UNIVERSITY OF TRENTO – ITALY

CIMeC – Center for Mind/Brain Sciences

Doctoral School in Cognitive and Brain Sciences
31st cycle

Ph.D. Thesis

Attentional Mechanisms in Natural Scenes

Supervisor:
Marius V. Peelen

Ph.D. Candidate:
Elisa Battistoni

Academic Year:
2017/2018

Attentional mechanisms in natural scenes

Ph.D. Thesis

Elisa Battistoni
31st cycle Ph.D. candidate

Doctoral School in Cognitive and Brain Sciences
Cognitive Neuroscience track

Center for Mind/Brain Sciences (CIMeC)
University of Trento

Supervisor:
Prof. Marius V. Peelen

Table of contents

Chapter 1:
General Introduction[1]

The visual analysis of the world around us is an incredibly complex neural process that allows humans to function appropriately within the environment. When one considers the intricacy of both the visual input and the (currently known) neural mechanisms necessary for its analysis, it is difficult not to remain enchanted by the fact that, even though the signal that hits the retina has a tremendous amount of simple visual features and that is ever-changing, ambiguous and incomplete, we experience the world around us in a very easy, stable and straightforward manner[2]. So much effort has been put into the study of vision, and despite the enormous scientific advances and important findings, many questions still need answers.

During my years spent as Ph.D. student, I investigated some questions related to the topic of top-down attentional mechanisms in natural scenes. Top-down attentional mechanisms are at the basis of all the different processing stages that define our visual search behavior, as defined by Eimer: preparation, guidance, selection and identification (Eimer, 2014). This definition well summarizes all the various topics that will be investigated in the following chapters: the preparation phase in Chapter 2 and 3, the guidance and selection phase in Chapter 4, and the identification phase in Chapter 5.

In the following pages, I will (briefly and broadly) introduce the reader to the issues related to the experiments in the following chapters, which will each have its own, more detailed, introduction.

---

1 This work contains parts of a journal article that has been published elsewhere: Battistoni, E., Stein, T., and Peelen, M.V. (2017). Preparatory attention in visual cortex. Ann. N. Y. Acad. Sci. *1396*, 92–107.

2 Here, of course, I refer to healthy humans with normal or corrected-to-normal vision.

1. Naturalistic Vision

The neural analysis of the visual input is a computationally highly complex process, characterized by a hierarchical sequence of stages that involve the analysis of progressively more complex visual features, and that interact with each other via forward, lateral and feedback connections. The intricacy of this analysis process becomes unimaginable when considering the real scenes that we face everyday in life. Our daily-life visual environments, such as city streets and living rooms, contain a multitude of objects, which tend to continuously change depending on factors like distance, position, lighting (illumination, shading), and to appear incomplete because of other occluding objects. Furthermore, many objects share similar visual features, including targets and non-target objects.

Despite this overwhelming amount of constantly-changing visual information, we are remarkably efficient at processing natural scenes (Biederman, 1972; Biederman et al., 1974; Potter, 1976; Thorpe et al., 1996; Henderson and Hollingworth, 1999; VanRullen and Thorpe, 2001; Li et al., 2002; Fei-Fei et al., 2007; Greene and Oliva, 2009a; Thorpe, 2009; Wolfe et al., 2011a; Peelen and Kastner, 2014). The main reason for such efficiency can be linked to our evolution and daily-life experience. Specifically, natural stimuli are thought to have an advantage of processing over artificial ones (Li et al., 2002; VanRullen et al., 2005): they require less attention and cognitive resources because the visual system has adapted, at evolutionary, developmental and behavioral timescales, to real-world scenes, their objects and their statistical regularities (Simoncelli and Olshausen, 2001; Braun, 2003; Felsen and Dan, 2005; Hasson et al., 2010; De Cesarei et al., 2017; Kaiser et al., 2018). One result of such adaptation can be observed in the finding that the human brain can rapidly extract many types of information from a single glance of a scene: information about the background and context (Oliva and Torralba, 2001; Torralba et al., 2006; Greene and Oliva, 2009b; Castelhano and Heaven, 2010), about other objects in the scene (Mack and Eckstein, 2011; Pereira and Castelhano, 2014; Koehler and Eckstein, 2017), and depth information (Sherman et al., 2011). All these scene properties help observers to predict and guide attention to the likely location (Eckstein et al., 2006; Neider and Zelinsky, 2006; Greene and Oliva, 2009b; Malcolm and Henderson, 2010; Castelhano and Heaven, 2011; Mack and Eckstein, 2011; Wolfe et al., 2011b; Pereira and Castelhano, 2014; Koehler and Eckstein, 2017) and size (Eckstein et al., 2017) of target objects, whose selection is the ultimate goal of visual search.

2. Visual search

Visual search is one of the most commonly performed visual behaviors in humans, and it can be thought of as the process leading to the selection and identification of objects that are relevant to current goals. Searching arises as a consequence of the limited computational resources of our visual system: we need to search because we cannot simultaneously identify all the objects in the visual field (Wolfe et al., 2011b; Wolfe and Horowitz, 2017). Despite the fact that most of the human cortex is dedicated to the direct or indirect analysis of the visual input, our visual system does not have the capacity to process all the visual input at the same time (Tsotsos, 1990). Fortunately, our sensory systems are equipped with a mechanism, also known in the literature as selective attention, that prioritizes incoming sensory input that is relevant for current behavioral goals (top-down attention) or made relevant through its saliency (bottom-up attention), filtering out what is irrelevant and distracting. In other words, selective attention is the mechanism through which our brain overcomes the computational problem of the limited processing resources: it solves the competition among stimuli by favoring one stimulus over the others, which then are lost (Desimone and Duncan, 1995).

Moving top-down attention onto objects in the visual field is what characterizes our visual search behavior. Target selection is the result of a multi-stage process starting with a “parallel” (or “spatially-global”) identification of likely target-related features across the visual field, followed by a second stage in which attention is serially moved onto the locations of objects containing those features until the target is found (Wolfe et al., 1989; Treisman and Sato, 1990; Hochstein & Ahissar, 2002; Ahissar et al., 2009). In other words, feature-based processes guide the allocation of spatial attention onto likely target objects (Wolfe et al., 1989; Treisman and Sato, 1990; Wolfe, 1994; Cave, 1999; Eimer, 2014).

The “global-to-local” pattern of attentional selection proposed in the context of the Feature Integration Theory (FIT; Treisman & Gelade, 1980), Guided Search Model (Wolfe, 1994) and Biased Competition Model (Desimone & Duncan, 1995), has also been formulated in the Reverse Hierarchy Theory by Hochstein and Ahissar in the context of conscious vision and perceptual learning (Hochstein & Ahissar, 2002). Specifically, they describe that fast and automatic feedforward (bottom-up) processes following stimulus onset lead to the “vision at a glance” percept in high-level areas. This percept consists of a generalized and categorical representation of a scene, which identifies a “forest before trees”, and that it is comparable to the “spatially-global” activation of features mentioned above. In a second stage called “vision with scrutiny”, feedback processes move along the reverse hierarchy and focus resources on specific low-level neuronal populations, allowing feature binding and conscious perception of specific items (e.g., trees). This stage would correspond

to location-specific attentional movements on items during the visual search process.

Direct neuronal evidence for such “feature-to-location” (or “global-to-local” or “vision at a glance – to – vision with scrutiny”) process has been provided for search tasks involving simple visual features and artificial displays, demonstrating that a spatially-global feature-based modulation precedes a spatially-specific enhancement of target objects (Hopf et al., 2004). However, it remains to be proven whether such progression extends to more complex tasks in real-world scenes. Chapter 4 will address this issue by employing multivariate pattern analysis (MVPA) on magnetoencephalography (MEG) data.

3. Attentional templates

A large body of research has characterized the effects of attention on neural activity evoked by a visual stimulus, as reviewed elsewhere (Reynolds and Chelazzi, 2004; Maunsell and Treue, 2006; Reynolds and Heeger, 2009; Buschman and Kastner, 2015). However, attention also includes a preparatory phase, before stimulus onset, in which the attended dimension is internally represented. More specifically, attentional mechanisms in visual search are not only engaged at the moment the eyes hit a scene: they often start before, when search goals are established. For example, when crossing a road, we decide to look out for (i.e., attend to) cars before physically looking in both directions to inspect the scene for the presence of cars. This intuitive concept of *preparatory attention* (also referred to as *attentional set*, *attentional template*, or *search image*) was described by William James as “The image in the mind *is* the attention; the *preperception* […] is half of the perception of the looked-for thing” (James, 1890), and it has subsequently played an important role in theories of attention (Duncan, 1989; Tinbergen, 1960; Bundesen, 1990; Wolfe et al., 1989; Treisman, 2006). Specifically, when we determine a target object that we want to find, we establish a preparatory attentional template, which can be thought of as an internal representation describing the object of interest.

Such preparatory attention acts through the selective pre-activation of the visual cortex, before stimulus onset and thus in the absence of visual input (Desimone and Duncan, 1995; Luck et al., 1997; Desimone, 1998; Kastner et al., 1999). Specifically, it is observed in those regions (or neurons) that represent the task-relevant stimulus, ranging from low-level feature representations in the early visual cortex to category representations in the high-level visual cortex, therefore operating across all levels of the visual hierarchy (Fig. 1; for a review on preparatory attention in visual cortex, see Battistoni et al., 2017).

Preparatory attention-related activity in visual cortex has a clear purpose: it serves to bias processing in favor of attended items (Fig. 1; Desimone and Duncan, 1995; Desimone, 1998; Kaiser

et al., 2016), such that items that are attended gain a competitive advantage (proportional to the degree to which they match the internal attentional template) and are prioritized for spatial attentional selection and further object processing (Wolfe et al., 1989; Wolfe, 1994; Desimone and Duncan, 1995). Notably, this preparatory activity is causally related to subsequent attentional selection and behavioral performance, as shown by transcranial magnetic stimulation (TMS) studies (Romei et al., 2010; Reeder et al., 2015b). In the context of real-world visual search, preparing to look for cars and people in scenes led to the pre-activation of category-specific neural patterns in the Object-Selective Cortex (OSC, within the Lateral Occipital Cortex, or LOC; Peelen and Kastner, 2011; Fig. 1). This preparatory activity, which can be thought of as the establishment and temporary maintenance of category-based attentional templates, was crucial for, and causally related to, subsequent behavioral performance (Peelen and Kastner, 2011; Reeder et al., 2015b).

However, what are the object characteristics that these category-based preparatory activity patterns code? To be most effective in guiding attention to targets, preparatory attention would need to be directed to those representations that optimally distinguish the target from the distractors rather than simply activate the representation of the target (Duncan and Humphreys, 1989; Navalpakkam and Itti, 2007; Scolari and Serences, 2009; Becker et al., 2010, 2013). Thus, one would predict that preparatory attention effects in the visual cortex for the same target differ as a function of the expected distractor set (e.g., scene context). For example, different shape features are diagnostic of the presence of people in a forest as compared with a desert. In relation to this kind of adjustment, scene context also provides information about the likely visual appearance of objects at different locations in the scene (e.g., as a function of depth): therefore, in order to increase its effectiveness and optimize the attentional template, preparatory mechanisms could code the expected target's distance, for example by scaling the size of its pre-activated representation.

Some research suggested that the template-defining features were high-level category-diagnostic object parts, such as the wheel of a car, and the legs, arms or torso of a person (Peelen and Kastner, 2011; Reeder and Peelen, 2013)[3]. Notably, these studies used complex and cluttered scenes, where many objects shared simple low-level features (e.g., orientation, color,..). Given that targets could not be discriminated from distractors from simple features, it is possible that participants might have used a strategy based on high-level characteristics (i.e., templates based on category-diagnostic object parts) because a lower-level strategy (i.e., templates based on low-level features, such as orientation) would have been highly ineffective in terms of both timing and accuracy. Importantly, such low-level strategy was not implausible: the results by Peelen and Kastner (2011) showed that some participants did employ it, and furthermore, the behavioral performance

3 The notion of "category-diagnostic object parts" is associated to the idea of "feature detectors" (Evans and Treisman, 2005) characterized by intermediate level representations useful for classification (Ullman et al., 2002).

of these participants was poor compared to those adopting a high-level strategy (Peelen and Kastner, 2011). Therefore, it remains to be established whether, in scenes where targets do not share low-level features (orientation) with distractors, observers can engage search templates based on low-level features. For example, if an observer had to search for a person in the desert or in a field, where there are no distractors matching the vertical orientation of a person, would the template still consist of category-diagnostic parts, or low-level features (i.e., orientation)? This question will be addressed in Chapter 2.

Related to the question of strategy adopted in real-world search tasks, is the question of whether expecting a nearby target vs. a distant target leads to a modification of the size of the category-based attentional template (which, hypothetically, could be characterized by a big-sized representation for nearby targets and a small-sized representation for distant targets). Imagine that you have to look for a friend who is on the beach far away, will your template represent a small person, or this representation will be invariant to the expected distance? Chapter 3 will cover the issue of the relative size of attentional templates in naturalistic visual search.

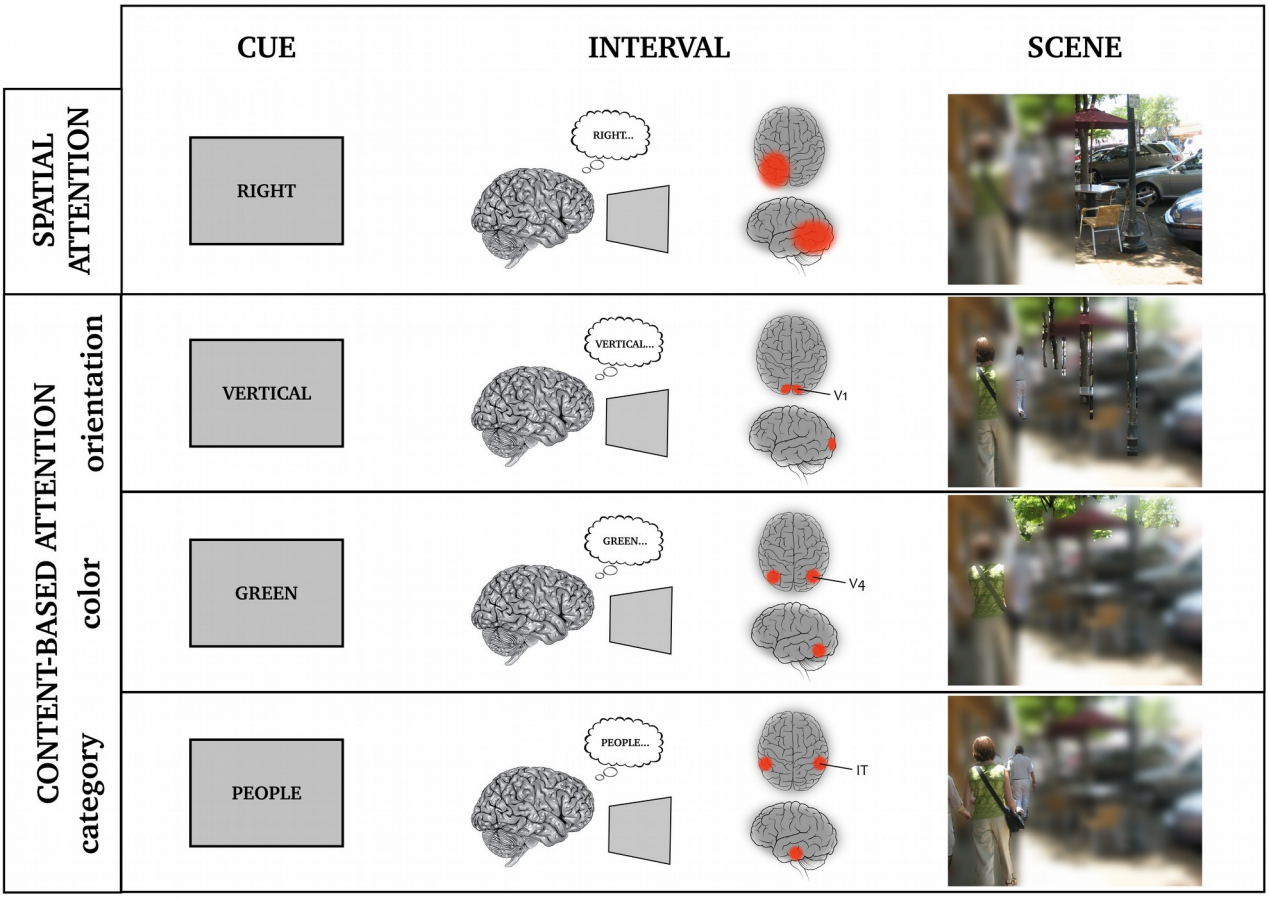

Figure 1[4]. Schematic overview of the involvement of attentional templates in visual search, in temporal sequence (from left to right) and based on different goals (different rows). Goals are prompted by words indicating the task-relevant

4 Figure 1 is adapted from Battistoni, E., Stein, T., and Peelen, M.V. (2017). Preparatory attention in visual cortex. Ann. N. Y. Acad. Sci. *1396*, 92–107.

dimension (left column). Preparatory attentional templates are established and maintained until scene onset through the activation of task-relevant representations in visual cortex (middle column). Visual processing is biased toward items that match the attentional templates (right column).

4. Invariant object recognition and size-constancy

The visual search process ends when a target is selected and identified: after a global activation of likely candidate objects, spatially selective attention is serially deployed to the location of each candidate object in order to collect and bind their features into a representation for later recognition (or identification) processes (Wolfe and Cave, 1999; Hochstein & Ahissar, 2002; Wolfe, 2003; Ahissar et al., 2009; Wolfe et al., 2011b). However, the visual perception of real-world objects and scenes is never static. This constant flux of ever-changing visual information implies that we (almost) never see an object in the exact same position and with the exact same characteristics. Nonetheless, despite of the different signals that hit the retina, we are able to recognize and categorize objects in different conditions of lighting, occlusion and distance, quickly and effortlessly. But how is this constant perception achieved? This problem, known in the literature as *invariant object recognition*, has been one of the most important issues in vision science (DiCarlo and Cox, 2007; DiCarlo et al., 2012; Gauthier and Tarr, 2016; Contini et al., 2017; Peelen and Downing, 2017). Despite all the research that has been done in the past 30 years, many questions still remain, among which the central problem of invariance. Specifically, what are the neural mechanisms that underlie the invariant perception of objects despite changes in appearance associated with variations of lighting, distance, view, and so on?

A particular interesting theme within this field concerns the invariant perception of the size of objects despite changes in distance. Specifically, even though the size of the image projected on the retina by an object is inversely proportional to its distance, and therefore constantly changes as we and the objects around us move, we perceive the object to be the same. In the literature, the perceptual mechanism characterized by a rescaling of an object's size as a function of distance is known as size constancy (Sperandio and Chouinard, 2015), and it is fundamental in allowing us to recognize an object despite the changes in retinal size that are associated with changes in distance.

In Chapter 5, I will investigate the temporal dynamics of size constancy in natural scenes using an MEG decoding approach. Linking this to Chapter 4, which investigates the time course of spatial attentional allocation onto target objects in natural scenes, it is plausible to hypothesize that brain activity patterns should encode information on the size of the target before allocating attention to the target (since target selection implies at least a coarse form of identification). Therefore, Chapter 5 will address the question of the temporal evolution of size-constancy

mechanisms, and whether they appear before the time in which spatial attention is allocated onto the target.

Notably, Chapter 3 is also linked to the topic of size-constancy: another way to think of the question (i.e., whether the size of category-based preparatory templates changes as a function of expected target's distance) is whether attentional templates are characterized by size-invariance.

5. Conclusions

To summarize, the general theme around which the thesis is centered is top-down attentional mechanisms in natural scenes. I will address only briefly bottom-up attention in Chapter 4, but the present lack of in-depth scrutiny does not intend to lessen the importance in everyday life of attentional capture by saliency-based mechanisms.

In Chapter 2 and 3 I will investigate the characteristics of attentional templates employed during the preparatory phase of a naturalistic visual search task; Chapter 4 will address the time course of spatial attention during a naturalistic visual search task using MEG decoding; Chapter 5 will explore the time course of size-constancy mechanisms in real-world scenes; finally, in Chapter 6 I will draw the conclusions, highlight points to improve and delineate some questions for future research.

Chapter 2:
Investigating the influence of distractor context expectations on attentional templates in natural scenes

1. Introduction

Daily-life environments are characterized by an uncountable amount of visual information. The limited computational resources of our brain (Tsotsos, 1990) constrain the quantity of input that can be processed at any given time, thus allowing us to not be overwhelmed and act appropriately in the world. Attention is the key neural mechanism to these selective processes. It enables us to efficiently prioritize and select only a small amount of information that is relevant for current behavioral goals and ignore what is irrelevant and distracting.

One method that has been largely employed to study attentional selection is the paradigm of visual search (Treisman and Gelade, 1980; Wolfe et al., 1989; Treisman and Sato, 1990; Wolfe, 1994; Cave, 1999; Eimer, 2014). In this framework, the concept of "attentional template" has gained a pivotal role (Duncan and Humphreys, 1989; Carlisle et al., 2011; Olivers et al., 2011; Eimer, 2014; Battistoni et al., 2017). When we determine something to look for, we establish an "offline" top-down attentional template (also referred to as "search image", "search template" or "attentional set"), which can be conceived of as an internal representation of the sought-after object (Duncan and Humphreys, 1989) and it is associated to increases of neuronal activity (in absence of visual stimulation) in the areas coding for the attended attribute (for a review, see Battistoni et al., 2017). When the visual stimulus appears, this pre-activated representation biases "online" visual processing resources guiding attention to template-matching items, even involuntarily (Folk et al., 1992, 1993; Desimone and Duncan, 1995; Folk and Remington, 1998; Woodman et al., 2007; Reeder and Peelen, 2013; Reeder et al., 2015a).

But what features define these internal templates? Towards what features is attention guided to? At least two accounts have been proposed on this matter: a feature similarity account and a relational account. Seemingly in contrast, they are not mutually exclusive and depending on the task at hand, observers can change search strategy (Harris et al., 2013; Becker et al., 2014), and, therefore, they can be reconciled. The *feature similarity account* (or feature detector theory; Folk and Remington, 1998) proposes that attention is based on and guided to the target's physical feature values, increasing the response gain in neurons coding features similar to the sought-after

target features (Treue and Martínez Trujillo, 1999; Martinez-Trujillo and Treue, 2004; Maunsell and Treue, 2006; Anderson and Folk, 2010), which can be very specific (e.g., a particular shade of red; Navalpakkam and Itti, 2006), or categorically broader (e.g., redness; Wolfe, 1994). Crucially, within this framework, the physical features defining the template are assumed to be represented in isolation from the characteristics of the surrounding context. Tuning attention to specific or broad target feature values may be an efficient attentional selection strategy when the features of the context in which the target will appear are unknown. But what if the features of the context are known in advance to the start of the search task? In this case, it would be reasonable to think that a more efficient selection strategy would take into account these contextual features, and consequently shape the attentional template accordingly, in a way that it will be most efficient for target detection. This strategy is at the core of the second account describing the contents of attentional templates, and is formalized by the theory of *relational target template* proposed first by Becker (Becker, 2010; Becker et al., 2010; Becker, 2013; Becker et al., 2013; Harris et al., 2013; Becker, 2014; Becker et al., 2014; Bravo and Farid, 2016; Schönhammer et al., 2016; Geng et al., 2017). The relational account postulates that attentional guidance and stimulus selection are determined by feature relationships. More specifically, when the characteristics of the upcoming search context are known, and the distractors are characterized by specific features, the template will be based on the relationship between the features of the target and the features of the distractors, maximizing their difference. For example, when looking for an orange target item among yellow items, the template will consist of a representation coding for “redder”; whereas, when searching for an orange item among red distractors, the template will feature “yellower” (Becker et al., 2010). These results were recently extended to more complex stimuli and visual search conditions (Bravo and Farid, 2016).

However, simplified displays such as those typically employed in visual search studies like those reviewed above, poorly resemble our visual experience in real life. The scenes that we encounter are usually cluttered with objects that share many low-level visual features, making it difficult to distinguish targets from distractors based on basic characteristics like orientation, color, or simple shapes. Such features normally vary in the environment as a function of lighting, perspective, occlusion, and distance, further increasing the problem of defining and detecting an object based on basic low-level features. Visual search in real life is most often performed at the category-level of an object, where exemplars can be defined by a wide range of low-level features. Despite this apparently insurmountable complexity, studies have demonstrated that humans are very skilled in detecting familiar object categories in natural scenes (Potter, 1976; Thorpe et al.,

1996; Li et al., 2002; Peelen and Kastner, 2014). The reasons for such efficiency are various. Visual and attention systems have developed and evolved to optimally performs real-world tasks like selecting objects in the environment (Barlow, 1961; Felsen and Dan, 2005; Wolfe et al., 2011b). Real-world scenes provide a visual context that constrains not only the interpretation and low-level features of objects, but also their possible location, by guiding attention to areas that are most likely to contain the target (Torralba et al., 2006; Wolfe et al., 2011b; Peelen and Kastner, 2014; Wolfe and Horowitz, 2017). Consistent and familiar arrangements of objects lead to perceptual grouping processes that facilitate target detection (Kaiser et al., 2014). Lastly, even though the low-level features of the exemplars of a category can vary, at the same time exemplars share some category-diagnostic features, upon which search templates are likely based (Reeder and Peelen, 2013).

Therefore, despite the importance of studies employing basic stimuli, which have allowed fundamental findings in the field of visual attention, when considering all the factors that are unique to real-world scenes, it appears difficult to draw parallels between visual processes in basic stimuli and visual processes in real scenes.

In this study, we aimed to investigate whether context-based templates could be established in naturalistic visual search tasks. More specifically, whether participants could adopt different search strategies by adjusting the features of the category-based attentional template depending on the expected distractor context. Previous studies on real-world search found that attentional templates are implemented in a higher-level visual area known as object-selective cortex (OSC; Peelen and Kastner, 2011), that they likely consist of category-diagnostic object parts, and tend to be orientation-invariant (Reeder and Peelen, 2013). Interestingly, participants who based their search strategy on a template consisting of low-level features, activating mostly low visual areas, tended to perform poorly in the task (Peelen and Kastner, 2011; Reeder et al., 2015b). Therefore, from these studies it would appear that in real-world search observers tend to adopt a template based on specific high-level category-diagnostic features. However, the distractors' context was not explicitly manipulated, leaving open the possibility that observers can, in fact, adjust the template when given the opportunity to form clear expectations about the characteristics of the distractors' context. Since the scenes contained a large amount and variety of distractors, observers could not form expectations about the context, and because the target could not be found with a low-level template, they might have been forced to adopt a high-level category-based template. As an example, if an observer had to search for a person in the desert, a search strategy based on a high-level template would be efficient, but would it not be much more efficient and economic (in terms of cognitive resources expenditure) just looking for something vertical? Of course, such a low-level

strategy would be detrimental if observers had to search for a person in a forest full of trees, which would match the low-level template, capture attention, and slow down the search process. Analogously, when searching for a car in a forest, a template representing “something horizontal” might be efficient as well as resources-saving. Notwithstanding, since such real-world tasks are usually performed daily and humans are highly skilled at them, it is possible that observers do not establish low-level templates because category-based templates are the default, automatic strategy that has stem from experience, and switching to a low-level template would, actually, be more time consuming and require a higher expense of cognitive resources.

In order to address whether, in a naturalistic visual search task, the characteristics of the templates could be shaped by expectations on the upcoming search context, we run two experiments in which we manipulated the relation of target-distractors orientation (Experiment 1) and the clutter of distractors (Experiment 2) in natural scenes. Importantly, a given target-distractor relation was kept constant within a session in order to implicitly encourage participants to adopt a specific template. In Experiment 1, one session was characterized by target and distractors with the same orientation (“same-orientation” session, in which participants looked for cars among horizontal distractors and for people among vertical distractors), and the other session had target and distractors with different orientation (“different-orientation” session, in which they searched for cars among vertical distractors and for people among horizontal distractors). In Experiment 2 we also manipulated the clutter of distractors, by drastically decreasing the number of distractors in the different-orientation session. We hypothesized that when the orientation of target and distractors did not match (in the different-orientation session), the most efficient and economic (in terms of cognitive resources) template would be based on low-level characteristics, such as the orientation of the target (vertical for people, horizontal for cars). This strategy would be especially efficient in Experiment 2, where the decreased number of distractors in the different-orientation session should have further encouraged participants to instantiate a low-level template. When target-distractor orientations did match (in the same-orientation session), we hypothesized that a low-level template based on target’s orientation would be no longer efficient, because it would match the orientation of the distractors. Thus, in this case, we expected participants to instantiate a higher-level template, likely based on category-diagnostic object parts (Reeder and Peelen, 2013). To note, we did not choose a full two-by-two design because of practical reasons (the number of trials would have become too large) and because we were specifically interested in comparing two conditions that we thought would maximally differ in the type of template employed.

Importantly, in both experiments, on a subset of trials participants also performed a prime

task, where participants saw, instead of the scenes, car and person silhouettes (which could be upright or rotated by 90°, and had to be ignored) and a dot appearing at the location of either silhouette (i.e., at the location of the template-matching silhouette or at the location of the silhouette that did not match the attentional template). This task served to probe the characteristics of the attentional template that was formed during the naturalistic visual search task.

We expected that the hypothesized pattern of search strategies (a low-level, orientation-based, template in the different-orientation session, and a higher-level, category-based, template in the same-orientation session) would be reflected in a smaller (or absent, or reversed) validity[1] effect in the rotated silhouette condition in the different-orientation session (compared to the same-orientation session). Specifically, we thought that if participants adopted a low-level template (vertical for people, horizontal for cars) in the different-orientation session, then their attention would be captured toward the rotated silhouette of the opposite category – which, because of its rotation, would match the orientation-based low-level template. In other words, we expected attention to be captured by the orientation of the silhouette, rather than its category, in the different-orientation session.

2. Materials and methods

2.1. Participants

Forty undergraduate and graduate students from the University of Trento took part in the two experiments for monetary compensation or course credits. Twenty participants were assigned to Experiment 1 (4 males; aged 19-32 years, mean age M = 22.6, SD = 3.6 years). After a preliminary inspection of their behavioral performance, one participant was removed from the analyses because of the poor accuracy in the prime task (43%). Twenty different participants took part in Experiment 2 (2 males; aged 21-36 years, mean age M = 23.6, SD = 3.8 years). All participants had normal or corrected-to-normal vision and provided written informed consent to take part in the experiments. All participants received monetary compensation (€8/session). The experiments were conducted in accordance with the Declaration of Helsinki and approved by the Ethical Committee of the University of Trento.

1 Valid trials were those in which the location of the dot corresponded to the location of the template-matching silhouette; invalid trials were those in which the dot appeared at the location of the silhouette that did not match the template. In Reeder & Peelen (2013), the “validity effect” was represented by shorter RTs in valid trials than in invalid trials.

2.2. Stimuli

Stimuli were presented on a 19-inch Philips 109P4 monitor with a screen resolution of 1600 x 1200 pixels and a monitor frame rate (refresh frequency) of 85Hz. Stimulus presentation was controlled with MATLAB 8.0 using the Psychtoolbox (Kleiner et al., 2007).

All stimuli were displayed on a grey background (RGB values: 148, 148, 148). The fixation point, a black plus ("+"), and letter cues (black 25-point uppercased Arial font), were presented at the center of the screen. The distance between the screen and eyes of participants was controlled by using a chinrest positioned at 55 cm from the monitor.

Natural scenes were grey-scaled and reduced to 427 (horizontal) x 320 (vertical) pixels, subtending 10.1 x 7.5 degrees of visual angle. Silhouettes were 136 x 136 pixels black exemplars of cars and people, subtending 3.2° of visual angle in height and width.

Masks had the same size of the scenes and were created by superimposing a naturalistic texture to white noise generated at different spatial frequencies, resulting in grey-scaled textures.

The dot stimulus was a black circle with a diameter of 7.5 pixels (0.2° of visual angle).

Natural scenes, masks, silhouettes, and dots were placed at a distance of 40 pixels from fixation.

