Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search

Sage E. P. Boettcher Department of Experimental Psychology, University of Oxford, Oxford, UK

Dejan Draschkow Department of Psychology, Johann Wolfgang Goethe-Universität, Frankfurt, Germany

Eric Dienhart Department of Psychology, Johann Wolfgang Goethe-Universität, Frankfurt, Germany

Melissa L.-H. Võ Department of Psychology, Johann Wolfgang Goethe-Universität, Frankfurt, Germany

The arrangement of the contents of real-world scenes follows certain spatial rules that allow for extremely efficient visual exploration. What remains underexplored is the role different types of objects hold in a scene. In the current work, we seek to unveil an important building block of scenes—anchor objects. Anchors hold specific spatial predictions regarding the likely position of other objects in an environment. In a series of three eye tracking experiments we tested what role anchor objects occupy during visual search. In all of the experiments, participants searched through scenes for an object that was cued in the beginning of each trial. Critically, in half of the scenes a target relevant anchor was swapped for an irrelevant, albeit semantically consistent, object. We found that relevant anchor objects can guide visual search leading to faster reaction times, less scene coverage, and less time between fixating the anchor and the target. The choice of anchor objects was confirmed through an independent large image database, which allowed us to identify key attributes of anchors. Anchor objects seem to play a unique role in the spatial layout of scenes and need to be considered for understanding the efficiency of visual search in realistic stimuli.

Introduction

Our surroundings are filled with regularities—pots appear on stoves, mailboxes can be found outside, and there are only a limited number of objects that we would expect to find floating in the air. Much like in a sentence or passage, the consistencies within a scene can help us generate predictions and therefore speed up processing (Bar, 2009; Biederman, Mezzanotte, & Rabinowitz, 1982; see Figure 1). "Scene grammar" refers to the regularities that are common to our surroundings (Draschkow & Võ, 2017; Võ & Wolfe, 2013a, 2015). When scene grammar is disrupted, processes such as object recognition (Biederman et al., 1982; Davenport & Potter, 2004), visual search (Cornelissen & Võ, 2016; Võ & Henderson, 2009; Võ & Wolfe, 2013b), memorization (Draschkow, Wolfe, & Võ, 2014; Josephs, Draschkow, Wolfe, & Võ, 2016), and scene construction (Draschkow & Võ, 2017) are less efficient.

When scene grammar is left intact, it can be used to guide visual search in naturalistic environments. A large body of evidence has shown that eye movements during visual search are guided by both the physical properties of an image (Bruce & Tsotsos, 2009; Itti & Koch, 2000, 2001) as well as features of the target object (Castelhano & Heaven, 2010; Malcolm & Henderson, 2009, 2010). Although models that consider both of these inputs do fairly well at predicting eye movements during visual search in random displays (Adeli, Vitu, & Zelinsky, 2016), understanding how humans search through naturalistic scenes is more nuanced. Wolfe and colleagues (Wolfe, Alvarez, Rosenholtz, Kuzmova, & Sherman, 2011) set out to define the visual set size of complex scenes—a metric common to basic visual search research—by annotating the nontarget objects within a particular scene. They found that visual search within scenes was much more

Citation: Boettcher, S. E. P., Draschkow, D., Dienhart, E., & Võ, M. L.-H. (2018). Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search. *Journal of Vision*, *18*(13):11, 1–13, https://doi.org/10.1167/18.13.11.

https://doi.org/10.1167/18.13.11 Received February 21, 2018; published December 18, 2018 ISSN 1534-7362

Figure 1. In both language as well as scenes, the "grammar" of the input allows us to fill in the missing information (ball). Licensed under Creative Commons CC0 on pixabay.com, this image is cleared for public use.

efficient than search for arbitrary objects on a blank background and proposed that scene-guided attention effectively eliminated most regions from the "functional set size." This guidance has repeatedly been shown to drive eye movements during visual search (Henderson, Malcolm, & Schandl, 2009; Torralba, Oliva, Castelhano, & Henderson, 2006; Võ & Henderson, 2010; Wolfe, Võ, Evans, & Greene, 2011; for a review see Malcolm et al., 2016).

Understanding the particulars of scene grammar is a crucial step in understanding how we are able to perform these efficient visual searches within scenes. One aspect of scene grammar that has received relatively little attention is the differential role that objects play within scenes. Greene (2013, 2016) extensively investigated the relationship between scenes and the objects within them by using a database of annotated scenes. The study identified objects that may be diagnostic, or telling, of a scene (e.g., a sandcastle is only found on a beach). By calculating the mutual information between all objects within a scene and the scene itself, the author identified objects that are the most "informative" about a particular scene. These results suggest that not all objects are created equal on the level of scene identification, and that some objects carry more information than others.

Neuroimaging work has additionally shown that when compared to objects without a strong contextual link, diagnostic objects will elicit activity in the parahippocampal place area (PPA)—a region of the brain shown to be selectively activated by scenes (Bar & Aminoff, 2003). Although the activation in the PPA is greater in full scenes compared to diagnostic objects (Henderson, Larson, & Zhu, 2008), the processing of diagnostic objects certainly differs from objects without strong associations. It is this dissociation between types of objects that implies scene grammar may be further informed by characteristics of objects.

Importantly, diagnostic objects inform the *what* of scene grammar, leading to a cascade of predictions concerning the type of scene as well as other highly probable objects that could be found within. However, diagnostic objects do not necessarily provide predictions concerning *where* other objects are likely to be positioned in the scene. Behaviorally, such spatial predictions are extremely relevant for visual search. Mack and Eckstein (2011) provided evidence for the guiding properties of individual objects in a mobile eye tracking study. In this experiment, participants were asked to search for objects in a room with objects chaotically placed on table tops. Target objects could either be positioned in close proximity or far away from a cue (target: headphones, cue: mp3 player), and researchers found that targets were found faster when they appeared in close proximity to the cue, and relevant cue objects were fixated more often than irrelevant distractors.

The spatial predictions that objects contain will likely vary in their strength. Knowing where the soap is will not necessarily aid our search for the toothbrush, but the sink could serve as a strong cue for both of these items. Additionally, while objects like toothbrushes and toothpastes often co-occur, they do so with a less predictable spatial relationship. Moreover, we would not use the toothpaste to guide our search to the toothbrush, since that would entail first searching for the toothpaste, adding an extra search. We propose a preliminary distinction between *anchor objects* and *local objects* (similar to Draschkow & Võ, 2017), and the focus of the current study will be on defining the role of anchor objects within scenes. In our framework, so-called anchors are a subsection of objects that we propose hold a critical role in scene grammar. As a preliminary definition, anchor objects contain a high amount of spatial information about other objects—*local objects*—within the same scene. That is, anchor objects hold spatial predictions about local objects. Considering objects across many scenes, anchor objects will co-occur, are found in close proximity, and are in a spatially consistent arrangement with the local objects they predict. Moreover, it is likely that local objects will cluster around anchor objects. Together, these regularities in the environment contribute to spatial priors, which are associated with a particular subset of objects—anchors. It is also important to note that anchor objects are generally large—and therefore are easily detectable even in the visual periphery—and static—not moved often in daily life, which contributes to the reliability and predictability of anchors. Although multiple anchors may carry the same categorical prediction—fridge, stove, and sink all predict kitchen—these objects hold fundamentally different spatial predictions for other items within the scene. From this it follows that the same object cannot be strongly associated with two anchors within the same room category.

