Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com

# Diagnostic parts are not exclusive in the search template for real-world object categories

Marcel Wurth [a], Reshanne R. Reeder [a, b, *]

[a] Department of Experimental Psychology, Otto-von-Guericke University, Magdeburg, Germany
[b] Center for Behavioral Brain Sciences, Magdeburg, Germany

ARTICLE INFO

*Keywords:*
Visual attention
Visual working memory
Visual representation

ABSTRACT

Visual search can be aided by a search template: a preparatory representation of relevant target features. But which features are relevant in complex, real-world category search? Previous research suggests that this template must be flexible to account for variations in naturalistic stimulus properties such as size and occlusion, and that shapes of diagnostic parts of objects are a likely candidate. Here, in three experiments, we systematically evaluated the contribution of diagnostic object parts and whole object shape to the category-level search template. Our hypothesis was that features that better match the active search template will capture attention during search more strongly than partially-matching features. Results showed that while whole objects captured attention reliably and globally across the visual field, diagnostic parts failed to do so in all three experiments. This suggests that whole object shape is a necessary component of the category-level search template.

## 1. Introduction

Classical visual search theories predict search will be easy when a target is identifiable by one simple feature that is easily distinguishable from distractors, and difficult when a target is composed of multiple features shared by distractors (Duncan & Humphreys, 1989; Treisman & Gelade, 1980; Wolfe, 1994). But imagine a search in which the target is poorly defined and always changing, identifiable by no single feature, surrounded by many similar distractors, and presented in a cluttered, complex environment. You might think that finding your target in this context would be impossible, but such search tasks are solved easily everyday by billions of people, including toddlers (Hasegawa & Miyashita, 2002), often in a matter of milliseconds and with remarkably low error rates (Thorpe, Fize, & Marlot, 1996; Zelinsky, 2008). Such is the nature of real-world search.

There is a reason why we succeed in real-world search on a regular basis: targets, distractors, and context are all highly trained over a lifetime of experience (Peelen & Kastner, 2014). Nevertheless, the information-richness of an everyday scene is enormous (Malcolm, Groen, & Baker, 2016), while the capacity of the human brain to process this is limited (Desimone & Duncan, 1995). One mechanism that can be used to manage the challenges of visual search is the preparatory search template: a top-down "attentional set" containing information that is relevant to the current search (Duncan & Humphreys, 1989). Activating a template enhances the representation of relevant features in preparation for search (Reeder, Hanke, & Pollmann, 2017), which ultimately aids target detection during the search process (Malcolm & Henderson, 2009, 2010).

So what are the necessary contents of a real-world template? Low-level features are not helpful when looking for a member of a broad category – for example, a human body – in an ever-changing, natural environment. In line with this, Reeder and Peelen (2013: Experiment 1) found that colors and textures are not part of the search template for category search. In such cases, the template must be flexible to include all relevant members of the target category (see Bravo & Farid, 2012). Previous experimental research indicates that object shape may be an important feature (Reeder & Peelen, 2013; Reeder, van Zoest, & Peelen, 2015). However, because natural objects can appear under varied circumstances, the shapes in the template cannot be rigid (e.g., not view- or orientation-specific). There is evidence that real-world object parts presented in spatially scrambled configurations (e.g., an eye

[*] Corresponding author at: Otto-von-Guericke-Universität, Institut für Psychologie, 39106 Magdeburg, Germany.
*Email address:* reshanne.reeder@gmail.com (R.R. Reeder)

https://doi.org/10.1016/j.actpsy.2019.03.006
Received 26 November 2018; Received in revised form 13 February 2019; Accepted 18 March 2019
Available online xxx
0001-6918/ © 2019.

next to a mouth) are processed earlier in the visual hierarchy than spatially intact configurations (Lerner, Hendler, Ben-Bashat, Harel, & Malach, 2001), which suggests that these parts are processed like simple features (Treisman & Gelade, 1980). This led to the hypothesis that a spatially flexible collection of parts dominates the template for real-world category search (see Reeder & Peelen, 2013: Experiment 4). However, an object "part" can be any fragment of the image (e.g., one half of the original image, a single pixel). The parts that are stored in an effective template must still contain category-diagnostic information – so what classifies a part as "diagnostic"?

Using computer simulations, Ullman, Vidal-Naquet, and Sali (2002) found that object features of "intermediate complexity" (median = 11%, standard deviation = 16% of the original image) were optimal for classification compared to simpler or more specific features. Fragments of images could be classified as belonging to a car or a human face with fewer training images if they contained parts that are commonly shared by the whole class of object (e.g., the eyes of a face), whereas large image fragments (e.g., the eyes, nose, and mouth) are highly specific to the individual image and cannot be used efficiently to classify other objects from the same class, and simpler image fragments (e.g., a dimple) could potentially have features in common with objects from the other class. In this case, as in the current study, "object parts" are not defined by low-level Gestalt principles but rather by the ability to classify image fragments as belonging to one basic category or the other. Applying this to human research, it has been found that removing such diagnostic parts as the eyes, mouth, or limbs from an animal image makes it significantly more difficult to categorize (Delorme, Richard, & Fabre-Thorpe, 2010).

It is clear from these studies that diagnostic parts are a necessary component of a flexible, real-world, category-level search template. Nevertheless, this does not mean that the template is optimally composed of an exclusive collection of parts. There is evidence from monkey single-cell recordings that some neurons in IT (a region that processes object form) respond selectively to various views of whole objects, and there is behavioral evidence that humans naturally learn to represent a global shape of animate and inanimate object categories that is resistant to variations in viewpoint or changes to local parts (see Logothetis & Sheinberg, 1996). Within the computational literature, Chen et al. (2014) argued that the most accurate and flexible categorical representations should contain various combinations of the "root" (the holistic representation) and its parts. Stemming from this, we hypothesize that the most accurate and flexible search template will contain information about both whole and parts. This, however, has not yet been tested experimentally.