2.3. Scene stimuli (naturalistic visual search task)

A hundred and eight basic naturalistic scenes were selected from the web. These real-world pictures depicted outdoor scenes with natural surroundings and country roads. Each basic scene was manually edited using the image processing tool GIMP (https://www.gimp.org) to create the naturalistic scene stimuli. Starting from a basic scene, two types of scenes were created: one with horizontally-oriented distractors, and one with vertically-oriented distractors. Examples of horizontal distractors were bushes, benches, picnic tables, vases, guardrails, fences, rocks, horizontally-oriented bird flocks, clouds, planes, boats, beach chairs. Examples of vertical distractors were street lamps, traffic lights, road signs, trees, lighthouses, vertically-positioned surf boards. On average, 7 distractors were added in a scene, and they were semantically consistent with the basic scene. From each of these two scenes (one with horizontally-oriented distractors and one with vertically-oriented distractors), three further scenes were created: one with a person, one with a car, and one with a person and a car. The exemplars of person and car were kept constant within a basic scene, as well as their size. This was done to ensure the most control and validity over the stimuli created, so as to avoid any possible confound in the stimuli. Hence, a total of 108 x 8 scenes were created (864): for each basic scene, four scenes were created with horizontally-oriented distractors (one empty, one with a car, one with a person, one with both a car and a person), and four scenes

had vertically-oriented distractors (one empty, one with a car, one with a person, one with both a car and a person). In order to meet the criteria of presenting a scene only once in the experiment, to obtain the number of scenes required, the original 864 scenes were horizontally-mirrored to create additional 864 novel scenes. The combination of session (same-orientation vs. different-orientation of target and distractors) and scene type (mirrored vs. non-mirrored) was counterbalanced across participants.

2.3.1. Experiment 1: "Distractors' Orientation"

In the different-orientation session, target and distractors had different orientations: car targets were presented among vertically-oriented distractors, and person targets were presented among horizontally-oriented distractors. According to our hypothesis, such target-distractors configuration should have driven participants to adopt a basic low-level template, that is a template representing "something vertical" for people and a template representing "something horizontal" for cars. Our intuition was that such search strategy would have been both efficient and less resource-consuming. In the same-orientation session, target and distractors had the same orientation: car targets were presented among horizontally-oriented distractors, and person targets were presented among vertically-oriented distractors. We expected that in this target-distractors setting, a low-level template would have been inefficient and detrimental to the performance since it would have captured attention towards the distractors, which would have matched the low-level template. Thus, we hypothesized that, in the same-orientation session, participants would adopt a higher-level template, likely based on category-diagnostic object parts.

2.3.2. Experiment 2: "Distractors' Orientation and Clutter"

in Experiment 2 we manipulated two aspects of the distractor context: their features (horizontally- vs. vertically- oriented), and their clutter (high vs. low clutter). Specifically, participants performed a same-orientation high-clutter session, and a different-orientation low-clutter session. The aim of this manipulation was to further prompt participants to adopt a low-level template in the different-orientation low-clutter session. In the same-orientation high-clutter session we used the scenes employed in Experiment 1 in the same-orientation session, which had been created since the beginning with many distractors (7 on average). This also provided the opportunity to replicate the findings of Experiment 1. For the scenes of the different-orientation low-clutter session, new scenes were created by drastically reducing the number of distractors in the scenes employed in the different-orientation session in Experiment 1 (leaving, on average, 3 distractors per scene).

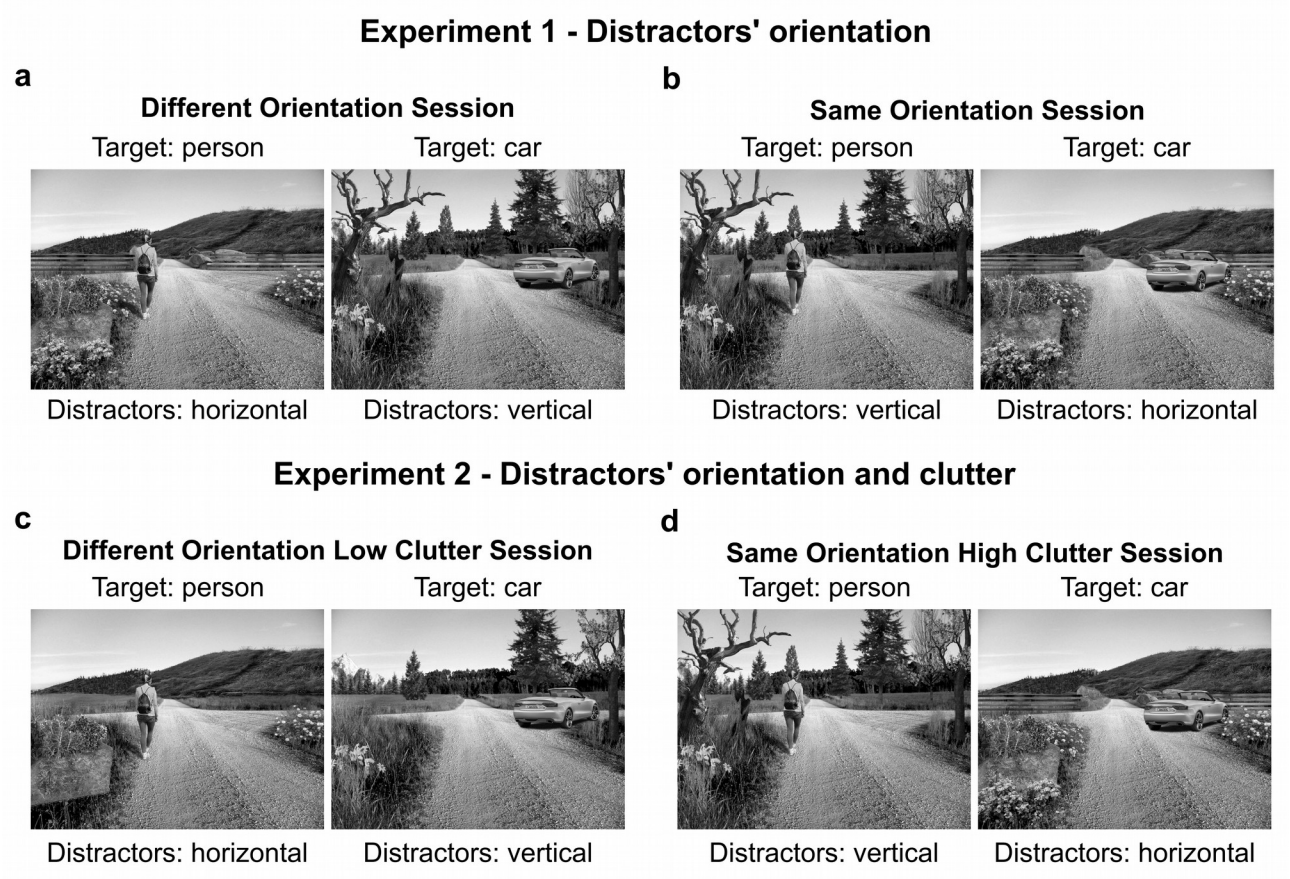

Figure 1. Examples of stimuli used in the naturalistic visual search tasks of the two experiments. In Experiment 1, the orientation of distractors was manipulated. (a) In the different-orientation session, participants looked for people among horizontally-oriented distractors (left scene) and looked for cars among vertically-oriented distractors (right scene). (b) In the same-orientation session, they looked for people among vertically-oriented distractors (left scene) and looked for cars among horizontally-oriented distractors (right scene). In Experiment 2 both distractors' orientation and distractors' clutter were manipulated. (c) In the different-orientation low-clutter session, participants searched for people among few horizontally-oriented distractors (left scene) and searched for cars among few vertically-oriented distractors (right scene). (d) In the same-orientation high clutter session, they searched for people among many vertically-oriented distractors (left scene) and searched for cars among many horizontally-oriented distractors (right scene).

2.4. Silhouette stimuli (prime task)

The stimuli used in the prime task were black silhouettes of cars and people (144 different exemplars for each object category) on grey background (the same hue of the display in which the stimuli appeared). Some of them came from a previous study (Reeder and Peelen, 2013), and some were manually created with GIMP starting from isolated cars or people.

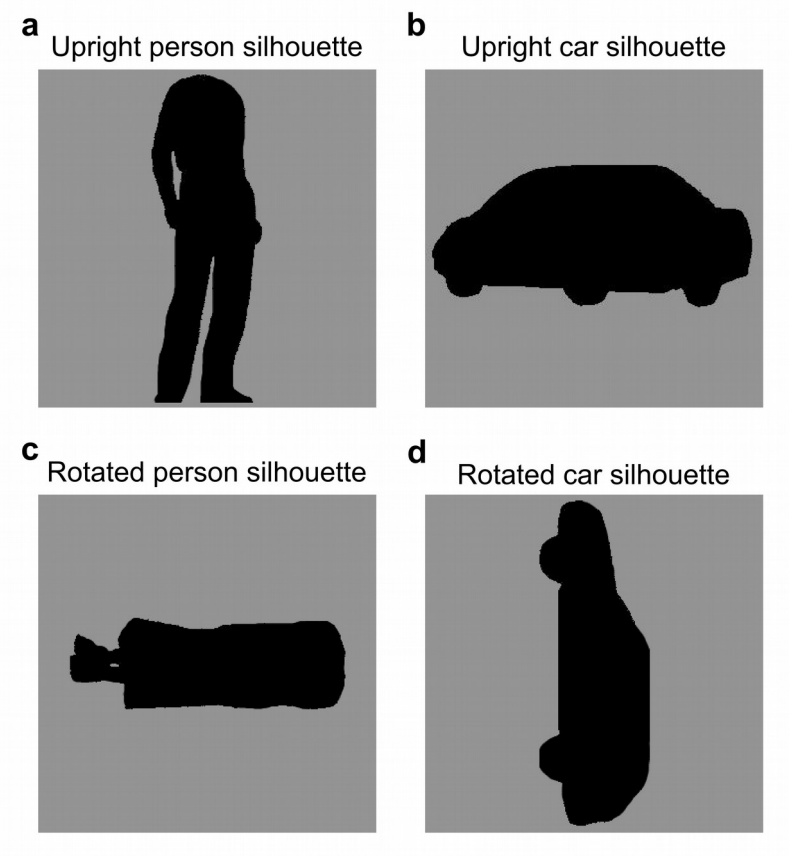

Figure 2. Examples of stimuli used in the prime task. They comprised (a) a set of upright person silhouetted, (b) a set of upright car silhouettes, (c) a set of 90°-rotated person silhouettes, and (d) a set of 90°-rotated car silhouettes.

2.5. Experimental procedure

Both experiments consisted of two sessions of 45 minutes each. Each participant took part in both sessions on separate days (taking place within a week), and the order of the two sessions was counterbalanced across participant (i.e., the same-orientation session was the first one for participant 1 and the second one for participant 2; the different-orientation session was the second one for participant 1 and the first one for participant 2; and so on). Each session consisted of 9 blocks of 64 trials each. In each block, the naturalistic visual search task made up three-fourths (i.e., 48) of trials, to ensure that participants would actively prepare to detect the cued object category. One-fourth of trials in a block (i.e., 16) had the prime task. In half of the trials with the prime task (i.e., 8) the orientation of the silhouette was upright; in the other half (i.e., 8) the orientation of the silhouettes was rotated clockwise by 90°. The order in which the visual search task and the prime task appeared was randomized, therefore subjects could not predict what task they had to perform on any subsequent trial. To ensure that participants could establish some form of template (and that the prime task would not probe a non-existing representation), the first 3 trials in each block did

not contain the prime task, but only the naturalistic visual search task. At the beginning of the first session, each participant completed one practice block in order to familiarize with the tasks.

2.5.1. Naturalistic visual search task

In the visual search task, each trial started with a fixation point (500ms); which was replaced by a letter cue (for English speakers, "C" for "car", and "P" for "person"; for Italian speakers, "M" for "macchina", and "P" for "persona"; 500ms), followed by the fixation point (1000ms). Then, two scenes were presented on either sides of fixation (67ms; the combination of the scenes in each trial could be one of the following: (1) scene with car on the left, scene with person on the right; (2) scene with person on the left, scene with car on the right; (3) scene with car and person on the left, scene with no car and no person on the right; (4) scene with no car and no person on the left, scene with car and person on the right)[2]. Notably, the two scenes had congruent distractor context (either both vertical or both horizontal in Experiment 1; either both vertical with high clutter, vertical with low clutter, horizontal with high clutter, or horizontal with low clutter in Experiment 2), and could not be variations of the same basic scene. An empty screen appeared after the scenes, the duration of which was manipulated with a staircase procedure: the duration of the empty screen in the first 5 trials in each block was fixed at 100ms; then, if the average accuracy in the visual search task (in the current block, the average accuracy of the search trials up to that point) was higher than 75%, the empty screen duration decreased by 20ms; whereas, if the accuracy was lower than 75%, the duration of the empty screen was increased by 20ms. The minimum duration of the empty screen could be 10ms, the maximum duration 300ms. This was done to ensure that the difficulty of the search task was more or less balanced across the two sessions. After the empty screen, two masks appeared at the location where the two scenes were presented (350ms); a fixation point followed (1660ms), then the feedback ("+0" for incorrect responses; "+1" for correct responses; 500ms). In these trials, the task of the participant was to indicate in which scene (left or right) the cued object category was present (by pressing on the keyboard one of the two keys "z" or "m", for left or right target, respectively).

2.5.2. Prime task: trial sequence

The first events of the trials with the prime task matched the trials with the visual search task: they started with a fixation point (500ms), followed by the letter cue (500ms) and the fixation point

2 The conditions (3) and (4) were added to ensure that the presence of an object category (e.g. a car) in one scene (e.g. the one to the left of the fixation point) did not imply the presence of the other object category (e.g., a person) in the other scene (i.e., the one to the right). This ensured that searching within a scene was exhaustive.

(1000ms). Instead of the scenes, two silhouettes were presented on either sides of fixation: one silhouette of a car on one side, and one silhouette of a person on the other side (67ms). A fixation point was presented briefly (50ms), followed by the appearance of a dot on one side of fixation (at the location of one of the two silhouettes, 100ms). A fixation point followed (1660ms), then the feedback ("+0" for incorrect responses; "+1" for correct responses; 500ms). In this task, participants were instructed to ignore the silhouette and press a key to indicate whether the dot appeared to the left or right of fixation ("z" for left, "m" for right").

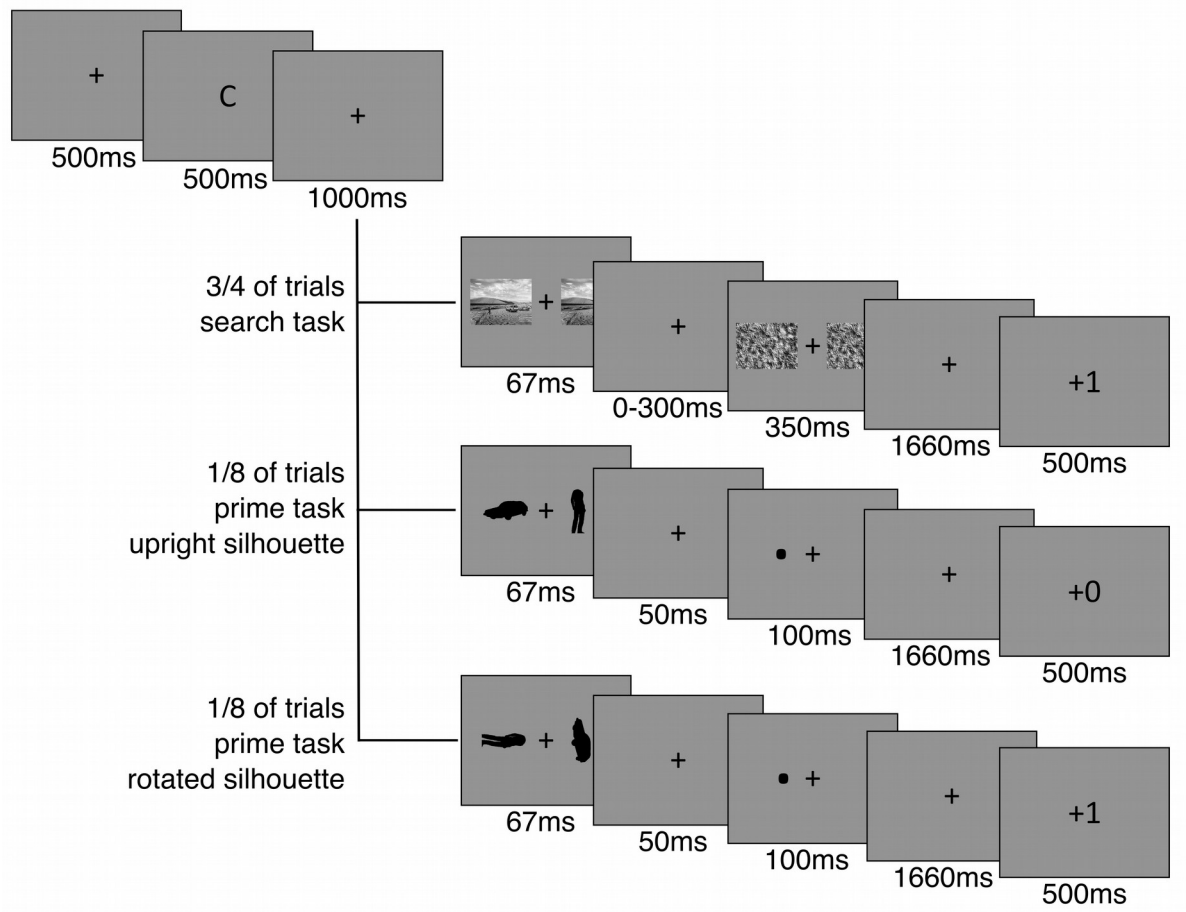

Figure 3. Schematic outline of the experimental procedure used in Experiment 1 and Experiment 2. Each block consisted of 64 trials. 75% (¾ in each block) of trials included a naturalistic visual search task, where participants indicated whether the cued target category was present in the scene to the left or right of fixation. 25% (¼ in each block) of trials consisted of a prime task, in which participants indicated whether the dot appeared to the left or right of the fixation point. Half of the prime task trials had upright silhouettes (⅛ in each block), the other half had 90°-rotated silhouettes (⅛ in each block). The first sequence showing an example of prime task, with upright silhouettes, also illustrates an example of valid trial: the dot appears at the location of the car silhouette (the category cued as search target at the beginning of the trial is car). The second sequence of prime task, with rotates silhouettes, is also an example of invalid trial (the dot appears at the location of the silhouette that does not match the cued target category).

2.6. Analysis

First, we tested whether in the naturalistic visual search task there was a difference in behavioral performance across the two sessions. Since in the same-orientation session (Experiment 1) and the same-orientation high-clutter session (Experiment 2) the orientation of the distractors matched the orientation of the targets, we hypothesized that finding the target in these circumstances would be more difficult than in the different-orientation session (Experiment 1) and in the different-orientation low-clutter session (Experiment 2).

To investigate the main question of whether expectations about the distractors' context orientation could influence the characteristics of the attentional template, we analyzed RTs and response accuracy in the prime task as a function of target-distractor orientation (different-orientation session vs. same-orientation session for Experiment 1; different-orientation low-clutter session vs. same-orientation high-clutter session for Experiment 2), silhouette orientation (upright vs. rotated), and validity of dot position (valid vs. invalid trials). Valid trials were those in which the dot appeared at the location of the template-matching silhouette; invalid trials were those in which the dot appeared at the location of the silhouette that did not match the template (Fig. 3).

Non-given responses or too-slow responses (i.e., responses not given before the end of the fixation screen lasting 1660ms) were considered incorrect responses. Incorrect responses were not included in the analysis of RTs.

3. Results

3.1. Naturalistic Visual Search task results (Experiment 1 and 2)

To check whether search difficulty in Experiment 1 and 2 was manipulated successfully across sessions, we analyzed the performance of participants in the naturalistic visual search task. Figure 4 illustrates RTs and response accuracy of the two sessions in each experiment.

In Experiment 1, the RTs of the different-orientation session were not significantly different than the RTs of the same-orientation session (two-tailed t-test, $t(18) = 1.4$, $p = 0.17$, Cohen's $d = 0.07$). Response accuracy in the same-orientation session was significantly different than the response accuracy in the different-orientation session (two-tailed t-test, $t(18) = 3.4$, $p =< 0.003$, Cohen's $d = 0.11$), with lower accuracy in the same-orientation session (mean = 71%, SD = 5%) than in the different orientation session (mean = 74%, SD = 6%).

In Experiment 2, in the same-orientation high-clutter session the RTs were not significantly different than the RTs in the different-orientation low-clutter session (two-tailed t-test, $t(19) =$

2.02; p = 0.057, Cohen's d = 0.17). Response accuracy in the same-orientation high-clutter session was significantly different than the accuracy in the different-orientation low-clutter session (two-tailed t-test, t(19) = 9.82; p < 0.001, Cohen's d = 0.33), with higher accuracy in the different-orientation low-clutter session (mean = 78%, SD = 4%) than in the same-orientation high-clutter session (mean = 71%, SD = 5%).

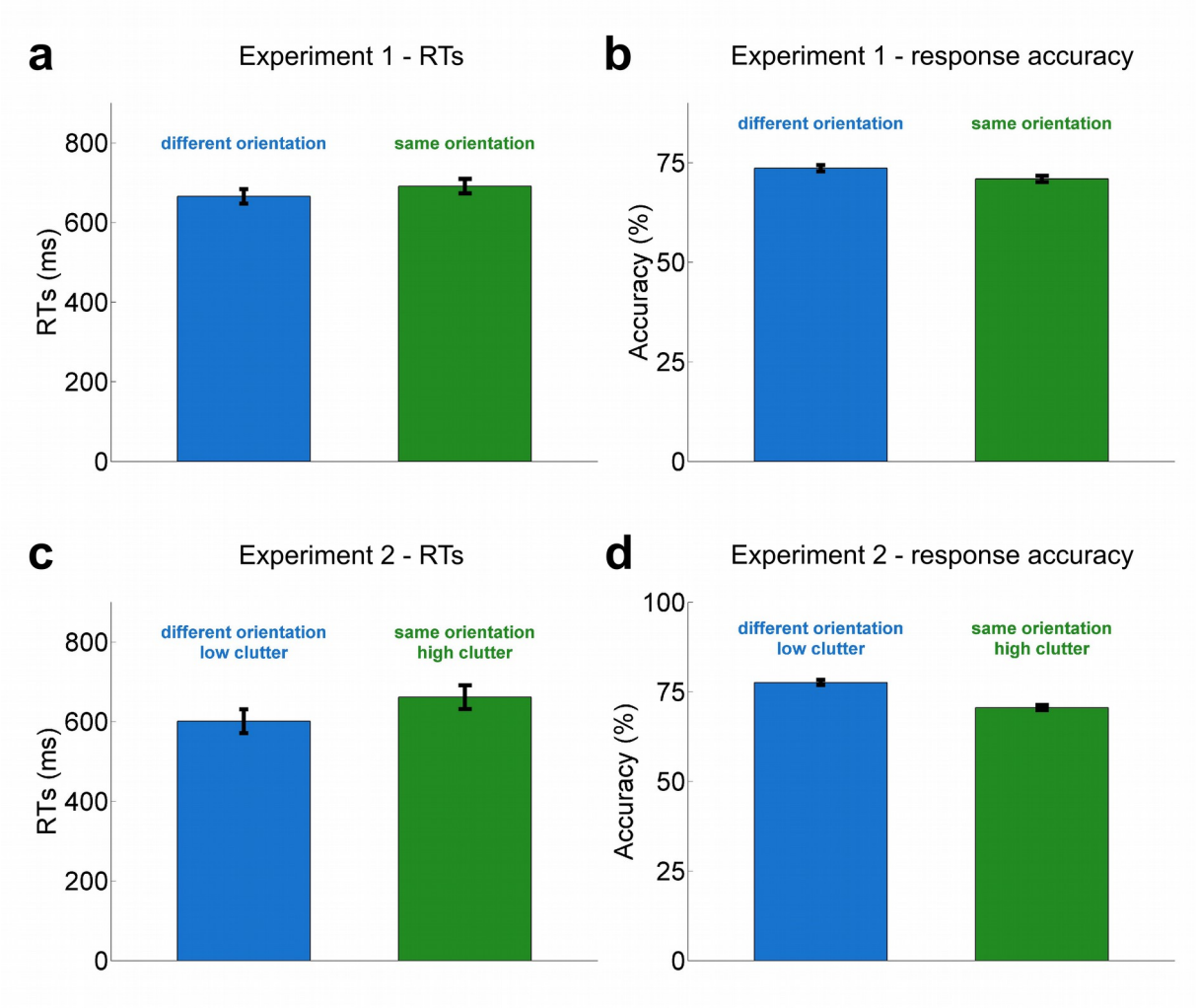

Figure 4. Results of the naturalistic visual search task in Experiment 1 and Experiment 2. Panel (a) and (b) show RTs and response accuracy in Experiment 1. Panel (c) and (d) illustrate RTs and response accuracy in Experiment 2. Bars represent the mean across subjects; error bars represent within-subjects SEM.

3.2. Prime task results in Experiment 1 ("distractors' orientation")

In order to investigate whether attentional templates were influenced by expectations on the orientation of distractors in scenes, we analyzed data in the prime task trials as a function of session (target-distractor orientation), silhouette orientation, and validity (Fig. 5). We performed a repeated measures ANOVA with factors: (1) target-distractor orientation (different-orientation session vs.

same-orientation session), (2) silhouette orientation (upright vs. rotated), and (3) validity (valid vs. invalid), on both RTs and response accuracy.

The three-way ANOVA on RTs revealed a significant main effect of validity ($F(1,18) = 23.5$; $p < 0.001$, Cohen's $d = 0.1$), where RTs in valid trials (mean = 424ms, SD = 44ms) were shorter than RTs in invalid trials (mean = 438ms, SD = 45ms). No effect of target-distractor orientation ($F(1,18) = 0.9$), no effect of silhouette orientation ($F(1,18) = 1.6$), and no interactions were found.

The three-way ANOVA on response accuracy revealed a significant main effect of target-distractor orientation ($F(1,18) = 14.7$; $p < 0.001$, Cohen's $d = 0.08$), with lower accuracy in the different-orientation session (mean = 93%, SD = 5%) than in the same-orientation session (mean accuracy = 95%, SD = 5%); a significant main effect of silhouette orientation ($F(1,18) = 8.7$; $p < 0.01$, Cohen's $d = 0.08$), with higher accuracy in the rotated silhouette condition (mean = 95%, SD = 5%) than in the upright silhouette condition (mean = 93%, SD = 6%); and a significant main effect of validity ($F(1,18) = 10.2$; $p < 0.01$, Cohen's $d = 0.17$), with higher accuracy in the valid condition (mean = 96%, SD = 4%) than in the invalid condition (mean = 92%, SD = 8%). No interactions were present.

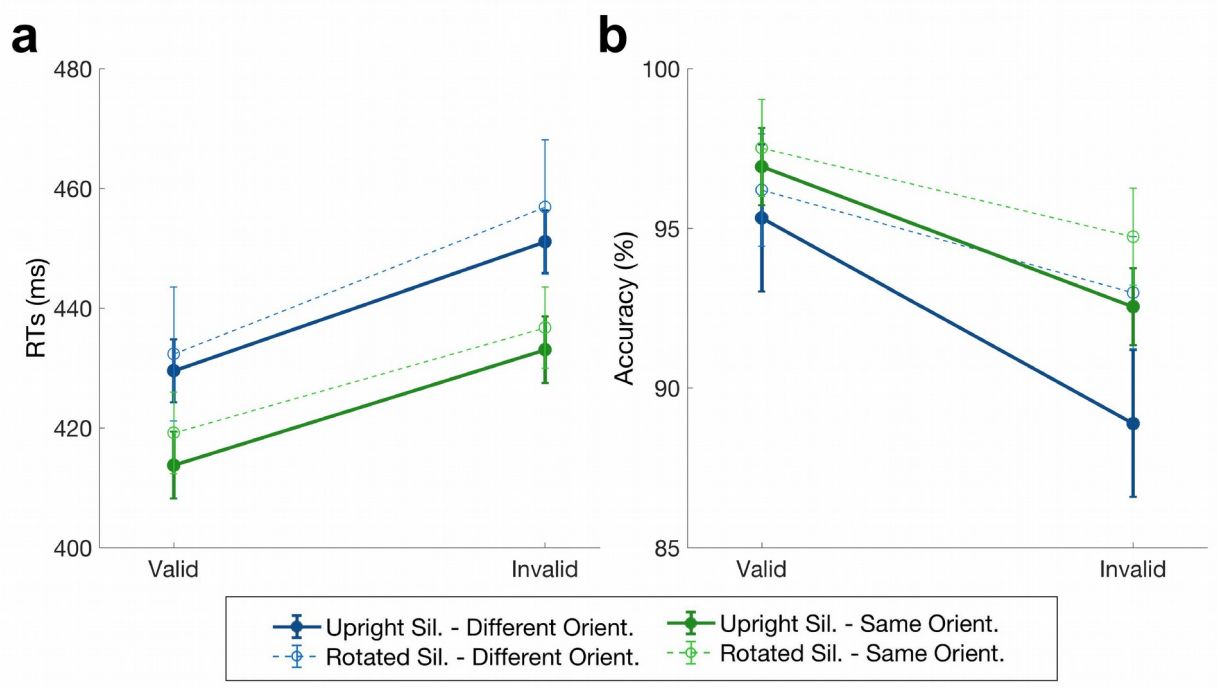

Figure 5. Results of the prime task in Experiment 1. Panel (a) illustrates RTs; panel (b) shows response accuracy. Circles represent means, error bars are illustrated as +1 and -1 within-subjects SEM with respect to the mean. In the legend, "Sil." is the abbreviation of "Silhouette", "Orient." is the abbreviation of "Orientation".

3.3. Prime task results in Experiment 2 ("distractors' orientation and clutter")

In Experiment 2, in order to investigate whether attentional templates were influenced by expectations on the orientation and clutter of distractors in scenes, we analyzed data in the prime task trials as a function of session (target-distractor orientation), silhouette orientation, and validity (Fig. 6). We performed a repeated measures ANOVA with factors: (1) target-distractor orientation and clutter (different-orientation low-clutter session vs. same-orientation high-clutter session), (2) silhouette orientation (upright vs. rotated), and (3) validity (valid vs. invalid), on both RTs and response accuracy.

The three-way ANOVA on RTs revealed a significant main validity effect ($F(1,19) = 14.12$; $p < 0.01$, Cohen's $d = 0.09$), with shorter RTs in the valid condition (mean = 398ms, SD = 67ms) than in the invalid condition (mean = 424ms, SD = 80ms); and a significant interaction between silhouette orientation and validity ($F(1,19) = 5.95$; $p < 0.05$, Cohen's $d = 0.12$). Post-hoc t-tests revealed that the validity effect (invalid-valid) was significant for both upright (two-tailed t-test, $t(19) = 3.69$, $p < 0.001$, corrected for multiple comparisons) and rotated (two-tailed t-test, $t(19) = 3.21$, $p < 0.01$, corrected for multiple comparisons) silhouettes, but it was larger in the upright silhouette condition than in the rotated silhouette condition (two-tailed t-test, $t(19) = 2.44$, $p < 0.05$). No main effect of target-distractor orientation and clutter ($F(1,19) = 0.13$), no main effect of silhouette orientation ($F(1,19) = 0.16$), and no other interactions were found. The ANOVA on response accuracy showed a significant main effect of silhouette orientation ($F(1,19) = 5.35$; $p < 0.05$, Cohen's $d = 0.05$), with higher accuracy in the rotated silhouette condition (mean = 93%, SD = 5%) than in the upright silhouette condition (mean = 92%, SD = 6%); and a significant main effect of validity ($F(1,19) = 29.58$; $p < 0.0001$, Cohen's $d = 0.33$), where response accuracy was higher in the valid condition (mean = 97%, SD = 3%) than in the invalid condition (mean = 88%, SD = 8%). No main effect of target-distractor orientation and clutter ($F(1,19) = 2.63$), and no interactions were found.