In a series of three experiments, we provide first evidence for the role of anchor objects in guiding search through scenes. To do so we have generated three-

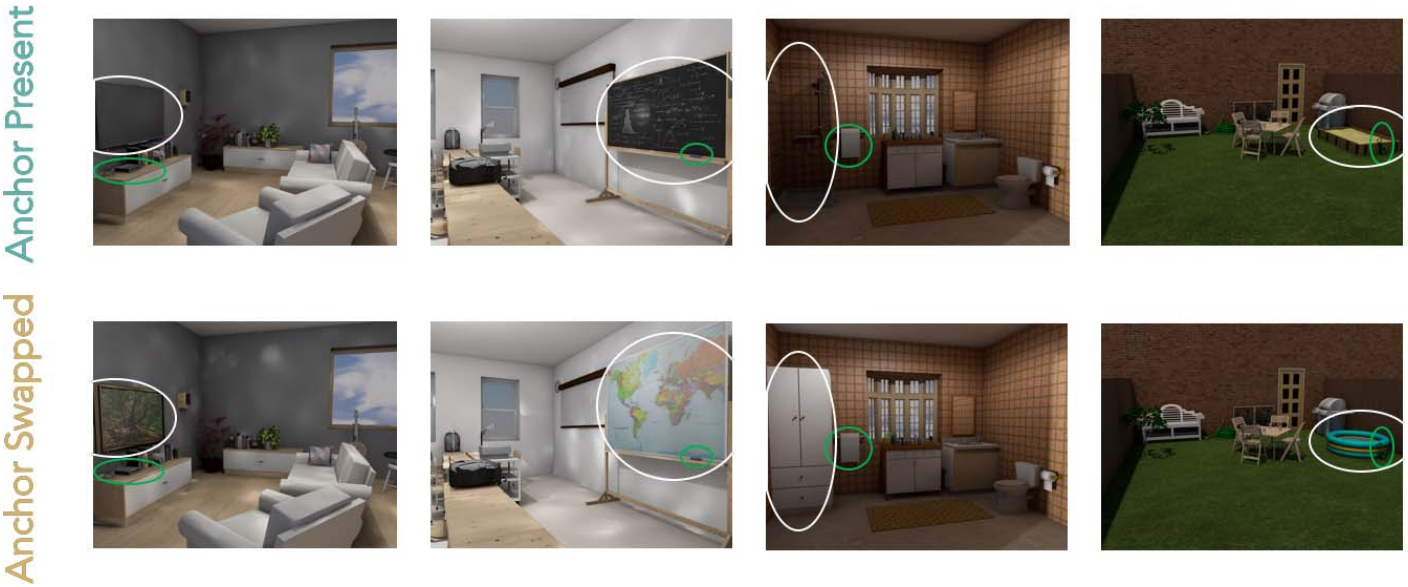

Figure 2. Examples of the 3-D rendered scenes used in Experiments 1–3. Targets are circled in green, and anchors—or their swapped counterparts—are circled in white. The top row shows the anchor-present trials (from left to right: television, blackboard, shower, sand box) and the bottom row shows the swapped images (from left to right: picture, map, cupboard, swimming pool).

dimensional (3-D) images of scenes in which a critical anchor object (e.g., the shower) was swapped out for a similar, semantically consistent surrogate object (e.g., a cupboard), which did not serve as an anchor for the current target (e.g., a towel). In Experiment 1, participants searched freely for these targets; in Experiment 2 participants were given a short preview of the target-absent scene before receiving the target probe and beginning their search with a gaze contingent window; and in Experiment 3, the preview was extended. To anticipate our results, there is a consistent effect of anchors on eye movements, and when participants' search is restricted to a gaze-contingent window, this proves to have a significant effect on response times as well. Finally, we used a large scene database to formally operationalize our definition of anchor objects. Using this independent set of stimuli, we show that the objects chosen in the current experiment show more anchor-like features with relation to the targets compared to their swapped counterparts.

Experiment 1: Unrestricted search through scenes

In Experiment 1, we wish to establish the role of anchor objects during visual search. Anchor objects, as we have defined them, are generally large stable objects within a space, and they necessarily hold a great degree of spatial information regarding other objects contained in the same space. We hypothesize that these anchors will aid visual search. In particular, we hypothesize that anchors exert their influence on both the time to locate the target as well as the time to decide whether the object is truly the target. These should become evident in the according eye-movement parameters.

Methods

Participants

Twelve participants (six females, mean age = 22.7, range = 19–26) were recruited at the Goethe University Frankfurt Psychology pool. All had normal or corrected-to-normal vision, passed the Ishihara Color Test, gave informed consent, and were volunteers receiving course credit.

Stimulus materials

The stimulus material was created and rendered using ArchiCAD software version 18 (Graphisoft, Munich, Germany). In total 64 images were created depicting 32 unique scenes (both indoor and outdoor). Each image contained an item that would be later used as a target in the search (e.g., towel in the bathroom). Each target had a particular anchor associated with it (e.g., shower). In half of the scenes this anchor was swapped out of the scene for another item that remained semantically related to the scene itself but no longer served as an anchor for the target item (e.g., cabinet). The swapped anchors were chosen to be as similar as possible in size and shape. Target objects appeared in the same location in both conditions. All images had a resolution of 1280×960 pixels and the bottom-up saliency of the anchor objects

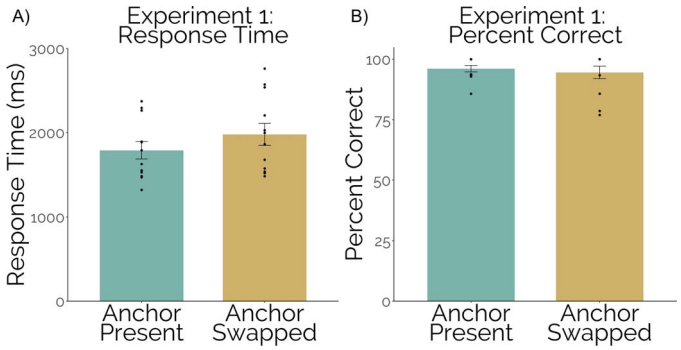

Figure 3. Mean reaction times (A) and percent correct responses (B) as a function of anchor presence. There is no significant difference between trials in which the anchor was present or swapped as measured by response time and percent correct. Error bars represent the standard error of the mean; dots represent the individual subject means.

compared to their swapped counterparts was assessed using the MATLAB (MathWorks, Natick, MA) Saliency Toolbox (Walther & Koch, 2006). The rank of the saliency peaks assigned to the anchors did not differ significantly between the present and swapped conditions. Participants searched through half of the 64 scenes such that they did not search through the same scene twice. Each observer saw 16 swapped trials and 16 present anchor trials. This was counterbalanced for every two participants. Initial analysis indicated that one of the scenes proved too challenging for the participant. That is, in this scene a maximum of one observer accurately located the target in at least one of the conditions. Therefore, this scene was excluded from the analysis. An example of several scenes can be seen in Figure 2 (for all stimuli see Supplementary Figure S1). The same stimuli will be used for Experiments 1 through 3.