In our previous studies (Reeder et al., 2015; Reeder & Peelen, 2013) we used a novel design that combined visual search and "contingent attention capture" (Folk, Remington, & Johnston, 1992) that allowed us to investigate different possible features of the naturalistic search template without changing the search task (which would invariably change the search template). Specifically, subjects were cued to search for object categories (cars, people) on every trial. On the majority of trials, the cue was followed by a search task – however, this task was only included to ensure that subjects activated a category-level search template following the cue. The critical condition occurred on a minority of trials intermixed with the search task – in which subjects were required to indicate whether a briefly presented dot probe appeared on the left or right of fixation. The dot was always preceded by features of search targets (e.g., textures, shapes), but subjects were instructed to ignore these, and to respond as quickly and accurately as possible to the location of the dot. We hypothesized that viewed features presented on dot-probe trials that matched the active template would capture attention involuntarily (as indicated by faster responses to a subsequent dot-probe on the same side of fixation as template-matching features), even if they were task-irrelevant and presented in search-irrelevant locations (Seidl-Rathkopf, Turk-Browne, & Kastner, 2015; Wyble, Folk, & Potter, 2013). In other words, any attention capture by such irrelevant items must be due to their matching the template rather than being used explicitly to improve search performance.

In the current study, we present three experiments that test the efficacy of parts versus wholes in capturing attention during real-world category search (see Figs. 1 and 2). Experiment 1 is a direct replication of Experiment 4 of Reeder and Peelen (2013), only with a larger sample size (25 compared to 13). We hypothesized that higher power associated with a larger *N* (as calculated by a power analysis) could potentially bring out small differences in efficacy between parts and wholes to capture attention. In Experiment 2, we presented capture stimuli in search-irrelevant locations to determine whether capture differences between parts and wholes could be due to subjects activating an inflexible, spatially rigid search template. Finally, Experiment 3 was conducted to determine whether capture differences could be attributed to some parts being less diagnostic than others; to control for this, we presented collections of four object parts and compared them to whole objects as capture stimuli. In all experiments, we found capture effects for wholes but not parts, suggesting parts alone are not a sufficient search template for real-world object categories.

## 2. Materials and methods

### *2.1. Subjects*

90 students and faculty of Otto-von-Guericke University, Magdeburg, were recruited for this study (29 for Experiment 1, 30 for Experiment 2, and 31 for Experiment 3). Nine subjects took part in more than one experiment. All subjects had normal or corrected-to-normal vision, received psychology credits or money as reimbursement for their participation, and provided written, informed consent prior to experimentation. These measures conformed to the Declaration of Helsinki and were approved by the research ethics committee of Otto-von-Guericke University.

Outlier exclusion was strict to control for possible confounds in subject performance. Subjects were excluded if their mean visual search accuracy was below 75%, as in the previous study. This was to ensure that subjects could activate an appropriate category-level search template. Subjects were also excluded if their mean dot-probe detection accuracy was below 90%. In the dot-probe task (see Section 2.3 Procedure), subjects simply had to respond on which side of fixation a dot appeared, and accuracy lower than 90% would suggest a lack of attention or inability to understand the task rather than normal human error. Furthermore, high dot-probe accuracy was necessary for sufficient power to analyze reaction times (RT; only reported for accurate trials) due to a low number of dot-probe trials per run (16, as in the previous study). Following these exclusions, subjects were lastly excluded if their mean dot-probe RT was slower than two standard deviations from the group mean. This was done to ensure that subjects did not consciously deliberate about their responses based on the capture stimuli that appeared prior to the dot.

Because of the strict exclusion criteria and to ensure an adequate number of subjects per experiment, we continued to run subjects in each experiment until a final number of 25 suitable subjects per experiment was reached. We determined that 25 subjects would provide adequate power in each experiment following an