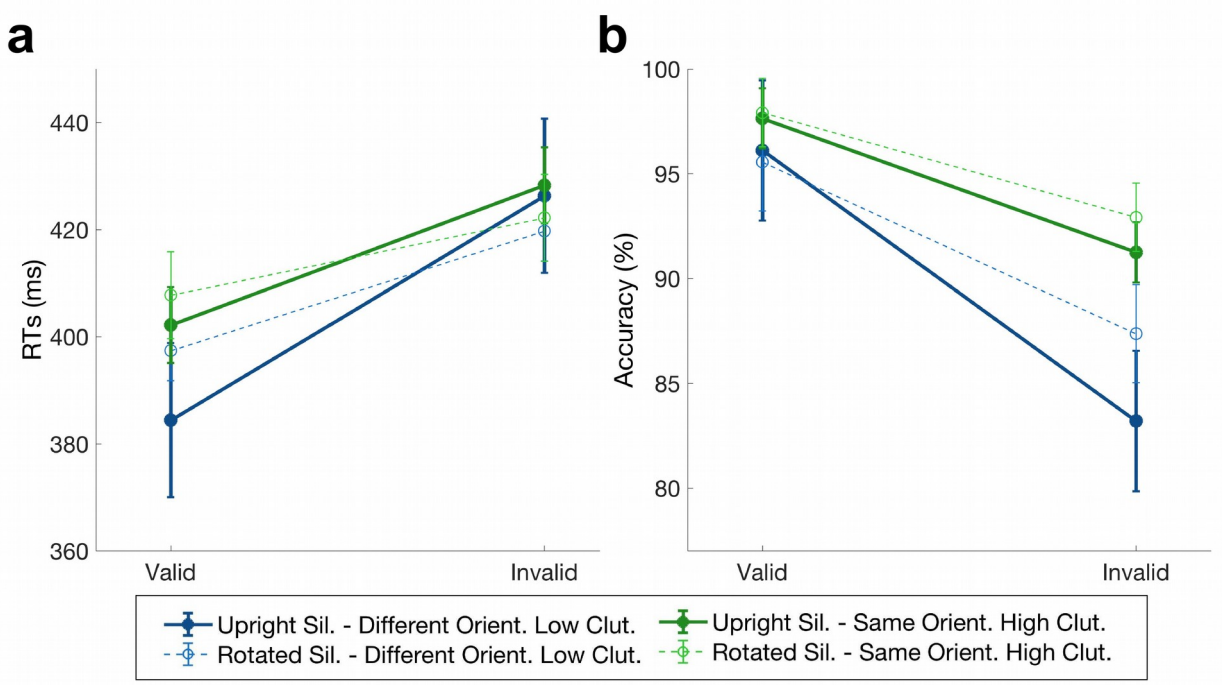

Figure 6. Results of the prime task in Experiment 2. Panel (a) illustrates RTs; panel (b) shows response accuracy. Circles represent means, error bars are illustrated as +1 and -1 within-subjects SEM with respect to the mean. In the legend, "Sil." is the abbreviation of "Silhouette", "Orient." is the abbreviation of "Orientation", "Clut." is the abbreviation of "Clutter".

4. Discussion

In this study we tried to determine whether preparatory attentional templates adopted in naturalistic visual search tasks could be adjusted based on expectations on the upcoming distractors' orientation (Experiment 1) and distractors' orientation and clutter (Experiment 2) in scenes.

The results of the search task show that in both experiments participants tended to perform better in the session in which the orientation of target and distractors did not match, as shown by the response accuracy being higher in the different-orientation session (in Experiment 1, compared to the same-orientation session) and different-orientation low-clutter session (in Experiment 2, compared to the same-orientation high-clutter session). Small effect sizes characterized both experiments, therefore we do not state that the different-orientation sessions were easier. It is likely that the staircase procedure, introduced to roughly homogenize the difficulty across sessions, contributed to such small effect sizes. To note, however, the effect size in Experiment 2 was slightly larger than in Experiment 1: this suggests the possibility that a lower amount of distractors in the different-orientation low-clutter session facilitated target detection.

Concerning the prime task, in both experiments, we found a validity effect both in reaction

times (shorter RTs in the valid condition than RTs in the invalid condition) and response accuracy (higher in the valid condition), therefore replicating the results on the basic "attentional capture effect" found by Reeder and Peelen (2013) and Reeder et al. (2015a). Interestingly, in both experiments we find that participants were more accurate in detecting the location of the dot when the silhouettes were rotated, compared to when they were upright. This might suggest that, overall, in the rotated silhouette condition attention was not captured as much as in the upright silhouette condition, allowing participants to make more accurate responses. In support of this interpretation, in the RTs of Experiment 2 we find an interaction between silhouette orientation and validity, with a larger validity effect in the upright condition than in the rotated condition. It is important to note that since this interaction was not present in Experiment 1, we cannot conclude with absolute certainty that rotated silhouettes captured attention to a less degree than upright silhouettes.

Taken together, these results suggest that observers tend to adopt a high-level category-based template when searching for cars and people in scenes, even in situations where adopting low-level templates might seem the most efficient and resources-saving strategy[3]. Therefore, the present study indicates that in naturalistic visual search, a template based on features similarity (Folk & Remington, 1998) is preferred to a relational target template that is shaped by the expected distractor's context (Becker, 2010).

Two possible explanations are available to interpret this pattern of results. First, it is possible that, by default, observers always establish a category-based template, because of the extensive experience gained throughout life. Being highly skilled at this, if the high-level strategy was automatic, then switching to a different strategy based on low-level characteristics, even if apparently effortless per se, might actually be more time consuming and demanding. On the other hand, it is also possible that the design of our study was not capable of eliciting or detecting a relational template strategy. For example, it is possible that our stimuli did not lead participants to form a specific template. In order to check this issue, we run an analysis on natural scenes (see Supplementary Materials) in which we tested the "verticality" and "horizontality" of each scene by measuring their directional gradients. We found that scenes with vertical distractors and scenes with horizontal distractors differed in terms of both their overall horizontality and verticality. However, it is also important to note that given the complexity of natural scenes, the analysis of gradients might be inappropriate, and a more complex analysis would be better suited. Alternatively, there might have been other variables linked to the scenes that we not took into account and that could explain

3 However, there is evidence that more specific templates are better in guiding search than more broader templates (Wolfe et al., 2004; Vickery et al., 2005), and that specific templates require less attentional resources.

the results. For example, it is possible that by making the search context “extreme” (e.g. desert vs. forest when searching for people), participants might have more easily adopted a low-level template. Furthermore, there is the possibility that simply participants did not have enough experience with our scenes and time to develop a relational template. Another potential flaw of the design is that we looked for a modulation of an effect that is already very subtle: the difference in RTs between valid and invalid trials is quite small, even if very consistent across subjects. Therefore, it might be that by employing a different approach, a relational template in naturalistic search could be detectable.

Studies on visual search have postulated that a mechanism known as “visual marking” (Watson & Humphreys, 1997) is actively and flexibly adopted to ignore old (already inspected) items, through top-down attentional inhibition mechanisms. In general, within this framework, one could consider expected items to gain the status of “old” items, and therefore to be actively deprioritized. The current study cannot disentangle whether participants were able to form specific expectations concerning the distractor’s context, and therefore it is not possible to state whether visual marking contributed to the results. Further research would be needed to investigate whether visual marking operates in such complex natural scenes, and whether such mechanism is influenced by the orientation and clutter of distractors.

In conclusion, our results suggest that in naturalistic visual search, observers prefer to adopt a strategy based on high-level, category-based template, even when the instantiation of a low-level template would seem more advantageous. However, our study is not conclusive on this matter, because other factors might have contributed to the absence of the hypothesized effect.

5. Supplementary Materials

We run a control analysis on natural scenes in order to check whether we manipulated the orientation of the distractors successfully. Specifically, we looked at whether natural scenes with vertical distractors had more verticality than natural scenes with horizontal distractors, and *vice versa*.

To this end, we calculated the directional gradients (Gx for x-direction, and Gy for y-direction) for each scene with vertical distractors (but no targets; n = 108) and each scene with horizontal distractors (but no targets; n = 108). Then, for each scene, a "verticality" measure was defined as the ratio of the y-gradient to the x-gradient (Gy/Gx), and a "horizontality" measure as the ratio of the x-gradient to the y-gradient (Gx/Gy).

Statistics were then computed to test whether scenes with vertical distractors had more verticality than scenes with horizontal distractors, and whether scenes with horizontal distractors had more horizontality than scenes with vertical distractors.

A one-tailed t-test revealed that scenes with vertical distractors had more verticality ($t(107) = 4.99$, $p < 0.0001$); however, scenes with horizontal distractors did not have more horizontality than scenes with vertical distractors ($t(107) = 0.65$, $p = 0.25$).

Chapter 3:
On the mechanisms of size constancy in natural vision: are attentional templates influenced by expected target distance?

1. Introduction

In everyday life, as we and the objects around us move, the image that is projected on the retina continuously changes. Despite this constant flux, we perceive stability in the world around us. One of the fundamental neural mechanism allowing this stability is perceptual size constancy (for a review, see Sperandio and Chouinard, 2015). This process operates by rescaling the size of an object as a function of its distance, enabling us to experience familiar objects as having constant size regardless of their distance and their retinal size. For example, when we watch a train departing from a station, even though its retinal size decreases, we do not perceive it as getting smaller, just more distant (for a more thorough discussion on the size-constancy mechanisms, see Chapter 5).

Interestingly, behavioral studies have shown that when we search for objects in scenes, we take into account their likely relative retinal size (Eckstein et al., 2017; Wolfe, 2017). Specifically, with a quick glance at a scene, the visual system can extract contextual information to infer the likely size of the target in relation to other objects (their size and position) and other cues such as depth, and guide attention to possible target objects that match the appropriate computed size. In their study, Eckstein et al. (2017) found that participants often missed targets whose size was inconsistent with the rest of the scene, even when such targets were very large and salient. This result indicates that, when searching for objects in scenes, the brain prioritizes those items that are at a spatial scale that is consistent with the surrounding context. They claimed that this strategy allowed to decrease false positives during search, that is, to rapidly discard those objects that visually resembled the target but were inappropriately sized. In line with this proposal, Sherman et al. (2011) found that "depth guidance" in scenes could reduce the set of candidate targets based on objects' relative size (Sherman et al., 2011). In sum, during real-world visual search tasks, it appears that we do not rely our search on an absolute, invariant representation of perceived size, but on a representation whose size is relative to the other objects in the scene.

These findings seem to be at odds with the evidence that the visual system preferentially represents the perceived size of objects (for a brief review, see Chouinard and Ivanowich, 2014), and more generally with size-constancy. How can be they reconciled? It is possible that the brain flexibly prioritizes objects according to their retinal size, while concurrently representing their

perceived size. But when, during the process of search, the size-invariant stored representation is replaced by a relative-size representation? Could it be during the preparatory phase, given the appropriate circumstances (e.g. expectations about target's distance)? From several studies investigating the characteristics of attentional templates, we now know that templates can be flexibly adjusted based on target-nontarget feature relations and expectations about the upcoming search context to optimize the selection process (Becker, 2010; Becker et al., 2010; Becker, 2013; Becker et al., 2013; Harris et al., 2013; Becker, 2014; Becker et al., 2014; Bravo and Farid, 2016; Schönhammer et al., 2016; Geng et al., 2017). So, is it possible that the size of preparatory attentional templates is shaped by expected target distance in scenes? Or are they characterized by size invariance, reflecting higher-order object representations in ventral ventral occipitotemporal cortex? In other words, do attentional templates represent an object's perceived size or retinal size? The two outcomes are equally plausible, and mutually exclusive. On one hand, it is possible that the size of templates is influenced by the expected target distance: if we expect distant targets the template could have a smaller size relative to when we expect near targets, in which case the template could be larger. This outcome would be in accordance with the suggestion by Eckstein et al. (2017), according to which a plausible brain mechanism for the context-target's size strategy could be reflected in baseline increases of those neurons that are tuned to the likely target sizes (that is, in preparatory attentional mechanisms; for a review, see Battistoni et al., 2017). On the other hand, it is possible that the size of templates remains unchanged by the expected target's distance. In support of this outcome, some studies have found that templates are characterized by some invariance, specifically concerning object's orientation (Reeder and Peelen, 2013), and size (Bravo and Farid, 2009). Since invariance might extend to size information, it is possible that size constancy is present already at the level of preparatory attentional templates, which would be in line with the idea that preparatory templates reflect higher-order object representations.

In this study, we sought to determine whether the size of attentional templates was influenced by expected target distance-size[1]. Participants searched for cars and people in natural scenes; crucially, these targets had either a big size and were positioned in the foreground (near), or had a small size and were positioned in the background (distant). Participants performed two different sessions: in one, only near-big targets were presented; in the other one, only distant-small targets were shown. This manipulation was intended to lead them to form expectations about the

1 I will refer to "target distance-size" because an object's distance and retinal size are inversely proportional, and thus cannot be disentangled: as the distance between the eyes and the object increases, its retinal size decreases. Unless specified, when I will use the word "size", I will refer to "retinal size".

distance and size of the targets. The size of attentional templates was probed by having them perform, on a subset of trials, a prime task in which small or big silhouettes of cars and people were presented. Participants were instructed to ignore these silhouettes and to detect a dot that could be presented at the location of either silhouette (Fig. 3). This paradigm, similarly to the previous chapter, was adapted from studies showing that participants were faster at detecting a dot presented at the position of the template-matching silhouette because their attention was captured by it (Reeder and Peelen, 2013; Reeder et al., 2015a). Therefore, concerning the current study, if the size of attentional templates was influenced by expected target's distance-size, then participants would be faster in detecting a dot that appeared at the location of the silhouette whose category matched the target template and whose size was consistent to the expected target size. To note, however, is that not finding such effect would not imply that the template is size invariant: the current study was designed to investigate whether attentional templates coded target size, but not whether they were size invariant. Therefore, if the expected results will not be observed, further study will be needed to determine whether templates are characterized by size invariance.

2. Materials and methods

2.1. Participants

Thirty healthy undergraduate and graduate students from the University of Trento participated in the experiment. All participants (25 women; aged 19-40 years, mean age M = 23.3 years, SD = 3.8 years) had normal or corrected-to-normal vision and provided written informed consent to take part in the study. Twenty-seven participants received monetary compensation (€8/session); three participants received course credits. The experiment was conducted in accordance with the Declaration of Helsinki and approved by the Ethical Committee of the University of Trento.

2.2. Stimuli

Stimuli were presented on a 19-inch Philips 109P4 monitor with a screen resolution of 1024 x 768 pixels and a monitor frame rate (refresh frequency) of 100Hz. Stimulus presentation was controlled with MATLAB 8.0 using the Psychtoolbox (Kleiner et al., 2007).

All stimuli were displayed on a grey background (RGB values: 148, 148, 148). The fixation point, a black plus ("+"), and letter cues (black 25-point uppercased Arial font), were presented at the center of the screen. The distance between the screen and eyes of participants was controlled by using a chinrest positioned at 55 cm from the monitor.

Natural scenes were grey-scaled and reduced to 427 (horizontal) x 320 (vertical) pixels (the original resolution was 640 x 480, then divided by 1.5; 640/1.5 = 427, 480/1.5 = 320), subtending 15.8 x 11.7 degrees of visual angle.
Silhouettes were black exemplars of cars and people on grey background (see Section 2.4. for more details). Masks had the same size of the scenes and were created by superimposing a naturalistic texture to white noise generated at different spatial frequencies, resulting in grey-scaled textures.
The dot stimulus was a black circle with a diameter of 7.5 pixels (0.3° of visual angle).
Natural scenes, masks, silhouettes, and dots were placed at a distance of 40 pixels from fixation.

2.3. Scene stimuli (naturalistic visual search task)

Three hundred seventy-eight scenes were selected from stimuli used in previous experiments and from Google Images. One hundred sixty-two scenes had targets in the foreground/near, and therefore they had big sizes: 54 scenes with cars, 54 scenes with people, 54 scenes with cars and people. One hundred sixty-two scenes had targets in the background/distant, and therefore they had small sizes: 54 scenes with cars, 54 scenes with people, 54 scenes with cars and people. Fifty-four scenes without cars or people were also selected. In order to increase the number of available scenes, each of these 378 scenes was horizontally-mirrored to create additional 378 novel scenes, and for each scene a copy was created, for a total of 1512 scenes. The scenes with near/big targets (648) were used in the Near Targets session, the scenes with distant/small targets (648) were used in the Distant Targets session, and the scenes without cars or people (216) were used in both the Near Targets session and in the Distant Targets session. Therefore, in each session a total of 864 scenes was used.

To ensure the validity of this manipulation, we measured the size of the target in each scene (quantified with the number of pixels on the vertical axis for person targets and on the horizontal axis for car targets), so that we were certain that the biggest target in the Distant Targets session was smaller than the smallest target in the Near Targets session. Figure 1 shows examples of scenes used in this task.

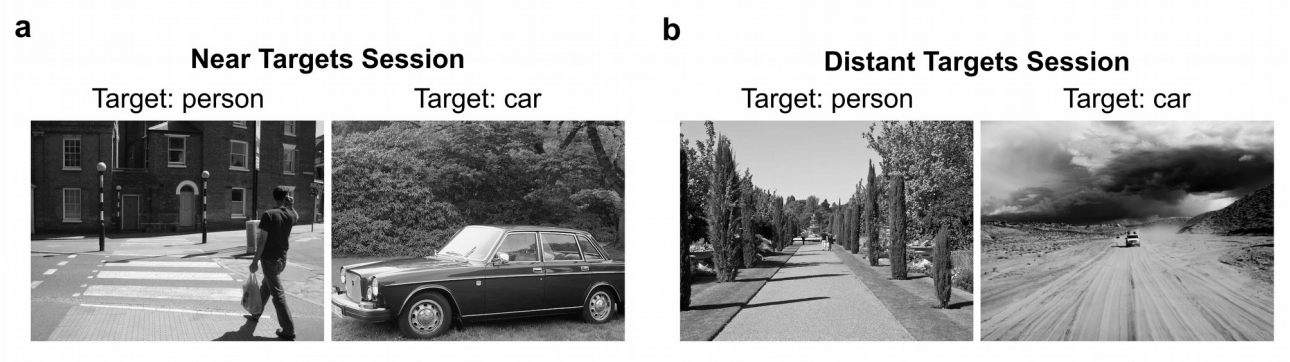

Figure 1. Examples of scene stimuli used in the naturalistic visual search task. (a) In the Near Targets session, targets in scenes had a big size and were generally located in the foreground. (b) In the Distant Targets session, targets had a small size and usually situated in the background.

2.4. Silhouette stimuli (prime task)

The stimuli used in the prime task were black silhouettes of cars and people on grey background. For each object category (car, person) and size (small, big), 144 different exemplars were selected, for a total of 576 silhouettes. Silhouettes were selected among those used in a previous study, and some were created with GIMP (https://www.gimp.org) starting from isolated exemplars of cars and people found in the Internet. These silhouettes were then manually edited with GIMP and MATLAB in order to match the average size of cars and people in scenes. In order to fulfil this, the size of all car and person targets in scenes was measured; specifically, pixels on the y-axis for persons and pixels on the x-axis for cars. The average size of big person targets was 360 pixels on the y-axis (mean: mean y-px in "person near big" scene, and mean of y-px of person in "car+person near big" scene), the average size of small person targets was 78 y-pixels (mean: mean y-px in "person far small" scene, and mean of y-px of person in "car+person far small" scene), of big car targets was 431 x-pixels (mean: mean x-px in "car near big" scene, and mean of x-px of car in "car+person near big" scene), and of small car targets was 84 x-pixels (mean: mean x-px in "car far small" scene, and mean of x-px of car in "car+person far small" scene). The size of the silhouettes was then adapted to approximately match these sizes, which measurement was 362 y-pixels for big person silhouettes, 79 y-pixels for small person silhouettes, 438 x-pixels for big car silhouettes, and 101 x-pixels for small car silhouettes. to create the final silhouette stimuli, these black silhouettes were placed within a rectangular grey background (RGB values: 148, 148, 148) with a resolution 640 x 480 pixels (not altering the size of the silhouettes). Silhouette stimuli were then reduced to 320 (=640/2) x 192 (=480/2.5) pixels to approximately match the size of the scenes. Figure 2 illustrates some examples of silhouette stimuli.

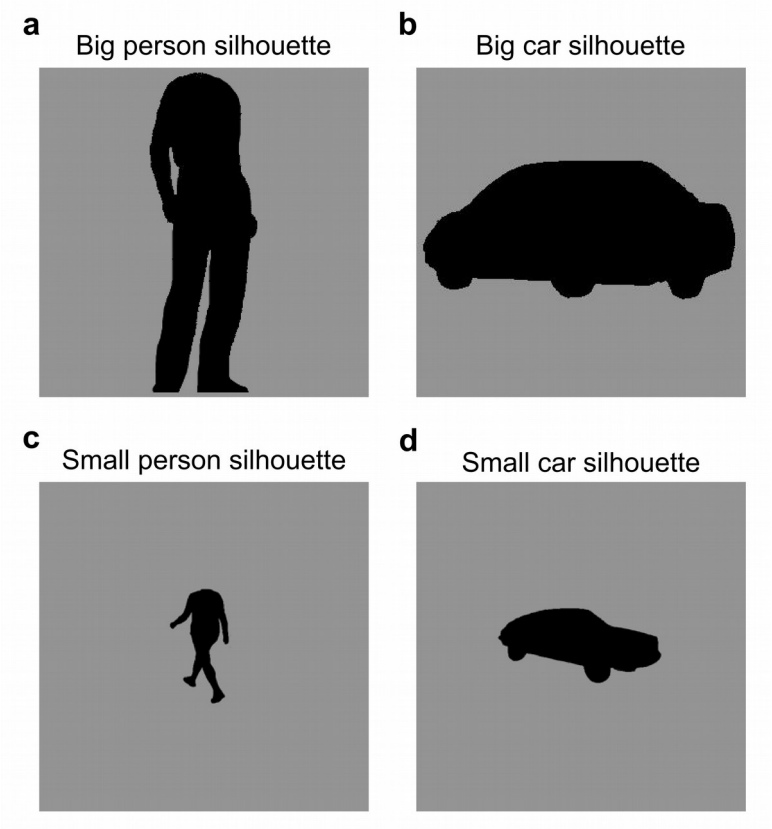

Figure 2. Examples of silhouette stimuli used in the prime task. (a) Big person silhouette; (b) big car silhouette; (c) small person silhouette; (d) small car silhouette.

2.5. Experimental procedure

The experiment consisted of two sessions of 45 minutes each, a "Near Targets" and a "Distant Targets" session. Each participant completed both sessions on separate days (the second session was completed within a week from the first session). The order of the session was counterbalanced across participants, so that even participants performed first the Distant Targets session and then the Near Targets session, and odd participants completed the Near Targets session as first and the Distant Targets session as second. Each session consisted of 9 blocks of 64 trials each. In each block, the naturalistic visual search task made up three-fourths (i.e., 48) of trials, to ensure that participants would actively prepare to detect the cued object category. One-fourth of trials in a block (i.e., 16) had the prime task. In half of the trials with the prime task (i.e., 8) the silhouettes were small; in the other half (i.e., 8) the silhouettes were big. The order in which the visual search task and the prime task appeared was randomized, therefore subjects could not predict what task they had to perform on any subsequent trial. To ensure that participants could establish some form of template (and that the prime task would not probe a non-existing representation), the first 3 trials in each block did not contain the prime task, but only the naturalistic visual search task. At the

beginning of the first session, each participant completed one practice block in order to familiarize with the tasks.

2.5.1. Naturalistic visual search task

In the visual search task, each trial started with a fixation point (500ms); which was replaced by a letter cue (for English speakers, “C” for “car”, and “P” for “person”; for Italian speakers, “M” for “macchina”, and “P” for “persona”; 500ms), followed by the fixation point (1000ms). Then, two scenes were presented on either sides of fixation (67ms; the combination of the scenes in each trial could be one of the following: (1) scene with car on the left, scene with person on the right; (2) scene with person on the left, scene with car on the right; (3) scene with car and person on the left, scene with no car and no person on the right; (4) scene with no car and no person on the left, scene with car and person on the right). The two scenes could not be variations (i.e., mirrored-versions or copies) of the same scene. An empty screen appeared after the scenes, the duration of which was manipulated with a staircase procedure: the duration of the empty screen in the first 5 trials in each block was fixed at 100ms; then, if the average accuracy in the visual search task (in the current block, the average accuracy of the search trials up to that point) was higher than 75%, the empty screen duration decreased by 20ms; whereas, if the accuracy was lower than 75%, the duration of the empty screen was increased by 20ms. The minimum duration of the empty screen could be 10ms, the maximum duration 300ms. This was done to ensure that the difficulty of the search task was more or less balanced across the two sessions. After the empty screen, two masks appeared at the location where the two scenes were presented (350ms); a fixation point followed (1660ms), then the feedback (“+0” for incorrect responses; “+1” for correct responses; 500ms). The task of the participant was to indicate in which scene (left or right) the cued object category was present (by pressing on the keyboard one of the two keys “z” or “m”, for left or right target, respectively).

2.5.2. Prime task: trial sequence

The first events of the trials with the prime task matched the trials with the visual search task: they started with a fixation point (500ms), followed by the letter cue (500ms) and the fixation point (1000ms). Instead of the scenes, two silhouettes were presented on either sides of fixation: one silhouette of a car on one side, and one silhouette of a person on the other side (67ms). The two silhouettes had congruent size (i.e., either both big or both small). A fixation point was presented briefly (50ms), followed by the appearance of a dot on one side of fixation (at the location of one of the two silhouettes, 100ms). A fixation point followed (1660ms), then the feedback (“+0” for

incorrect responses; "+1" for correct responses; 500ms). Participants were instructed to ignore the silhouette and press a key to indicate whether the dot appeared to the left or right of fixation ("z" for left, "m" for right").

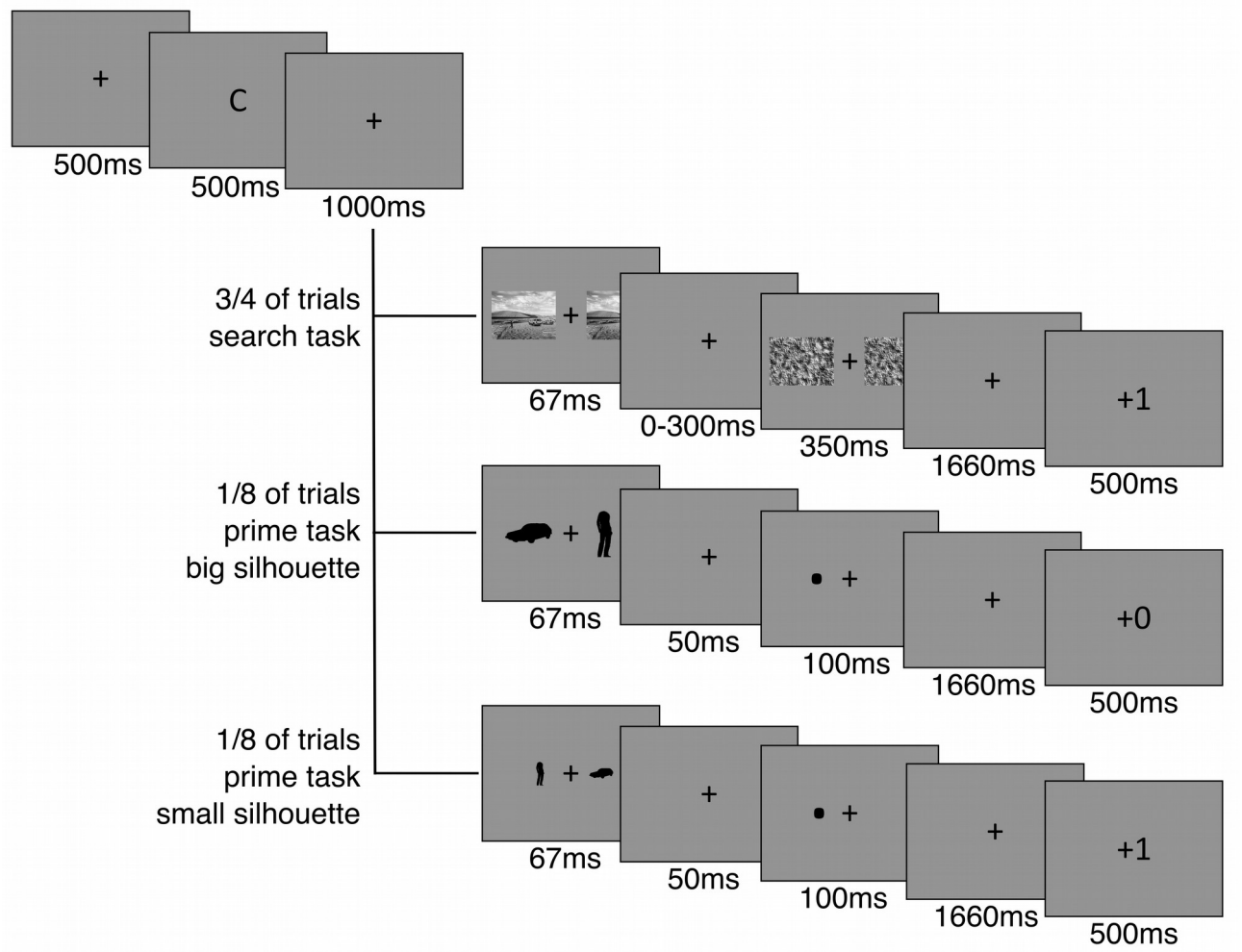

Figure 3. Schematic outline of the experimental procedure. Each block consisted of 64 trials. 75% (¾ in each block) of trials included a naturalistic visual search task, where participants indicated whether the cued target category was present in the scene to the left or right of fixation. 25% (¼ in each block) of trials consisted of a prime task, in which participants indicated whether the dot appeared to the left or right of the fixation point. Half of the prime task trials had big silhouettes (⅛ in each block), the other half had small silhouettes (⅛ in each block).

2.6. Analysis

First, we analyzed RTs and response accuracy from the naturalistic visual search task to test whether there was a difference in performance between the Near Targets session and the Distant Targets session. We expected participants to perform better in the Near Targets session because targets were bigger in size (compared to the Distant Targets session) and in the foreground (as opposed to the

background in the Distant Targets session), and hence likely to be more easily detected. In the RTs analysis, only trials with correct responses were included.

Next, to address the main question of whether expectations about the distance-size of targets influences the characteristics (i.e., size) of attentional templates, we analyzed RTs and response accuracy from prime task trials. Data was analyzed as a function of (1) validity and (2) consistency. The valid condition consisted of trials in which the dot appeared at the location of the silhouette whose category matched the target template; the invalid condition consisted of trials in which the dot appeared at the opposite location (i.e., at the location of the silhouette whose category did not match the template). The consistent condition comprised trials in which big silhouettes appeared in the near targets session and small silhouettes appeared in the distant targets session (i.e., silhouettes and targets had congruent size); RTs and accuracy was averaged across the conditions near target - big silhouette and distant target - small silhouette. The inconsistent condition was made up of trials in which big silhouettes appeared in the distant targets session and small silhouettes appeared in the near targets session (i.e., silhouettes and targets had incongruent size); RTs and accuracy was averaged across the conditions near target - small silhouette and distant target - big silhouette. Figure 4 shows examples of the four conditions. Non-given responses or too-slow responses (i.e., responses not given before the end of the fixation screen lasting 1660ms) were considered as incorrect. Incorrect responses were not included in the analysis of RTs.

In the current study, if participants adopt a template whose size is congruent with the expected target distance-size, we expect an interaction between validity and consistency in both RTs and response accuracy: we predict a smaller validity effect in the inconsistent size condition than in the consistent size condition because attention would be captured to a less degree to silhouettes whose size is inconsistent with the expected target size, allowing them to be more accurate and therefore showing a smaller difference in accuracy between invalid and valid trials.