Procedure

Trials began with a centrally located fixation cross. When participants were ready to start the trial, they were required to look at the fixation cross and press the space bar; thereafter they were presented with a target cue word for 750 ms. Immediately following the cue, participants could begin their search. All trials were target-present, and when participants were confident they had found the target, they were asked to fixate it and press the space bar. The experiment then continued to the next trial.

Eye tracking apparatus and analysis

The stimuli were presented on a 24-in. monitor with a refresh rate of 128 Hz. Participants were positioned in a chinrest 65 cm from the screen, and eye movements were recorded with the Eyelink-1000+ desktop mount (SR Research, Ontario, Canada) at 1000 Hz. Interest areas for the anchor objects were defined as the smallest polygon that contained both the swapped and present version of the anchor. Target interest areas were also polygons. Saccades and fixations were extracted from raw gaze data during recording by the Eyelink parser. Velocity and acceleration thresholds were set to the Eyelink default values of 30°/s and 8000°/s^2 respectively. For all analyses, we used the ez (Lawrence, 2013), lsr (Navarro, 2015), and StatCheck (Epskamp & Nuijten, 2016) packages in R. Figures were generated using ggplot2 (Wickham, 2009).

Results

Behavioral

Figure 3 shows the effect of the swapped anchor on observers' behavior. There was no significant effect of the swapped anchor on either the response time, $t(11) = 1.77$, $p = 0.10$, $d = 0.51$, or the accuracy, $t(11) = 0.75$, $p = 0.47$, $d = 0.22$. However, in both cases, the numerical difference is in the expected direction—with targets in scenes containing anchor objects found faster and more accurately.

Eye tracking

For a more fine-grained investigation of search behavior we computed the decision time, the time of the first fixation on both the anchor as well as the target, the anchor-target transition time, and the percent of the scene that is covered during the search. Decision time was defined by the time between first fixation of the target and time of response. The time to first fixation was measured from the onset of the scene until observers first fixated the interest area (either the anchor or the target). The anchor-target transition time—computed as the time between fixating the anchor and the target—will lead to further understanding of the amount of guidance the anchor has relative to the target. Trials in which the anchor was not fixated were not included in this measure. Scene coverage is a common measure indicating what percentage of an area is covered during search and was calculated by assuming a 2° circle around each fixation. The total area of the fixations was then calculated and divided by the total area of the search display. Once again, we will be comparing trials with an anchor associated with the target to trials in which the anchor has been swapped.

Decision time, $t(11) = 0.65$, $p = 0.53$, $d = 0.18$, time of first fixation (anchor: $t[11 = 1.16$, $p = 0.15$, $d = 0.45$; target: $t[11] = 1.78$, $p = 0.10$, $d = 0.52$), and the anchor-target transition time, $t(11) = 1.19$, $p = 0.26$, $d = 0.34$, were not significantly different between the two trial types. However, participants covered significantly less of the display when the anchor was present compared with when the anchor was swapped, $t(11) = 4.00$, $p = 0.002$, $d = 1.15$ (Figure 4).

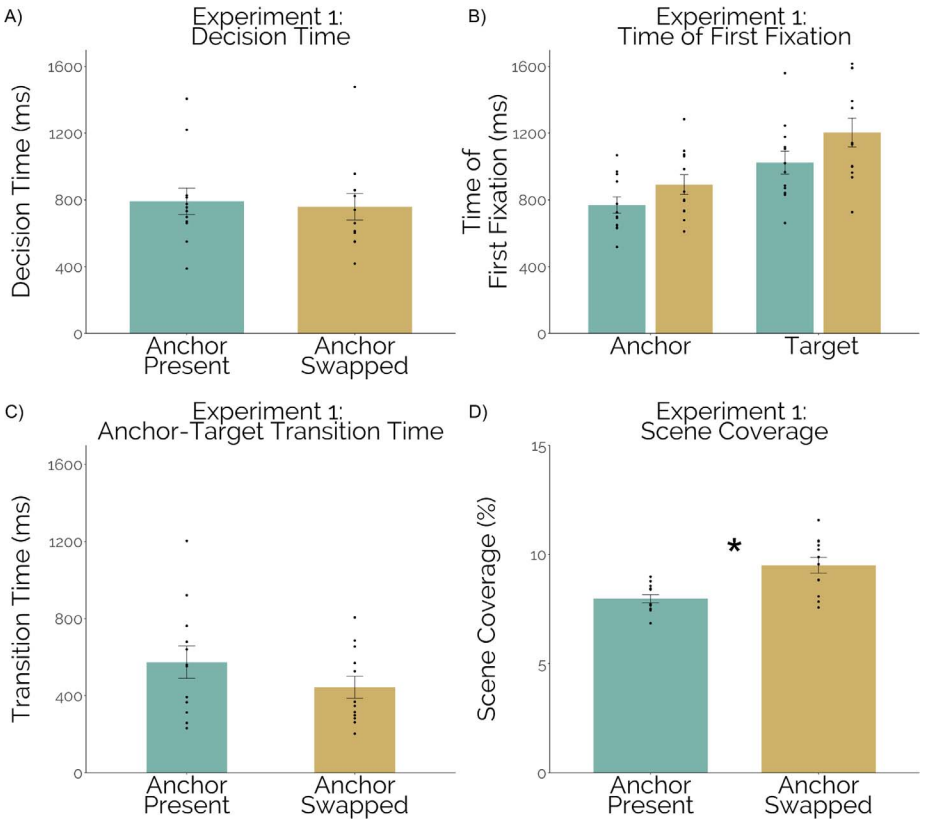

Figure 4. Mean decision time (A), time of first fixation (B), transition time (C), and scene coverage (D) as a function of anchor presence. Decision time, time of first fixation, and anchor-target transition time are statistically equivalent for trials in which the anchor was swapped. However, observers covered significantly less of the scene when the anchor was present compared to swapped. Error bars represent the standard error of the mean; dots represent the individual subject means.