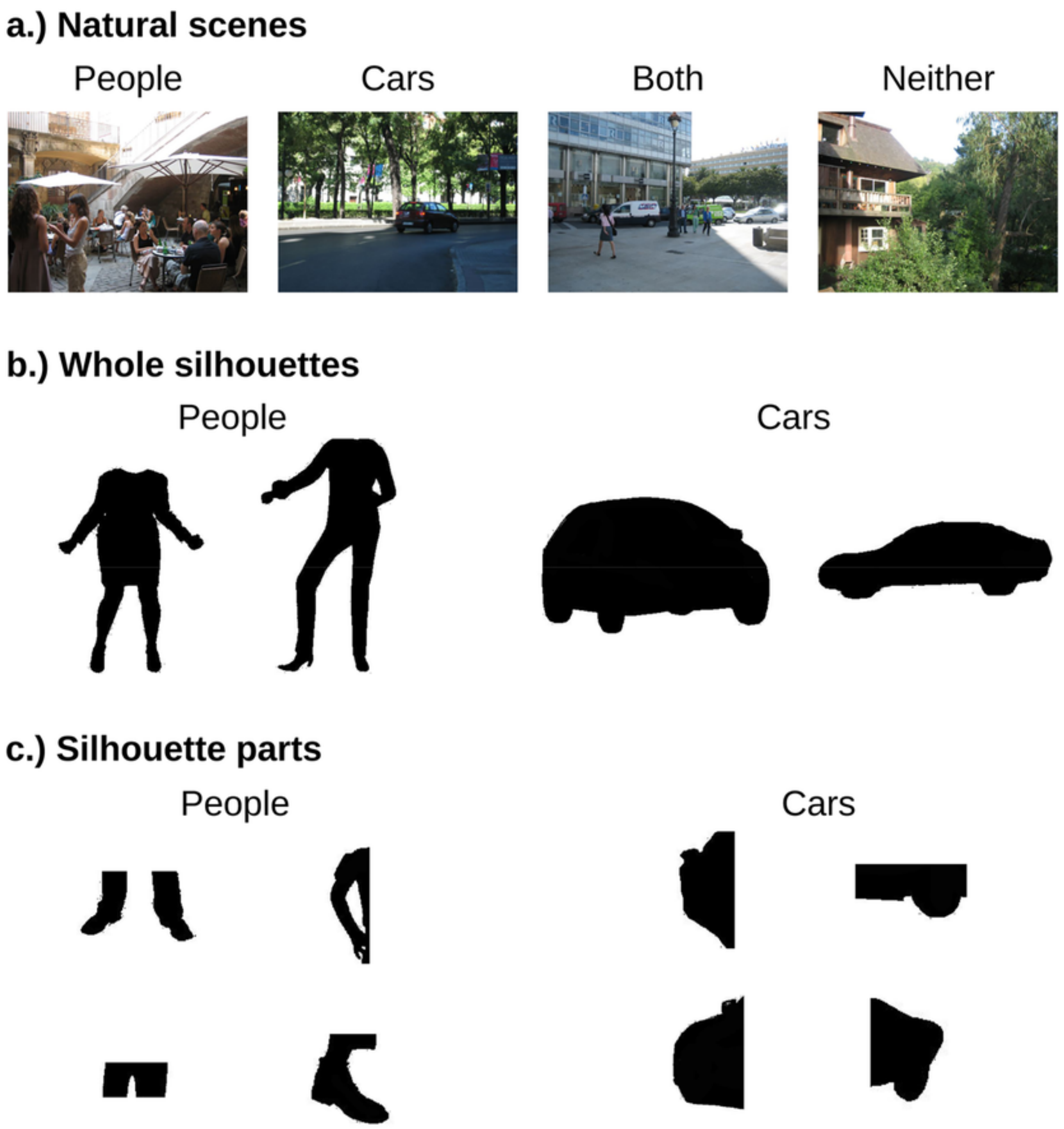

**Fig. 1.** All stimuli are taken from Reeder and Peelen (2013). a.) examples of search stimuli. b.) examples of whole silhouette capture stimuli. c.) examples of silhouette part capture stimuli. In Experiment 2, only one part from each category was shown in isolation on either side of fixation. In Experiment 3, four parts from each category appeared equally spaced in a 2 × 2 grid as depicted here. Images are not shown to scale.

analysis run in GPower (Faul, Erdfelder, Buchner, & Lang, 2009). From the results of Experiment 4 of Reeder and Peelen (2013), a power analysis (power (1-β) set at 0.95 and α = 0.05, two-tailed) indicated that to find a reliable difference in RT between consistent and inconsistent dot-probe trials on which silhouettes of object parts appeared, we would need to run 22 subjects. Although a sample size of 12 was found to be adequate based on a power analysis run on the capture effects of whole silhouettes, we used a sample size that could reveal any true capture effect by parts on their own because we specifically sought to target potential RT differences between parts and wholes.

Four subjects were excluded from Experiment 1 (two due to low search accuracy, and one each due to low dot-probe accuracy and slow dot-probe RT); five subjects were excluded from Experiment 2 (three due to low search accuracy, one due to low dot-probe accuracy, and one due to slow dot-probe RT); and six subjects were excluded from Experiment 3 (three due to low search accuracy, two due to both low search accuracy and low dot-probe accuracy, and one due to slow dot-probe RT). 75 subjects contributed to the final results; 25 in Experiment 1 (age range = 18–30 years, mean age = 21.64 years, 5 men, 2 left-handed), 25 in Experiment 2 (age range = 19–30 years, mean age = 22.35 years, 4 men, 1 left-handed), and 25 in Experiment 3 (age range = 19–27 years, mean age = 22.05 years, 6 men, 1 left-handed).

### 2.2. *Stimuli*

All stimuli were presented on a 24-in. Samsung with a 1920 × 1080 screen resolution and a 60 Hz refresh frequency (Samsung Electronics Co., Ltd., Suwon, South Korea). Stimuli were created in Python using PsychoPy functions (Peirce, 2008). The fixation cross and the letter cues “A” and “M” were text stimuli with a height of 31 pixels (0.8 cm) for the fixation and 70 pixels (1.8 cm) for the letter cues, presented in the center of the screen in Times New Roman font. Subjects viewed all stimuli from a free-viewing distance of approximately 57 cm.

#### 2.2.1. *Search stimuli*

Search stimuli were natural scene photographs used in Reeder and Peelen (2013; see Fig. 1a). 960 total scenes were used, 240