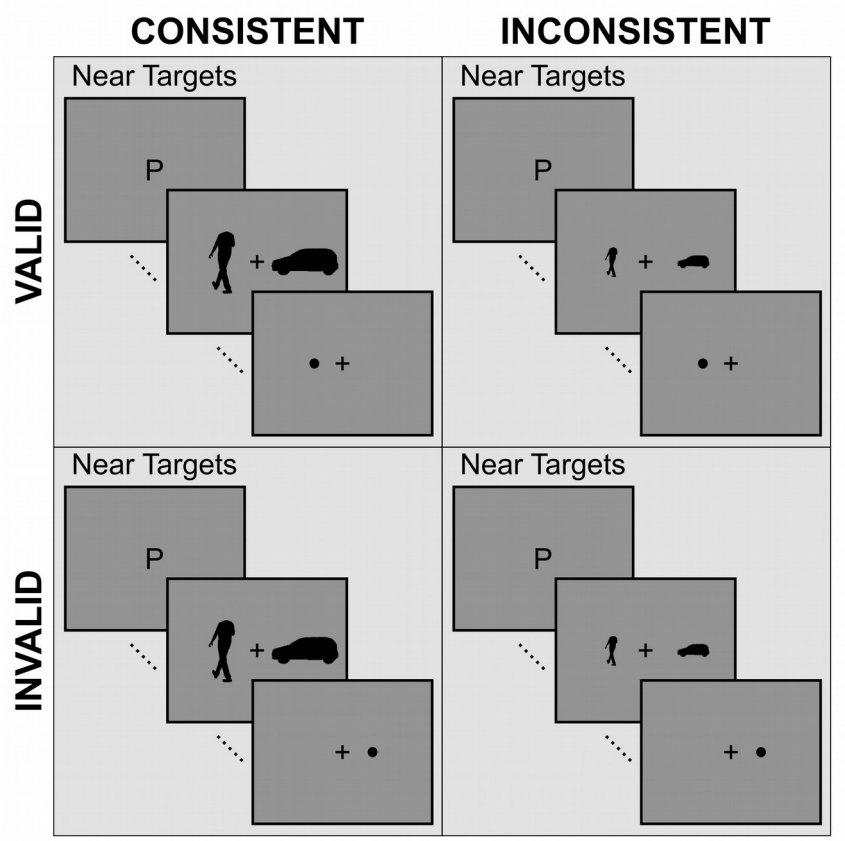

Figure 4. Examples of the four conditions.

3. Results

3.1. Naturalistic Visual Search task

A two-tailed t-test on the RTs in the distant targets session (mean = 622 ms, SD = 105 ms) and RTs in the near targets session (mean = 559 ms, SD = 96 ms) revealed that participants were significantly slower in the distant targets session than in the near targets session ($t(29) = 3.25$, $p < 0.01$, Cohen's $d = 0.20$). A two-tailed t-test on response accuracy in the distant targets session (mean = 74.5%, SD = 6.5%) and in the near targets session (mean = 87.9%, SD = 4.2%) showed that participants were significantly more accurate in the near targets session ($t(29) = 13.38$, $p < 0.0001$, Cohen's $d = 0.58$). Figure 5 illustrates the results.

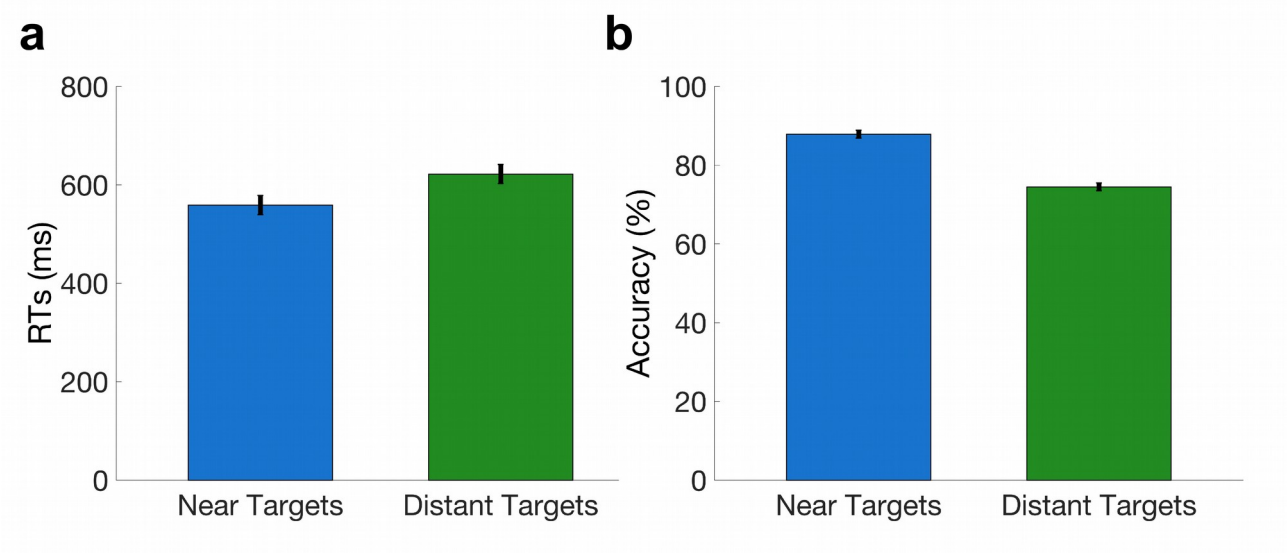

Figure 5. Results of the naturalistic visual search task. (a) mean RTs and (b) mean response accuracy; error bars represent within-subjects SEM.

3.2. Prime task

To test whether the size of attentional templates changed as a function of expected target distance-size, we performed a two-way repeated-measures ANOVA on both RTs and response accuracy, with factors (1) validity of dot position (valid vs. invalid) and (2) consistency of silhouette's and target's size (consistent vs. inconsistent). The ANOVA on RTs revealed a significant main effect of validity ($F(1,29) = 73.12$, $p < 0.0001$; Cohen's $d = 0.15$), with shorter RTs in the valid condition (mean = 380ms, SD = 50ms) than in the invalid condition (mean = 410ms, SD = 50ms). There was no main effect of consistency ($F(1,29) = 0.94$), and no interaction between consistency and validity ($F(1,29) = 1.09$). The ANOVA on response accuracy revealed a significant main effect of validity ($F(1,29) = 38.91$, $p < 0001$, Cohen's $d = 0.34$), with higher accuracy in the valid condition (mean = 98%, SD = 3%) than in the invalid condition (mean = 90%, SD = 8%). It also revealed a significant main effect of consistency ($F(1,29) = 9.50$, $p < 0.01$, Cohen's $d = 0.05$), with higher accuracy in the inconsistent condition (mean = 95%, SD = 5%) than in the consistent condition (mean = 93.7%, SD = 5.7%). A significant interaction between validity and consistency was present ($F(1,29) = 9.88$, $p < 0.01$, Cohen's $d = 0.07$), where there was a bigger difference between valid and invalid conditions in the consistent condition than in the inconsistent condition (two-tailed t-test, $t(29) = 3.14$, $p < 0.01$). Figure 6 illustrates RTs and response accuracy of the four conditions.

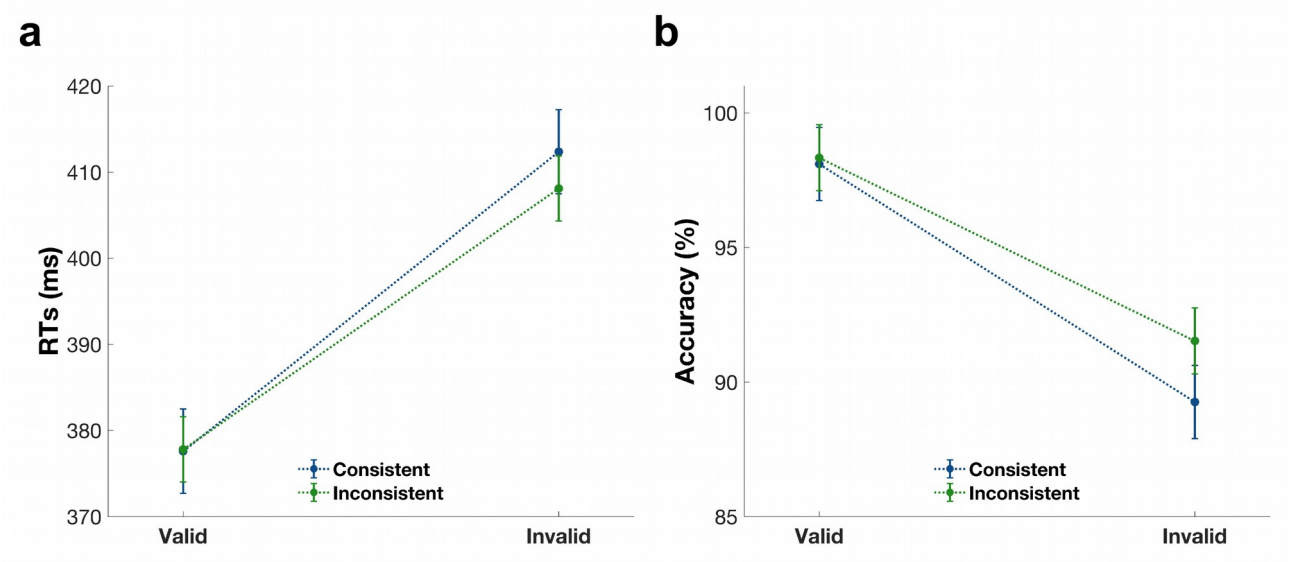

Figure 6. Results of the prime task: (a) Reaction times (RTs), (b) response accuracy. Circles represent means, error bars are illustrated as +-1 within-subjects SEM with respect to the mean.

4. Discussion

This study aimed to determine whether the size of attentional templates was influenced by expected target's distance-size in scenes. On the majority of trials, participants performed a naturalistic search task in which the targets (cars or people) were placed either in the foreground or in the background, and therefore had a big or small size, respectively. The distance-size of targets was blocked within a session, allowing participants to form expectations on their size. On a subset of trials, participants performed a prime task in which, instead of the scenes, silhouettes of cars and people were presented, followed by a dot that could match the position of either silhouette. Importantly, silhouettes could be either both big or both small. Participants were instructed to ignore the silhouette and indicate the position of the dot.

First of all, this study replicates previous findings on the basic attentional capture effect (Reeder and Peelen, 2013; Reeder et al., 2015a), as shown by the result of a significant validity effect (Fig. 6a). Specifically, participants were faster in detecting a dot that was presented at the location of the template-matching silhouette than when it was presented at the location of the silhouette that did not match the template. This effect can be directly linked to the finding that attention is captured toward the items that correspond to the current content of working memory (Downing, 2000). Concerning the present results, the difference in RTs is due to the fact that the matching between the template and the silhouette provided an advantage in detecting a dot that was presented immediately after at that location (i.e., in valid trials). In invalid trials, participants

were slower because they had to disengage their attention from the location of the template-matching silhouette and move it to the opposite location, to be able to make a response.

Despite the presence of a significant validity effect in RTs, we did not find an effect of consistency nor an interaction between validity and consistency (Fig. 6a). This indicates that the speed of dot detection did not change as a function of the congruency between silhouette size and expected target size. Therefore, RT results do not support the hypothesis that the size of the attentional template is influenced by expected target size. The absence of the expected effects in RTs might be due to the fact that the validity effect in RTs is already a small effect on its own, like in previous results (Reeder and Peelen, 2013), and it is possible that further modulations would go undetected. Alternatively, the design of this study might have not been adequate to reveal whether the template was shaped by the expected target distance.

The results on response accuracy highlighted a main effect of validity, with better accuracy in valid trials than invalid trials. Furthermore, participants were more accurate in inconsistent trials, that is, when the size of the silhouette did not match the size of the expected target. This is in line with expectations: attention was captured to a less degree to the silhouette whose size did not match the expected target size, allowing participants to be more accurate than in consistent trials, where the size of the silhouette matched the size of the expected target, leading to more attentional capture and therefore worse performance. Furthermore, the interaction between validity and consistency, highlighting a bigger difference in response accuracy between valid and invalid trials in the consistent condition than in the inconsistent condition, supports the previous result, and suggests that templates might be influenced by expected targets size. However, we should be careful in drawing conclusions from these results, because effect sizes were rather small. If the present findings were corroborated by future studies, then they would suggest that the preparatory templates are not size invariant: the current results indicate that the contents of attentional templates appear to match the (expected) retinal size of objects, not their perceived size, providing support to behavioral studies showing that, in naturalistic visual search, an object's retinal size (i.e. the physical size of the object relative to the other objects in the scene) is an important guiding attribute (Sherman et al., 2011; Eckstein et al., 2017; Wolfe, 2017).

Interestingly, the present results raise a few questions on the neural basis of this relative-size template. Many studies on size constancy and, more generally, on size, have consistently found that the visual system preferentially represents the perceived size of objects, not their retinal size (Murray et al., 2006; Sterzer and Rees, 2006; Fang et al., 2008; Liu et al., 2009; Cate et al., 2011; Konkle and Oliva, 2011; Schwarzkopf et al., 2011; Amit et al., 2012; Konkle and Oliva, 2012;

Sperandio et al., 2012; Pooresmaeili et al., 2013; Chouinard and Ivanowich, 2014; Gabay et al., 2016). All these studies were centered on stimulus-evoked neural activity, but evidence suggests that content-specific preparatory attentional mechanisms involve the activation of the same regions and patterns that are observed in stimulus-evoked responses (for a review, see Battistoni et al., 2017). Therefore, the results of this study might challenge the notion that the brain represents only the perceived size of objects in situations in which it is possible to form expectations on their likely retinal size. Given the present finding and studies showing that retinal size matters (Sherman et al., 2011; Eckstein et al., 2017; Wolfe, 2017), and considering the low temporal resolution of fMRI, it is possible that those studies did not capture the full temporal course of size-constancy mechanisms. More specifically, it is possible that the brain initially represents for a short time the retinal size of objects in scenes, and later in time feedback connections from higher-level regions "adjust" the activity in lower level areas, inhibiting neural activity that is not consistent with the stored knowledge regarding an object's perceived size. The temporal dynamics of size-constancy could therefore be explained within a predictive coding framework (see Chapter 5 for an investigation of the temporal dynamics of size-constancy).

In addition, the present results bring forward an important issue: if preparatory attentional templates contain information about object's retinal size, and not their perceived size, where in the visual system are they represented? It is possible that higher-order areas "shrink" the representation of cars and people in the object-selective cortex (OSC; in lateral-occipital areas), where they have been previously found to be represented (Peelen and Kastner, 2011).

Crucially, it will be important to replicate the present findings, and establish, with a different design (possibly including the condition in which objects in scenes have the same retinal size but different perceived size of silhouettes), wether templates are affectively size invariant.

In conclusion, this study suggests that preparatory attentional templates for real-world objects might code information related to their expected retinal size. This finding, like other studies, highlights that in visual search an object's retinal size matters, contrary to what might be expected from imaging studies highlighting that the visual system preferentially represents objects' perceived size. In fact, this apparent controversy might be simply due to the fact that fMRI studies did not allow to determine the full temporal resolution of size-constancy mechanisms. It is plausible to think that higher-order visual areas represent real-world size (Konkle & Oliva, 2012), but it is possible that lower-level areas might, for a short time, briefly code information related to an object's retinal size.

Chapter 4:
Spatial attention follows category-based attention during naturalistic visual search: evidence from MEG decoding[1]

1. Introduction

Top-down attentional selection serves to deal effectively with the large amount of visual information present in everyday environments. It does this by prioritizing processing of goal-relevant stimuli (e.g., cars when crossing a road) and ignoring goal-irrelevant stimuli (e.g., trees when crossing a road). To study this top-down selection mechanisms in the laboratory, many studies have used the visual search paradigm, in which participants are instructed to find simple stimuli, such as oriented bars or colored circles, amongst a variety of distractors (Wolfe and Horowitz, 2004). The use of these artificial displays allows for careful control over variables such as the specific position of targets and distractors and the features that distinguish the target from the distractors. This approach has been fruitfully used in M/EEG studies to reveal the temporal evolution of attentional selection in a variety of visual search paradigms. One of the findings from these studies is that the top-down selection of a peripheral target evokes a lateralized response over occipitotemporal and parietal areas, peaking around 200-300 ms after stimulus onset (Luck and Hillyard, 1994; Eimer, 1996; Hopf et al., 2000; Hickey et al., 2009).

However, visual search in simplified displays differs in many ways from visual search in real life. For example, naturalistic search is typically for a familiar object or object category (e.g., “cars”) rather than a visual feature. These target objects appear in scenes that are usually cluttered with a variety of distractors that share many low-level features with the target. Furthermore, the visual properties of target and distractor objects in real-world scenes vary as a function of lighting, perspective, occlusion, and distance.

Despite the apparent complexity of naturalistic search, search in natural scenes is surprisingly efficient (Thorpe et al., 1996; Wolfe et al., 2011b). There are several reasons for this efficiency. For example, real-world scenes provide a rich visual context that provides information about likely target features (e.g., objects that are far away appear smaller) and likely target locations (e.g., cars appear on roads). Furthermore, objects in natural scenes are positioned in

1 This work has been published elsewhere: Spatial attention follows category-based attention during naturalistic visual search: evidence from MEG decoding. Elisa Battistoni, Daniel Kaiser, Clayton Hickey, Marius V. Peelen. bioRxiv 390807; doi: https://doi.org/10.1101/390807

regular configurations, allowing for the grouping of objects into meaningful chunks (Kaiser et al., 2014). The many differences between artificial and naturalistic visual search highlights the importance of examining the temporal evolution of attentional selection in naturalistic conditions.

A recent magneto-encephalography (MEG) study from our group (Kaiser et al., 2016) took this approach, investigating the time course of object category processing in natural scenes as a function of task relevance. In this study, participants detected either cars or people in a large set of natural scenes. Importantly, the same set of scenes was shown in both tasks, such that objects (cars, people) appeared both as targets and as distractors. Data were analyzed using multivariate pattern analysis (MVPA), decoding the processing of within-scene objects using a classifier trained on data from a separate experiment in which isolated exemplars of cars and people were shown.

Averaged across conditions, the category of the objects present in scenes could be decoded from around 180ms after stimulus onset. Crucially, this early stage of decoding fully depended on the behavioral relevance of the object: early decoding was only possible when the object was the target of the search. These findings show that top-down attention quickly modulates the processing of object category, at around the time that object categories are first being extracted from scenes. However, unlike earlier M/EEG studies investigating attention in simplified displays, the study by Kaiser et al. was not designed to provide information about the spatial component of attentional selection.

In the present study, we closely followed the approach of Kaiser et al. but with significant changes that allowed for measurement of the spatial component of attentional selection. First, targets (cars, people) were presented either in the left or right hemifield so as to elicit lateralized processing. Second, our procedure included an independent experiment in which participants performed a simple detection task in an artificial, non-naturalistic display, reporting the presence of a cross in the left or right hemifield (Fig. 1). We used data from this experiment to train a multivariate classifier to categorize the deployment of spatial attention to the left or right with high temporal resolution. We subsequently used this classifier to detect the deployment of attention to the left or right in data from the main experiment in which participants detected examples of real-world objects in natural scenes.

The objects in the scenes varied as a function of behavioral relevance, allowing us to determine when the location of targets was better decoded than the location of distractors. Importantly, because participants were cued to search for either cars or people in each trial, the same scene stimuli appeared in both target-present and target-absent trials, allowing us to examine neural activity elicited by identical stimuli as a function of whether they currently served as target

or nontarget. To exclude the contribution of low-level visual priming, the attended category was symbolically cued, varied on a trial-by-trial basis rather than in blocks, and scenes did not repeat across trials. The cross-decoding approach furthermore allowed us to exclude the contribution of unintended confounds present in natural scenes and thus to relate the deployment of attention to naturalistic stimuli with the deployment of attention to carefully controlled artificial stimuli.

Our findings show that spatial attention is deployed to the target (relative to the distractor) from around 240ms after stimulus onset. Interestingly, information about target presence itself was available from 180ms after stimulus onset, at the same time as the category-based modulation observed by (Kaiser et al., 2016). We conclude that spatial attention follows category-based attention during naturalistic visual search.

2. Materials and methods

2.1. Participants

Data were acquired from 42 healthy participants with normal or corrected-to-normal vision (19 male, mean age M = 26.36 years, SD = 3.75 years). All participants gave informed consent and received monetary compensation. The experiment was conducted in accordance with the Declaration of Helsinki and approved by the Ethical Committee of the University of Trento. Because of a technical problem, no behavioral data was collected for the first three participants.

2.2. General experimental procedure

While recording MEG data, participants performed two experiments: a naturalistic visual search experiment in which they detected cars, people or trees in naturalistic scenes (Fig. 1a,c), and a physical salience experiment where they detected the presence of a cross that was made physically salient by converging line elements (Fig. 1b,d). The physical salience experiment was designed to isolate location-specific brain activity patterns, which were used as the training dataset for multivariate classifiers (see below). The full experimental session lasted 80 minutes. Stimuli were back-projected onto a translucent screen located 115cm from the participants. Stimulus presentation was controlled using Matlab 8.0 and the Psychtoolbox (Kleiner et al., 2007).

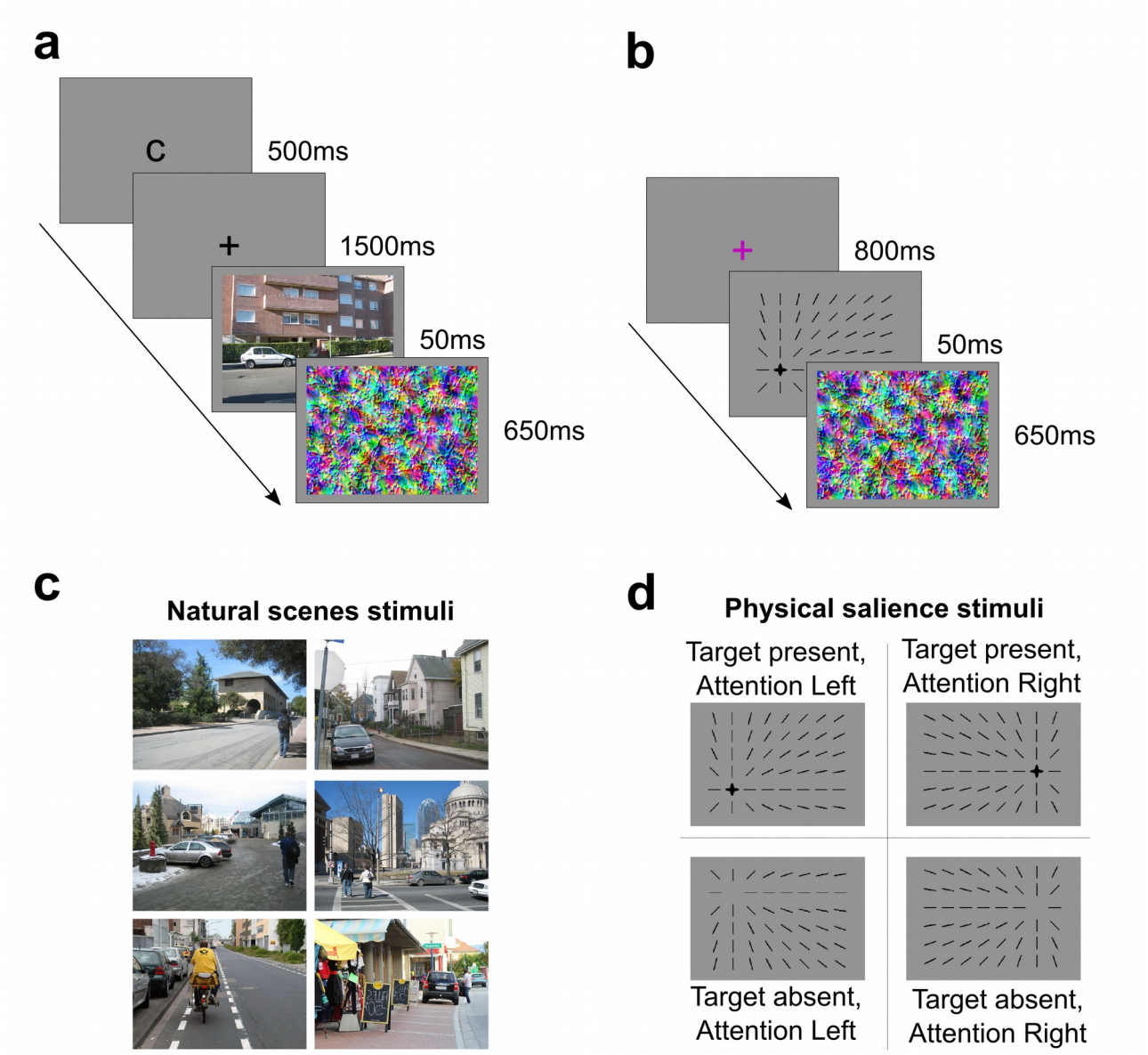

Figure 1. Experimental paradigms. Schematics of paradigms used in (a) naturalistic visual search experiment and (b) physical salience experiment. Example stimuli of (c) the naturalistic visual search experiment and (d) the physical salience experiment.

2.3. Naturalistic visual search experiment

In the naturalistic visual search experiment participants reported the presence or absence of a cued target category (cars, people, or trees) in briefly presented photographs of natural scenes by pressing one of two buttons. Participants performed 12 blocks of 48 trials each. The mapping of button to target presence and absence was counterbalanced across participants. As illustrated in Fig. 1a, a letter cue (500ms) displayed at the beginning of every trial indicated the target category (for English-speakers “C” indicated “car”, “P” “person” and “T” “tree”; for Italian-speakers, “M” indicated

“macchina”, “P” “persona” and “A” “albero”). After a fixation interval (a “plus” symbol; 1500ms), a natural scene was presented briefly (50ms), and followed by a perceptual mask (650ms). After an additional fixation interval (500ms), participants received feedback (displayed for 500 ms) consisting of 1 or 100 points for correct performance and 0 points for incorrect performance (points were converted to money at the end of the experiment). Trials were separated by a randomly jittered inter-trial interval (rectangular distribution; 1000ms to 2000ms). The average trial duration was 5.2s. The reward feedback manipulation (1 or 100 points) was employed to test a question regarding the effect of reward association on the processing of objects when these appear as distractors. The relevant trials in this context were those where participants searched for trees; these trials are not analyzed or further treated in the current paper. The trials of current interest were those in which participants were cued to detect either cars or people in photographs of real-world scenes that could include one or more exemplars of cars and people (Fig. 1c). Ninety-six scenes contained either cars or people, located on the left or right of the scene. An additional 48 scenes contained both categories (cars and people), where in 24 scenes the two categories appeared on the same side, and in the other 24 scenes they appeared on different sides. In total, the stimulus set consisted of 288 scenes. During the experiment, each scene was presented once in its original version and once flipped horizontally, leading to a total of 576 unique scenes. All pictures were reduced to 480 (vertical) x 640 (horizontal) pixels, subtending 13.5° x 10° of visual angle. Masks of the same size as the scenes (n = 576) were created by superimposing a naturalistic texture to white noise generated at different spatial frequencies, resulting in colored textures. All stimuli were presented centrally and displayed on a grey background.

2.4 Physical salience experiment

In the physical salience experiment, participants reported the presence or absence of a cross by pressing one of two buttons. The location of the cross was made salient by converging line elements (Fig. 1d) to mimic global contextual cues in natural scenes. Participants performed 2 blocks of 80 trials. The mapping of button to target presence and absence was counterbalanced across participants. Fig. 1b shows the trial structure. After a fixation interval (a pink “plus” symbol presented for 800ms), the line array was presented for 50ms, followed by a perceptual mask (650ms). Trials were separated by a randomly jittered inter-trial interval ranging from 2200ms to 3000ms. The perceptual mask, its timing, and the timing of the stimulus, were identical to those in the naturalistic visual search experiment. Stimuli consisted of 48 black lines on a grey background (displayed on 6 imaginary rows and 8 imaginary columns, each subtending about 1.5° of visual

angle), drawn within an area of 13.5° x 10° of visual angle (Fig. 1b). The lines made a position in the display salient; in half of the trials a black cross (the target, of size 1.5° x 1.5°) was presented at this location and in half it was absent. The position of the area within which the target could appear was counterbalanced across 8 possible locations: within the second column (i.e. on the left) or the seventh column (i.e. on the right), the target could appear in one of four positions (in the second, third, fourth or fifth row). All stimuli were displayed on a grey background.

2.5 MEG data acquisition and preprocessing

Neuromagnetic activity was recorded using a whole-head MEG system with 102 magnetometers and 204 planar gradiometers (Elekta Neuromag 306 MEG system, Helsinki, Finland). Data were acquired continuously (with online sampling rate of 1000 Hz) and band-pass filtered online between 0.1 and 300 Hz. Offline preprocessing was performed using MATLAB 8.0 and the Fieldtrip toolbox (Oostenveld et al., 2011). Data were epoched from -200 to 500 ms with respect to stimulus onset. No offline filter was applied to the data[2]. Based on visual inspection, and blind to condition, trials and channels containing artifacts (i.e., blinks, eye-movements, or unusually large peak-to-peak amplitudes) were discarded from subsequent analysis. All trials (correct and incorrect) were included in the analysis. Next, data were baseline corrected with respect to the pre-stimulus period (with baseline from -200ms to 0ms) and down-sampled to 100Hz to improve signal-to-noise ratio (Grootswagers et al., 2017). Data from rejected channels were interpolated based on the average of neighboring sensors of the same type.

2.6 MEG multivariate pattern analysis

All multivariate classification analyses were performed using MATLAB 8.0 and the CoSMoMVPA toolbox (Oosterhof et al., 2016). Single-trial classification was performed separately for every 10ms time bin of the evoked field data of all magnetometers; only data from magnetometers were used as these sensors offered more reliable classification performance than gradiometers in a comparable study (Kaiser et al., 2016). To increase the signal-to-noise ratio, 1000 synthetic trials were created for every condition in both the training and testing sets. Each synthetic trial was created by randomly selecting 5 trials and averaging across these trials. Trials were selected without replacement until the pool of trials was exhausted, such that each trial contributed to a roughly

2 In a previous version of the analysis we applied an offline high-pass filter. This revealed earlier attention effects than without the filter (reported here), possibly reflecting filtering artifacts (Acunzo et al., 2012). These early attention effects did not emerge in subsequent analyses (e.g., reverse cross-decoding, see footnote 3) and were thus deemed unreliable.

equal number of synthetic trials. Classification accuracy was evaluated by computing the percentage of correct predictions of the classifier. The decoding analysis was repeated for every possible combination of training and testing time, leading to a 50x50 points (i.e. 500 ms x 500 ms with 100Hz resolution) matrix of classification accuracies for every participant. Single-subject accuracy matrices were smoothed using a 3x3 time points averaging box filter (i.e. 30 x 30ms, for the training and testing times, respectively); single-subject accuracy matrix diagonals were smoothed with a 3-point (30 ms) boxcar filter. To determine time periods of significant above-chance classification, a threshold-free cluster enhancement procedure (Smith and Nichols, 2009) was used with default parameters. The multiple-comparisons correction was based on a sign-permutation test with null distributions created from 10,000 bootstrapping iterations and a significance threshold of $Z > 1.64$ (i.e., $p<0.05$, one-tailed).

2.6.1. Within-experiment decoding analyses. A within-experiment decoding procedure was employed to test whether Linear Discriminant Analysis (LDA) classifiers could reliably discriminate MEG activity patterns evoked by stimuli in the left vs. right hemifield. This procedure was performed once within the physical salience experiment and once within the naturalistic visual search experiment. To this end, each of the datasets was divided into two independent subsets of trials, one of which was used as training set and the other as testing set.