Discussion

Participants were not significantly faster in locating the target in the 3-D rendered scenes when the corresponding anchor was present compared to swapped. However, the eye movements reveal there was a difference in the amount of the scene searched. This indicates that participants had slightly different search strategies between the two conditions. This difference was not evident in the decision time, the time of first fixation, or the time between fixating the anchor and the target. Scene coverage encompasses a combination of guidance variables and thus provides a more sensitive measure of guidance. That is, the numerical differences present in a multitude of variables combine to reveal a significant difference in scene coverage.

It is worth noting that the response times in this experiment were relatively fast—participants took less than 2 s to find the target. Unfortunately, the nature of the stimuli we used in these experiments (3-D–generated scenes)—although they were easy to manipulate allowing us to swap anchors in and out of scenes—might have been too sparse in comparison to photographs of indoor scenes and therefore might not have allowed for enough RT variance. It is therefore likely that there was simply not enough time for the anchor to speed search. In the following two experiments, we will explore how visual search is conducted through these same scenes, but with participants' eye movements limited to a small gaze-contingent window to eliminate information from the periphery. Moreover, we provided the participants with a brief preview of the scene before search was performed under the restricted viewing conditions with the goal of observers activating the representation of the full scene, particularly the anchor objects, to guide search.

Experiment 2: Gaze contingent search with a short preview

Methods

Participants

Twelve new participants (seven females, mean age = 22.9, range = 18–27) were recruited at the Goethe University Frankfurt Psychology pool. All had normal

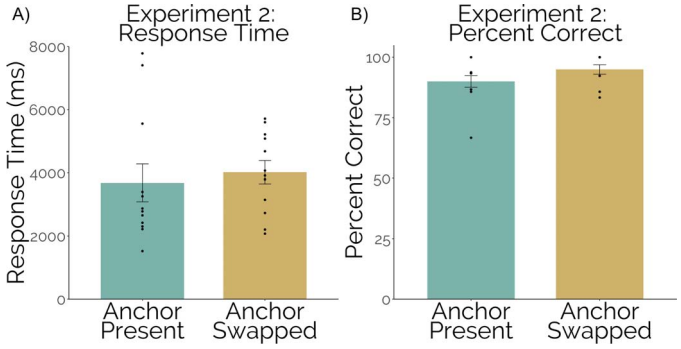

Figure 5. Mean reaction times (A) and percent correct responses (B) as a function of anchor presence. Once again, the response times and percent correct did not differ between the conditions. Error bars represent the standard error of the mean; dots represent the individual subject means.

or corrected-to-normal vision, gave informed consent, and were volunteers receiving course credit.

Procedure

Trials again began with a centrally located fixation cross that participants fixated and pressed the space bar when they wished to begin the trial. They were then presented with a target-absent scene preview for 250 ms. These target-absent scenes were generated by simply deleting the target from the 3-D models and rendering the scenes again. This preview was intended to provide observers with a template of the scene that they could use to guide their eye movements despite the restrictions of the gaze-contingent window. Immediately following this scene preview observers saw the target word for 750 ms. Participants then conducted their search through a gaze–contingent window of 6° diameter. We used a flash-preview moving window paradigm (Castelhano & Henderson, 2007; Võ & Henderson, 2010) to make search more difficult while simultaneously forcing participants to rely heavily on the brief preview of the full scene—notably absent of the target—to guide their search behavior. When they were confident they had found the object in question they looked at the target and pressed the space bar. The experiment then continued to the next trial. If observers are able to utilize the information extracted during the preview then present anchors should guide search more than their swapped counterparts.

Results

Behavioral

Participants again did not differ in their reaction times, $t(11) = 0.89$, $p = 0.39$, $d = 0.26$, or percent correct, $t(11) = 1.82$, $p = 0.10$, $d = 0.53$. However, the overall reaction times increased from Experiment 1, indicating that the gaze-contingent window did make the experiment more difficult (Figure 5).

Eye tracking

Again, there was no significant effect of the swapped anchor objects on the decision time—the time between first fixating the target and responding to it, $t(11) = 1.32$, $p = 0.21$, $d = 0.38$. Moreover, we did not find a significant difference in the first fixation time on the anchor or the target when comparing anchor present and swapped trials (anchor: $t[11] = 0.35$, $p = 0.73$, $d = 0.10$; target: $t[11] = 1.59$, $p = 0.14$, $d = 0.46$). However, similar to Experiment 1, we again see that observers searched through significantly less of the scene when the anchor was present compared to swapped, $t(11) = 3.48$, $p = 0.005$, $d = 1.0$. Additionally, now that participants' search was restricted to a gaze-contingent window, there was significantly less time between fixating the anchor and the target compared to the time between fixating the swapped anchor and the target, $t(11) = 2.69$, $p = 0.02$, $d = 0.78$ (Figure 6).

Discussion

We once again found that the state of the anchor affected how the search was conducted. In a replication of Experiment 1, we found that participants searched through significantly less of the scene when the correct anchor was present compared to swapped. Furthermore, after fixating the anchor, participants needed less time to find the target when the anchor was present compared to swapped. When observers had full access to the visual field in Experiment 1, we did not see a difference in this measure. Increasing search times by imposing a gaze-contingent window resulted in strengthening the effect of the anchor. Namely, we found the transition time between fixating the anchor and the target was significantly reduced when the anchor was informative. This indicates that the anchors contain guiding properties relative to the target items.

We hypothesized that with a preview of the scene followed by restricted viewing, the guidance imposed by the anchor would be made more useful and result in faster reaction times in the anchor present compared with the anchor swapped condition. This was not the case. We hoped that participants would gain enough information regarding the anchor during the scene preview. We used a preview time of 250 ms, which is more than enough of time to identify the category of a scene (Potter, 1976); however, it may not be enough time to discern a sufficient amount of objects—