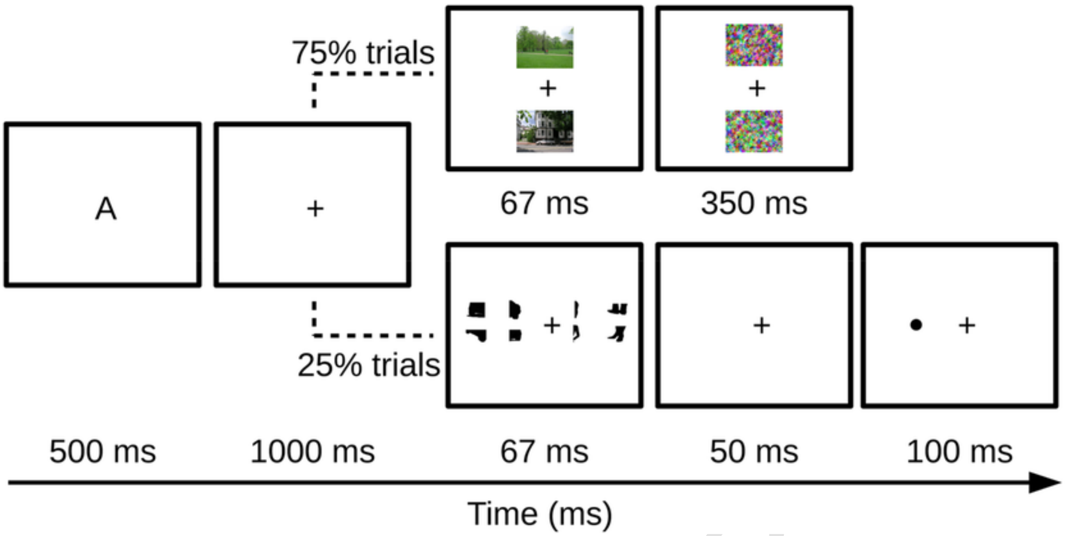

**Fig. 2.** A schematic of the search task (75% of trials) and dot-probe task (25% of trials). Here is an example of the search scenes appearing above and below fixation (Experiments 2 and 3) in the search task and collections of silhouette parts (Experiment 3) appearing in the dot-probe task. In Experiment 1, search scenes appeared to the left and right of fixation. In Experiments 1 and 2, single silhouette parts were shown on half of dot-probe trials.

containing cars but not people, 240 containing people but not cars, 240 containing both cars and people, and 240 containing neither cars nor people. No scene was viewed twice. Scenes contained objects in various natural viewpoints, levels of occlusion, distances from the observer, colors, sizes, genders (for people), and makes and models (for cars). Furthermore, one or multiple targets could appear in the same scene. Our goal with the variety of our images was to simulate natural everyday views as well as possible, and to encourage subjects to activate realistic templates for real-world search. Each scene had a size of 548 × 411 pixels, corresponding to 13.9 × 10.4 cm. Two scenes were presented to the left and right of fixation with 76 pixels (1.9 cm) separating the two images in Experiment 1, or above and below fixation with 261 pixels (6.6 cm) separating the two images in Experiments 2 and 3. The larger spatial separation in the latter experiments was to ensure that the scenes did not spatially overlap with the capture stimuli.

### 2.2.2. Capture stimuli

Capture stimuli were whole silhouettes of cars and people, and silhouettes of parts of cars and people, used in Reeder and Peelen (2013; see Fig. 1b). People were presented without heads to remain consistent with the previous study. 160 different images were used, each with 80 cars and 80 people in different perspectives and positions. Silhouette parts were created by cutting out sections of the total area of each of the whole silhouette images (range = 4.61%–24.19%, mean = 14.66% for cars, and range = 7.68%–23%, mean = 14.27% for people). A test designed to ensure that the part pictures were clearly discriminable as belonging to a car or person showed that subjects ($N = 8$) could correctly discriminate the object categories 97.1% of the time (Reeder & Peelen, 2013). Parts were scaled in a way that allowed them to be presented in the same sizes and locations as whole silhouettes. In Experiments 1 and 2, silhouettes were presented in one of three sizes, with the longest dimension of the image at 100, 180, or 200 pixels (2.5, 4.6, or 5.1 cm), determined randomly for each stimulus on a trial-by-trial basis. These were presented at one of three distances from the center of the screen to the center of the image at 130, 160 or 250 pixels (3.3, 4.1, or 6.4 cm), also determined randomly and independently for each image. This is in accordance with the stimulus parameters detailed in Reeder and Peelen (2013).

Part collections in Experiment 3 were 160 newly generated images (80 per image category) composed of four silhouette parts arranged in a 2 × 2 imaginary bounding box (see Fig. 1c). Images were randomly selected from the set of 80 single part images per category. Because there were only 80 single part images per category, parts were repeated across different part collections. Each part was randomly assigned one of the four locations in the collection, and no precise combination of parts and part locations was repeated. The total area covered by part collections was determined based on visibility of the individual parts, with each image randomly assigned to a size of 180 × 180, 200 × 200 or 280 × 280 pixels (4.6 × 4.6, 5.1 × 5.1, or 7.1 × 7.1 cm) on a trial-by-trial basis. Distance between the center of fixation and the center of the image could be 160, 200, or 250 pixels (4.1, 5.1, or 6.4 cm), randomly assigned to each stimulus separately. This kept part collections within the field of vision while remaining identifiable.

## 2.3. Procedure

Subjects performed two different intermixed tasks in the experiment (see Fig. 2). In the search task (75% of trials, 48 trials per block), subjects were required to indicate, with the directional arrow keys, which of two scenes presented to the left and right of fixation (Experiment 1) or above and below fixation (Experiments 2 and 3) contained a cued category — a car or a person. Cars and people were chosen as the search categories for three reasons: first, they were used in our previous study and we wanted to follow the methods as closely as possible; second, both are highly familiar and are viewed every day by most people; and finally, they are both highly variable categories that can contain myriad different features (colors, sizes, shapes) – and yet both are quickly and easily identified. This latter point highlights the importance of discovering how such information-rich and variable objects are represented at a categorical level in the search template. The letters “A” (for the German word *Auto*, which means car) or “M” (for the German word *Mensch*, which means human), presented prior to the scenes and intermixed an equal number of times within a block, cued subjects as to which category to look for. Each scene pair either contained cars in one image and people in the other, or cars and people in one image and neither in the other. Therefore, both the relevant and irrelevant category appeared on every trial, and the location of one category exemplar did not predict the location of the other category exemplar. Each scene type (car, person, both, neither) appeared an equal number of times on either side of fixation. A complete search trial consisted of a 500 ms fixation, followed by a 500 ms letter cue, then a 1000 ms fixation, two 67 ms search scenes followed by a 350 ms mask, and a final 1660 ms of fixation. Participants were instructed to use the arrow keys (right and left in Experiment 1, up and down in Experiments 2 and 3) to indicate in which scene the target category appeared.