2.6.2. Cross-decoding analyses. In the cross-decoding analysis, LDA classifiers were trained to discriminate between two conditions of interest in the physical salience experiment (MEG patterns evoked by left vs. right stimuli) and employed to discriminate between conditions in the independent naturalistic visual search experiment (MEG patterns evoked by left vs. right objects in natural scenes; Fig. 2a)[3]. This procedure was performed separately for each time point. Classifier testing was performed as a function of the behavioral relevance of objects in scenes, with identical scenes appearing in both target and distractor conditions across participants (Fig. 2b). The difference of the decoding time courses for target and distractor conditions was then tested against zero. It should be noted that the classifier trained on the physical salience experiment can use

3 The cross-decoding analysis was also performed in the reverse direction: training the classifier on the main experiment (separately for target and distractor trials) and testing on the physical salience experiment. This yielded similar results as the analysis reported here: The target-distractor difference on the diagonal emerged after 200 ms, with significant differences from 210ms to 250ms and from 280ms to 380ms. When averaging results of the two cross-decoding analyses, the target-distractor difference was significant ($p<0.05$, corrected for multiple comparisons) from 220ms to 390ms and from 430ms to 500ms.

activity patterns driven by both physical asymmetries and spatial attention shifts. Crucially, however, these can be disentangled in the naturalistic search experiment: in the main comparison, between target and distractor decoding, the same scenes are included as targets and distractors, thus eliminating the contribution of any physical asymmetries.

2.6.3. Searchlight analyses. To explore the approximate anatomical location of target and distractor processing, a sensor-space searchlight analysis was performed on consecutive 50ms time windows ranging from 0ms to 500ms post-stimulus. The cross-decoding procedure was performed across the scalp using sensor neighborhoods of 20 sensors each (Kaiser et al., 2016). Each of these neighborhoods was created by defining a neighborhood of 10 adjacent sensors in the left hemisphere that was symmetrically mirrored with corresponding sensors in the right hemisphere, resulting in bilaterally symmetric maps. The searchlight was performed for each 10ms time point, and the results of the individual time points within each 50ms window were averaged to obtain a single searchlight map for that window.

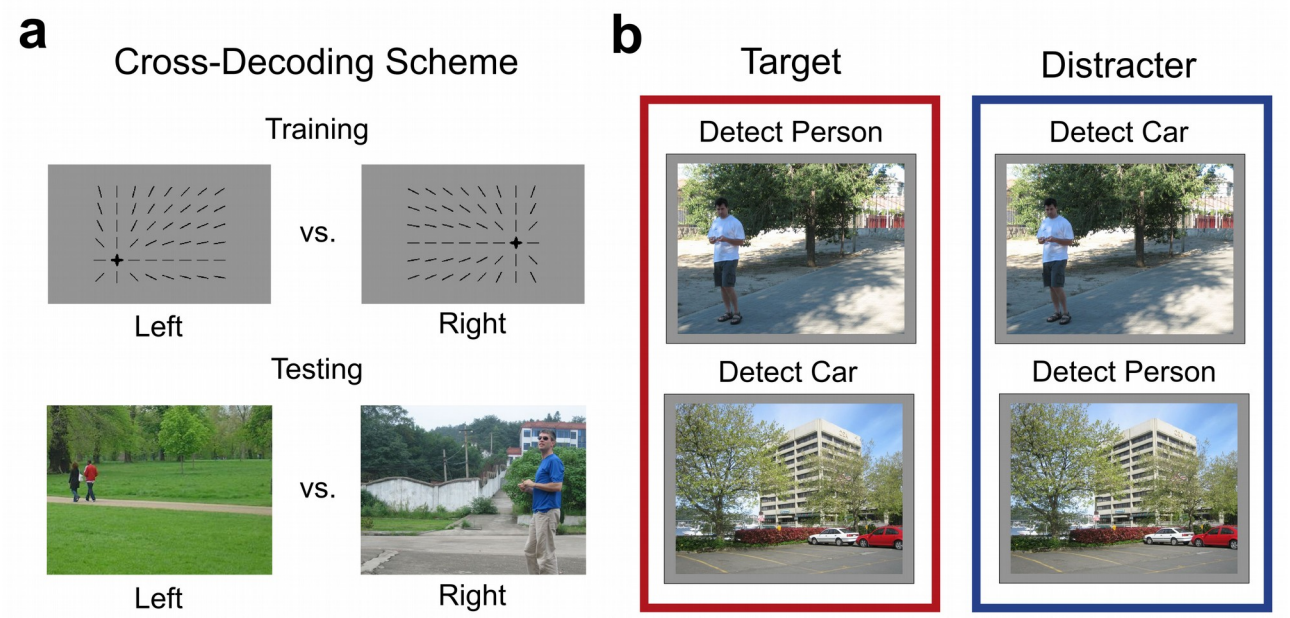

Figure 2. Analysis procedure. Using a cross-decoding approach (a), multivariate classifiers were trained on data from the salience experiment and tested on data from the naturalistic visual search experiment. Classifier testing was performed separately for target and distractor locations (b). Note that the same scenes could appear as target or as distractor, with the only difference being the top-down set of the participant on that trial.

3. Results

3.1. Behavioral results

Behavioral performance in the naturalistic search experiment showed that the task was sufficiently challenging, with an average response accuracy of around 75% (target present trials: 81%, SD = 10%; target absent trials: 68%, SD = 17%). The average RT was around 500 ms (target present trials: 444ms, SD = 73ms; target absent trials: 555ms, SD = 79ms).

In the physical salience experiment, response accuracy was around 58% (target present trials: 64%, SD = 20%; target absent trials: 52%, SD = 17%). The average RT was around 620 ms (target present trials: 585ms, SD = 190ms; target absent trials: 659ms, SD = 219ms).

3.2. Within-experiment decoding results

To determine that MEG activity patterns contained decodable information, and in this way ensure the feasibility of the cross-decoding procedure, in a first analysis we checked whether stimulus location within each experiment was decodable from the data. Within each experiment, multivariate classifiers were trained on a subset of trials to discriminate between left vs. right stimulus location, then tested on a different, independent subset of the data. Significant above-chance decoding was observed within each experiment (Fig. 3). Specifically, decoding within the physical salience experiment was reliable from 100 ms to 500 ms, reaching its maximum accuracy at 330 ms (Fig. 3a). Decoding within the naturalistic visual search experiment was reliable as well, ranging from 90ms to 150ms and from 190ms to 280ms, peaking at 230ms (Fig. 3b). These results highlight that MEG patterns contained information about the stimulus location.

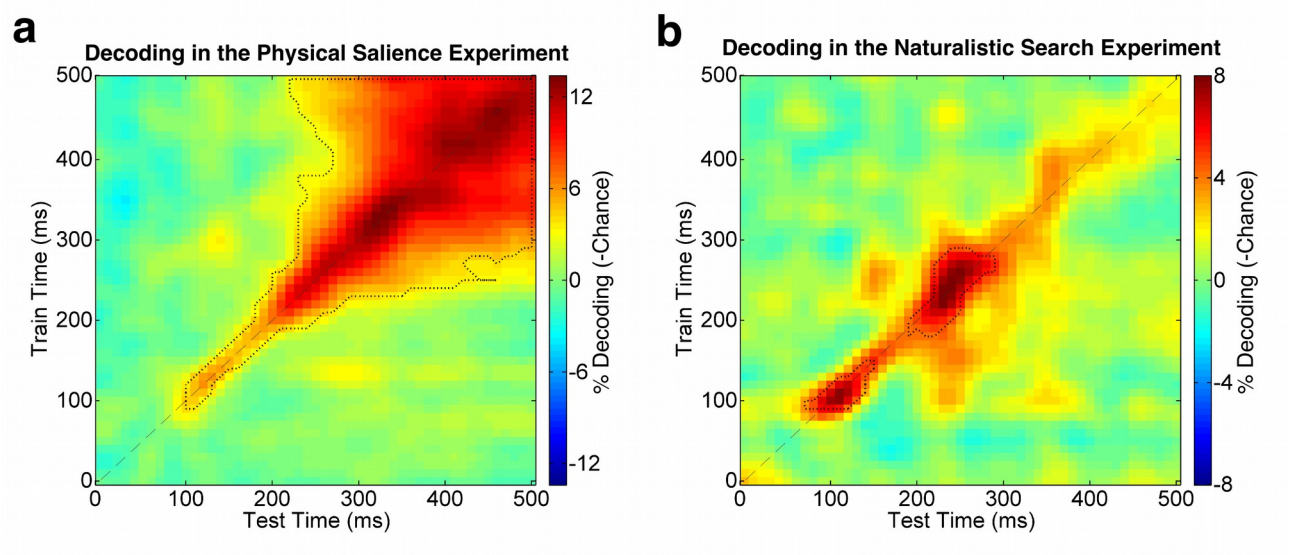

Figure 3. Within-experiment decoding results. Panel (a) shows the time-by-time decoding matrix within the physical salience experiment, panel (b) the resulting matrix within the naturalistic visual search experiment. The outlined areas highlight where decoding accuracy is significantly above chance ($p<0.05$, corrected for multiple comparisons). Decoding on the diagonal of the matrix in (a) is significantly above chance from 100ms to 500ms, reaching its maximum accuracy at 330ms post-stimulus. Decoding on the diagonal of the matrix in (b) is significantly above chance from 90ms to 150ms and from 190ms to 280ms, peaking at 230ms.

3.3.1. Cross-decoding results.

Multivariate classifiers were trained on MEG data from the physical salience experiment and tested on MEG data from the naturalistic search experiment (see Fig. 2). This cross-decoding, averaged across attention conditions (i.e., decoding in target and distractor scenes), was highly reliable from 50ms after stimulus onset, with a first peak at 100ms and a second peak at 260ms (Fig. 4a,b). This result provides evidence for a correspondence between the lateralized processing evoked by the artificial stimuli in the physical salience experiment and the objects in the natural scene experiment. It is worth noting that the decoding peaks of the overall time x time decoding matrix (Fig. 4a) fell on the diagonal, indicating that the temporal evolution of the evoked patterns was similar across the two experiments.

Having established that the location of objects in scenes can be reliably decoded from MEG activity patterns, we next asked when attention modulates this signal. To this end, we separately decoded the position of target objects and distractor objects in otherwise identical scenes. As illustrated in Figure 4c (right panel), decoding of target location was stronger and more reliable than decoding of distractor location. This is clearly illustrated in Figure 4d, which plots the diagonal of these matrices

and shows that significant differences emerged from 240ms to 320ms, from 340ms to 360ms, and from 480ms to 500ms.

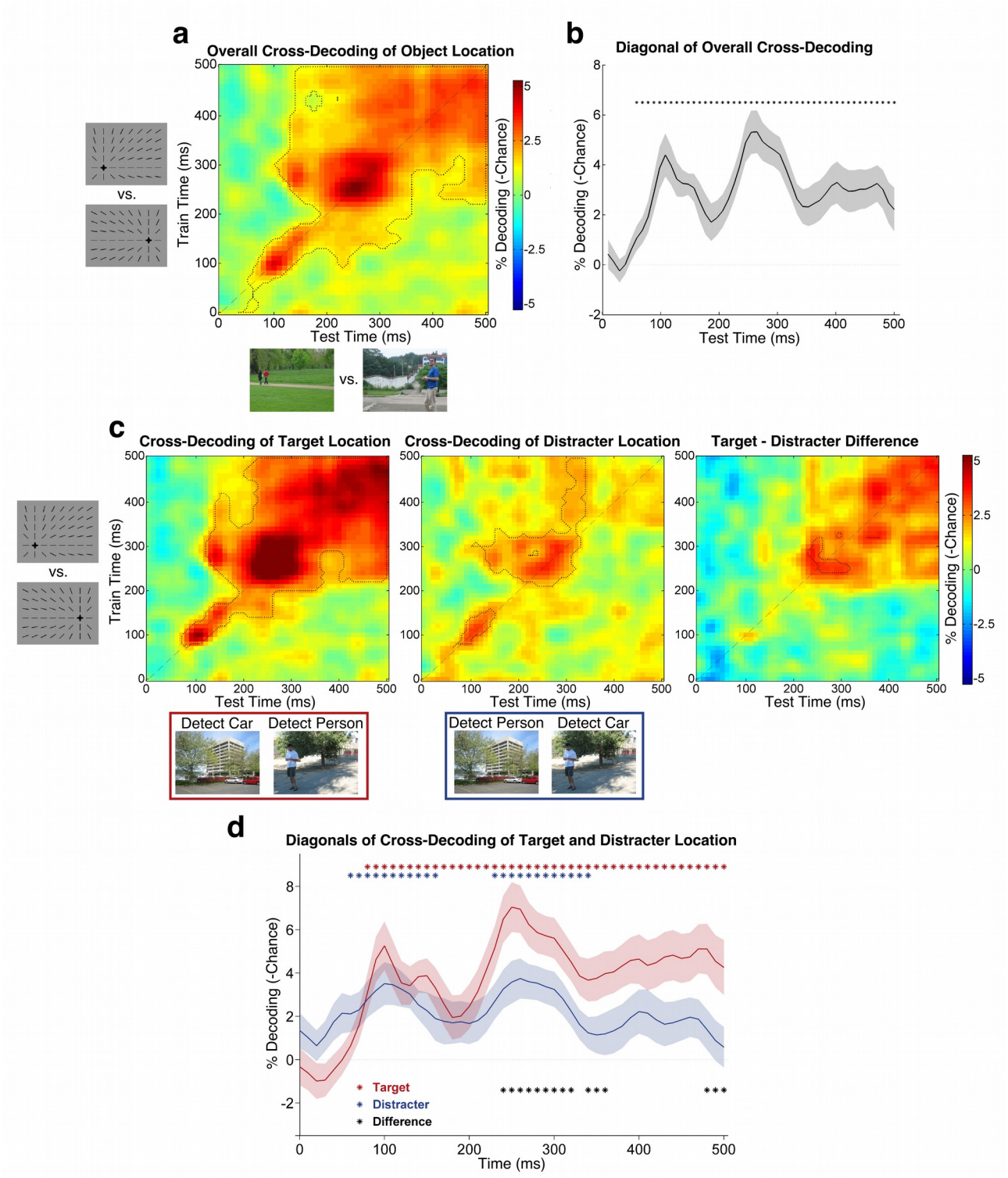

Figure 4. Results of the cross-decoding analysis: time-by-time matrices and time courses of target and distractor conditions. (a) Overall cross-decoding of object location, averaging across target and distractor conditions. The outlined area highlights where decoding accuracy is significantly above chance ($p<0.05$, corrected for multiple comparisons).

Panel (b) shows the diagonal of the overall cross-decoding matrix. This time course is significantly above chance from 50ms after stimulus onset as highlighted by black asterisks ($p<0.05$, corrected for multiple comparisons). Decoding accuracy was maximum at 100ms and 260ms. Panel (c) shows the time-by-time cross-decoding matrices of the target (left) and the distractor (center) conditions, and their difference (right). Panel (d) shows the time course of decoding target object location (red line) and distractor object location (blue line), reflecting the diagonals of the matrices shown in (c). Shaded coloured areas represent SEM. Target decoding on the diagonal was significantly above chance ($p<0.05$, corrected for multiple comparisons) from 70ms to 500ms, peaking at 100ms and 250ms (maximum at 250ms). distractor decoding on the diagonal was significant from 50ms to 150ms, from 220ms to 330ms, peaking at 100ms and 260ms (maximum at 260ms). Target-distractor difference decoding on the diagonal was significant from 240ms to 320ms, from 340ms to 360ms, and from 480ms to 500ms.

3.3.2. Searchlight results.

To explore the topography of these effects we performed a sensor-space searchlight analysis on consecutive time windows of 50ms each, from 0ms to 500ms post-stimulus. This analysis revealed the time course of the cross-decoding across the scalp, suggesting that the attention effect at 250 ms was primarily driven by lateral posterior sensors before moving more anteriorly (Fig. 5c).

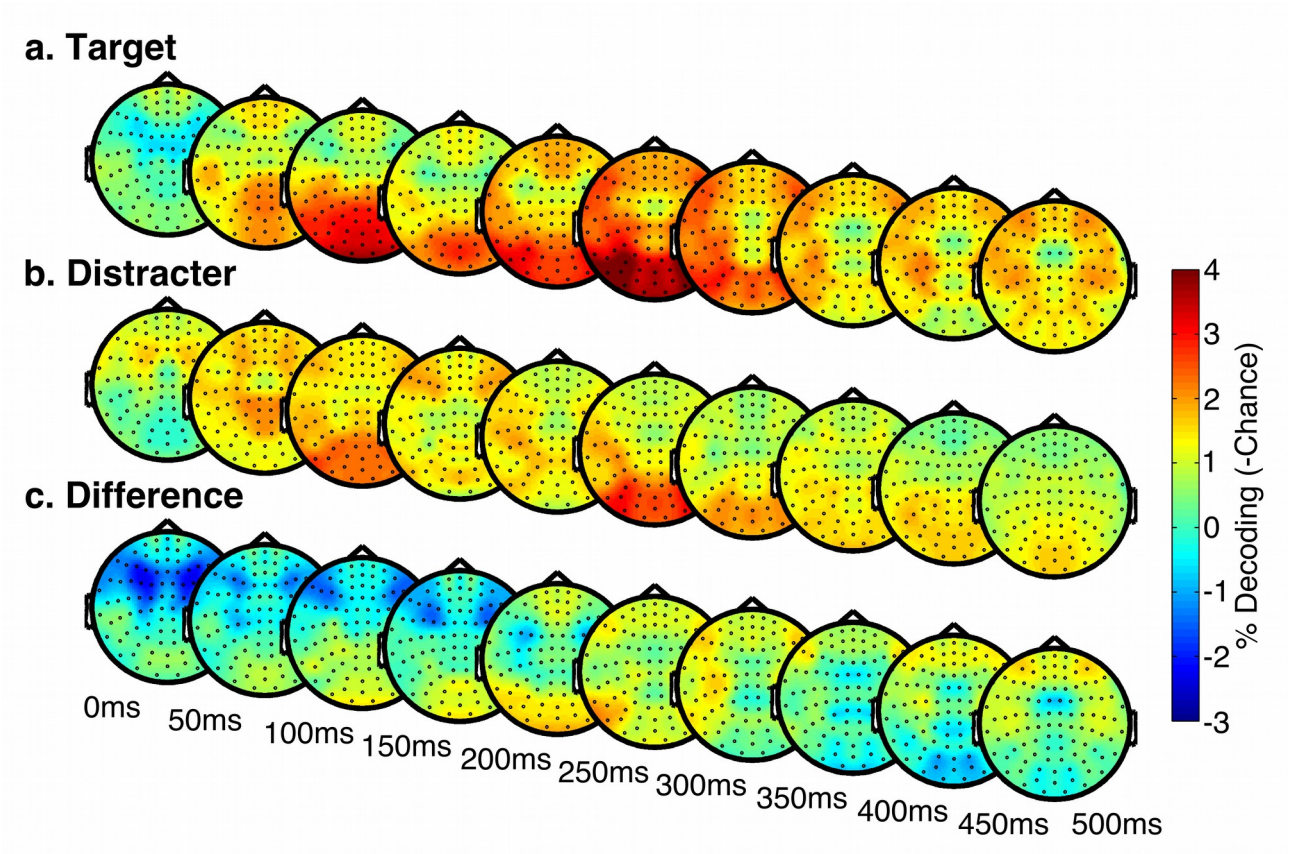

Figure 5. Results of the searchlight analysis. Topographical maps show the results of the cross-decoding searchlight analysis on consecutive time windows of 50ms each from 0ms to 500ms after stimulus onset, separately for the target condition (a), the distractor condition (b) and their difference (c).

3.4. Target-distractor decoding

The cross-decoding analysis provided evidence that spatial attentional selection starts at around 240ms after scene onset, which is later than category-specific attentional modulation in similar tasks, found from 180ms after onset (Kaiser et al., 2016). The presence of category-based attention at 180ms implies that the brain already differentiates target and distractor scenes at that time, thus before the spatial attention effects observed here. To test whether in the current study target and distractor scenes could similarly be differentiated at this time point, we ran an additional analysis within the naturalistic search experiment. In this analysis, we directly decoded the presence of a target (vs. a distractor) in scenes showing either cars or people. Because the only relevant aspect in this analysis was whether the objects were targets or distractors (i.e., matched or mismatched the category of the preceding cue) we averaged across category and location of the objects in the scene. Interestingly, targets could be distinguished from distractors from 180ms after scene onset (Figure 6). The peak was found at 400 ms, shortly before responses were made (mean RT=440 ms in target presence trials). These results indicate that target presence is detected before attention moves to its location.

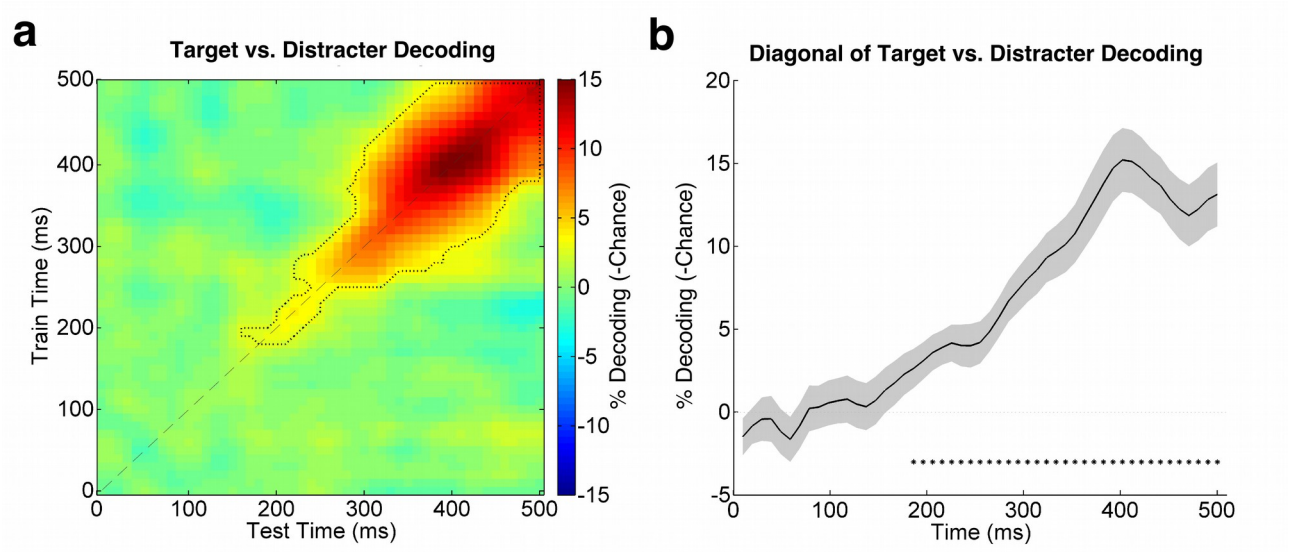

Figure 6. Results of the analysis decoding target vs distractor scenes. (a) Time-by-time matrix of decoding accuracy. The outlined area highlights where decoding accuracy is significantly above chance ($p<0.05$, corrected for multiple comparisons). Panel (b) shows the diagonal of the decoding matrix. This time course is significantly above chance from 180ms after stimulus onset as highlighted by black asterisks ($p<0.05$, corrected for multiple comparisons).

4. Discussion

The current study investigated the time course of attentional orienting in cluttered natural scenes using multivariate decoding of MEG data. We found that the location of objects in natural scenes can be decoded with high accuracy from MEG activity patterns from 50ms after scene onset. The effect of top-down attention on this decoding arose much later, starting at around 240 ms. Target presence itself (irrespective of location) could be decoded from 180 ms after scene onset. While the decoding of object locations at 50ms clearly reflects a stimulus-driven effect (i.e., presence vs absence of a foreground object), we can be confident that the effects at 180 ms and 240 ms reflect influences of top-down attention: First, the same set of scenes was used for targets and distractors, with target status being determined solely by the match between the scene category and the preceding symbolic cue. Second, we excluded the contribution of bottom-up priming effects because the target category varied unpredictably on a trial-by-trial basis. Taken together with the results of (Kaiser et al., 2016), our results indicate that spatial attentional selection follows spatially-global category-based attentional modulation.

The present results are consistent with previous M/EEG studies investigating visual search in artificial arrays. These studies showed that the attentional selection of a target evokes lateralized activity in posterior sensors between 200ms – 300ms after stimulus onset ("N2pc", e.g., Luck and Hillyard, 1994; Eimer, 1996). Our study indicates that spatial attentional selection in naturalistic search occurs at a similar latency (Fig 4d) and with a similar topography (Fig 5c). This demonstrates an important generalization of previous findings to more naturalistic conditions, despite the differences between artificial and naturalistic search (Wolfe et al., 2011b; Peelen and Kastner, 2014) and between univariate and multivariate analysis methods[4].

The current study complements a recent study that used similar methods to investigate the time course of top-down category-specific attentional modulations in scenes (Kaiser et al., 2016). There, decoding focused on object category processing, with classifiers trained to distinguish exemplars of people and cars and tested on scenes containing exemplars of these categories. Results showed that the category of objects present in scenes could be decoded from around 180ms after stimulus onset. Importantly, this effect was specific to the behaviorally-relevant category from its first emergence, with better decoding of target than distractor category already at 180ms. In other words, top-down attention modulated category-level processing as soon as category information was available.

4 Single sensors did not show reliable attention effects in the current study.

Our present results show that spatially-specific attention effects – starting at 240ms – emerge after this category-level modulation. This indicates that attention first modulates spatially-global category representations, followed by the spatial selection of the target. This sequence matches that observed in previous studies investigating search for simple features in artificial displays, showing that feature-based attentional modulation precedes spatially-selective enhancement (Hopf et al., 2004; Eimer, 2014). Our results thus support the idea that content-based guidance is not limited to low-level features but can be implemented at higher levels of the visual system as well (Wyble et al., 2013; Hickey et al., 2015; Battistoni et al., 2017; Wyble et al., 2018).

The spatial modulation observed here provides a neural correlate of behavioral findings of attentional capture by objects matching a top-down category-based attentional set (Reeder and Peelen, 2013; Reeder et al., 2015a). In these studies, participants searched for cars and people in natural scenes. On a subset of trials, two irrelevant stimuli appeared instead of the scenes. One of these stimuli was quickly followed by a dot that participants were instructed to detect. Results showed that participants were faster to detect the dot when it appeared at the location of a stimulus that shared mid-level features with the target category (e.g., a wheel of a car, or an arm attached to a torso), providing evidence for attentional capture. Importantly, the effect was also observed when the mid-level features appeared at locations that were never relevant to the search task. These findings demonstrate that category-based attention is spatially global and that it guides spatial attention to template-matching stimuli. In conjunction with (Kaiser et al., 2016), the current results provide novel insight into the temporal evolution of both these effects.

Interestingly, although spatial attention is captured by template-matching objects, the detection of familiar object categories in natural scenes may not require spatial attentional selection. For example, target-specific EEG responses in these tasks have been observed before 200 ms (Thorpe et al., 1996), more likely corresponding to the category-based modulation observed in Kaiser et al. (2016) than the spatial selection observed here. Similarly, in the current study response patterns evoked by target and distractor scenes differed from around 180 ms after stimulus onset, indicating that target features are detected before spatial attention moves to the target location. Behavioral studies have shown that participants may not be able to localize object categories in natural scenes that have nonetheless been detected (Evans and Treisman, 2005). Others have argued that the detection of familiar object categories may not even require spatial attention at all (Li et al., 2002; Stein and Peelen, 2017). These findings suggest that spatially-global category-based attention may be sufficient for detecting target-diagnostic features.

In daily life, however, the detection of category-diagnostic features is often not sufficient for guiding our behavior. Many situations require us to bind features to identify objects at finer levels. For example, we might need to distinguish our red car from our friend's green car, or to find our friend among other people. These tasks require spatial attention to bind features, as elegantly shown by work in neurological patients with parietal damage (Cohen and Rafal, 1991; Friedman-Hill et al., 1995). Thus, while not directly required in the current task, spatial selection may be an integral and obligatory aspect of top-down attention, even when directed to high-level categories (Wyble et al., 2013; Reeder et al., 2015a).

To conclude, the current study shows that spatial attentional selection of target objects in natural scenes occurs at around 240 ms after scene onset. This spatial modulation follows an earlier spatially-global categorical attention modulation that provides information about target presence from around 180 ms. Our results are in line with theories of visual search proposing that spatial attention is guided by feature-based selection (Treisman and Sato, 1990; Wolfe, 1994), and importantly generalize this idea to naturalistic search for familiar object categories in natural scenes.

Chapter 5:
MEG decoding as a tool to study the temporal dynamics of size constancy and distance perception in natural scenes

1. Introduction

A correct perception of the distance and size of an object is crucial for a reliable representation of the three-dimensional (3D) world surrounding us, and indispensable for behaviors such as navigation, reaching and grasping. Distance perception is determined by the integration of contextual information derived from monocular and binocular cues. Monocular cues provide distance information when viewing a scene with only one eye, and include, among others, occlusion, perspective, texture gradient and relative size. Binocular cues give information about depth when using both eyes, and include binocular disparity (also known as stereopsis or retinal disparity) and convergence. Size perception is regulated by size constancy mechanisms, which act by rescaling the size of an object as a function of perceived distance (Holway and Boring, 1941; Gruber, 1954; Andrews, 1964; Morgan, 1992; for a review, Sperandio and Chouinard, 2015). These rescaling mechanisms allow us to perceive the world as stable despite the continuously changing flux of visual input that hits the retina; therefore they can be regarded as one of the processes at the basis of invariant object recognition (DiCarlo et al., 2012).

The processes of size and distance perception have long been thought as theoretically inseparable, since changes in perceived size are influenced by changes in perceived distance and *vice versa*, as formally postulated by the size-distance invariance hypothesis or SDIH (Boring, 1940; Gilinsky, 1951; Epstein et al., 1961; Epstein, 1963; Kaufman et al., 2006; Qian and Yazdanbakhsh, 2015; Kim et al., 2016). The SDIH maintains that an object's perceived size is determined by the product of its perceived distance and some function of its retinal size; more formally:

$$S = D\ tan a$$

where S is the object's perceived size, D is the object's perceived distance, and a is the object's angular size (i.e., retinal size or visual angle subtended by the object). The SDIH is directly based on Emmert's perceptual law of apparent size (originally postulated in the context of

afterimages), which states that the perceived size of an object is proportional to its perceived distance (Emmert, 1881). Together with Euclid's principle of optical geometry, which postulates that the angular size of an object is inversely proportional to its distance, these laws fully describe the phenomenon of size constancy. For example, when we watch a train depart from a station, its retinal image size decreases as the distance increases (Euclid's law); however, even though the distance between the eyes and an object doubles and the retinal object size halves, our perception of the real size of the object remains unchanged: we do not perceive the train as smaller, just farther away (Emmert's law).

The close relationship between distance and size is also at the basis of many optical size illusions, such as the Ponzo illusion, the powerful real-world moon illusion, and several others (Rock and Kaufman, 1962; Kaufman and Rock, 1962; Dees, 1966; Ross, 1967; Fisher, 1968; McCready, 1986; Kaufman and Kaufman, 2000; Ross, 2000; Redding, 2002; Kaufman et al., 2007; Qian and Petrov, 2012; Weidner et al., 2014; Gregory, 2015; Sperandio and Chouinard, 2015). For example, in the Ponzo illusion (Fig. 1a) the two horizontal lines project the same image on the retina (physically equal retinal size). However, we commonly experience them as having different lengths: the upper line appears longer than the bottom line. Interestingly, this experience is not canceled or diminished after the observer takes physical measurements and is aware of the perceptual deception (Fig. 1b).

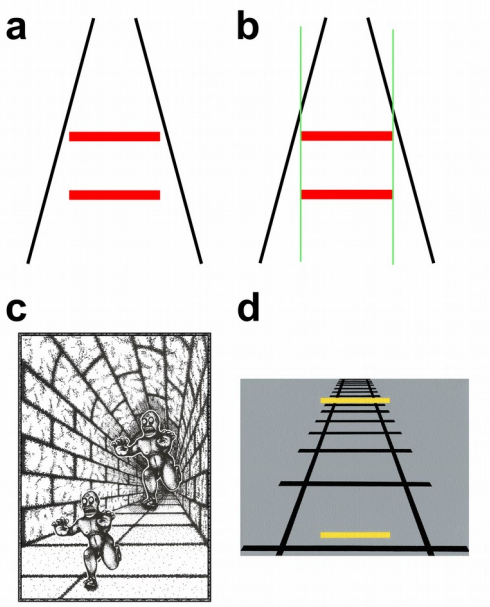

Figure 1. Examples of Ponzo Illusion. (a) The classical Ponzo illusion. Panel (b) highlights that even taking physical measurements of the length of the two horizontal lines does not diminish the illusion that the upper line is longer than the bottom line. Panel (c) and (d) show variations of the Ponzo illusion.

