Vision Research 51 (2011) 1192–1205

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Semantic guidance of eye movements in real-world scenes

Alex D. Hwang *, Hsueh-Cheng Wang, Marc Pomplun

Department of Computer Science, University of Massachusetts Boston, 100 Morrissey Blvd., Boston, MA 02125-3393, USA

ARTICLE INFO

Article history:
Received 24 August 2010
Received in revised form 10 March 2011
Available online 21 March 2011

Keywords:
Eye movements
Visual attention
Visual search
Scene inspection
Latent semantic analysis

ABSTRACT

The perception of objects in our visual world is influenced by not only their low-level visual features such as shape and color, but also their high-level features such as meaning and semantic relations among them. While it has been shown that low-level features in real-world scenes guide eye movements during scene inspection and search, the influence of semantic similarity among scene objects on eye movements in such situations has not been investigated. Here we study guidance of eye movements by semantic similarity among objects during real-world scene inspection and search. By selecting scenes from the LabelMe object-annotated image database and applying latent semantic analysis (LSA) to the object labels, we generated semantic saliency maps of real-world scenes based on the semantic similarity of scene objects to the currently fixated object or the search target. An ROC analysis of these maps as predictors of subjects' gaze transitions between objects during scene inspection revealed a preference for transitions to objects that were semantically similar to the currently inspected one. Furthermore, during the course of a scene search, subjects' eye movements were progressively guided toward objects that were semantically similar to the search target. These findings demonstrate substantial semantic guidance of eye movements in real-world scenes and show its importance for understanding real-world attentional control.

1. Introduction

In order to study the control of visual attention under fairly natural conditions, many researchers have analyzed observers' eye movements in real-world images during inspection or search tasks, in which visual attention and eye movements are tightly coupled (e.g., Findlay, 2004). In these tasks, various mechanisms of eye-movement control based on low-level visual features have been examined, including bottom-up (e.g., Bruce & Tsotsos, 2006; Itti & Koch, 2001; Parkhurst, Law, & Niebur, 2002) and top-down control of visual attention (e.g., Hwang, Higgins, & Pomplun, 2009; Peters & Itti, 2007; Pomplun, 2006; Zelinsky, 2008). These studies demonstrated that observers' attention is biased toward visually salient locations, e.g., high-contrast areas, during scene inspection and toward regions similar to the search target in visual search tasks. It is important to notice, though, that real-world scenes consist of objects representing not only low-level visual information but also higher-level, semantic data. However, factors such as object meanings, between-object relations, or conceptual semantic effects, which are naturally involved in real-world situations, were not considered in previous work.

Although there have been studies of contextual effects in visual search, in which eye movements were constrained by contextual spatial knowledge of the scene, e.g., information about the objects likely to be found (Neider & Zelinsky, 2006; Torralba, Oliva, Castelhano, & Henderson, 2006), and studies of primitive semantic effects based on co-occurrence of objects in terms of implicit learning (Chun & Jiang, 1998; Chun & Phelps, 1999; Manginelli & Pollmann, 2009), the contextual relations investigated in those experiments depended on the spatial distribution or the consistency of scene objects.

Scene consistency itself has been the subject of numerous studies. One line of research has focused on objects that are not semantically consistent with the scene gist, referred to as "semantic violations", such as an octopus in a farmyard, or a microscope in a kitchen (Biederman, Mezzanote, & Rabinowitz, 1982; Bonitz & Gordon, 2008; Henderson, Weeks, & Hollingworth, 1999; Joubert, Fize, Rousselet, & Fabre-Thorpe, 2008; Loftus & Mackworth, 1978; Stirk & Underwood, 2007; Underwood, Humphreys, & Cross, 2007). Another line of research has studied objects that are semantically consistent but located in unexpected places in the scene structure or in unusual orientations, referred to as "syntactic violations", e.g., a floating cocktail glass in a kitchen or a fire hydrant on top of a mailbox in a street scene (Becker, Pashler, & Lubin, 2007; Biederman et al., 1982; Gareze & Findlay, 2007; Võ & Henderson, 2009).

The mechanisms underlying the effects of semantic or syntactic violations on eye movements are still not well understood. There is a current debate on whether semantic inconsistence guides eye movements in a similar way as visual saliency does. For example, Biederman et al. (1982), Stirk and Underwood (2007), Underwood

* Corresponding author. Fax: +1 617 287 6433.
E-mail address: baquibul@gmail.com (A.D. Hwang).