In the dot-probe task (25% of trials, 16 trials per block), subjects were instructed only to respond to the location of a black dot that appeared on the left or right of fixation, using the left and right arrow keys. Capture stimuli (the silhouettes) appeared prior to the dot, but subjects were told to ignore these to the best of their ability. One car and one body silhouette appeared on each capture trial and appeared an equal number of times on the left and right of fixation. The location of the target-matching silhouette was tied to the same side of fixation as the dot-probe on half of trials (consistent trials). The target-matching silhouette appeared on the opposite side of fixation from the dot-probe on the other half of trials (inconsistent trials). In Experiments 1 and 2, capture stimuli were either both whole silhouettes or both silhouette parts, with these trials randomly mixed within a block. In Experiment 3, capture stimuli were either both whole silhouettes or both collections of silhouette parts. A capture trial started the same as a search trial, with a 500 ms fixation, followed by a 500 ms letter cue, and 1000 ms fixation. Because trials from the search task and dot-probe task were mixed, subjects did not know what type of stimuli would appear following the cue. This ensured that subjects would form a search template on every trial. Following the appearance of the two capture stimuli for 67 ms, there was a fixation for 50 ms, the dot-probe for 100 ms, and a final fixation of 1660 ms.

Trials were organized in ten 64-trial blocks, with the first block being a practice block that did not contribute to the analysis. Before each experiment, a short, slowed practice block was performed under the supervision of the experimenter, to ensure that each subject understood the task correctly.

### *2.4. Analysis*

Our main goal was to find out if subjects performed better on consistent dot-probe trials than on inconsistent dot-probe trials. A consistent trial is defined as a trial on which the cue-matching silhouette (e.g., the car silhouette following the "A" cue) appears on the same side of fixation as the dot-probe. An inconsistent trial is a trial on which the cue-matching silhouette appears on the other side of fixation from the dot-probe. We calculated RT from the onset of the dot-probe. Only correct trials were input into the RT analysis. Attention capture was defined as faster RT on consistent trials compared to inconsistent trials.

## 3. Results

### *3.1. Experiment 1*

Experiment 1 was a replication of Experiment 4 in Reeder and Peelen (2013). Search scenes, capture stimuli, and dot-probes were all presented to the left and right of fixation. Capture stimuli were either whole silhouettes or single silhouette parts. Our aim was to see if the earlier results could be replicated with a larger sample, and if the resulting increase in power could reveal a more detailed understanding of the role that diagnostic object parts play in the search template.

RTs were submitted to a 2 × 2 repeated-measures ANOVA, with consistency (consistent, inconsistent) and silhouette type (whole object, object part) as factors (see Fig. 3). There was a main effect of consistency ($F(1,24) = 20.918$, $p < 0.001$, $\eta_p^2 = 0.466$), indicating generally faster responses on consistent trials (mean = 436 ms, standard error (SE) = 11 ms) compared to inconsistent trials (mean = 453 ms, SE = 13 ms). There was no main effect of silhouette type ($F(1,24) = 0.050$, $p = 0.825$, $\eta_p^2 = 0.002$), but a significant interaction between consistency and silhouette type ($F(1,24) = 9.322$, $p = 0.005$, $\eta_p^2 = 0.280$). Paired-samples *t*-tests revealed that the consistency effect for whole silhouettes was significant ($t(24) = -4.984$, $p < 0.001$; consistent RT mean = 432 ms, SE = 12 ms; inconsistent RT mean = 457 ms, SE = 13 ms), whereas the consistency effect for silhouette parts was not ($t(24) = -1.848$, $p = 0.077$), although the results trended in the right direction (consistent RT mean = 441 ms, SE = 12 ms; inconsistent RT mean = 449 ms, SE = 13 ms).

The previous finding that diagnostic object parts capture attention in a comparable way to whole objects (Reeder & Peelen, 2013: Experiment 4) could not be replicated in a larger sample size. However, we were still able to find a reliable capture effect by whole objects, suggesting that whole object shape may be a necessary part of the search template. Nevertheless, it is possible that subjects were not using a flexible template for this task for whatever reason (perhaps German students rely on more rigid templates than Italian students). In our previous study, we found that the category-level template was not only flexible in terms of changes in viewpoint, but also in terms of location around the display. The representation of features without a particular spatial configuration (such as the representation of object parts) relies on rapid, spatially non-specific processing (e.g., Quinlan, 2003; Singh & Hoffman, 2001), so it is possible that the representation of parts in the template is also associated with a spatially global representation. Therefore, in Experiment 2, we presented capture stimuli and search stimuli in separate, non-overlapping locations. If the cap-