What happens in this and other visual illusions is well described by Bayesian accounts (Geisler and Kersten, 2002), which are directly linked to Helmholtz's claim that perception depends on the rules that the brain has learned about the world (Von Helmholtz, 1867). More specifically, the brain incorrectly interprets the converging lines as perspective cues, triggering the application of size-rescaling rules acquired from experience. This leads to the erroneous inference that the upper horizontal line is farther away than the bottom line, and consequently, that the upper line is longer than the bottom line. The perceptual rule that is applied in this case is that the physical properties of objects remain constant regardless changes in retinal information elicited by changes in viewing distance, perspective, and lighting (rule of perceptual constancy). The phenomenon of the incorrect application of perceptual rules is also explained by Gregory's theory of inappropriate constancy scaling (Gregory, 1963, 1968, 2015).

Several fMRI studies have looked for the neural basis of the Ponzo illusion (Murray et al., 2006; Fang et al., 2008; He et al., 2015). Surprisingly, they found that the subjective experience of size was associated to activity changes in the primary visual cortex (V1). Specifically, they measured brain activity while participants viewed two objects with same retinal size, one located in the foreground and one in the background in images conveying a perception of distance through perspective cues. They found that objects perceived as bigger (i.e. those in the background) activated the most anterior part of V1, where big stimuli are retinotopically represented. Objects perceived as smaller (i.e. those in the foreground) activated the most posterior part of V1, where small stimuli are retinotopically represented. Therefore, these unexpected results highlighted that activity in the primary visual area reflected the perceived size of an object, and not its retinal size. These findings were supported by other studies demonstrating that V1 carries information about perceived size (Murray et al., 2006; Sterzer and Rees, 2006; Fang et al., 2008; Liu et al., 2009; Schwarzkopf et al., 2011; Sperandio et al., 2012; Pooresmaeili et al., 2013; Chouinard and Ivanowich, 2014).

However, how is it possible that V1, which is the first cortical region that receives signals from the retina (and therefore information related to the physical size of objects), is modulated by perceived size? Electrophysiological measures in monkeys have revealed that several visual-occipital areas are involved in size-constancy mechanisms (Dobbins et al., 1998; Ni et al., 2014). In humans, neuroimaging and lesion studies have demonstrated the

contribution of several higher-order areas, among which parietal, dorsal and ventral regions, in size and distance perception and in the processing of spatial and contextual information (Gnadt and Mays, 1995; Berryhill and Olson, 2009; Berryhill et al., 2009; Preston et al., 2013; Costa et al., 2015). It is possible that this network projects feedback connections related to distance information to lower-level areas and therefore contribute to size-rescaling processes.

Concerning size, several studies showed that activity in ventral occipitotemporal cortex is modulated by perceived size, as determined by stored semantic information (Cate et al., 2011; Amit et al., 2012; Gabay et al., 2016). Specifically, large real-world objects tend to activate medial areas in the ventral occipitotemporal cortex, while small objects tend to activate more lateral areas (Konkle and Oliva, 2012; Troiani et al., 2014). This well-corroborated finding has led to the claim that real-world object size is an important organizing principle of high-level visual object representations, including scene-responsive and object-responsive regions (Julian et al., 2017).

In summary, many studies suggest that several ventral and parietal areas are involved in the representation of the perceived object size. The finding that even V1 represents an object's perceived size (the conscious perception of an object's size instead of the physical input from the retina) can be well explained within a predictive coding framework. Specifically, higher-level areas would project feedback connections to area V1 "explaining away" (i.e., inhibiting) neural activations of image feature that are not consistent with the stored knowledge related to the object's real world size.

However, the temporal dynamics with which rescaling processes take place are yet to be established. For example, does V1 initially represent the retinal size of objects, and afterwards their perceived size? When does the representation of perceived size emerge? And more generally, what are the temporal dynamics of size constancy? To address these questions, we employed a multivariate pattern analysis (MVPA) approach on magnetoencephalography (MEG) data to decode with high temporal resolution stimulus-related information from brain activity patterns. Participants watched big vs. small objects presented near vs. distant in natural scenes or in grey backgrounds, and were instructed to complete an unrelated oddball task in which they had to press a button whenever objects had a golden color. By decoding information related to the size of objects as a function of their distance and background type, we were able to determine the likely time course of perceived

size in natural scenes, and to highlight the validity of MEG as a tool to investigate the neural dynamics underlying size constancy.

2. Materials and methods

2.1. Participants

Seventeen healthy participants with normal or corrected-to-normal vision (10 male, mean age M = 24.2, SD = 3.8) took part in the MEG experiment. All participants provided written informed consent and received monetary compensation. The experiment was conducted in accordance with the Declaration of Helsinki and approved by the Ethical Committee of the University of Trento.

2.2. Experimental procedure

While recording MEG data, participants performed an oddball task on objects in natural scenes (Fig. 2). The experiment consisted of 10 different blocks of 72 trials each. Five blocks had stimuli with natural scene backgrounds, and 5 blocks had stimuli with grey backgrounds. Blocks with stimuli with natural scene backgrounds and blocks with stimuli with grey backgrounds were presented in an interleaved manner (odd-numbered blocks had natural backgrounds, even-numbered blocks had grey backgrounds). Each stimulus contained an object (see Stimuli), and participants were instructed to press a button when the object had a golden color. Each trial started with a fixation point (a plus, "+", presented centrally in the screen), which lasted 800 ms, and was followed by a preview of the stimulus without the object (whose duration was randomly jittered from a rectangular distribution ranging from 800 ms to 1400 ms). Then, the stimulus with the object was presented for 200ms, and it was followed by an inter-trial-interval with a fixation point (2200 ms - 3000 ms, randomly jittered from a rectangular distribution). Average trial duration was 4700 ms, and average block duration was 5.65 minutes.

The experimental session lasted 60 minutes. Stimuli were back-projected onto a translucent screen located 115cm from the participants. Stimulus presentation was controlled using MATLAB 8.0 and the Psychtoolbox (Kleiner et al., 2007).

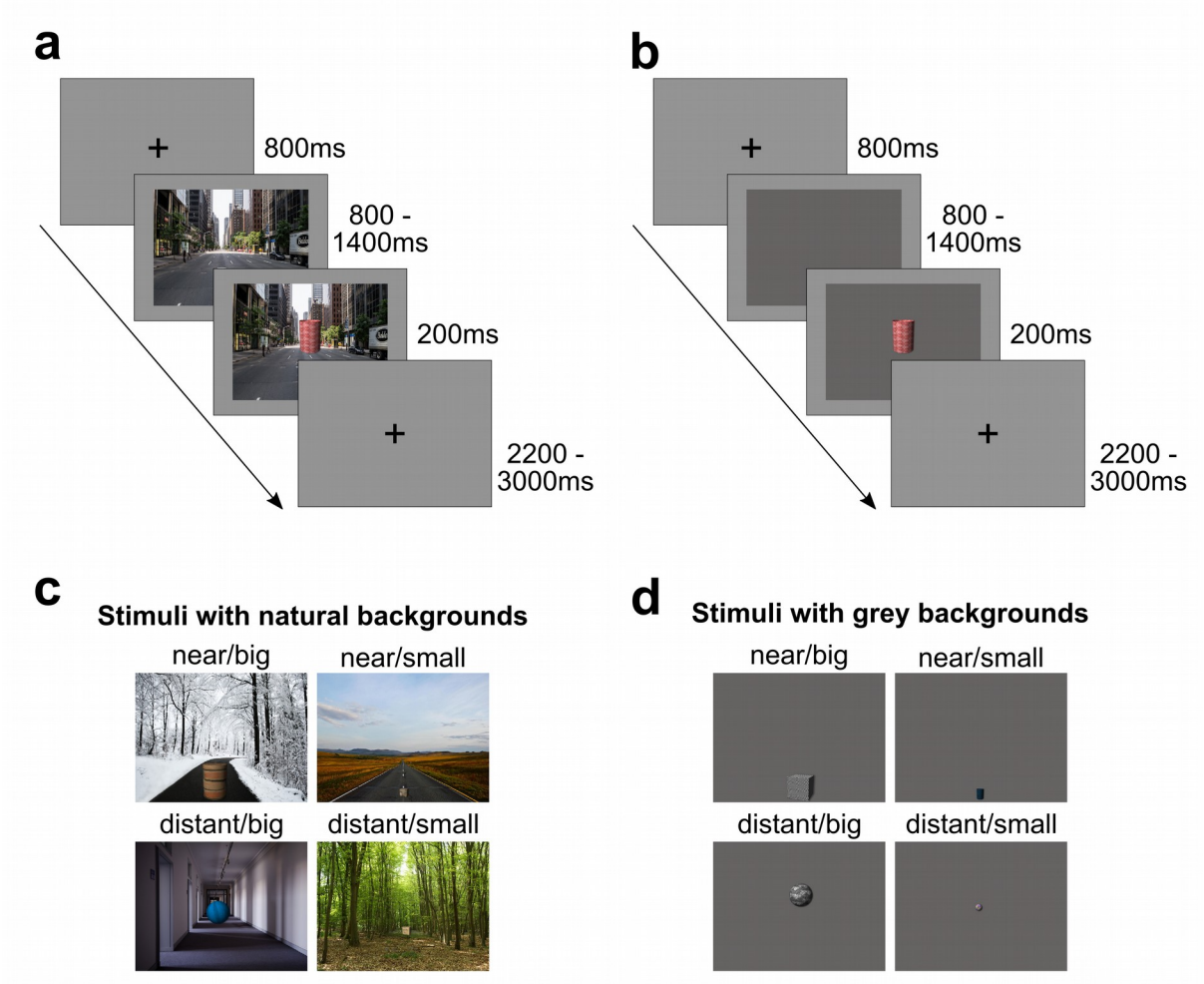

Figure 2. Schematics of the experimental paradigm. Trial sequences in blocks in which (a) stimuli had natural scene backgrounds and (b) stimuli had grey backgrounds. Examples of (c) stimuli used in natural background blocks, and (d) stimuli used in grey background blocks.

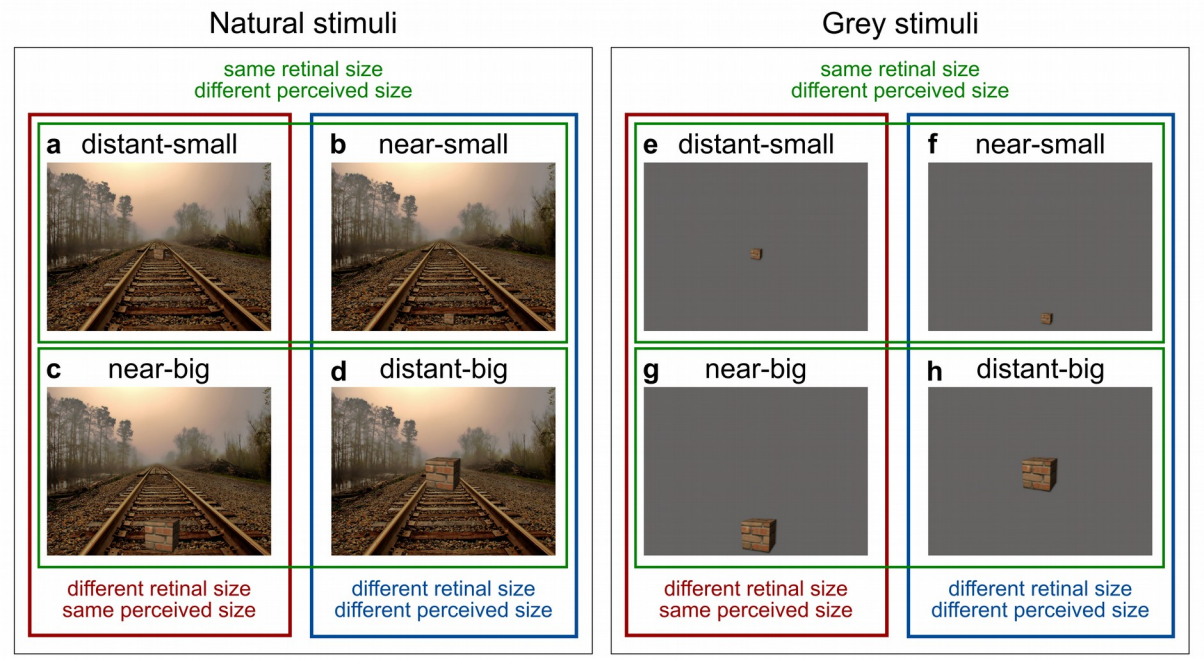

Figure 3. Examples of stimuli and experimental conditions. The two panels show the conditions for stimuli with natural backgrounds (left) and for stimuli with grey backgrounds (right). For each scene, four variations were created: (a) one with a distant small object, (b) one with a near small object, (c) one with a near big object, and (d) one with a distant big object. The size of small objects and big objects across the variations of a scene were constant. To create the grey stimuli, each the scene of each natural stimulus was removed and substituted with an empty grey background, so that object positions across natural and grey stimuli remained unvaried. Relevant for the analysis, different stimulus combinations lead to different conditions: (1) the "different retinal size - same perceived size" condition was realized with the stimuli (a) and (c); (2) the "different retinal size - different perceived size" condition with the stimuli (b) and (d); (3) two "same retinal size - different perceived size" conditions, one with the stimuli like (a) and (b), and one with the stimuli like (c) and (d). For the sake of clarity and consistency across stimulus types (natural vs. grey), the condition labels of grey stimuli were the same as those of natural stimuli. Of course, given the absence of contextual and perspective cues, we did not assume participants' percept to correspond to the conditions' labels.

2.3. Stimuli

Stimuli were created with the software GIMP (https://www.gimp.org). Ninety different natural scenes with depth cues (that provided a perception of depth and distance; Fig. 3) were selected from the Internet. Each scene was used to create four variations: one with a near big object, one with a distant big object, one with a near small object and one with a distant small object. Thus, the total number of stimuli consisting of a scene and an object amounted to 360. Additional 360 stimuli were created from each scene and object by

removing the natural background and replacing it with a grey background (RGB values: 102, 99, 99). The objects were ninety different exemplars of cubes (30), spheres (30) and cylinders (30) created in GIMP. Each object had two variations: one with a “big” size, and one with a “small” size; these size were determined manually by the experimenter in a way that when placed in the distance, the small object appeared to have the same perceived size of the near big object. Each object type was paired with a specific natural scene. The 10% of the stimuli (72) were created with golden objects: half of them had a natural background and the other half had a grey background; in one-third of them (24) the object was a cube, in one-third it was a cylinder, and in one-third it was a sphere. Each scene and object was presented only once throughout the experiment (the experiment had 720 trials, in each trial there was one of 720 stimuli). All stimuli were reduced to 480 (vertical) x 640 (horizontal) pixels, subtending 13.5° x 10° of visual angle, and were presented on a grey background (RGB values: 148, 148, 148).

2.4. MEG data acquisition and preprocessing

Neuromagnetic activity was recorded using a whole-head MEG system with 102 magnetometers and 204 planar gradiometers (Elekta Neuromag 306 MEG system, Helsinki, Finland). Data were acquired continuously (with online sampling rate of 1000 Hz) and band-pass filtered online between 0.1 and 300 Hz. Offline preprocessing was performed using MATLAB 8.0 and the Fieldtrip toolbox (Oostenveld et al., 2011).

Data were epoched from -300 to 500 ms with respect to stimulus onset (the scene with the object). No offline filter was applied to the data. No offline filter was applied to the data because this appeared to be the most reliable procedure in the previous experiment (Chapter 4), where the application of filters induced filtering artifacts.

Based on visual inspection, and blind to condition, trials and channels containing artifacts (i.e., blinks, eye-movements, or unusually large peak-to-peak amplitudes) were discarded from subsequent analysis. All trials (correct and incorrect) were included in the analysis. Data were baseline corrected with respect to the pre-stimulus period (with baseline from -200ms to 0ms) and down-sampled to 100Hz to improve signal-to-noise ratio (Grootswagers et al., 2017). Data from rejected channels were interpolated based on the average of neighboring sensors of the same type.

2.5. MEG multivariate pattern analysis

All multivariate classification analyses were performed using MATLAB 8.0 and the CoSMoMVPA toolbox (Oosterhof et al., 2016). Single-trial classification was performed separately for every 10ms time bin of the evoked field data of all magnetometers; only data from magnetometers were used as these sensors offered more reliable classification performance than gradiometers (Kaiser et al., 2016). To increase the signal-to-noise ratio, 1000 synthetic trials were created for every condition in both the training and testing sets. Each synthetic trial was created by randomly selecting 5 trials and averaging across these trials. Trials were selected without replacement until the pool of trials was exhausted, such that each trial contributed to a roughly equal number of synthetic trials. Classification accuracy was evaluated by computing the percentage of correct predictions of the classifier. The decoding analysis was repeated for every possible combination of training and testing time, leading to a 50 x 50 points (i.e. 500 ms x 500 ms with 100Hz resolution) matrix of classification accuracies for every participant. Single-subject accuracy matrices were smoothed using a 3 x 3 time points averaging box filter (i.e. 30 x 30ms, for the training and testing times, respectively); single-subject accuracy matrix diagonals were smoothed with a 3-point (30 ms) boxcar filter. To determine time periods of significant above-chance classification, a threshold-free cluster enhancement procedure (Smith and Nichols, 2009) was used with default parameters. The multiple-comparisons correction was based on a sign-permutation test with null distributions created from 10,000 bootstrapping iterations and a significance threshold of $Z > 1.64$ (i.e., $p < 0.05$, one-tailed; (Oosterhof et al., 2016)).

3. Results

3.1. Decoding of distance in natural and grey stimuli (within-experiment)

As postulated by the size-distance invariance hypothesis, the perception of distance is fundamental in size-constancy mechanisms, and it is triggered by monocular and binocular depth cues (see Introduction). Grey stimuli, which do not provide such bottom-up cues, should not elicit a representation of distance, but only of position (slightly lower vs. slightly upper in the visual field): that is, in grey stimuli the position of objects should not be associated to a specific distance (near vs. distant, respectively), given the absence of depth cues. On the contrary, natural stimuli should elicit a representation of both position and

distance. Therefore, we would expect higher decoding accuracy of object distance in natural stimuli than in grey stimuli[1].

Importantly, the decodability of object distance in natural scenes is essential for the emergence of any size-constancy process, and therefore crucial for the feasibility and validity of subsequent analyses. The absence of such effect would indicate that our stimuli failed to trigger a perception of distance, hence ruling out any possibility of size-constancy.

In this first preliminary analysis, we investigated whether distance (near vs. distant objects), was decodable in natural stimuli, and crucially, whether this classification was more accurate in natural stimuli than in grey stimuli (for examples of the stimuli used, see Fig. 3).

First, decoding was performed within natural stimuli: Linear Discriminant Analysis (LDA) classifiers were trained and tested on MEG activity patterns evoked by natural stimuli with near objects vs. natural stimuli with distant objects (regardless of shape). Then, decoding was conducted within grey stimuli, where classifiers were trained and tested on data elicited by grey stimuli with near objects vs. grey stimuli with distant objects (regardless of shape). Each dataset was divided into 10 independent chunks; classifiers were iteratively trained on 9 chunks and tested on 1 chunk, and the results were averaged. Decoding in both the within-natural and within-grey analysis was performed twice: once with stimuli with small objects and once with stimuli with big objects; the results were then averaged in order to eliminate the possible confounding factor of size. Next, to test whether decoding of distance in natural stimuli was more accurate than in grey stimuli, the results of the within-grey analysis were subtracted from the results of the within-natural analysis. Here, the difference between natural and grey was the most important comparison: because natural stimuli contained information on both object position and distance, and grey stimuli contained only information on object position, the difference isolated a representation of object distance from a representation of object position. Importantly, we can always only talk about representation of distance, not perception of distance, since we do not know the conscious perceptions of participants.

Figure 4 illustrates the time courses of the within-natural decoding and within-grey decoding.

1 To be noted, in order to simplify the reading, and given the way in which stimuli were created and coded, sometimes we will use the term “distance” to actually refer to the “position” of objects in grey stimuli; in grey stimuli we cannot truly talk about “distance” because of the absence of bottom-up depth cues necessary to give rise to a distance percept.

Significant above-chance decoding accuracy was observed within each condition. Specifically, decoding within the natural condition was reliable from 80ms to 500ms with respect to stimulus onset, reaching its maximum at 150ms (Fig. 4, red line). Decoding within the grey condition was also reliable from 80ms to 500ms, peaking at 140ms (Fig. 4, blue line). Decoding of distance within the natural condition was stronger and more reliable than within the grey condition from 140ms to 190ms.

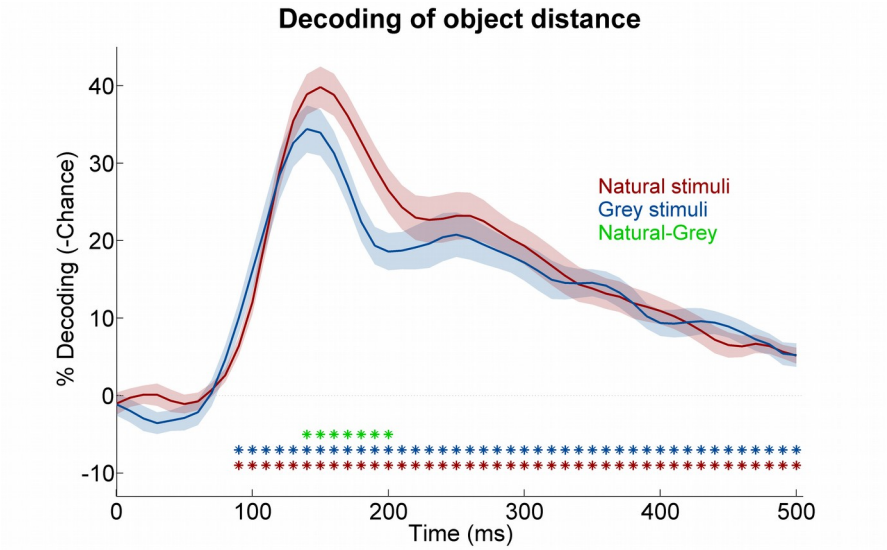

Figure 4. Results of the decoding analysis of near vs. distant objects separately for natural and grey stimuli. Lines represent the time courses. Shaded colored areas represent SEM. Asterisks indicate time points at which decoding is significantly above chance ($p < 0.05$, corrected for multiple comparisons). The time course of the decoding of distance in natural stimuli (Within-Natural, red line) is significant from 80ms to 500ms (maximum at 150ms). The time course of decoding of distance in grey stimuli (Within-Grey, blue line) is significant from 80ms to 500ms (maximum at 140ms). Natural-Grey difference decoding (green asterisks) is significant from 140ms to 190ms (green asterisks).

The results of this preliminary analysis show that object position can be decoded equally well in both grey and natural stimuli. The presence of a decoding difference between natural and grey stimuli shows that object distance can be decoded from brain patterns evoked by natural stimuli from 140ms after scene onset. This is an important result, as it demonstrates the presence of a basic distance effect (only present in natural stimuli), likely triggered by depth cues, which is necessary for the occurrence of size-constancy mechanisms.

3.2. Decoding of perceived size in natural and grey stimuli (within-experiment)

After establishing that MEG activity patterns elicited by natural stimuli contained information about distance (near vs. distant objects), we sought to determine when the representation of perceived object size emerged. For this purpose, we performed two series of analyses, one within natural stimuli and the other within grey stimuli. Each analysis was made up of two conditions. (1) In the different perceived size condition, classifiers were trained and tested on data evoked by stimuli with near-small objects vs. distant-big objects, which have different retinal size and different perceived size. (2) In the same perceived size condition, classifiers were trained and tested on data elicited by near-big objects vs. distant-small objects, which have different retinal size but same perceived size. Figure 3 illustrates examples of these conditions. Like the previous analysis, each dataset was divided into 10 independent chunks; classifiers were iteratively trained on 9 chunks and tested on 1 chunk, and the results were averaged. The results of the same perceived size condition were then subtracted from the results of the different perceived size condition (different-same condition) in order to highlight when two objects of different retinal size started being perceived as having the same size; or, in other words, the time at which size constancy mechanisms are initiated. Finally, the results of the different-same condition within the grey stimuli condition were subtracted from the results of the different-same condition within the natural stimuli condition.

The results showed that in the natural stimuli condition, decoding of different perceived size was significant from 90ms to 500ms (Fig. 5a, red line), and decoding of same perceived size was significant from 80ms to 400ms (Fig. 5a, blue line). The different-same difference decoding was significant from 260ms to 420ms (Fig. 5a, green asterisks). In the grey stimuli condition, decoding of different perceived size was significant from 80ms to 500ms (Fig. 5b, red line), and decoding of same perceived size was significant from 80ms to 500ms (Fig. 5b, blue line). The different-same difference decoding was significant from 160ms to 180ms, from 250ms to 310ms, and from 390ms to 410ms (Fig. 5b, green asterisks).

Notably, the different-same difference in natural stimuli was not significantly different from the different-same difference in grey stimuli (Fig. 5c).

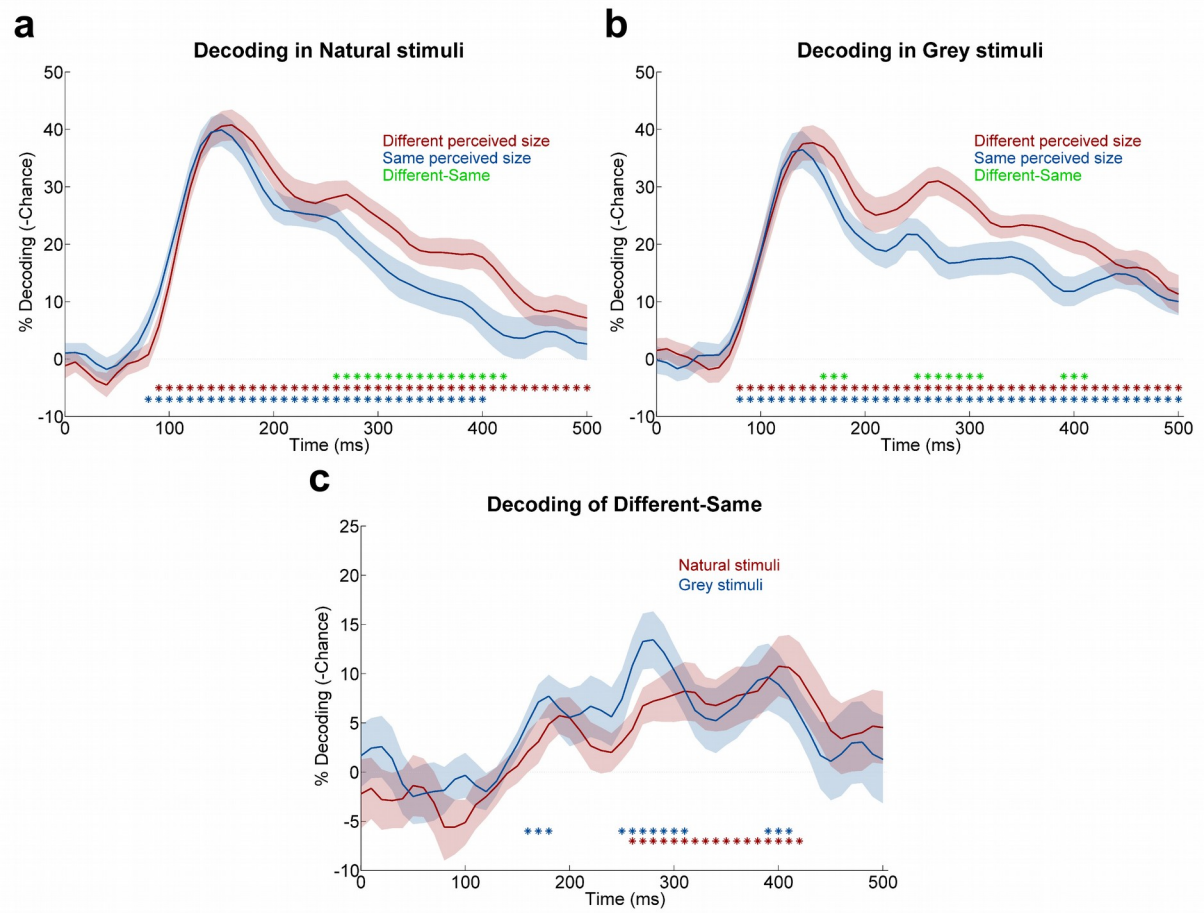

Figure 5. Time courses of perceived size in natural and grey stimuli. Shaded colored areas represent SEM. Asterisks highlight time points where decoding is significantly above chance ($p < 0.05$, corrected for multiple comparisons). (a) Decoding of different perceived size in natural stimuli (red line; near-small vs. distant-big objects) is significant from 90ms to 500ms (maximum at 160ms). Decoding of same perceived size in natural stimuli (blue line, near-big vs. distant-small objects) is significant from 80ms to 400ms (maximum at 150ms). The different-same difference decoding (green asterisks) is significant from 260ms to 420ms. (b) Decoding of different perceived size in grey stimuli (red line; near-small vs. distant-big objects) is significant from 80ms to 500ms (maximum at 150ms). Decoding of same perceived size in grey stimuli (blue line, near-big vs. distant-small objects) is significant from 80ms to 500ms (maximum at 140ms). The different-same difference decoding (green asterisks) is significant from 160ms to 180ms, from 250ms to 310ms, and from 390ms to 410ms. Panel (c) shows the different-same difference decoding separately for natural stimuli (significant from 260ms to 420ms; maximum at 400ms) and grey stimuli (significant from 160ms to 180ms, from 250ms to 310ms, from 390ms to 410ms; maximum at 280ms). The difference between these two time courses was not significant.

The difference between different perceived and same perceived size conditions in natural stimuli suggest that size constancy mechanisms are initiated around 260ms after scene onset (in other words, two objects with different retinal size but same perceived size,

start being perceived as having the same size around 260ms).
Surprisingly, we find a similar pattern of results within grey stimuli as well: it appears that they do trigger some size constancy mechanism, as demonstrated by the result that perceived size starts being represented around 160ms after stimulus onset. Furthermore, the absence of a decoding difference between natural and grey stimuli suggests that grey stimuli are as effective as natural stimuli in eliciting size constancy processes.

3.3. Cross-decoding of perceived size

Training and testing classifiers within a specific stimulus type, even though the two data sets are independent from each other, might not be optimal, because stimulus-specific factors might contribute to the decoding. Therefore, we next investigated the temporal dynamics of perceived size (size constancy mechanisms) using a cross-decoding procedure. This question was addressed in two complementary ways. First, we considered a condition similar to the Ponzo illusion, also employed in past fMRI studies (Murray et al., 2006; Fang et al., 2008). Specifically, in this situation the near and distant object have the same retinal size, but different perceived size due to size-constancy mechanisms (the distant object is perceived as having a bigger size than the near object). We will refer to this situation, in which two objects with same retinal size are perceived as having different size, as the Ponzo illusion condition.
Then, we considered a more typical situation in which a specific object is perceived as having the same size even though its retinal size varies as a function of distance (i.e., small retinal size when it is distant, big retinal size when it is near). We will refer to this situation, in which two objects with different retinal size are perceived as having the same size, as the typical condition. This analysis is the equivalent of the previous analysis (3.2.) but adopting a cross-decoding approach. These two conditions are exemplified in Figure 2.

3.3.1. Cross-decoding of perceived size in the Ponzo illusion condition

In the classical Ponzo illusion and in its variations (Murray et al., 2006; Fang et al., 2008), when the retinal size of a near object and a distant object is kept equal, size-rescaling mechanisms lead us to perceive the distant object as being bigger than the near object. Therefore, in this analysis we asked when near objects are more often classified as small than distant objects (or, when distant objects are more often classified as big than near objects). Another way to think of this analysis, is when a near object, having the same retinal size of a

distant object, starts being perceived as being smaller (or, when the distant object starts being perceived as being bigger).