0042-6989/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2011.03.010

et al. (2007), Becker et al. (2007), Bonitz and Gordon (2008) and Loftus and Mackworth (1978) found that inconsistent objects are often found earlier and detected more accurately, and they conclude that it might be the result of parafoveal or peripheral information processing that enables object identification. On the contrary, Henderson et al. (1999) and Võ and Henderson (2009) found no evidence for such extrafoveal analysis. Finally, Joubert et al. (2008) found a mixed result of lower detection rate and faster reaction time for scene-inconsistent objects than for scene-consistent ones. Despite these varying results, there seems to be consensus that after the identification of an object, semantically or syntactically inconsistent objects draw more attention, suggesting that the estimation of semantic or syntactic relations is simultaneously processed with object identification.

It should be noted that these observed effects on eye movements are based on a single object-scene relation (semantic or syntactic violation) that rarely occurs in the real-world. The above studies thus over-simplify high-level visual perception, making it problematic to apply their findings to common cases in which the conceptual relations among all scene objects contain no semantic or syntactic violations.

Recently, there have been many efforts to understand the role of conceptual semantic influence on attention using various experimental methods. Belke, Humphreys, Watson, Meyer, and Telling (2008) and Moores, Laiti, and Chelazzi (2003) used a visual search paradigm, in which the search target was verbally specified before a set of object drawings was displayed. By analyzing observers' response times and eye movements, these studies demonstrated that attention was preferentially attracted to those objects that were semantically similar to the target. Corresponding effects were obtained by Huettig and Altmann (2006) and Yee and Sedivy (2006) using the visual world paradigm. In their work, observers' eye-movement bias was analyzed while they were looking at multiple, well-segregated object drawings and listening to spoken object names. However, all of these studies were limited to briefly presented, simple search displays containing four to eight objects that were "intuitively" selected for their semantic relations – a scenario that drastically differs from any real-world situation. Moreover, these studies only demonstrate a generic tendency of semantic influence when activated by external, verbal stimuli.

While these previous findings point out the relevance of semantics to scene inspection and visual search, their contribution to our everyday control of visual attention is still unclear. For example, whenever we routinely inspect our real-world visual environment, is it possible that the semantic similarity among objects in the scene influences our visual scan paths (gaze transitions)? Conceivably, in order to quickly develop a semantic understanding of a given scene, observers may inspect semantically similar objects consecutively. If such effects exist, do they depend on the visual task, e.g., scene inspection or visual search, and do they vary for the same scene over time? Such *semantic guidance* has not been studied, most likely due to the difficulties of assigning eye fixations to objects in real-world scenes and due to the intricacy of defining semantic relations among objects. Moreover, a quantitative approach of assessing semantic guidance in eye-movement data is necessary.

Analyzing eye fixations on objects in scene images requires object segmentation and labeling. There have been numerous attempts to solve these problems automatically, ranging from global scene classification (Bosch, Munoz, & Marti, 2007; Grossberg & Huang, 2009; Le Saux & Amato, 2004; Rasiwasia & Vasconcelos, 2008) to local region labeling (Athanasiadis, Mylonas, Avrithis, & Kollias, 2007; Chen, Corso, & Wang, 2008; Li, Socher & Li, 2009). However, their results are still unsatisfactory compared to human performance in terms of segmentation and descriptive labeling.

Thanks to the LabelMe object-annotated image database (Russell, Torralba, Murphy, & Freeman, 2008) developed by the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), a large number of real-world scene images, which were manually segmented into annotated objects by human volunteers, are freely available. In this database, the locations of objects are provided as coordinates of polygon corners, and they are labeled with English words or phrases (see Fig. 1). Therefore, series of eye fixations on these scenes can be easily translated into sequences of visually inspected objects and their labels.

In order to estimate the effect of semantic similarities between objects purely based on visual scenes, the co-occurrence of objects in a large number of scene images and the importance of each object in the scene context – defined by its attributes such as size, location or luminance – would have to be carefully considered. For example, objects of frequent co-occurrence, close proximity, or similar shape could be considered as semantically similar. Unfortunately, analyzing a sufficient amount of scenes and computing semantic relations directly from the image data sources is impractical. It is important to notice, however, that semantic relations are formed at the conceptual rather than at the visual level and thus do not have to be derived from image databases. Consequently, any database that can generate a collection of contexts or knowledge might be used to represent the semantic similarity of objects.