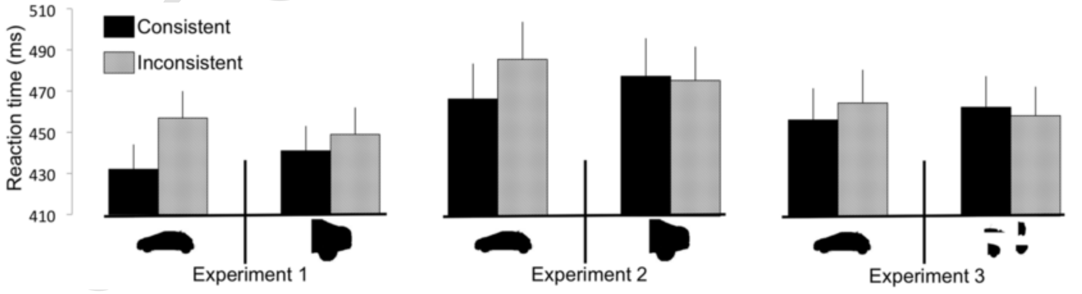

**Fig. 3.** Error bars show the standard error of the mean. RT results for consistent and inconsistent dot-probe trials are labeled by experiment. The silhouette shown below the bars represents whether the capture stimuli for those data were whole silhouettes (represented by a whole car silhouette here), single silhouette parts (represented by a single car part silhouette here), or collections of parts (represented by a collection of car parts here).

ture effect for whole silhouettes is impaired in the next experiment, it would suggest a reliance on a spatially focused search template. Alternatively, a replication of the pattern of results seen in Experiment 1 would provide evidence that even a flexible, spatially non-specific template is not exclusively composed of parts of objects.

### *3.2. Experiment 2*

Experiment 2 was the same as Experiment 1 with the exception that search scenes now appeared above and below fixation, with capture stimuli and dot-probes still appearing to the left and right. RT results were submitted to a 2 × 2 repeated-measures ANOVA with consistency and silhouette type as factors (see Fig. 3). This revealed a main effect of consistency ($F(1,24) = 20.096$, $p < 0.001$, $\eta_p^2 = 0.456$), again indicating faster RT on consistent trials (mean = 471 ms, SE = 17 ms) compared to inconsistent trials (480 ms, SE = 17 ms). There was again no main effect of silhouette type ($F(1,24) = 0.195$, $p = 0.663$, $\eta_p^2 = 0.008$), but a significant interaction between consistency and silhouette type ($F(1,24) = 13.638$, $p = 0.001$, $\eta_p^2 = 0.362$). Paired-samples *t*-tests revealed the same pattern as before, with whole silhouettes showing a significant consistency effect ($t(24) = -5.353$, $p < 0.001$; consistent RT mean = 466 ms, SE = 17 ms; inconsistent RT mean = 485 ms, SE = 18 ms), and silhouette parts showing none ($t(24) = 0.536$, $p = 0.597$; consistent RT mean = 477 ms, SE = 18 ms; inconsistent RT mean = 475 ms, SE = 16 ms). This demonstrates that the search template for object categories is spatially global, and optimally composed of information about both diagnostic parts and their configuration around the whole.

The argument arises that perhaps the reason for a lack of a capture effect with object parts is because it is possible that some object part images are less diagnostic than others (despite prior evidence of high general diagnosticity). Furthermore, a whole object contains multiple diagnostic parts that could all contribute to the capture effect. This could lead to a natural disadvantage for the silhouette part trials compared to the whole silhouette trials. To control for this, we conducted Experiment 3, which presented collections of four diagnostic parts (instead of one part in isolation) on silhouette part capture trials. If some parts are less diagnostic than others, or if stronger capture by whole objects is due to a summation of diagnostic part information, then part collections should prove to be a more equatable condition to the whole silhouettes.

### *3.3. Experiment 3*

Experiment 3 was the same as Experiment 2 except now a collection of four parts instead of one appeared on either side of fixation on silhouette part capture trials. A 2 × 2 repeated-measures ANOVA was conducted with consistency and silhouette type (whole object, part collection) as factors (see Fig. 3). This revealed no main effects (consistency: $F(1,24) = 0.557$, $p = 0.463$, $\eta_p^2 = 0.023$; silhouette type: $F(1,24) < 0.001$, $p = 0.994$, $\eta_p^2 < 0.001$), but a significant interaction between consistency and silhouette type ($F(1,24) = 8.472$, $p = 0.008$, $\eta_p^2 = 0.261$). Paired-samples *t*-tests revealed a significant consistency effect for whole silhouettes ($t(24) = -2.069$, $p = 0.049$), with consistent trials showing faster RT (mean = 456 ms, SE = 15 ms) than inconsistent trials (464 ms, SE = 16 ms). Conversely, part collections showed no consistency effect ($t(24) = 1.268$, $p = 0.217$), with consistent trials (mean = 462 ms, SE = 15 ms) showing no performance advantage compared to inconsistent trials (mean = 458 ms, SE = 14 ms).

These results indicate that the null effect for parts as seen in the previous experiments is not likely due to a lack of diagnostic information being present in the capture stimuli, but rather due to parts being a less accurate match to the template than whole silhouettes.

### *3.4. Cars versus people*

It could be argued that the lack of a capture effect for parts reported in these experiments is driven by a strong, natural holistic representation of conspecifics (Reed, Stone, Bozova, & Tanaka, 2003) compared to other objects. In other words, perhaps diagnostic parts are the favored template for cars (and generally other objects), but a holistic and configural representation of human bodies steers the results of the current study. This is unlikely, since the bodies in our study were headless, and research suggests that faces on bodies is what necessitates such effects (Brandman & Yovel, 2010; Yovel, Pelc, & Lubetzky, 2010). Nevertheless, to test for this, we analyzed the consistency effects in our experiments separately for cars and people. Because separating the results on this added dimension leads to few trials per condition (2 per experiment run), we analyzed all 75 subjects together to increase power and decrease noise.

