for upright silhouettes of people and cars, but not for color/texture patches extracted from those same objects. This is evidence that the search template in our study was predominantly composed of object shape. Experiment 2 was conducted to test the necessity for canonical orientation of whole object shape, and we found that objects can be inverted and still elicit a comparable capture effect to upright images, suggesting that the representations activated in the search template are orientation invariant. In Experiment 3, we presented cars and people rotated by 90° to rule out the possibility that searchers may prepare for targets based on low-level orientation features (i.e., preparing for cars and people by looking for horizontally oriented and vertically oriented objects, respectively). Results indicated that cars presented along a vertical plane and people presented along a horizontal plane nevertheless captured attention, providing further evidence for an orientation-invariant search template for object form. Following from these findings, we hypothesized that the search template likely consists of a collection of diagnostic object parts rather than representations of whole objects since this would allow searchers to prepare more flexibly for varied category exemplars in complex scenes. Experiment 4 confirmed this hypothesis, showing significant capture effects by various object parts (e.g., arms, feet, a car tire) that consisted of only about 15% of the pixels of the whole silhouette. Finally, in Experiment 5 we showed that silhouettes capture attention even when they are presented at locations that are irrelevant to the search task, indicating that the search template for this task is spatially global.

Results of Experiment 5 showed that attentional capture occurred even at locations that were fully irrelevant for the search task (i.e., subjects never had to detect objects at these locations). Despite this, we found strong category-specific attentional capture at these locations. This is evidence that the search template spreads to locations outside of where search is performed, suggesting a category-specific attentional bias across the visual field, as previously observed using neuroimaging methods (Peelen et al., 2009). This is reminiscent of the effects observed for feature-based attention (Serences & Boynton, 2007). Indeed, behavioral work has shown that familiar object categories can be detected in the absence of focal attention (Li et al., 2002), similar to low-level features (Treisman & Gelade, 1980). To account for these behavioral results, Treisman (2006) suggested that subjects "may be set to sense, in parallel, a highly overlearned vocabulary of features that characterize a particular semantic category" when searching for familiar object categories in natural scenes. In support of this account, a recent study (Korjoukov et al., 2012) found that subjects were capable of detecting object categories in briefly presented natural scenes, while they were much worse at grouping sets of features into a coherent whole, which suggests that more focused attention is needed to perform perceptual grouping than to detect familiar object categories.

Altogether, the results of the current study suggest that the template for object category search in natural scenes is made up of view-invariant, category-diagnostic shapes of object parts represented globally across the visual field. Such a template likely reflects the ways in which the visual and attentional systems optimally deal with the complexities of search in the real world; natural scenes are cluttered and highly detailed, targets may share many low-level features with nontargets, and features may vary widely even within the target category. The key to accurate detection is forming a search template that is flexible enough to account for versatile targets, but specific enough to eliminate similar nontargets. For example, forming a representation that encompasses all vertically oriented objects may lead to incorrect detection of a lamp post as a person, while forming a holistic representation of a person in a prototypical standing posture is only diagnostic of a subset of people, in which case people shown in other stances may be overlooked. But a template composed of spatially nonspecific shapes of body parts is sufficiently diagnostic of people in many different views and locations, and at the same time eliminates the risk of lower-level (e.g., orientation-based) identification errors. This is consistent with the computational results of Ullman et al. (2002) and behavioral results of Delorme et al. (2010), which both concluded that variable objects are optimally classified by such diagnostic part features.

The current study explored the components of the search template for the detection of people and cars; however, in daily life we also search for objects at more specific or general levels, such as looking for one's own car versus looking for any kind of vehicle, respectively. It would be interesting for future studies to use the current paradigm to directly compare the contents of the search template for these different levels of search. We expect that if the exact identity of the target is known prior to search (e.g., a red scarf), or if a certain feature dimension reliably discriminates the target from distractors (e.g., the color red), then search based on low-level features (e.g., color) may be most efficient for that task (Pomplun, 2006; Vickery et al., 2005; Wolfe et al., 2004). Similarly, for object categories for which surface features are diagnostic of their presence (e.g., trees, for which the color green is a consistent feature across exemplars), such features are likely part of the search template for these categories. For search tasks that are broader than those in the current study (e.g., detecting vehicles rather than cars), a shape-based template may not always be advantageous for target detection, as members of the same superordinate category may share few shape features. In this case, search preparation could be specified at higher levels of the visual processing hierarchy, possibly at the conceptual level (Wyble, Folk, & Potter, 2012). Collectively, these findings provide evidence for a "flexible template," which may change depending on task demands (Bravo & Farid, 2009, 2012). That the template may depend on the specific task demands also highlights the value of using natural scenes because it is hard to mimic the particularities of the real world using artificial search arrays (Wolfe, Alvarez, Rosenholtz, Kuzmova, & Sherman, 2011). It may therefore also be beneficial for future studies to investigate whether the current results extend to active search tasks that involve eye movements (Findlay & Gilchrist, 2003).