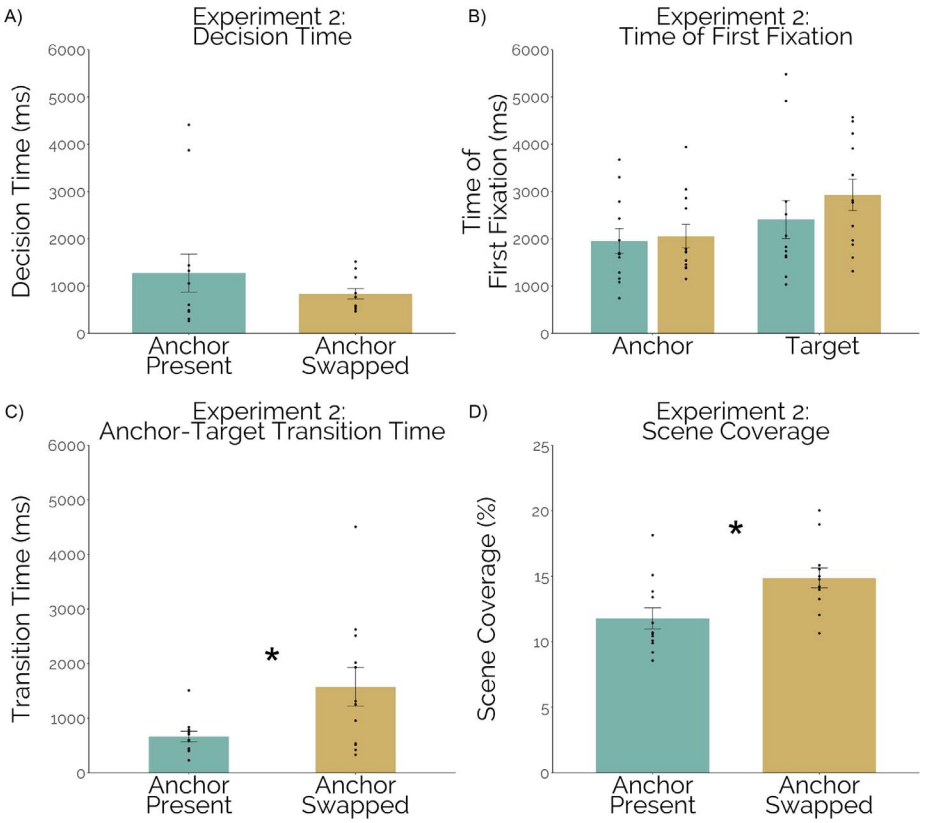

Figure 6. Mean decision time (A), time of first fixation (B), transition time (C), and scene coverage (D), as a function of anchor presence. There was no significant difference between anchor-present and anchor-swapped trials found in decision time or time of first fixation. Observers did spend less time between fixating the anchor and the target when the anchor was present, and once again observers covered significantly less of the scene when the anchor was present compared to swapped. Error bars represent the standard error of the mean; dots represent the individual subject means.

including the anchor—within the scene. If subjects could not consistently discern the anchor during the preview, it is logical that they would not necessarily show faster time to first fixation or reaction times. However, if participants came across the anchor during the gaze-contingent search, they showed significantly faster transition time to the target due to the spatial prediction provided by the anchor. In the next experiment, we extended the amount of time of the preview to 750 ms allowing the subjects to gain a more robust representation of the scene before beginning their limited window search.

Experiment 3: Gaze-contingent search with a longer preview

Methods

Participants

Twelve new participants (nine females, mean age = 25.1, range = 21–33) were recruited at the Goethe University Frankfurt Psychology pool. All had normal or corrected-to-normal vision, gave informed consent, and were volunteers receiving course credit.

Procedure

The trial procedure was exactly the same as Experiment 2 with one exception. In Experiment 3, participants were given 750 ms to preview the scene rather than 250 ms. This preview was once again followed by the target cue, which was again followed by a gaze-contingent search of the scene.

Results

Behavioral

With a longer scene preview, we found a significant difference between the reaction times in the two conditions, $t(11) = 2.5$, $p = 0.03$, $d = 0.73$. Once again, there was no difference between the accuracy in the anchor present and the anchor swapped conditions, $t(11) = 0.21$, $p = 0.84$, $d = 0.06$ (Figure 7).

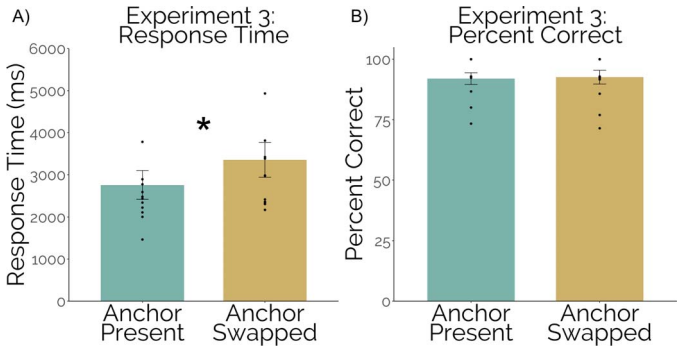

Figure 7. Mean reaction times (A) and percent correct responses (B) as a function of anchor presence. Observers were significantly faster at reporting the target when the anchor was present compared to swapped there was no significant difference in the accuracy. Error bars represent the standard error of the mean; dots represent the individual subject means.

Eye tracking

The results of the eye movements in Experiment 3 (Figure 8) were a complete replication of Experiment 2. There was no significant effect of anchor presence on decision time, $t(11) = 0.373$, $p = 0.72$, $d = 0.11$, and only a trend for the time of first fixation on the anchor or target (anchor: $t[11] = 0.40$, $p = 0.69$, $d = 0.12$; target: $t[11] = 1.81$, $p = 0.09$, $d = 0.52$). Once again, we found significant effects on the anchor-target transition time, $t(11) = 2.26$, $p = 0.04$, $d = 0.65$, and on scene coverage, $t(11) = 2.23$, $p = 0.04$, $d = 0.64$.

Discussion

In Experiment 3 we extended the preview time of the scene to 750 ms. In doing so, we observe a significant effect of the anchor presence on the reaction times. Previously we were using a preview time of 250 ms, which may not have been enough time to make full use of the anchors within the room. With more time during the preview, observers gained knowledge of the structure of the room and some of the key anchors within it. Therefore, when they were subsequently cued with the search target, the knowledge of the relevant anchor could be used to guide their search. This was not possible when the anchor was swapped.

In addition to the behavioral effects, in Experiment 3 we replicated the eye tracking effects from Experiment