To this end, we performed a standard cross-decoding analysis (in which classifiers were trained on data from grey stimuli, and tested on data from natural stimuli) and a reverse cross-decoding analysis (where classifiers were trained on data from natural stimuli and tested on data from grey stimuli). In both standard and reverse cross-decoding, classifiers were trained on different retinal size (big vs. small objects, regardless of position and shape), and tested on different distance (near vs. distant objects, regardless of size and shape). To determine whether classification was limited to a specific training-testing direction (standard vs. reverse), we tested their difference by subtracting the results of the reverse analysis from the results of the standard analysis.

Results showed that decoding accuracy in the standard cross-decoding condition was significantly above chance from 160ms to 190ms, and from 240ms to 350ms (maximum at 170ms; Fig. 7, red line). Decoding in the reverse cross-decoding condition was reliable as well, reaching significance from 260ms to 270ms, and from 320ms to 500ms (maximum at 410ms). The comparison between the two decoding directions (standard-reverse) revealed no significant difference.

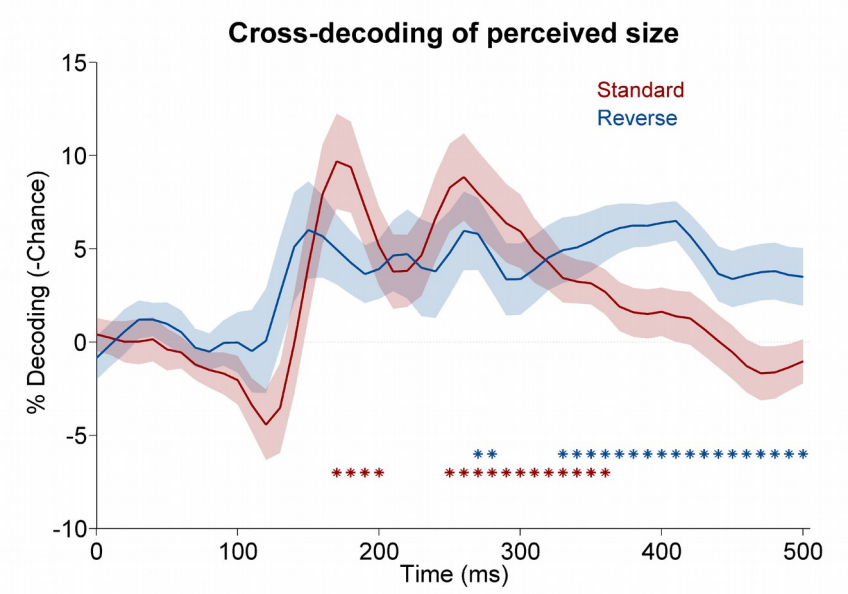

Figure 7. Results of the cross-decoding analysis of perceived size in the Ponzo illusion: time courses of standard and reverse cross-decoding conditions. Shaded colored areas reflect SEM; asterisks indicate where decoding is significantly above chance ($p < 0.05$, corrected for multiple comparisons). Decoding in the standard condition was

significant from 160ms to 190ms, and from 240ms to 350ms; peaking at 170ms. Decoding in the reverse condition was significant from 260ms to 270ms, and from 320ms to 500ms; peaking at 410ms. No difference between these two diagonals is present.

The reliability of decoding in both conditions suggest that, in a Ponzo illusion situation, two objects with same retinal size were perceived as having different sizes. The results suggest that in natural stimuli, this rescaling process appears around 160ms, while in grey stimuli emerged about 100ms later (around 260ms). One could speculate that this delay might be due to a purely "downward" (top-down) rescaling (as opposed to a more "upward" or bottom-up initiated rescaling in natural stimuli), which would take more time to be initiated because it does not come from fast feedforward signals related to perspective cues.
The absence of a difference between standard and reverse cross-decoding indicates that our grey stimuli triggered size-rescaling processes to a similar degree of natural stimuli, in line with the results of the previous analysis. This is another surprising result, because in the reverse cross-decoding classifier testing is performed in grey stimuli with objects at different positions and not containing distance information per se.

3.3.2. Cross-decoding of perceived size in typical conditions

Another way to investigate the temporal dynamics of size-constancy, and more relatable to everyday situations, is when a near-big object and distant-small object (where the object would be the same) start being perceived as having the same size (Fig. 3). In this analysis we go back to the question addressed in analysis 3.2., but employing a cross-decoding approach. In other words, when do perceived object size effect emerge (because of distance)?
The cross-decoding analysis was performed once in a standard direction (training on data from grey stimuli and testing on data from natural stimuli), and once in a reverse direction (training on data from natural stimuli and testing on data from grey stimuli). In both analyses, classifiers were trained on MEG activity patterns evoked by different retinal size (big vs. small objects, regardless of shape and distance). Then, they were tested as a function of perceived size: in the different perceived size condition, on the near-small vs. distant-big classification; in the same perceived size condition, on the near-big vs. distant-small classification. Then, the difference between the two condition (different perceived - same perceived size) was tested against zero in order to extract the time at which size-constancy is

initiated (i.e., when two objects of different retinal size, located at different distances from the observer (specifically, a near-big object and a distant-small object), start being perceived as having the same size).

We expected decoding accuracy in the different perceived size condition to be significantly above chance at all time points, because the two compared objects had different retinal size and different perceived size. Whereas, in the same perceived size condition, we expected decoding accuracy to initially rise at early time points (because objects have different retinal size) but to decrease in time (because objects start to be perceived as having the same size). Thus, we hypothesized the presence of a difference between these two conditions (different perceived - same perceived) at late time points.

Furthermore, the difference of the difference (i.e., Different-Same in Standard - Different-Same in Reverse) was evaluated to understand whether the effect was limited to a specific decoding direction, and therefore whether it was stimulus-specific (i.e., specific to natural stimuli or whether it extended to grey stimuli).

Figure 6 illustrates the results. In the standard cross-decoding (Fig. 6a), as expected, we find decoding in the different perceived size condition to be sustained in the whole time interval (specifically, it is significant from 140ms to 470ms, peaking at 170ms; Fig. 6a, red line). Contrary to our hypothesis, decoding in the same perceived size condition is reliable throughout the whole time interval as well, but for shorter intervals (from 100ms to 140ms, from 370ms to 380ms, and from 440ms to 500ms, with maximum at 130ms; Fig. 6a, blue line). Crucially, a difference between these two conditions rises at 170ms (specifically, from 170ms to 190ms, and from 240ms to 320ms; Fig. 6a, green asterisks), suggesting that size-rescaling mechanisms start to act around 170ms. In the reverse cross-decoding (Fig. 6b), we find the decoding in the different perceived size condition to be significantly above chance from 100ms to 500ms (Fig. 6b, red line), while decoding in the same perceived size condition does not reach significance (Fig. 6b, blue line). The comparison of the two conditions pointed out a significant difference from 170ms to 190ms, and from 320ms to 430ms (Fig. 6b, green asterisks), suggesting that grey stimuli elicit size-rescaling processes as well. Next, we tested whether there was a difference between the two decoding procedures (standard vs. reverse) in the time course of perceived size (as reflected by the difference between the different perceived size and the same perceived size conditions). This allowed to assess again whether grey stimuli elicited size constancy mechanisms. We found no significant difference between

the standard and reverse cross-decoding analyses, and the two time courses tended to resemble each other (Fig. 6c).

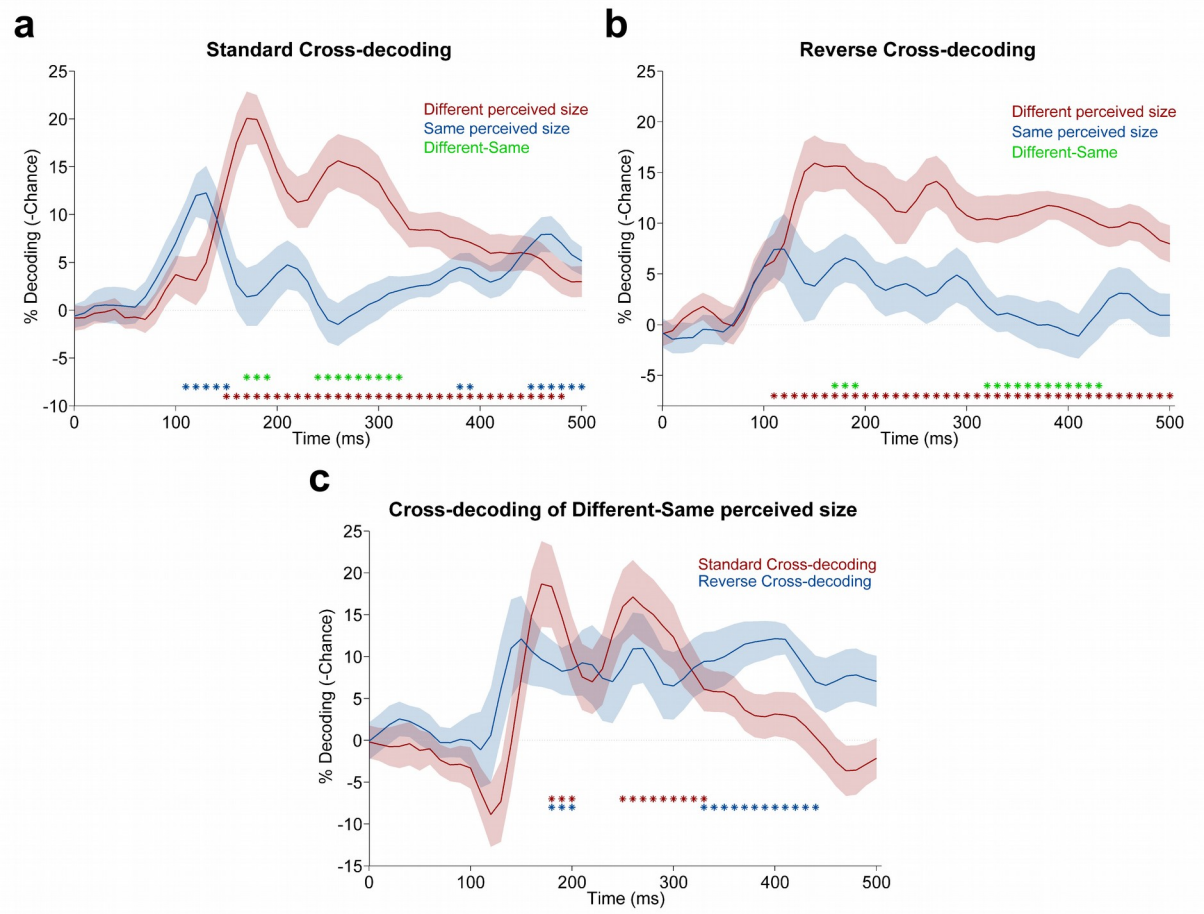

Figure 6. Time course of the standard and reverse cross-decoding of perceived size in typical situations. Shaded colored area represent SEM. Asterisks indicate where decoding accuracy is significantly above chance ($p < 0.05$, corrected for multiple comparisons). (a) Standard cross-decoding: classification in the different perceived size condition (red line) is significant from 140ms to 470ms, peaking at 170ms; decoding in the same perceived size condition (blue line) is significant from 100ms to 140ms, from 370ms to 380ms, and from 440ms to 500ms (maximum at 130ms); the difference (different-same, green asterisks) is significant from 170ms to 190ms, and from 240ms to 320ms. (b) Reverse cross-decoding: classification in the different perceived size condition (red line) is significant from 100ms to 500ms, peaking at 150ms; decoding in the same perceived size condition (blue line) is not significant (maximum at 120ms); the difference (different-same, green asterisks) is significant from 170ms to 190ms, and from 320ms to 430ms. (c) No difference between standard and reverse cross-decoding procedures was observed.

The above results show that in natural stimuli, two objects with different retinal size

(one big, one small) and with different positions (near and distant, respectively), start being perceived as having the same size around 170ms, as shown by the difference between the time courses of decoding in the different perceived size condition and the same perceived size condition (Fig. 6a). Surprisingly, also in grey stimuli, which do not provide perspective/depth cues and hence distance information, there is a difference between the time courses of decoding of different perceived size and same perceived size conditions (Fig. 6b). Furthermore, the comparison of the time course of perceived size in natural and grey stimuli (as shown by the results from the standard and reverse cross-decoding analysis, respectively; Fig. 6c) highlighted no significant difference between them, and a similarity in the temporal dynamics. Taken together, these results show that size-rescaling processes in natural stimuli begin around 170ms, and that such rescaling processes are present also in grey stimuli. In addition, the absence of a difference between the two decoding directions provides further support to the general finding that grey stimuli triggered some size constancy mechanisms, at least to a similar degree of our natural stimuli.

4. Discussion

The study of perceptual size constancy has its origins more than three centuries ago, when it was described by Renè Descartes in his Dioptrics (Descartes, 1637; Gregory, 2015). Since then, a lot of progress has been made: theories have been developed, and some of its neural mechanisms have been unraveled. However, much work is yet to be done to fully understand how the brain achieves an invariant perception of size for familiar objects. Recent fMRI studies have provided incontrovertible evidence of the neural basis of perceived size. The primary visual cortex (V1) is the first cortical area that computes signals coming from the eyes. Since the retinal image size of an object is inversely proportional to its distance (Euclid's law), it would be plausible to expect activity in V1 to be directly related to retinal size. However, the aforementioned fMRI studies actually found its activity to be related to the perceived size of an object (Murray et al., 2006; Fang et al., 2008; Sperandio et al., 2012; for a review, see Sperandio and Chouinard, 2015). This was a surprising result, because it challenged the notion that V1 activity simply reflected information related to feedforward processing, as it was commonly assumed at the time. Instead, this V1 effect was described to be related to feedback connections from higher-level areas, or from lateral connections

coming from adjacent areas computing information related to the context in which the objects were embedded, such as depth and perspective.

Because of the fMRI methodological limitations, one aspect that these studies could not address were the temporal dynamics of these perceptual rescaling processes. Furthermore, rescaling processes can be investigated at least in two ways. One simulating a situation in which two identical objects, with same retinal size, that are positioned at different distances in an image are being perceived as having different sizes (specifically, the nearest object is perceived as being smaller than the farthest object). The other simulating a situation in which a near object (having a big retinal size) and a distant object (having a small retinal size) are perceived as being the same object (i.e. having the same perceived size). The previous fMRI studies adopted only the foremost method, also known as Ponzo Illusion (Murray et al., 2006; Fang et al., 2008). We believe that these two methods are complementary in the study of size constancy, and in order to fully explain this phenomenon, both should be addressed.

In the present study, we employed a MEG decoding approach to investigate the temporal dynamics of size constancy. Participants performed an oddball task unrelated to the dependent variables of interest while recording MEG data. Objects with different shapes were embedded in natural scenes or blank grey backgrounds, and could be positioned either in the foreground (near) or in the background (distant), and have a big or small size. We performed several analysis within stimulus type (within-natural or within-grey), where we used the within-grey results as a control condition to test whether effects within natural stimuli were limited to them or generalized to grey stimuli (our initial hypothesis was that effects in natural stimuli should not emerge to the same extent in grey stimuli). In the cross-decoding procedures, in which classifier training and testing were performed with different stimulus types, we used the reverse decoding (train on natural, test on grey) as a control condition, and we expected effects to be stronger in the standard condition (train on grey, test on natural).

We run a first preliminary analysis in order to ensure that information about objects' distance could be decoded from MEG activity patterns, a necessary condition for rescaling processes to arise (as postulated by the size-distance invariance hypothesis, SDIH, at the basis of size constancy). We found decoding of object position (slightly lower for near objects and slightly upper for distant objects) to be reliable in both natural and grey stimuli; however,

decoding of object distance was more reliable in (and restricted to) natural stimuli. With the second analysis we observed that decoding of perceived size was significantly above chance in both natural and grey stimuli, with, surprisingly, no difference between them. The third analysis, investigating the temporal dynamics of perceived size in two ways, revealed not only the time course of perceived size in natural stimuli, but also the similarity in temporal patterns of size rescaling between the two stimulus types.

Taken together, two overall major results stand out. First, size-rescaling processes in natural scenes arise around 160ms after scene onset, as shown in Fig. 6 (red line, standard cross-decoding) and Fig. 7a (green asterisks, reflecting the difference between the different perceived size and the same perceived size conditions). Second, our results suggest the presence of size-rescaling mechanisms also in grey stimuli (as shown in Fig. 5, 6, and 7), which contain no bottom-up depth cues and therefore no distance information. This is surprising because, by definition, rescaling mechanisms are triggered by depth information, and grey stimuli do not provide it. Therefore, we asked whether such results could be due to other variables. For example, in the experiment, we did not record eye movements, and no fixation point was displayed in the stimulus (although the "+" presented before and after the presentation of the stimulus, and the verbal instructions, should have encouraged participants to keep their gaze on the center of the screen); therefore we had no way to know where they fixated when the stimulus was presented. A post-experiment scrutiny of the stimuli revealed that near objects tended to be closer to the center of the stimulus scene than distant objects (Supplementary Materials, 1. Stimulus analysis). Therefore, if participants fixated the center of the screen, because of foveal magnification there should be more signal/information (i.e. better decoding) for near objects than distant objects. The presence of a processing difference between near and distant objects (near > distant) could be a reason for the supposed/apparent rescaling mechanisms in grey stimuli. To clarify this issue, we ran a further analysis (Supplementary Materials, Fig. 1) in which we compared the classification of big vs. small objects as a function of their distance (in the near condition, classifiers were trained and tested on the discrimination near-big vs. near-small; in the distant condition, they were trained and tested on distant-big vs. distant-small), separately for grey and natural stimuli. We found no difference between the classification of near vs. distant objects, neither for natural nor grey stimuli, therefore ruling out the possible explanation in which our results would be due to such differences (the objects are equally strongly represented in both

locations, so the effects that we find are unlikely to be due to more general differences between far and near positions; e.g., because of fixation point). This control analysis thus provides support to the overall finding that our grey stimuli triggered some rescaling processes. But how can we explain this?

Given the absence of depth information by contextual cues, no bottom-up information on distance was present in grey stimuli; however, seeing the same objects placed in natural stimuli might lead participants to infer that lower vs. upper objects in grey stimuli are also near vs. distant, automatically activating a scene frame and generalizing some aspect of prior experience to grey stimuli. Alternatively, maybe these results highlight some more fundamental property of the visual system, linking location and size. Specifically, it could be possible that, by default, the visual system links certain locations with certain distances and real-world sizes (Kaiser and Cichy, 2018; Kaiser et al., 2018). Within a predictive coding framework, a distance representation might be inferred from stored knowledge acquired in a lifetime, either related to the fact that objects located slightly upper in the visual field tend to be distant and objects that are placed slightly lower in the visual field tend to be nearer, or that near objects subtend a larger image on the retina than distant objects. Finally, it is also possible that this effect simply reflects where participants fixated. Further research would be needed to clarify this issue.

In conclusion, the present results show that rescaling processes underlying size constancy act rather early in time, around 160-170ms after scene onset. Interestingly, these processes appear to operate with similar temporal dynamics also in blank stimuli without depth-related information, suggesting that experience might give rise to associations between object locations and size. However, some experimental lacks of this study (e.g. not recording eye movements, and not controlling the position of objects relative to the center of the stimuli), make it non-conclusive in establishing with certainty the temporal dynamics of rescaling processes and their extension to gray stimuli. Therefore, a second study will be needed to replicate the present findings. Despite these drawbacks, this study demonstrates that MEG decoding can be an important and valid tool in the study of the temporal dynamics of perceptual size constancy.

5. Supplementary Materials

1. Stimulus analysis

As a check, we tested whether near objects and distant objects were equidistant from the fixation point. To this end, we considered only stimuli with grey background (since stimuli with natural background were created just by replacing the background of grey stimuli). A one-tailed t-test revealed that the distance from the center of distant objects (mean = 74 px, SD = 45.1 px) was significantly larger ($t(179) = 5.67$, $p < 0.0001$) than the distance from the center of near objects (mean = 51 px, SD = 42.6 px).

2. Decoding control

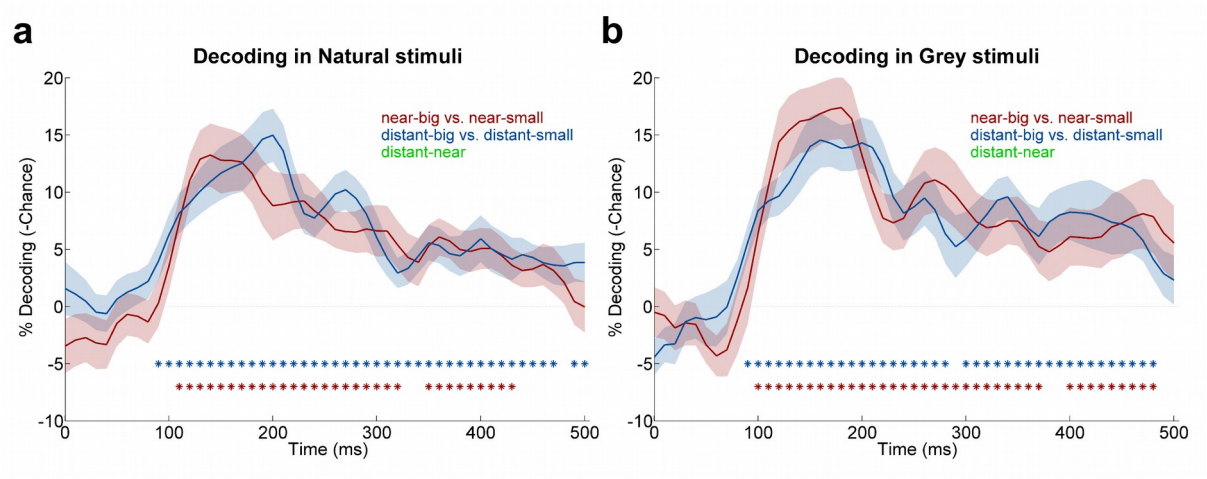

Figure 1 SM: Results of the control analysis. Time course of retinal size decoding as a function of distance and stimulus type. A control decoding analysis was performed separately within natural stimuli and within grey stimuli. Within each stimulus type, two analyses were run. First, classifiers were trained and tested on the big-near vs. small-near classification. Second, classifiers were trained and tested on the distant-big vs. distant-small classification. Then, the difference between these conditions (distant-near) was tested against zero. In terms of analysis parameters, we adopted the same procedure of the analysis described in the Methods section, paragraph 2.6. Specifically, classifier training and testing was performed on independent subsets of the data (data was divided into 10 chunks, classifiers were trained on 9 chunks and tested on 1 chunk iteratively). (a) Within natural stimuli, decoding of retinal size (big vs. small) for distant objects (blue line) was significant from 90ms to 470ms, and from 490ms to 500ms (maximum at 200ms). Decoding of retinal size for near objects (red line) was significant from 110ms to 320ms, and from 350ms to 430ms (peak at 140ms). No difference was found between the distant and near condition. (b) Within grey stimuli, decoding of retinal size (big vs. small) for distant objects (blue line) was significant from 90ms to 280ms, and from 300ms to 480ms (peak at 200ms). Decoding of retinal

size for near objects (red line) was significant from 100ms to 370ms, and from 400ms to 480ms (maximum at 180ms). No significant retinal size decoding difference was found between the conditions (distant-near). In addition, the comparison of the retinal size decoding difference as a function of distance between the two stimulus types revealed no significant difference.

3. Behavioral analysis

Behavioral performance was analyzed for target present trials (i.e., trials with golden colored objects) with stimuli with grey background vs. stimuli with natural scenes. Too-slow responses were considered as incorrect. In the RT analysis, only correct response trials were used.

A two-tailed t-test was applied on the difference between RT in grey stimuli (mean = 540ms, SD = 102) and RT in natural stimuli (mean = 537ms, SD = 155ms); it revealed no significant difference ($t(16) = -0.13$).

A two-tailed t-test was applied on the difference between response accuracy in grey stimuli (mean = 87%, SD = 12%) and response accuracy in natural stimuli (mean = 87%, SD = 9%), and no difference was found ($t(16) = 0.24$).

Chapter 6:
General Discussion and Conclusions[1]

This thesis addressed several topics related to naturalistic vision: the characteristics of attentional templates when preparing to search for objects in scenes; the temporal course of spatial attention guidance; and finally, the temporal dynamics of size-constancy mechanisms in real-world scenes.

In the following pages I will briefly summarize and discuss the results, draw the conclusions, and delineate questions for future research.

1. The characteristics of preparatory attentional templates in real-world visual search

Top-down preparatory attentional mechanisms code relevant target-defining features, biasing processing toward template-matching items once a search scene appears (Battistoni et al., 2017). However, research suggested that, in order to be most effective in terms of attentional guidance, templates could represent those features that optimally distinguish targets from distractors (Becker, 2010). Chapter 2 investigated whether such type of templates, optimally-tuned to the expected features of the context, could be established also for real-world visual search. Specifically, we aimed to determine whether expectations on the distractors' context influenced the characteristics of preparatory templates. We asked whether expecting scenes where targets and distractors had orthogonally different orientations led participants to establish attentional templates based on low-level features, compared to scenes in which targets and distractors had the same orientations and for which we expected participants to adopt a template based on category-diagnostic features, like previously shown (Reeder and Peelen, 2013). We found no evidence for a modulation of the characteristics of preparatory attentional templates as a function of the expected relation between the orientation of targets and distractors. Orientation-based, low-level templates were not engaged even when the degree of clutter of distractors in scenes was drastically reduced. This pattern of results suggests that when searching for objects in scenes, observers tend to adopt templates based on high-level category-diagnostic features, regardless of the expected relation between the features of the target and the features of the distractors. This suggests that relational target templates (Becker, 2010) do not seem to be engaged in real-world visual search tasks. It is possible that, a life-

1 This chapter contains part of journal papers published elsewhere: (1) Battistoni, E., Stein, T., & Peelen, M. V. (2017). Preparatory attention in visual cortex. *Annals of the New York Academy of Sciences*, *1396*(1), 92-107. (2) Spatial attention follows category-based attention during naturalistic visual search: evidence from MEG decoding. Elisa Battistoni, Daniel Kaiser, Clayton Hickey, Marius V. Peelen. bioRxiv 390807; doi: https://doi.org/10.1101/390807

long experience of seeing natural scenes, leads us to form templates based on category-diagnostic features that are always automatically activated regardless of contextual expectations. Therefore, even in situations in which a low-level template would seem to be the most efficient and less demanding strategy in terms of cognitive resources (e.g., intuitively, it would appear less costly to pre-activate just "vertical" than the arms, torso and other features diagnostic of a person), observers still adopt category-based templates. Activating a low-level set of features might actually be more costly than pre-activating the learned set of category-diagnostic features. In support of this hypothesis, several studies have shown that the more detailed templates are, the more efficient attentional selection is (Vickery et al., 2005; Schmidt and Zelinsky, 2009; Maxfield and Zelinsky, 2012; Wu et al., 2013).

To note, our findings would need to be replicated. Like it was found by Reeder and Peelen (2013), the validity effect is characterized by a very small, but consistent across participants, difference between RTs of different experimental conditions. This leaves open the possibility that a further modulation might be so small to go unnoticed. Therefore, to conclude with certainty that the attentional templates in real-world visual search are not influenced by distractors' context expectations, other studies will need to address this question, likely with a different approach or by making the difference between targets and distractors more extreme.

The purpose of the experiment in Chapter 3 was to determine whether expectations about the distance of targets were coded at a template level. Specifically, we adopted a paradigm similar to the experiments in Chapter 2, and we hypothesized that if templates represent the distance of expected targets, then their size would not be constant, but it would reflect the expected target's retinal size. The RTs results did not show an effect of expected target distance on the size of category-based preparatory attentional templates. However, response accuracy results suggested the presence of an effect: there was a bigger difference in response accuracy between valid and invalid trials in the consistent size condition (where the size of the silhouettes was consistent to the size of the expected target) than in the inconsistent size condition. It is possible that in the inconsistent size condition, attention was less captured towards the silhouette that matched the template because the size was not consistent, allowing participants to be more accurate on those trials, and therefore exhibiting a smaller difference of response accuracy between valid and invalid trials. This suggests that participants established small-sized templates when targets in scenes were distant (in the background), and big-sized templated when targets were near (in the foreground).

However, based on the present data, we cannot conclude with certainty that participants scaled the size of templates as a function of expected target distance: first of all, the effect was present only in accuracy data but not in RTs data; secondly, the effect size were very small.

Similarly to Chapter 2, one reason for the lack of effect in RTs might be linked to the properties of the validity effects: since the difference in RTs between valid and invalid conditions is already small – even though consistent – further modulations might be too little to be statistically significant. Therefore, before drawing final conclusions on whether templates code the expected distance-size of targets, it is important for future research to investigate more this topic.

2. The temporal dynamics of object processing in natural scenes

In daily life, attention is often directed to high-level object attributes, such as when we look out for cars before crossing a road. Previous work using MEG decoding investigated the influence of such category-based attention on the time course of object category representations (Kaiser et al., 2016). Attended object categories were more strongly represented than unattended categories from 180ms after scene onset. In Chapter 4, we used a similar approach to determine when, relative to this category-level modulation, attention is spatially focused on the target. Results showed that the location of both target and distracter objects could be accurately decoded shortly after scene onset (50 ms). However, the emergence of spatial attentional selection – reflected in better decoding of target location than distracter location – emerged only later in time (240 ms). Target presence itself (irrespective of location and category) could be decoded from 180 ms after stimulus onset.

Combined with the earlier work, these results indicate that naturalistic category search operates through an initial spatially-global modulation of category processing that then guides attention to the location of the target. This “feature- to location- based selection” (Hopf et al., 2004), also referred to as “global to local” process (Campana et al., 2016), has been proposed in classical theories of attentional selection, among which Guided Search (Wolfe et al., 1989; Wolfe, 1994) and Reverse Hierarchy Theory (RHT; Hochstein & Ahissar, 2002; Ahissar et al., 2009), and demonstrated for simple stimuli in artificial displays (Treisman and Sato, 1990; Cave, 1999; Hopf et al., 2004; Eimer, 2014; Campana et al., 2016).

The studies by Kaiser et al. (2016) and by Battistoni et al. (2018) provide strong evidence that such classical findings extend to more complex stimuli and tasks. Interestingly, together they highlight at least two stages that characterize our visual search behavior in natural scenes. Initially, at around 50ms, just briefly after the image hits the retina, MEG activity patterns encode information on the location of objects in scenes (regardless of behavioral relevance). At 180ms,

object category information is more strongly represented for target than distracters, suggesting that, at this time, top-down category-based attentional mechanisms are already engaged (the first emergence of category information is modulated by attention). This information then guides the allocation of attention to the location of the target object, as evidenced by the result that at 240 ms attention is spatially focused on the target. It is plausible to speculate that this attentional allocation paves the ways for further visual and semantic processing (Wolfe and Cave, 1999; Wolfe, 2003; Wolfe et al., 2011b).

3. Size-constancy and object processing

Size-constancy mechanisms are fundamental in everyday life, as they are one of the mechanisms that allow us to achieve an invariant perception of the objects around us. In Chapter 5, we investigated the time course of the perception of size, distance, and of the emergence of size invariance, in stimuli with grey backgrounds and stimuli with natural backgrounds using MEG decoding. From the analyses, three overall results stood out.

First, a representation of position (slightly toward the bottom of the stimulus vs. slightly toward the top of the stimulus) was decodable from MEG activity patterns elicited by objects in both grey stimuli and natural stimuli from 80ms after stimulus onset. Importantly, grey stimuli did not provide bottom-up depth cues, and therefore should not have elicited a representation of distance (because, by definition, distance perception is triggered by depth and perspective cues). Given this feature of grey stimuli, by comparing the two time courses (natural vs. grey) we were able to isolate the time at which a representation of distance (triggered by depth cues) emerged for objects in natural scenes (140ms, Fig. 4 in Chapter 5).

Second, in all the analyses in which it was possible to compare the time course of perceived size between grey and natural stimuli (Fig. 5C, Fig. 6, Fig. 7C, in Chapter 5), we found no significant difference between them. This general result suggests that grey stimuli triggered a perception of size constancy to a similar degree of natural stimuli. A possible explanation could be that, given the nature of the experiment, the visual system tends to link certain positions with certain distances (slightly bottom as near, slightly top as distant; Kaiser and Cichy, 2018; Kaiser et al., 2018).