Finally, an interesting avenue for future research is to explore individual differences in the contents of the search template, and how these relate to differences in search performance. There are considerable individual differences in the ability to detect object categories in briefly presented real-world scenes (e.g., Peelen & Kastner, 2011), which are stable over time, and are only partly explained by general traits such as intelligence (Huang, Mo, & Li, 2012). These differences have been found to relate to differences in self-reported search strategy, with good searchers reporting to use a more general strategy (e.g., preparing for cars and people at multiple angles) and poor searchers reporting to use a more specific strategy (e.g., preparing for cars and people in prototypical orientations; Peelen & Kastner, 2011). It may be of interest to apply the current paradigm to studies of individual differences in the template, revealing the most effective strategy for a given search task by comparing the templates of good and poor searchers. This avenue may be particularly useful when applied to professions in which search is of high practical relevance, such as radiology, airport security, emergency services, and the military.

General conclusions

Here we presented a novel attentional capture paradigm to explore the contents of the search template for familiar object categories in real-world scenes. The results of the current study indicate that such a template is made up of spatially global, view-invariant shapes of diagnostic object parts. Our paradigm can be adopted to explore the templates for various search tasks and individual strategies. These investigations could be of use to professionals whose search performance and efficiency are of high practical importance.

Keywords: natural scenes, visual search, object category, top-down attention

Acknowledgments

We thank Timo Stein and Wieske van Zoest for helpful comments and discussions. M. V. P. was supported in part by the fondazione Cassa di Risparmio di Trento e Rovereto.

Commercial relationships: none.

Corresponding author: Marius Peelen.
Email: marius.peelen@unitn.it.
Address: Center for Mind/Brain Sciences, University of Trento, Rovereto, Italy.

References

Bravo, M. J., & Farid, H. (2009). The specificity of the search template. *Journal of Vision*, *9*(1):34, 31–39, http://www.journalofvision.org/content/9/1/34, doi:10.1167/9.1.34. [PubMed] [Article]

Bravo, M. J., & Farid, H. (2012). Task demands determine the specificity of the search template. *Attention Perception & Psychophysics, 74*(1), 124–131, doi:10.3758/s13414-011-0224-5.

Castelhano, M. S., & Heaven, C. (2010). The relative contribution of scene context and target features to visual search in scenes. *Attention Perception & Psychophysics, 72*(5), 1283–1297, doi:10.3758/APP.72.5.1283.

Castelhano, M. S., Pollatsek, A., & Cave, K. R. (2008). Typicality aids search for an unspecified target, but only in identification and not in attentional guidance. *Psychonomic Bulletin & Review*, *15*(4), 795–801, doi:10.3758/PBR.15.4.795.

Delorme, A., Richard, G., & Fabre-Thorpe, M. (2010). Key visual features for rapid categorization of animals in natural scenes. *Frontiers in Psychology, 1*, 21, doi:10.3389/fpsyg.2010.00021.

Downing, P. E. (2000). Interactions between visual working memory and selective attention. *Psychological Science, 11*(6), 467–473.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96*(3), 433–458.

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.

Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing*. Oxford: Oxford University Press.

Folk, C. L., Leber, A. B., & Egeth, H. E. (2002). Made you blink! Contingent attentional capture produces a spatial blink. *Perception & Psychophysics, 64*(5), 741–753.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology. Human Perception and Performance, 18*(4), 1030–1044.

Foulsham, T., & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual salience in scene perception? *Perception*, *36*(8), 1123–1138.

Hannus, A., van den Berg, R., Bekkering, H., Roerdink, J. B. T. M., & Cornelissen, F. W. (2006). Visual search near threshold: Some features are more equal than others. *Journal of Vision*, *6*(4):15, 523–540, http://www.journalofvision.org/6/4/15, doi:10.1167/6.4.15. [PubMed] [Article]

Huang, L., Mo, L., & Li, Y. (2012). Measuring the interrelations among multiple paradigms of visual attention: An individual differences approach. *Journal of Experimental Psychology. Human Perception and Performance, 38*(2), 414–428.

Hwang, A. D., Higgins, E. C., & Pomplun, M. (2009). A model of top-down attentional control during visual search in complex scenes. *Journal of Vision*, *9*(5):25, 1–18, http://www.journalofvision.org/content/9/5/25, doi:10.1167/9.5.25. [PubMed] [Article]

Korjoukov, I., Jeurissen, D., Kloosterman, N. A., Verhoeven, J. E., Scholte, H. S., & Roelfsema, P. R. (2012). The time course of perceptual grouping in natural scenes. *Psychological Science, 23*(12), 1482–1489, doi:10.1177/0956797612443832.