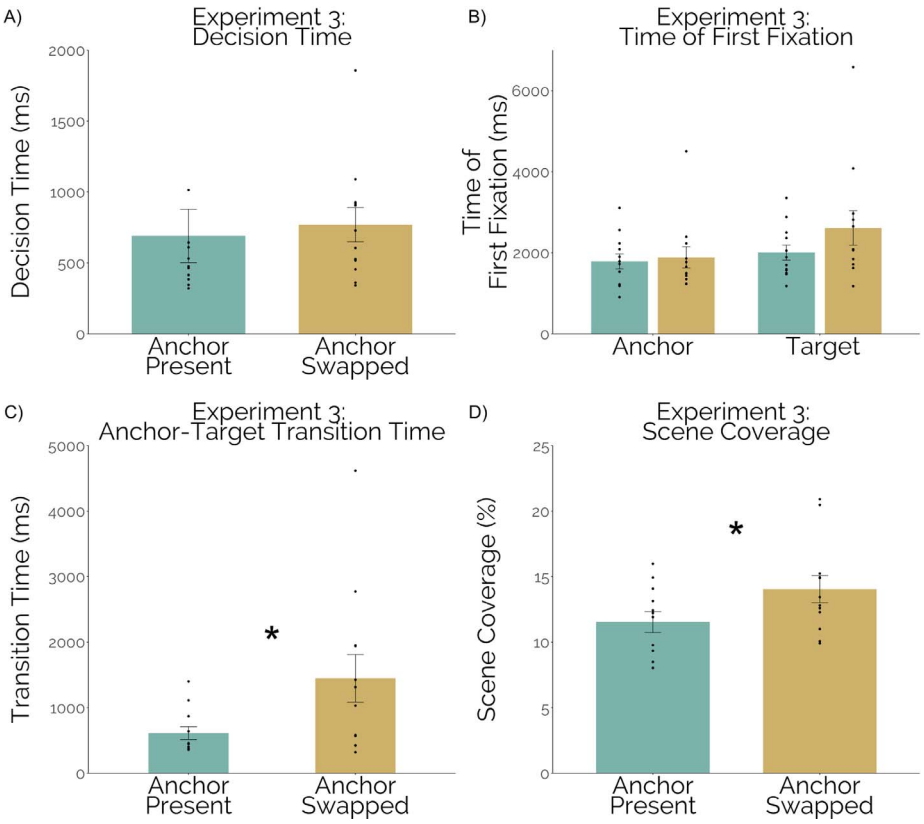

Figure 8. Mean decision time (A), time of first fixation (B), transition time (C), and scene coverage (D), as a function of anchor presence. In a replication of Experiment 2 we found no significant difference between anchor-present and anchor-swapped trials found in decision time. Observers spent less time between fixating the anchor and the target when the anchor was present, and once again observers covered significantly less of the scene when the anchor was present compared to swapped; dots represent the individual subject means.

2. We once again found that observers were faster to fixate the target after having fixated the anchor, when the anchor was not swapped. Moreover, participants covered less of the scene during the search in anchor present trials. We did not, however, find an effect of anchor presence on decision time in any of the three experiments. Although we cannot draw strong conclusions from this null effect, it does hint at the fact that an anchor object may not play a critical role in the recognition of the target.

Data driven operationalization of "anchor" concept

The stimuli for the three experiments were generated based on a theoretical notion of what constitutes an "anchor" object, but the actual selection of the stimulus was intuition-driven. In an attempt to provide a more objective measure of "anchorness," we operationalized the constraints of what an anchor actually is. We then validated this formulation using a large, annotated scene data base.

Methods

Stimuli

We used the complete SUN database (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010) with object annotations edited by human observers using the LabelMe Toolbox (Russell, Torralba, Murphy, & Freeman, 2008). In a first preprocessing step, we extracted a subset of scenes that contained the targets from the stimuli used in the current study. Note that not all targets were present in the scene database, leaving us with 23 target categories. In order to match objects in the database with the object in our study, we renamed synonymous word labels to fit the labels of the objects we used in our study (e.g., "fireplace utensils" to "firetools," "washing machine" to "washer," "night table" to "night stand," etc.). Further, we removed labels of annotations that were not of central interest to our theoretical notion of object-to-object relationships, such as ceilings, windows, floors, doors, pipes, etc. Finally, we equated divergent spelling, plurals, and redundant synonyms.

Analysis and results

The main aim of the analysis was to quantify the spatial relationship between objects in the database and validate the anchor selection in this study in a formalized fashion. Our approach aims at providing measures that reflect the strength of the "anchorness" of an object.

Crucially, in order to provide an independent measure, all metrics presented here are calculated on the scene information of the SUN database and not the stimuli used in the study. The values extracted from the SUN scenes were then separated according to the conditional assignment (anchor vs. anchor-swapped) employed in the three experiments, which ensures that all metrics are blind to the scene structures used in our study, thus providing a generalization of scene statistics from the SUN database to our stimulus set. Below we have outlined four critical components of an anchor.

Object pair frequency (*OPF*) reflects the frequency of object cooccurrences and is calculated by counting how often object pairs occur in a scene category (i.e., the object corresponding to the target identity used in our study appeared together with the anchor [normal or swapped] in all scenes which contained the target). This value is then divided by the total number of scenes in which the target was present; therefore, an OPF of 0.5 means that the anchor was present in 50% of the scenes that contained that target. Figure 9a depicts the average OPF extracted from the SUN scenes separated for both present and swapped anchor labels. For this stimulus set we would not expect that the present anchors appear more often in the same scene with the target since we chose swapped anchors that remained semantically consistent with the scene. Indeed, using an independent samples t test we find there is no significant difference between swapped and present anchors, $t(28.5) = 0.32$, $p = 0.75$, $d = 0.11$.

Object mean distance (*OMD*) reflects the mean distance between object pairs across scenes in the database—normalized for the size of the image. This measure captures the close spatial relationship between anchors and their associates. If anchor objects consistently appear closer to targets than the swapped objects, this should be reflected in the OMD. The value is inverted such that smaller distances are represented by larger values. This can be seen in Figure 9b. There was a significant difference between present and swapped anchors, $t(29.8) = 3.7$, $p < 0.001$, $d = 1.25$.

Object spatial vertical variance (*OSV*) is defined as the average spatial variance across scenes between object pairs. That is, on average how consistently did two objects appear across scenes with a similar arrangement (i.e., the toothbrush above the sink). The inverse variance for swapped versus normal anchors can be seen in Figure 9c, and shows a significantly higher value for the present anchor compared to the swapped, $t(25.73) = 2.05$, $p = 0.05$, $d = 0.73$. Again, inverse variance is used here in order to represent smaller variance with larger values. We have concentrated on vertical spatial variance as this most accurately captures the common deviations among

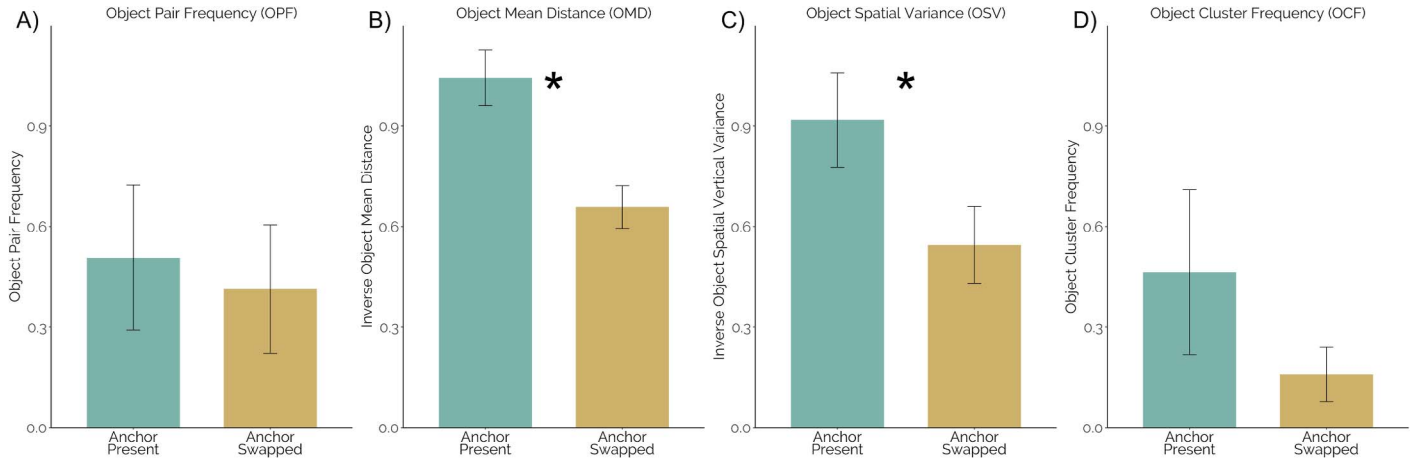

Figure 9. Four key metrics of anchorness. OPF (A) refers to the frequency of object co-occurrence. The OMD (B) is the inverse of the mean distance between the anchor and the target item across multiple scenes. The OSV (C) refers to the inverse variance of the vertical position of the target relative to the anchor. The OCF (D) is the frequency in which an object is the largest item within a cluster.

objects. That is, toothbrushes are almost always found above the sink; however, the degree to the left or right may vary across scenes.