We conducted a 2 × 2 × 2 repeated-measures ANOVA with category (cars, people), consistency (consistent, inconsistent), and silhouette type (whole objects, object parts) as factors. There was no main effect of category ($F(1,74) < 0.001$, $p = 0.996$, $\eta_p^2 < 0.001$), and no interaction between category or any other factor (all *F*s < 1.388, all *p*s > 0.242, all $\eta_p^2 < 0.019$). These results suggest that the reported capture effects cannot be attributed to a dominant holistic representation of people compared to cars.

## 4. General discussion

The results of the three experiments presented here provide a critical examination of the contribution of diagnostic object parts and whole object shape to the search template for real-world category search. Whole object silhouettes, presented as task-irrelevant distractors, captured attention across all experimental variations, both spatially specific and global. This suggests that whole object shape is represented efficiently and flexibly in preparation to search for object categories that appear in naturalistic contexts. The results are far less promising for diagnostic object parts — when presented at task-irrelevant locations, they do not capture attention at all, and if presented in task-relevant locations, only a weak trend emerges. The current study therefore favors a template composed of both diagnostic parts and the whole.

In a direct replication of Experiment 4 from Reeder and Peelen (2013), the capture effect for object parts was far weaker than for whole objects, and did not reach statistical significance. One basic axiom of statistical testing might lend an explanation for the difference in results between the current Experiment 1 and its counterpart from 2013: smaller samples tend to produce more extreme results (Huber, 2011). It is possible that with a smaller sample size of 13, we recruited a disproportionate number of subjects who happened to have a high disposition toward an exclusively part-based template (also see Reeder, 2017). Contreras, Rubio, Peña, and Santacreu (2010) remarked that people tend to fall on a continuum between relying more on whole objects and relying more on segments for visual search. The likelihood to recruit a sample that is skewed to either side of that spectrum is bigger for a sample size of 13 than it is for a sample size of 25. Another possibility is that parts can capture attention to some extent because they are still a necessary part of the search template, so the difference in the capture effects between silhouette parts and wholes is not as robust as, say, the difference between texture patches and wholes. Larger samples, experiment replication, and task variation can therefore all contribute to a more accurate assessment of the contribution of diagnostic parts to the contents of the category-level template.

The results of Experiments 2 and 3 provide evidence that, when presented at search-irrelevant locations, object parts do not even show a trend toward a capture effect. This supports the idea that parts alone do not make an efficient template, and are perhaps easier to inhibit in the latter two experiments. Our results are in contrast to Ullman et al. (2002), who showed computationally that parts of moderate complexity should actually be an ideal candidate to identify objects at the category level. However, the algorithms used in that study did not operate under the same limitations in time and capacity as our very human subjects did. The futility of adding more object parts in Experiment 3 and the trend toward capture effects for parts in Experiment 1 indicate further that it is not a lack of diagnostic information, but rather inefficiency of template matching, that prevents object parts from capturing attention.

One reason we initially hypothesized that the category-level template for real-world objects is composed of parts rather than wholes, is because we found capture effects for whole object silhouettes that were not only upright, but also inverted and rotated (Reeder & Peelen, 2013: Experiment 2 and 3). Because a canonical viewpoint of object features is not a necessary aspect of the template, we proposed that it would be easier to activate a spatially non-specific collection of parts rather than many variations of whole objects in the template. However, there is another possibility: in the real world, we may encounter whole object shapes that are inverted or rotated: just think of a car that has been turned upside down in an accident, or a person performing a handstand or leaning against a wall. We can still easily recognize these objects; however, the parts are always presented in a sensible relation to one another – even if a car is upside down, the wheels are attached to the lower body, with the doors appearing between the hood and the wheels. The chance of beholding a car with the doors on the ground and the wheels above the hood is almost non-existent. It is more likely, as an adaptation to real life probabilities, that proper spatial relations between parts (regardless of viewpoint) is a necessary part of the template. Therefore, our results point toward a viewpoint- and location-flexible template that nevertheless requires configural information to remain intact. One line of future research would be to present collections of parts in meaningful layouts (e.g., by removing the torso from images of people), and to evaluate whether configural information without the body can benefit a capture effect by object parts.

The fact that parts do not show a significant consistency effect in this study does not mean that the search template can never be composed of diagnostic parts in isolation. Specific attributes of the scenes we used might have enhanced the value of configural representations as a template. The search was broad (cars and people appeared in all sorts of sizes, colors, positions), swift (search scenes were presented for 67 ms), and included various levels of target occlusion (there are no deliberate occlusions, and many scenes showed entire objects); these are all properties that might favor a template composed of a coarse representation of the whole object shape, rather than the exclusive representation of diagnostic parts. The latter might be a good template for a specific search that requires the representation of more detailed part information (e.g., detecting sports cars). Parts in isolation would also be a more appropriate template if a search required subjects to detect predominantly incomplete or occluded targets. Furthermore, longer presentation times for search scenes would have allowed subjects to attend smaller or richer details of objects, which could in turn bias the template more heavily toward diagnostic parts. Finally, the results should not be generalized to other stimuli too readily — both people and cars are special with regards to their commonality and the distinctiveness in the shapes of their parts. More exotic or uncommon stimuli would likely require a more specific shape template for detection. There are also many object categories for which shape itself is not diagnostic, one example being the proverbial apples and oranges (for which color-based templates may be more diagnostic).

## 5. Conclusions

The results of this study indicate that the search template for object categories presented in natural scenes is composed of both whole object shape and diagnostic parts. Shapes of diagnostic object parts alone are not an adequate template, contrary to previous hypotheses.

## Author contributions

RRR developed and designed the experiment. MW and RRR both contributed to data collection, data analysis and interpretation, and writing the manuscript. Both authors approved the final version of the manuscript.

## Acknowledgements

This research was funded by an Open Research Area grant [DFG PO 548/16-1].