Third, size-rescaling mechanisms in natural stimuli arise around 160-170ms after scene onset (Fig. 6, Fig. 7A, in Chapter 5). Since a coarse form of object identification is necessary for attention to be allocated on that object (as shown by the results of Chapter 4), one could hypothesize that information about the object's size might be included in the representation guiding

attention to the object. Following this logic, it is possible that a representation of object's size could appear before allocating attention to the target. Our results seem to be in line with this idea: in Chapter 5 we find size-rescaling processes (which index the presence of a representation of size) around 160-170ms, while in Chapter 4 we find that spatial attention is focused on the target around 240ms. Interestingly, several studies have shown that an invariant perception of real-world objects is achieved within 150-200ms after stimulus onset (Thorpe et al., 1996; for a review, see DiCarlo et al., 2012). Therefore, our results appear to be consistent with other research on invariant object recognition, supporting that size and object invariance tend to be computed before 200ms.

4. Closing remarks and questions for future research

This thesis constitutes a collection of evidence on some fundamental top-down attentional mechanisms acting in real-world visual search: specifically, the processing stages of preparation, guidance, selection, and identification (Eimer, 2014).

Concerning the preparatory phase, we propose that attentional templates in real-world visual search tasks are based on category-diagnostic features even when other lower-level strategies would be effective as well. Furthermore, we showed that such templates code the expected target size/distance, suggesting that there is no size invariance at the level of preparatory attentional templates.

In the context of the attentional guidance and selection stage, we demonstrate that attention spatially focuses on targets around 240ms, following category-based attentional modulations appearing at 180ms after scene onset. Moreover, we showed that size-constancy mechanisms appear before 200ms post-scene. This is in line with the expectation that a coarse identification of an object, including its size, should be computed before spatially focusing attention onto the target. This spatially global-to-local pattern provides evidence in support of theories of attentional selection (Wolfe, 1994) and it is in line with the Reverse Hierarchy Theory (Hochstein & Ahissar, 2002): *vision at a glance* happens before 200ms and includes activations of spatially-global target-category representations and computations of its size; then, *vision with scrutiny* allows to allocate spatial attention onto the target for further behavioral processing.

The findings outlined in this thesis raise some possible questions that future research could address:

- Is the timing of spatial attentional selection influenced by expected target size? Expecting a big-near target but then being presented with a small-distant target delays the timing of attentional allocation onto the target?

- Are the results on the timing of size-constancy processes extendable to even more naturalistic conditions, with cars and people instead of simple objects?
- How do size-variant attentional templated interact with size-constancy mechanisms? Future experiments could combine manipulations of attentional templates and manipulations of same vs. different perceived size to further deepen our knowledge on size-constancy mechanisms.
- Finally, virtual reality headsets could really improve the ecological validity of studies on naturalistic visual search, therefore it would be interesting if future research will being to employ such devices.

To conclude, this thesis contains studies that help to improve our understanding of top-down attentional processes engaged in real-world visual search, and paves the way for questions that future research could address.

References

Acunzo, D.J., Mackenzie, G., and van Rossum, M.C.W. (2012). Systematic biases in early ERP and ERF components as a result of high-pass filtering. J. Neurosci. Methods *209*, 212–218.

Ahissar, M., Nahum, M., Nelken, I., & Hochstein, S. (2009). Reverse hierarchies and sensory learning. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 364(1515), 285-299.

Amit, E., Mehoudar, E., Trope, Y., and Yovel, G. (2012). Do object-category selective regions in the ventral visual stream represent perceived distance information? Brain Cogn. *80*, 201–213.

Anderson, B.A., and Folk, C.L. (2010). Variations in the magnitude of attentional capture: testing a two-process model. Atten. Percept. Psychophys. *72*, 342–352.

Andrews, D.P. (1964). Error-correcting perceptual mechanisms. Q. J. Exp. Psychol. *16*, 104–115.

Barlow, H. (1961). Possible principles underlying the transformations of sensory messages. In Sensory Communication, W. Rosenblith, ed. (MIT Press), pp. 217–234.

Battistoni, E., Kaiser, D., Hickey, C., and Peelen, M.V. (2018). Spatial attention follows category-based attention during naturalistic visual search: evidence from MEG decoding. BioRxiv 390807.

Battistoni, E., Stein, T., and Peelen, M.V. (2017). Preparatory attention in visual cortex. Ann. N. Y. Acad. Sci. *1396*, 92–107.

Becker, S.I. (2010). The role of target-distractor relationships in guiding attention and the eyes in visual search. J. Exp. Psychol. Gen. *139*, 247–265.

Becker, S.I. (2013). Why you cannot map attention: A relational theory of attention and eye movements. Aust. Psychol. *48*, 389–398.

Becker, S.I. (2014). Guidance of Attention by Feature Relationships: The End of the Road for Feature Map Theories? In Current Trends in Eye Tracking Research, (Springer, Cham), pp. 37–49.

Becker, S.I., Folk, C.L., and Remington, R.W. (2010). The role of relational information in contingent capture. J. Exp. Psychol. Hum. Percept. Perform. *36*, 1460–1476.

Becker, S.I., Folk, C.L., and Remington, R.W. (2013). Attentional capture does not depend on feature similarity, but on target-nontarget relations. Psychol. Sci. *24*, 634–647.

Becker, S.I., Harris, A.M., Venini, D., and Retell, J.D. (2014). Visual search for color and shape: when is the gaze guided by feature relationships, when by feature values? J. Exp. Psychol. Hum. Percept. Perform. *40*, 264–291.

Berryhill, M.E., and Olson, I.R. (2009). The representation of object distance: evidence from neuroimaging and neuropsychology. Front. Hum. Neurosci. *3*, 43.

Berryhill, M.E., Fendrich, R., and Olson, I.R. (2009). Impaired Distance Perception and Size Constancy Following Bilateral Occipitoparietal Damage. Exp. Brain Res. Exp. Hirnforsch. Exp. Cerebrale *194*, 381–393.

Biederman, I. (1972). Perceiving real-world scenes. Science *177*, 77–80.

Biederman, I., Rabinowitz, J.C., Glass, A.L., and Stacy, E.W. (1974). On the information extracted from a glance at a scene. J. Exp. Psychol. *103*, 597–600.

Boring, E.G. (1940). Size Constancy and Emmert's Law. Am. J. Psychol. *53*, 293–295.

Braun, J. (2003). Natural scenes upset the visual applecart. Trends Cogn. Sci. *7*, 7–9.

Bravo, M.J., and Farid, H. (2009). The specificity of the search template. J. Vis. *9*, 34.1-9.

Bravo, M.J., and Farid, H. (2016). Observers change their target template based on expected context. Atten. Percept. Psychophys. *78*, 829–837.

Bundesen, C. (1990). A theory of visual attention. Psychol. Rev. *97*, 523–547.

Buschman, T.J., and Kastner, S. (2015). From behavior to neural dynamics: An integrated theory of attention. Neuron *88*, 127–144.

Campana, F., Rebollo, I., Urai, A., Wyart, V., and Tallon-Baudry, C. (2016). Conscious Vision Proceeds from Global to Local Content in Goal-Directed Tasks and Spontaneous Vision. J. Neurosci. *36*, 5200–5213.

Carlisle, N.B., Arita, J.T., Pardo, D., and Woodman, G.F. (2011). Attentional templates in visual working memory. J. Neurosci. Off. J. Soc. Neurosci. *31*, 9315–9322.

Castelhano, M.S., and Heaven, C. (2010). The relative contribution of scene context and target features to visual search in scenes. Atten. Percept. Psychophys. *72*, 1283–1297.

Castelhano, M.S., and Heaven, C. (2011). Scene context influences without scene gist: Eye movements guided by spatial associations in visual search. Psychon. Bull. Rev. *18*, 890–896.

Cate, A.D., Goodale, M.A., and Köhler, S. (2011). The role of apparent size in building- and object-specific regions of ventral visual cortex. Brain Res. *1388*, 109–122.

Cave, K.R. (1999). The FeatureGate model of visual selection. Psychol. Res. *62*, 182–194.

Chouinard, P.A., and Ivanowich, M. (2014). Is the primary visual cortex a center stage for the visual phenomenology of object size? J. Neurosci. Off. J. Soc. Neurosci. *34*, 2013–2014.

Cohen, A., and Rafal, R.D. (1991). Attention and Feature Integration: Illusory Conjunctions in a Patient with a Parietal Lobe Lesion. Psychol. Sci. *2*, 106–110.

Contini, E.W., Wardle, S.G., and Carlson, T.A. (2017). Decoding the time-course of object recognition in the human brain: From visual features to categorical decisions. Neuropsychologia *105*, 165–176.

Costa, T.L., Costa, M.F., Magalhães, A., Rêgo, G.G., Nagy, B.V., Boggio, P.S., and Ventura, D.F. (2015). The role of early stages of cortical visual processing in size and distance judgment: a transcranial direct current stimulation study. Neurosci. Lett. *588*, 78–82.

De Cesarei, A., Loftus, G.R., Mastria, S., and Codispoti, M. (2017). Understanding natural scenes: Contributions of image statistics. Neurosci. Biobehav. Rev. *74*, 44–57.

Dees, J.W. (1966). Moon illusion and size-distance invariance: An explanation based upon an experimental artifact. Percept. Mot. Skills *23*, 629–630.

Descartes, R. (1637). The Dioptrics. In Discourse on Method, Optics, Geometry, and Meteorology, p.

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. Philos. Trans. R. Soc. B Biol. Sci. *353*, 1245–1255.

Desimone, R., and Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. Annu. Rev. Neurosci. *18*, 193–222.

DiCarlo, J.J., and Cox, D.D. (2007). Untangling invariant object recognition. Trends Cogn. Sci. *11*, 333–341.

DiCarlo, J.J., Zoccolan, D., and Rust, N.C. (2012). How does the brain solve visual object recognition? Neuron *73*, 415–434.

Dobbins, A.C., Jeo, R.M., Fiser, J., and Allman, J.M. (1998). Distance Modulation of Neural Activity in the Visual Cortex. Science *281*, 552–555.

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

Duncan, J. (1989). Boundary conditions on parallel processing in human vision. Perception *18*, 457–469.

Duncan, J., and Humphreys, G.W. (1989). Visual search and stimulus similarity. Psychol. Rev. *96*, 433–458.

Eckstein, M.P., Drescher, B.A., and Shimozaki, S.S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. Psychol. Sci. *17*, 973–980.

Eckstein, M.P., Koehler, K., Welbourne, L.E., and Akbas, E. (2017). Humans, but Not Deep Neural Networks, Often Miss Giant Targets in Scenes. Curr. Biol. CB *27*, 2827-2832.e3.

Edwards, W. (1950). Emmert's law and Euclid's optics. Am. J. Psychol. *63*, 607–612.

Eimer, M. (1996). The N2pc component as an indicator of attentional selectivity. Electroencephalogr. Clin. Neurophysiol. *99*, 225–234.

Eimer, M. (2014). The neural basis of attentional control in visual search. Trends Cogn. Sci. *18*, 526–535.

Emmert, E. (1881). Grossenverhaltnisse der nachbilder. Klin. Monatsbl. Augenheilkd. *19*, 443–450.

Epstein, W. (1963). Attitudes of judgment and the size-distance invariance hypothesis. J. Exp. Psychol. *66*, 78–83.

Epstein, W., Park, J., and Casey, A. (1961). The current status of the size-distance hypotheses. Psychol. Bull. *58*, 491–514.

Evans, K.K., and Treisman, A. (2005). Perception of objects in natural scenes: is it really attention free? J. Exp. Psychol. Hum. Percept. Perform. *31*, 1476–1492.

Fang, F., Boyaci, H., Kersten, D., and Murray, S.O. (2008). Attention-dependent representation of a size illusion in human V1. Curr. Biol. CB *18*, 1707–1712.

Fei-Fei, L., Iyer, A., Koch, C., and Perona, P. (2007). What do we perceive in a glance of a real-world scene? J. Vis. *7*, 10.

Felsen, G., and Dan, Y. (2005). A natural approach to studying vision. Nat. Neurosci. *8*, 1643–1646.

Fisher, G.H. (1968). Illusions and Size-Constancy. Am. J. Psychol. *81*, 2–20.

Folk, C.L., and Remington, R. (1998). Selectivity in distraction by irrelevant featural singletons: evidence for two forms of attentional capture. J. Exp. Psychol. Hum. Percept. Perform. *24*, 847–858.

Folk, C.L., Remington, R.W., and Johnston, J.C. (1992). Involuntary covert orienting is contingent on attentional control settings. J. Exp. Psychol. Hum. Percept. Perform. *18*, 1030–1044.

Folk, C.L., Remington, R.W., and Johnston, J.C. (1993). Contingent attentional capture: A reply to Yantis (1993). J. Exp. Psychol. Hum. Percept. Perform. *19*, 682–685.

Friedman-Hill, S.R., Robertson, L.C., and Treisman, A. (1995). Parietal contributions to visual feature binding: evidence from a patient with bilateral lesions. Science *269*, 853–855.

Gabay, S., Kalanthroff, E., Henik, A., and Gronau, N. (2016). Conceptual size representation in ventral visual cortex. Neuropsychologia *81*, 198–206.

Gauthier, I., and Tarr, M.J. (2016). Visual Object Recognition: Do We (Finally) Know More Now Than We Did? Annu. Rev. Vis. Sci. *2*, 377–396.

Geisler, W.S., and Kersten, D. (2002). Illusions, perception and Bayes. Nat. Neurosci. *5*, 508–510.

Geng, J.J., DiQuattro, N.E., and Helm, J. (2017). Distractor probability changes the shape of the attentional template. J. Exp. Psychol. Hum. Percept. Perform. *43*, 1993–2007.

Gilinsky, A.S. (1951). Perceived size and distance in visual space. Psychol. Rev. *58*, 460–482.

Gnadt, J.W., and Mays, L.E. (1995). Neurons in monkey parietal area LIP are tuned for eye-movement parameters in three-dimensional space. J. Neurophysiol. *73*, 280–297.

Greene, M.R., and Oliva, A. (2009a). The Briefest of Glances: The Time Course of Natural Scene Understanding. Psychol. Sci. *20*, 464–472.

Greene, M.R., and Oliva, A. (2009b). Recognition of natural scenes from global properties: seeing the forest without representing the trees. Cognit. Psychol. *58*, 137–176.

Gregory, R.L. (1963). DISTORTION OF VISUAL SPACE AS INAPPROPRIATE CONSTANCY SCALING. Nature *199*, 678–680.

Gregory, R.L. (1968). Perceptual illusions and brain models. Proc. R. Soc. Lond. B Biol. Sci. *171*, 279–296.

Gregory, R.L. (2015). Eye and brain: The psychology of seeing (Princeton university press).

Grootswagers, T., Wardle, S.G., and Carlson, T.A. (2017). Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. J. Cogn. Neurosci. *29*, 677–697.

Gruber, H.E. (1954). The Relation of Perceived Size to Perceived Distance. Am. J. Psychol. *67*, 411–426.

Harris, A.M., Remington, R.W., and Becker, S.I. (2013). Feature specificity in attentional capture by size and color. J. Vis. *13*.

Hasson, U., Malach, R., and Heeger, D.J. (2010). Reliability of cortical activity during natural stimulation. Trends Cogn. Sci. *14*, 40–48.

He, D., Mo, C., Wang, Y., and Fang, F. (2015). Position shifts of fMRI-based population receptive fields in human visual cortex induced by Ponzo illusion. Exp. Brain Res. *233*, 3535–3541.

Henderson, J.M., and Hollingworth, A. (1999). High-Level Scene Perception. Annu. Rev. Psychol. *50*, 243–271.

Hickey, C., Di Lollo, V., and McDonald, J.J. (2009). Electrophysiological indices of target and distractor processing in visual search. J. Cogn. Neurosci. *21*, 760–775.

Hickey, C., Kaiser, D., and Peelen, M.V. (2015). Reward guides attention to object categories in real-world scenes. J. Exp. Psychol. Gen. *144*, 264–273.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. Neuron, 36(5), 791-804.

Holway, A.H., and Boring, E.G. (1941). Determinants of Apparent Visual Size with Distance Variant. Am. J. Psychol. *54*, 21–37.

Hopf, J.M., Luck, S.J., Girelli, M., Hagner, T., Mangun, G.R., Scheich, H., and Heinze, H.J. (2000). Neural sources of focused attention in visual search. Cereb. Cortex N. Y. N 1991 *10*, 1233–1241.

Hopf, J.-M., Boelmans, K., Schoenfeld, M.A., Luck, S.J., and Heinze, H.-J. (2004). Attention to Features Precedes Attention to Locations in Visual Search: Evidence from Electromagnetic Brain Responses in Humans. J. Neurosci. *24*, 1822–1832.

James, W. (1890). The principles of psychology, Vol. 2. NY, US: Henry Holt and Company.

Julian, J.B., Ryan, J., and Epstein, R.A. (2017). Coding of Object Size and Object Category in Human Visual Cortex. Cereb. Cortex N. Y. N 1991 *27*, 3095–3109.

Kaiser, D., and Cichy, R.M. (2018). Typical visual-field locations facilitate access to awareness for everyday objects. BioRxiv 297523.

Kaiser, D., Oosterhof, N.N., and Peelen, M.V. (2016). The Neural Dynamics of Attentional Selection in Natural Scenes. J. Neurosci. Off. J. Soc. Neurosci. *36*, 10522–10528.

Kaiser, D., Moeskops, M.M., and Cichy, R.M. (2018). Typical retinotopic locations impact the time course of object coding. NeuroImage *176*, 372–379.

Kaiser, D., Stein, T., and Peelen, M.V. (2014). Object grouping based on real-world regularities facilitates perception by reducing competitive interactions in visual cortex. Proc. Natl. Acad. Sci. U. S. A. *111*, 11217–11222.

Kastner, S., Pinsk, M.A., De Weerd, P., Desimone, R., and Ungerleider, L.G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. Neuron *22*, 751–761.

Kaufman, L., and Kaufman, J.H. (2000). Explaining the moon illusion. Proc. Natl. Acad. Sci. *97*, 500–505.

Kaufman, L., and Rock, I. (1962). The moon illusion. Sci. Am. *207*, 120–132.

Kaufman, L., Kaufman, J.H., Noble, R., Edlund, S., Bai, S., and King, T. (2006). Perceptual distance and the constancy of size and stereoscopic depth. Spat. Vis. *19*, 439–457.

Kaufman, L., Vassiliades, V., Noble, R., Alexander, R., Kaufman, J., and Edlund, S. (2007). Perceptual distance and the moon illusion. Spat. Vis. *20*, 155–175.

Kim, S., Carello, C., and Turvey, M.T. (2016). Size and distance are perceived independently in an optical tunnel: Evidence for direct perception. Vision Res. *125*, 1–11.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., and Broussard, C. (2007). What's new in psychtoolbox-3. Perception *36*, 1–16.

Koehler, K., and Eckstein, M.P. (2017). Beyond scene gist: Objects guide search more than scene background. J. Exp. Psychol. Hum. Percept. Perform. *43*, 1177–1193.

Konkle, T., and Oliva, A. (2011). Canonical visual size for real-world objects. J. Exp. Psychol. Hum. Percept. Perform. *37*, 23–37.

Konkle, T., and Oliva, A. (2012). A real-world size organization of object responses in occipitotemporal cortex. Neuron *74*, 1114–1124.

Li, F.F., VanRullen, R., Koch, C., and Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. Proc. Natl. Acad. Sci. U. S. A. *99*, 9596–9601.

Liu, Q., Wu, Y., Yang, Q., Campos, J.L., Zhang, Q., and Sun, H.-J. (2009). Neural correlates of size illusions: an event-related potential study. Neuroreport *20*, 809–814.

Luck, S.J., Chelazzi, L., Hillyard, S.A., and Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. J. Neurophysiol. *77*, 24–42.

Luck, S.J., and Hillyard, S.A. (1994). Spatial filtering during visual search: evidence from human electrophysiology. J. Exp. Psychol. Hum. Percept. Perform. *20*, 1000–1014.

Mack, S.C., and Eckstein, M.P. (2011). Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. J. Vis. *11*, 1–16.

Malcolm, G.L., and Henderson, J.M. (2010). Combining top-down processes to guide eye movements during real-world scene search. J. Vis. *10*, 4.1-11.

Martinez-Trujillo, J.C., and Treue, S. (2004). Feature-based attention increases the selectivity of population responses in primate visual cortex. Curr. Biol. CB *14*, 744–751.

Maunsell, J.H.R., and Treue, S. (2006). Feature-based attention in visual cortex. Trends Neurosci. *29*, 317–322.

Maxfield, J.T., and Zelinsky, G.J. (2012). Searching Through the Hierarchy: How Level of Target Categorization Affects Visual Search. Vis. Cogn. *20*, 1153–1163.

McCready, D. (1986). Moon illusions redescribed. Percept. Psychophys. *39*, 64–72.

Morgan, M.J. (1992). On the scaling of size judgements by orientational cues. Vision Res. *32*, 1433–1445.

Murray, S.O., Boyaci, H., and Kersten, D. (2006). The representation of perceived angular size in human primary visual cortex. Nat. Neurosci. *9*, 429–434.

Navalpakkam, V., and Itti, L. (2006). Top-down attention selection is fine grained. J. Vis. *6*, 1180–1193.

Navalpakkam, V., and Itti, L. (2007). Search Goal Tunes Visual Features Optimally. Neuron *53*, 605–617.

Neider, M.B., and Zelinsky, G.J. (2006). Scene context guides eye movements during visual search. Vision Res. *46*, 614–621.

Ni, A.M., Murray, S.O., and Horwitz, G.D. (2014). Object-centered shifts of receptive field positions in monkey primary visual cortex. Curr. Biol. CB *24*, 1653–1658.

Oliva, A., and Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. Int. J. Comput. Vis. *42*, 145–175.

Olivers, C.N.L., Peters, J., Houtkamp, R., and Roelfsema, P.R. (2011). Different states in visual working memory: when it guides attention and when it does not. Trends Cogn. Sci. *15*, 327–334.

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. *2011*, 156869.

Oosterhof, N.N., Connolly, A.C., and Haxby, J.V. (2016). CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. Front. Neuroinformatics *10*, 27.

Peelen, M.V., and Downing, P.E. (2017). Category selectivity in human visual cortex: Beyond visual object recognition. Neuropsychologia *105*, 177–183.

Peelen, M.V., and Kastner, S. (2011). A neural basis for real-world visual search in human occipitotemporal cortex. Proc. Natl. Acad. Sci. *108*, 12125–12130.

Peelen, M.V., and Kastner, S. (2014). Attention in the real world: toward understanding its neural basis. Trends Cogn. Sci. *18*, 242–250.

Pereira, E.J., and Castelhano, M.S. (2014). Peripheral guidance in scenes: The interaction of scene context and object content. J. Exp. Psychol. Hum. Percept. Perform. *40*, 2056–2072.

Pooresmaeili, A., Arrighi, R., Biagi, L., and Morrone, M.C. (2013). Blood oxygen level-dependent activation of the primary visual cortex predicts size adaptation illusion. J. Neurosci. Off. J. Soc. Neurosci. *33*, 15999–16008.

Potter, M.C. (1976). Short-term conceptual memory for pictures. J. Exp. Psychol. [Hum. Learn.] *2*, 509–522.

Preston, T.J., Guo, F., Das, K., Giesbrecht, B., and Eckstein, M.P. (2013). Neural representations of contextual guidance in visual search of real-world scenes. J. Neurosci. Off. J. Soc. Neurosci. *33*, 7846–7855.

Qian, J., and Petrov, Y. (2012). StarTrek illusion--general object constancy phenomenon? J. Vis. *12*, 15.

Qian, J., and Yazdanbakhsh, A. (2015). A Neural Model of Distance-Dependent Percept of Object Size Constancy. PLoS ONE *10*.

Redding, G.M. (2002). A test of size-scaling and relative-size hypotheses for the moon illusion. Percept. Psychophys. *64*, 1281–1289.

Reeder, R.R., and Peelen, M.V. (2013). The contents of the search template for category-level search in natural scenes. J. Vis. *13*, 13–13.

Reeder, R.R., van Zoest, W., and Peelen, M.V. (2015a). Involuntary attentional capture by task-irrelevant objects that match the search template for category detection in natural scenes. Atten. Percept. Psychophys. *77*, 1070–1080.

Reeder, R.R., Perini, F., and Peelen, M.V. (2015b). Preparatory Activity in Posterior Temporal Cortex Causally Contributes to Object Detection in Scenes. J. Cogn. Neurosci. *27*, 2117–2125.

Reynolds, J.H., and Chelazzi, L. (2004). Attentional modulation of visual processing. Annu. Rev. Neurosci. *27*, 611–647.

Reynolds, J.H., and Heeger, D.J. (2009). The normalization model of attention. Neuron *61*, 168–185.

Rock, I., and Kaufman, L. (1962). The Moon Illusion, II: The moon's apparent size is a function of the presence or absence of terrain. Science *136*, 1023–1031.

Romei, V., Gross, J., and Thut, G. (2010). On the Role of Prestimulus Alpha Rhythms over Occipito-Parietal Areas in Visual Input Regulation: Correlation or Causation? J. Neurosci. *30*, 8692–8697.

Ross, H.E. (1967). Water, fog and the size-distance invariance hypothesis. Br. J. Psychol. Lond. Engl. 1953 *58*, 301–313.

Ross, H.E. (2000). Cleomedes (c. 1st century AD) on the celestial illusion, atmospheric enlargement, and size-distance invariance. Perception *29*, 863–871.

Schönhammer, J.G., Grubert, A., Kerzel, D., and Becker, S.I. (2016). Attentional guidance by relative features: Behavioral and electrophysiological evidence. Psychophysiology *53*, 1074–1083.

Schwarzkopf, D.S., Song, C., and Rees, G. (2011). The surface area of human V1 predicts the subjective experience of object size. Nat. Neurosci. *14*, 28–30.

Scolari, M., and Serences, J.T. (2009). Adaptive Allocation of Attentional Gain. J. Neurosci. *29*, 11933–11942.

Sherman, A.M., Greene, M.R., and Wolfe, J.M. (2011). Depth and Size Information Reduce Effective Set Size for Visual Search in Real-World Scenes. J. Vis. *11*, 1334–1334.

Schmidt, J., and Zelinsky, G.J. (2009). Search guidance is proportional to the categorical specificity of a target cue. Q. J. Exp. Psychol. 2006 *62*, 1904–1914.

Simoncelli, E.P., and Olshausen, B.A. (2001). Natural image statistics and neural representation. Annu. Rev. Neurosci. *24*, 1193–1216.

Smith, S.M., and Nichols, T.E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage *44*, 83–98.

Sperandio, I., and Chouinard, P.A. (2015). The Mechanisms of Size Constancy. Multisensory Res. *28*, 253–283.

Sperandio, I., Chouinard, P.A., and Goodale, M.A. (2012). Retinotopic activity in V1 reflects the perceived and not the retinal size of an afterimage. Nat. Neurosci. *15*, 540–542.

Stein, T., and Peelen, M.V. (2017). Object detection in natural scenes: Independent effects of spatial and category-based attention. Atten. Percept. Psychophys. *79*, 738–752.

Sterzer, P., and Rees, G. (2006). Perceived size matters. Nat. Neurosci. *9*, 302–304.

Thorpe, S.J. (2009). The Speed of Categorization in the Human Visual System. Neuron *62*, 168–170.

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. Nature *381*, 520–522.
Tinbergen, L. (1960). The Natural Control of Insects in Pinewoods. Arch. Néerl. Zool. *13*, 265–343.

Torralba, A., Oliva, A., Castelhano, M.S., and Henderson, J.M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol. Rev. *113*, 766–786.

Treisman, A. (2006). How the deployment of attention determines what we see. Vis. Cogn. *14*, 411–443.

Treisman, A.M., and Gelade, G. (1980). A feature-integration theory of attention. Cognit. Psychol. *12*, 97–136.

Treisman, A., and Sato, S. (1990). Conjunction search revisited. J. Exp. Psychol. Hum. Percept. Perform. *16*, 459–478.

Treue, S., and Martínez Trujillo, J.C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. Nature *399*, 575–579.

Troiani, V., Stigliani, A., Smith, M.E., and Epstein, R.A. (2014). Multiple object properties drive scene-selective regions. Cereb. Cortex N. Y. N 1991 *24*, 883–897.

Tsotsos, J.K. (1990). Analyzing Vision at the Complexity Level. Behav. Brain Sci. *13*, 423–445.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. Nature neuroscience, 5(7), 682.

VanRullen, R., and Thorpe, S.J. (2001). Is it a Bird? Is it a Plane? Ultra-Rapid Visual Categorisation of Natural and Artifactual Objects. Perception *30*, 655–668.

VanRullen, R., Reddy, L., and Fei-Fei, L. (2005). Binding is a local problem for natural objects and scenes. Vision Res. *45*, 3133–3144.

Vickery, T.J., King, L.-W., and Jiang, Y. (2005). Setting up the target template in visual search. J. Vis. *5*, 81–92.

Von Helmholtz, H. (1867). Handbuch der physiologischen Optik (Voss).

Watson, D. G., & Humphreys, G. W. (1997). Visual marking: prioritizing selection for new objects by top-down attentional inhibition of old objects. *Psychological review*, *104*(1), 90.

Weidner, R., Plewan, T., Chen, Q., Buchner, A., Weiss, P.H., and Fink, G.R. (2014). The moon illusion and size-distance scaling--evidence for shared neural patterns. J. Cogn. Neurosci. *26*, 1871–1882.

Wolfe, J.M. (1994). Guided Search 2.0 A revised model of visual search. Psychon. Bull. Rev. *1*, 202–238.

Wolfe, J.M. (2003). Moving towards solutions to some enduring controversies in visual search. Trends Cogn. Sci. *7*, 70–76.

Wolfe, J.M. (2017). Visual Attention: Size Matters. Curr. Biol. *27*, R1002–R1003.

Wolfe, J.M., and Cave, K.R. (1999). The psychophysical evidence for a binding problem in human vision. Neuron *24*, 11–17, 111–125.

Wolfe, J.M., and Horowitz, T.S. (2004). What attributes guide the deployment of visual attention and how do they do it? Nat. Rev. Neurosci. *5*, 495–501.

Wolfe, J.M., and Horowitz, T.S. (2017). Five factors that guide attention in visual search. Nat. Hum. Behav. *1*, 0058.

Wolfe, J.M., Horowitz, T.S., Kenner, N., Hyle, M., and Vasan, N. (2004). How fast can you change your mind? The speed of top-down guidance in visual search. Vision Res. *44*, 1411–1426.

Wolfe, J.M., Cave, K.R., and Franzel, S.L. (1989). Guided search: an alternative to the feature integration model for visual search. J. Exp. Psychol. Hum. Percept. Perform. *15*, 419–433.

Wolfe, J.M., Alvarez, G.A., Rosenholtz, R., Kuzmova, Y.I., and Sherman, A.M. (2011a). Visual search for arbitrary objects in real scenes. Atten. Percept. Psychophys. *73*, 1650–1671.

Wolfe, J.M., Vo, M.L.-H., Evans, K.K., and Greene, M.R. (2011b). Visual search in scenes involves selective and non-selective pathways. Trends Cogn. Sci. *15*, 77–84.

Woodman, G.F., Luck, S.J., and Schall, J.D. (2007). The role of working memory representations in the control of attention. Cereb. Cortex N. Y. N 1991 *17*, i118–i124.

Wu, R., Scerif, G., Aslin, R.N., Smith, T.J., Nako, R., and Eimer, M. (2013). Searching for something familiar or novel: ERP correlates of top-down attentional selection for specific items or categories. J. Cogn. Neurosci. *25*, 719–729.

Wyble, B., Folk, C., and Potter, M.C. (2013). Contingent attentional capture by conceptually relevant images. J. Exp. Psychol. Hum. Percept. Perform. *39*, 861–871.

Wyble, B., Hess, M., Callahan-Flintoft, C., and Folk, C. (2018). Conceptual content in images triggers rapid shifts of covert attention. BioRxiv 259929.