Object cluster frequency (*OCF*) was derived from a clustering algorithm that was applied to each scene. Implemented in the pamk() function of the R package fpc (Hennig, 2018), partitioning around medoids with estimation of number of clusters uses the average silhouette width to minimize the average distance between nodes within a cluster while simultaneously maximizing the average distance between nodes of different clusters (Calinski & Harabasz, 1974). Once the optimal clusters are found, we extracted the largest object from each cluster from each scene. The OCF is a measure of the frequency of cluster "wins" divided by the total number of scenes for a particular object. The OCF reflects how often objects cluster around a particular anchor. The more scenes in which that anchor appears as the largest object in a cluster, the higher the OCF for that object. Again, we attempted to choose stimuli that were matched for size and semantically related to the scene; in doing so, we see that the OCF measure between swapped and present anchors did not differ significantly, $t(24.03) = 1.12$, $p = 0.25$, $d = 0.37$. The average OCF is shown in Figure 9d for both swapped and normal objects.

Discussion

The definition of anchors that we have outlined here captures some of the most important aspects of anchor objects that could explain their role in visual search. That is, anchors appear often (OPF) and in close proximity (OMD) to the smaller objects we are often searching for. Moreover, these anchor objects and local objects appear with a predictable spatial layout (OSV)—suggesting that the location of the anchor also provides a reliable prediction for the location of the local object. Finally, it seems that objects tend to cluster around these large anchor objects (OCF).

Using a large database of scenes, we confirmed that the anchors in the current study had a lower mean distance and spatial variance with the corresponding target objects compared with the swapped anchors. It is unsurprising—and even reassuring—that these particular stimuli did not differ in the OPF or the OCF since these are measures of semantic relation and general clustering, respectively. Since we attempted to control for these characteristics between the swapped and present anchors, and rather manipulate the amount of spatial information that the critical item contained regarding the target, the current results confirm our choice of stimuli. Taken together, the anchor objects chosen for this study adhere to a formalized definition of anchor-target pairs. Moving forward, we would suggest using similar methods when defining anchor objects for future studies.

General discussion

We have defined anchor objects as objects that hold specific spatial information regarding other objects within a scene. This can be operationalized through the frequency in which objects appear together, the distance between objects, as well as the variance of the spatial location. Finally, it is useful to consider how

objects cluster within scenes when defining anchor objects. We have formally confirmed the stimuli chosen for a series of three experiments through a large database of scenes.

Throughout the three experiments anchor objects consistently affected visual search. In all three experiments, participants searched through significantly less of the scene when the critical anchor object was related to the target. Coverage is a variable that encompasses several measures of guidance. Although the difference in coverage was not always borne out in reaction times, our analysis indicated that participants were numerically faster to fixate both the target and the anchor when the anchor was not swapped. Once again, these differences did not reach significance; however, they were present in all three experiments and may help to explain some of the differences in the coverage analysis.

When we restricted the visual field of view through a gaze-contingent window (Experiments 2 and 3) we found that the time between fixating the anchor and the target was significantly longer when the anchor was swapped compared to present. We hypothesize that under normal viewing conditions (Experiment 1) the search performance was near ceiling and thus anchors could not contribute additional beneficial guidance. It is important to note that we used 3-D rendered scenes, which, although they were designed to mimic real scenes, do not compare to the complexity of the real world. This is exemplified through the particularly fast reaction times in Experiment 1—less than 2 s. Under more natural constraints (e.g., with more clutter and complexity) we would expect the anchors to play are larger role as demonstrated in Experiments 2 and 3.

In Experiments 1 and 2 we failed to find any effect of the anchor object on response time. In Experiment 1, this may be driven by the limited variance of the relatively fast reaction times. In Experiment 2, it is likely that observers were not given enough time to consistently integrate bottom-up scene information and target template before entering into the search (see Võ & Henderson, 2010). In a concrete example, imagine a kitchen scene in which you are required to find a pan. If the preview of this scene does not allow for enough time to discern the stove, this anchor is not useful in guiding your search. However, once the stove is fixated, the search is guided to the pan (anchor – target transition time). If the initial representation of the scene *does* include the stove, one would benefit from this guidance immediately upon starting the search. With more time to preview the scene we find an effect on response time in Experiment 3. Once again, it is necessary to point out how our experimental design differs from the real world. Observers were necessarily searching through unfamiliar scenes for their targets. In our daily lives, this is rarely the case. Most of the visual searches we are conducting are within environments that are well known to us such as our homes or offices. Therefore, although we may spend time searching for the remote control in the living room, we have robust representation of the general layout of the room, such that a search for the coffee table would be trivial. It follows, that we may use our knowledge of these "anchor objects" to guide our search for the remote control. How and to what degree we use anchor objects in *familiar* scenes is an open question and would require additional experiments.

In Experiment 3, we found significant effects of anchor presence on the amount of the scene covered, the anchor-target transition time, and the response time. Participants previewed the scenes for 750 ms before they began a gaze-contingent search through the scene. On average, the anchor was fixated during this preview window on 27% of the trials. Although 750 ms is enough time to make up to two eye movements, it is important to note that participants' gaze during this time did not differ significantly between anchor-present and -swapped trials—that is, both present and swapped anchors were fixated at the same rate. Since observers were unaware of the target of the search during the scene preview, it is likely they used this time to quickly explore the layout of the scene.

We have demonstrated the additional role of anchor objects in efficiently guiding visual search and have established a need for considering these objects and their role in scene grammar (e.g., Draschkow & Võ, 2017). However, it remains unclear how these anchors are being utilized in guiding our search. At what stage during search do anchors exhibit control over attention allocation? For instance, the search template may be modulated by the associated anchor objects. If we are searching for the chalk in the classroom, it is widely accepted that we hold some representation of "chalk"—small, cylindrical, white, object—in memory as we conduct our search. It has been shown that the fidelity of that search template is an important component in visual search (Hout & Goldinger, 2015). However, in most of the previous research on search templates, experimenters only consider the features of the target itself. Given the modulating role of anchor objects during real-world search demonstrated in the current study, it could be interesting to consider whether—and if so, how—anchor objects are incorporated into search templates. Therefore, perhaps when we are searching through the kitchen for a batch of freshly baked cookies, we are additionally holding the features of the stove in memory as well.