For the present study, we chose the linguistics-based computational method referred to as latent semantic analysis (LSA; Landauer & Dumais, 1997) to serve as a quantitative measure of semantic similarity between objects. LSA is a theory and method for extracting and representing the contextual usage-meaning of words by statistical computations applied to a large corpus of text. The basic premise in LSA is that the aggregate contexts in which a word does or does not appear provide a set of mutual constraints to deduce the word's meaning (Landauer, Foltz, & Laham, 1998). A high-dimensional '*semantic space*' is established from the text corpus, and terms (which are usually words) and documents (which are often collections of words) can be represented as vectors in this space. The semantic similarity between two terms, one term and one document, or two documents can be calculated as the cosine value of the angle between the two corresponding vectors in semantic space. The greater the cosine value, the higher is the semantic similarity. Since annotated objects in LabelMe have descriptive text labels, their semantic similarity can be estimated by calculating cosine values for the labels of object pairs. In this

Fig. 1. The LabelMe object-annotated image database (http://labelme.csail.mit.edu/).

study, we used the LSA@CU text/word latent semantic analysis tool developed by the University of Colorado at Boulder for LSA computation.

Equipped with above tools, we conducted two experiments to study two everyday visual activities - scene inspection (Experiment 1) and scene search (Experiment 2). For each recorded eye fixation during scene inspection, we generated a semantic saliency map. These maps were similar to feature-wise saliency maps used in visual feature guidance analysis (e.g., Hwang et al., 2009; Peters & Itti, 2007; Pomplun, 2006; Zelinsky, 2008), but were entirely based on the semantic similarities between the currently fixated object and the other objects in the scene. Here, saliency was defined as the corresponding LSA cosine value. If observers' immediate gaze transitions between objects are guided by the objects' semantic similarities, then these saliency maps should predict the next saccade target at an above-chance level. We measured these effects of *transitional semantic guidance* using the Receiver Operating Characteristic (ROC). Similarly, in the scene search experiment, we additionally measured *target-induced semantic guidance* by computing saliency as the LSA cosine between the search target and the label of each non-target scene object, followed by an identical ROC analysis. Several control analyses were conducted to exclude confounds and ensure that actual semantic guidance was measured.

2. Experiment 1

2.1. Method

2.1.1. Participants

Ten subjects participated in Experiment 1. All of them were students at the University of Massachusetts Boston, aged between 19 and 40 years old, with normal or corrected-to-normal vision. Each participant received a $10 honorarium.

2.1.2. Apparatus

Eye movements were tracked and recorded using an SR Research EyeLink-II system with a sampling rate of 500 Hz. After calibration, the average error of visual angle in this system is 0.5°. Stimuli were presented on a 19-inch Dell P992 monitor. Its refresh rate was 85 Hz and its resolution was 1024 × 768 pixels. Subjects' responses were entered using a game-pad.

2.1.3. Materials

A total of 200 photographs (1024 × 768 pixels) of real-world scenes, including landscapes, home interiors, and city scenes, were selected from the LabelMe database (http://labelme.csail.mit.edu/, downloaded on March 10, 2009) as stimuli (see Fig. 2 for an example scene). Objects in each scene were annotated with polygon coordinates defining the outline of the object shape, and they were labeled with English words. When displayed on the screen, the photographs covered 40° × 30° of visual angle. Each scene contained an average of 53.03 ± 38.14 labeled objects (in the present work, '±' always indicates a mean value and its standard deviation), and the median object number per image was 40. On average, labeled objects covered 92.88 ± 10.52% of the scene area.

2.1.4. Procedure

Subjects were instructed to inspect the scenes and memorize them for subsequent object recall tests (see Fig. 2a). After the five-second presentation of each scene, an English word was shown and subjects were asked whether the object indicated by the word had been present in the previously viewed scene. Subjects had to respond within three seconds by pressing a button on the game-pad. If they were unable to make the decision within that period, the trial would time out and the next trial would begin.

2.2. Data analysis

2.2.1. Computing semantic similarity based on LSA

In a nutshell, LSA similarity computation can be described as follows: First, an occurrence matrix is constructed from a large corpus of text, where each row typically stands for a unique word, each column stands for a document (''word-by-document matrix'', Landauer et al., 1998) and each cell contains the frequency with which the word occurred in the document. Subsequently, each cell frequency is normalized by an information-theoretic measure. However, it is computationally inefficient to operate with this very high-dimensional matrix. Therefore, a form of factor analysis called Singular Value Decomposition (SVD; see Berry, Dumais, & Obrien, 1995) is applied to reduce the matrix to a lower-dimensional vector space called '*semantic space*'. Previous empirical testing showed that optimal results are usually obtained with a number of dimensions ranging between 100 and 300 (Berry, Drmac, & Jessup, 1999; Jessup & Martin, 2001; Lizza & Sartoretto, 2001). LSA has the nice property that it can still estimate the semantic similarity of two words that never co-occur in the same document (Jones & Mewhort, 2007; Landauer & Dumais, 1997).