Lewis, M. B., & Ellis, H. D. (2003). How we detect a face: A survey of psychological evidence. *International Journal of Imaging Systems and Technology*, *13*(1), 3–7, doi:10.1002/Ima.10040.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences USA, 99*(14), 9596–9601.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence-intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476–490.

MacLeod, C., Mathews, A., & Tata, P. (1986). Attentional bias in emotional disorders. *Journal of Abnormal Psychology, 95*(1), 15–20.

Malcolm, G. L., & Henderson, J. M. (2010). Combining top-down processes to guide eye movements during real-world scene search. *Journal of Vision, 10*(2):*4*, 1–11, http://www.journalofvision.org/

content/10/2/4, doi:10.1167/10.2.4. [PubMed] [Article]

Malcolm, G. L., & Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes: evidence from eye movements. *Journal of Vision, 9*(11):8, 1–13, http://www.journalofvision.org/content/9/11/8, doi:10.1167/9.11.8. [PubMed] [Article]

Peelen, M. V., Fei-Fei, L., & Kastner, S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature*, *460*(7251), 94–97, doi:10.1038/nature08103.

Peelen, M. V., & Kastner, S. (2011). A neural basis for real-world visual search in human occipitotemporal cortex. *Proceedings of the National Academy of Sciences USA, 108*(29), 12125–12130, doi:10.1073/pnas.1101042108.

Pomplun, M. (2006). Saccadic selectivity in complex visual search displays. *Vision Research, 46*(12), 1886–1900, doi:10.1017/j.visres.2005.12.003.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision, 77*(1–3), 157–173, doi:10.1007/S11263-007-0090-8.

Schmidt, J., & Zelinsky, G. J. (2009). Search guidance is proportional to the categorical specificity of a target cue. *The Quarterly Journal of Experimental Psychology, 62*(10), 1904–1914, doi:10.1080/17470210902853530.

Schwarzbach, J. (2011). A simple framework (ASF) for behavioral and neuroimaging experiments based on the psychophysics toolbox for MATLAB. *Behavior Research Methods, 43*(4), 1194–1201, doi:10.3758/s13428-011-0106-8.

Seidl, K. N., Peelen, M. V., & Kastner, S. (2012). Neural evidence for distractor suppression during visual search in real-world scenes. *Journal of Neuroscience, 32*(34), 11812–11819, doi:10.1523/JNEUROSCI.1693-12.2012.

Serences, J. T., & Boynton, G. M. (2007). Feature-based attentional modulations in the absence of direct visual stimulation. *Neuron*, *55*(2), 301–312, doi:10.1016/j.neuron.2007.06.015.

Stein, T., Sterzer, P., & Peelen, M. V. (2012). Privileged detection of conspecifics: Evidence from inversion effects during continuous flash suppression. *Cognition*, *125*(1), 64–79, doi:10.1016/j.cognition.2012.06.005.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522, doi:10.1038/381520a0.

Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition, 14*(4-8), 411–443, doi:10.1080/13506280500195250.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136, doi:0010-0285(80)90005-5.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5*(7), 682–687, doi:10.1038/nn870.

Underwood, G., Foulsham, T., van Loon, E., & Underwood, J. (2005). Visual attention, visual saliency, and eye movements during the inspection of natural scenes. In J. Mira & J. R. Álvarez (Eds.), *Artificial intelligence and knowledge engineering applications: A bioinspired approach* (pp. 459–468). Berlin Heidelberg: Springer-Verlag.

Vickery, T. J., King, L. W., & Jiang, Y. (2005). Setting up the target template in visual search. *Journal of Vision, 5*(1):8, 81–92, http://www.journalofvision.org/content/5/1/8, doi:10.1167/5.1.8. [PubMed] [Article]

Wolfe, J. M., Alvarez, G. A., Rosenholtz, R., Kuzmova, Y. I., & Sherman, A. M. (2011). Visual search for arbitrary objects in real scenes. *Attention, Perception, & Psychophysics, 73*(6), 1650–1671, doi:10.3758/s13414-011-0153-3.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology. Human Perception and Performance, 15*(3), 419–433.

Wolfe, J. M., Horowitz, T. S., Kenner, N., Hyle, M., & Vasan, N. (2004). How fast can you change your mind? The speed of top-down guidance in visual search. *Vision Research, 44*(12), 1411–1426, doi:10.1016/j.visres.2003.11.024.

Wyble, B., Folk, C., & Potter, M. C. (2012). Contingent attentional capture by conceptually relevant images. *Journal of Experimental Psychology. Human Perception and Performance,* doi:10.1037/a0030517.

Yang, H., & Zelinsky, G. J. (2009). Visual search is guided to categorically-defined targets. *Vision Research, 49*(16), 2095–2103, doi:10.1016/j.visres.2009.05.017.