Keywords: visual search, scene grammar, eye movements, anchors, predictions

Acknowledgments

This work was supported by Deutsche Forschungsgemeinschaft (DFG) grant VO 1683/2-1 and by SFB/TRR 135 project C7 to MLV. We wish to thank Maximilian Scheuplein and Marvin Schröder for valuable help with data collection.

Commercial relationships: none.
Corresponding author: Sage E. P. Boettcher.
Email: sage.boettcher@psy.ox.ac.uk.
Address: University of Oxford, Department of Experimental Psychology, Brain & Cognition Lab, Oxford Center for Human Brain Activity, Oxford, UK.

References

Adeli, H., Vitu, F., & Zelinsky, G. J. (2016). A model of the superior colliculus predicts fixation locations during scene viewing and visual search. *Journal of Neuroscience*. Retrieved from http://www.jneurosci.org/content/early/2016/12/30/JNEUROSCI.0825-16.2016

Bar, M. (2009). The proactive brain: Memory for predictions. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *364*(1521), 1235–1243, http://doi.org/10.1098/rstb.2008.0310.

Bar, M., & Aminoff, E. (2003). Cortical analysis of visual context. *Neuron*, *38*(2), 347–358. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12718867

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*(2), 143–177. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7083801

Bruce, N. D. B., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, *9*(3):5, 1–24, http://doi.org/10.1167/9.3.5. [PubMed] [Article]

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics: Theory and Methods*, *3*(1), 1–27, http://doi.org/10.1080/03610927408827101.

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

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 753–763, http://doi.org/10.1037/0096-1523.33.4.753.

Cornelissen, T. H. W., & Võ, M. L.-H. (2016). Stuck on semantics: Processing of irrelevant object-scene inconsistencies modulates ongoing gaze behavior. *Attention, Perception & Psychophysics*, http://doi.org/10.3758/s13414-016-1203-7.

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

Draschkow, D., & Võ, M. L.-H. (2017). Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific Reports*, *7*(1), 16471, http://doi.org/10.1038/s41598-017-16739-x.

Draschkow, D., Wolfe, J. M., & Võ, M. L.-H. (2014). Seek and you shall remember: Scene semantics interact with visual search to build better memories. *Journal of Vision*, *14*(8):10, 1–18, http://doi.org/10.1167/14.8.10. [PubMed] [Article]

Epskamp, S., & Nuijten, M. B. (2016). *statcheck: Extract statistics from articles and recompute p values*. Retrieved from http://cran.r-project.org/package=statcheck

Greene, M. R. (2013). Statistics of high-level scene context. *Frontiers in Psychology*, *4*, 777, http://doi.org/10.3389/fpsyg.2013.00777.

Greene, M. R. (2016). Estimations of object frequency are frequently overestimated. *Cognition*, *149*, 6–10, http://doi.org/10.1016/j.cognition.2015.12.011.

Henderson, J. M., Larson, C. L., & Zhu, D. C. (2008). Full Scenes produce more activation than close-up scenes and scene-diagnostic objects in parahippocampal and retrosplenial cortex: An fMRI study. *Brain and Cognition*, *66*(1), 40–49, http://doi.org/10.1016/j.bandc.2007.05.001.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*(5), 850–856, http://doi.org/10.3758/PBR.16.5.850.

Hennig, C. (2018). fpc: Flexible procedures for clustering. *R package*. Retrieved from https://cran.r-project.org/package=fpc%0D%0A

Hout, M. C., & Goldinger, S. D. (2015). Target templates: The precision of mental representations affects attentional guidance and decision-making in visual search. *Attention, Perception & Psychophys-*

ics, *77*(1), 128–149, http://doi.org/10.3758/s13414-014-0764-6.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10–12), 1489–1506. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10788654

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203, http://doi.org/10.1038/35058500.

Josephs, E. L., Draschkow, D., Wolfe, J. M., & Võ, M. L.-H. (2016). Gist in time: Scene semantics and structure enhance recall of searched objects. *Acta Psychologica*, *169*, 100–108, http://doi.org/10.1016/j.actpsy.2016.05.013.

Lawrence, M. (2013). ez: Easy analysis and visualization of factorial experiments. *R Package Version*. Retrieved from https://scholar.google.com/scholar?cluster=31099283308244089 3&hl=en&oi=scholarr

Mack, S. C., & Eckstein, M. P. (2011). Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *Journal of Vision*, *11*(9):9, 1–16, http://doi.org/10.1167/11.9.9. [PubMed] [Article]

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

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

Malcolm, G. L., & Henderson, J. M. (2010). Combining top-down processes to guide eye movements during real-world scene search. *Journal of Vision*, *10*(2):4, 1–11, http://doi.org/10.1167/10.2.4. [PubMed] [Article]

Navarro, D. J. (2015). *Learning statistics with R: A tutorial for psychology students and other beginners*. Adelaide, Australia: University of Adelaide. Retrieved from https://cran.r-project.org/web/packages/lsr/index.html

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(5), 509–522. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1003124

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

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766–786, http://doi.org/10.1037/0033-295X.113.4.766.

Võ, M. L.-H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, *9*(3):24, 1–15, http://doi.org/10.1167/9.3.24. [PubMed] [Article]

Võ, M. L.-H., & Henderson, J. M. (2010). The time course of initial scene processing for eye movement guidance in natural scene search. *Journal of Vision*, *10*(3):14, 1–13, http://doi.org/10.1167/10.3.14. [PubMed] [Article]

Võ, M. L.-H., & Wolfe, J. M. (2013a). Differential electrophysiological signatures of semantic and syntactic scene processing. *Psychological Science*, *24*(9), 1816–1823, http://doi.org/10.1177/0956797613476955.

Võ, M. L.-H., & Wolfe, J. M. (2013b). The interplay of episodic and semantic memory in guiding repeated search in scenes. *Cognition*, *126*(2), 198–212, http://doi.org/10.1016/j.cognition.2012.09.017.

Võ, M. L.-H., & Wolfe, J. M. (2015). The role of memory for visual search in scenes. *Annals of the New York Academy of Sciences*, *1339*, 72–81, http://doi.org/10.1111/nyas.12667.

Walther, D. & Koch, C. (2006), Modeling attention to salient proto-objects. *Neural Networks*, *19*, 1395–1407.

Wickham, H. (2009). *Ggplot2: elegant graphics for data analysis*. New York: Springer. http://doi.org/10.1007/978-0-387-98141-3

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

Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, *15*(2), 77–84, http://doi.org/10.1016/j.tics.2010.12.001.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 3485–3492). San Francisco, CA: IEEE. http://doi.org/10.1109/CVPR.2010.5539970