## References

Brandman, T., Yovel, G., 2010. The body inversion effect is mediated by face-selective, not body-selective, mechanisms. Journal of Neuroscience 30 (31), 10534–10540 https://doi.org/10.1523/JNEUROSCI.0911-10.2010.

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

Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A., 2014. Detect what you can: Detecting and representing objects using holistic models and body parts. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1971–1978 https://doi.org/10.1109/CVPR.2014.254.

Contreras, M.J., Rubio, V.J., Peña, D., Santacreu, J., 2010. On the robustness of solution strategy classifications. Journal of Individual Differences 31 (2), 68–73 https://doi.org/10.1027/1614-0001/a000012.

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

Desimone, R., Duncan, J., 1995. Neural mechanisms of selective visual attention. Annual Review of Neuroscience 18, 193–222 https://doi.org/10.1146/annurev.ne.18.030195.001205.

Duncan, J., Humphreys, G.W., 1989. Visual search and stimulus similarity. Psychological Review 96 (3), 433–458 https://doi.org/10.1037/0033-295X.96.3.433.

Faul, F., Erdfelder, E., Buchner, A., Lang, A.-G., 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods 41, 1149–1160 https://doi.org/10.3758/BRM.41.4.1149.

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 https://doi.org/10.1037/0096-1523.18.4.1030.

Hasegawa, I., Miyashita, Y., 2002. Categorizing the world: Expert neurons look into key features. Nature Neuroscience 5 (2), 90–91 https://doi.org/10.1038/nn0202-90.

Huber, P.J., 2011. Robust statistics. In: Lovric, M. (Ed.), International encyclopedia of statistical science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1248–1251 https://doi.org/10.1007/978-3-642-04898-2_594.

Lerner, Y., Hendler, T., Ben-Bashat, D., Harel, M., Malach, R., 2001. A hierarchical axis of object processing stages in the human visual cortex. Cerebral Cortex 11 (4), 287–297 https://doi.org/10.1093/cercor/11.4.287.

Logothetis, N.K., Sheinberg, D.L., 1996. Visual object recognition. Annual Review of Neuroscience 19 (1), 577–621 https://doi.org/10.1146/annurev.ne.19.030196.003045.

Malcolm, G.L., Groen, I.I., Baker, C.I., 2016. Making sense of real-world scenes. Trends in Cognitive Sciences 20 (11), 843–856 https://doi.org/10.1016/j.tics.2016.09.003.

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 https://doi.org/10.1167/9.11.8.

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 https://doi.org/10.1167/10.2.4.

Peelen, M.V., Kastner, S., 2014. Attention in the real world: Toward understanding its neural basis. Trends in Cognitive Sciences 18 (5), 242–250 https://doi.org/10.1016/j.tics.2014.02.004.

Peirce, J.W., 2008. Generating stimuli for neuroscience using PsychoPy. Frontiers in Neuroinformatics 2 (10) https://doi.org/10.3389/neuro.11.010.2008.

Quinlan, P.T., 2003. Visual feature integration theory: Past, present, and future. Psychological Bulletin 129 (5), 643–673 https://doi.org/10.1037/0033-2909.129.5.643.

Reed, C.L., Stone, V.E., Bozova, S., Tanaka, J., 2003. The body-inversion effect. Psychological Science 14 (4), 302–308 https://doi.org/10.1111/1467-9280.14431.

Reeder, R.R., 2017. Individual differences shape the content of visual representations. Vision Research 141, 266–281 https://doi.org/10.1016/j.visres.2016.08.008.

Reeder, R.R., Hanke, M., Pollmann, S., 2017. Task relevance modulates the cortical representation of feature conjunctions in the target template. Scientific Reports 7 (1), 4514 https://doi.org/10.1038/s41598-017-04123-8.

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) https://doi.org/10.1167/13.3.13.

Reeder, R.R., van Zoest, W., Peelen, M.V., 2015. Involuntary attentional capture by task-irrelevant objects that match the search template for category detection in natural scenes. Attention, Perception, & Psychophysics 77 (4), 1070–1080 https://doi.org/10.3758/s13414-015-0867-8.

Seidl-Rathkopf, K.N., Turk-Browne, N.B., Kastner, S., 2015. Automatic guidance of attention during real-world visual search. Attention, Perception, & Psychophysics 77 (6), 1881–1895 https://doi.org/10.3758/s13414-015-0903-8.

Singh, M., Hoffman, D.D., 2001. Part-based representations of visual shape and implications for visual cognition. In: Advances in psychology. Vol. 130, pp. 401–459, North-Holland https://doi.org/10.1016/S0166-4115(01)80033-9.

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

Treisman, A., Gelade, G., 1980. A feature integration theory of attention. Cognitive Psychology 12, 97–136 https://doi.org/10.1016/0010-0285(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 https://doi.org/10.1038/nn870.

Wolfe, J.M., 1994. Guided Search 2.0 a revised model of visual search. Psychonomic Bulletin & Review 1 (2), 202–238 https://doi.org/10.3758/BF03200774.

Wyble, B., Folk, C., Potter, M.C., 2013. Contingent attentional capture by conceptually relevant images. Journal of Experimental Psychology: Human Perception and Performance 39 (3), 861–871 https://doi.org/10.1037/a0030517.

Yovel, G., Pelc, T., Lubetzky, I., 2010. It's all in your head: Why is the body inversion effect abolished for headless bodies?. Journal of Experimental Psychology: Human Perception and Performance 36 (3), 759–767 https://doi.org/10.1037/a0017451.

Zelinsky, G.J., 2008. A theory of eye movements during target acquisition. Psychological Review 115 (4), 787–835 https://doi.org/10.1037/a0013118.