Every term, every document, and every novel collection of terms (''pseudo-document'') has a vector representation in the semantic space. Thus, the pair-wise semantic similarity between any of them can be calculated as the cosine value of the angle between the two corresponding vectors, with greater cosine value indicating greater similarity. Table 1 shows examples of LSA cosine values for various object labels used in LabelMe scene image ''Dining20'' (see Fig. 3) in terms of the reference object label ''FORK''. This label has, for instance, a higher cosine value (greater semantic similarity) with ''TABLE TOP'' (0.43) than with ''SHELVES'' (0.09).

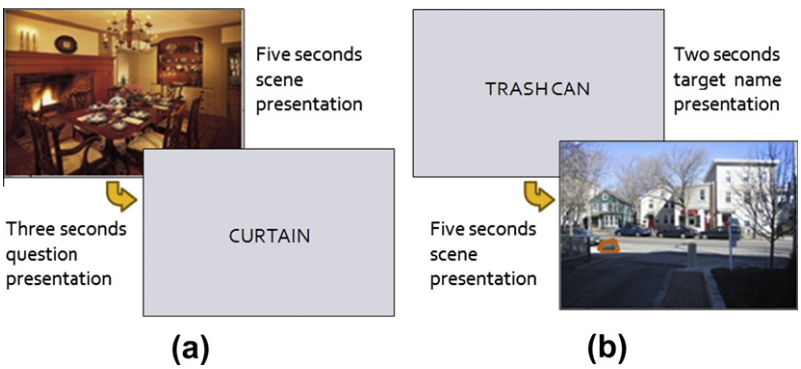

Fig. 2. Examples of the experiment procedures. (a) Scene inspection task and (b) scene search task. The target object is marked for illustrative purpose.

Table 1
Sample LSA cosine values.

Label 1	Label 2	Cosine
–	–	–
FORK	TABLE TOP	0.43
FORK	PLATE	0.34
FORK	CANDLESTICKS	0.27
FORK	FIRE PLACE	0.17
FORK	SHELVES	0.09
–	–	–

This difference indicates that in the text corpus, "FORK" and "TABLE TOP" occur in more similar contexts than do "FORK" and "shelves", which is plausible since, for example, forks are used for eating food on table tops rather than on shelves. The important feature of LSA is that it can quantify higher-level conceptual semantic similarity, regardless of any geometrical relation, functional relation or visual relation.

Since the images in the LabelMe database were annotated by many different contributors, objects are not always labeled consistently. For instance, the same object could be labeled "computer screen" in one image and "monitor" in another. In this study, we preserved the original object labels as much as possible by minimizing any kind of label modification – only spelling mistakes were corrected. Since LSA represents each term or document as a vector in semantic space, inconsistent but appropriate labels (synonyms) are mapped onto similar vectors. Therefore, the semantic similarity between synonyms is typically very high. While measuring the agreement among contributors in a meaningful way would require data beyond those offered by the LabelMe database, a study by Russell et al. (2008) suggests a high level of agreement. In their study, WordNet (see Fellbaum, 1998) was used to unify different contributors' object descriptions. They found only a small increase in the number of returned labels for several object queries before and after applying WordNet, indicating good consistency of the labels in the database.

To compute semantic similarity for each pair of object labels in our materials, a web-based LSA tool, LSA@CU (http://lsa.colorado.edu), developed at the University of Colorado at Boulder, was used. This tool was set to create a semantic space from "general readings up to 1st year college" and 300 factors (dimensions). Based on this space, we computed semantic similarity as the LSA cosine value, ranging between 0 and 1, for each object label compared to all other objects' labels for the same image. LSA cosine values are sometimes slightly smaller than 0 because of the high-dimensional space computation; we rounded negative values to zero. The average semantic similarity value for the pairs of labels in our materials was 0.245 ± 0.061.

The reason for choosing LSA as our semantic similarity measure is that it is one of the fundamental and widely used approaches for estimating semantic relationships at the conceptual level based on semantic analysis among words, sentences, or documents. It would be difficult to reach a consensus about how semantic similarity should be measured, and LSA is just one possible approach that may capture relationships between words at the conceptual level. It is thus important to keep in mind that in the present work, we define "semantic similarity" to be similarity as measured by LSA.

2.2.2. Constructing semantic saliency maps for gaze prediction

We introduce the term "semantic saliency map" to refer to a saliency map purely based on the LSA cosine between the label of a given object (the currently fixated object during inspection or the target object during search) and the labels of other objects in the scene. The semantic saliency maps were normalized so that the total volume under them was one. In the current study, semantic saliency maps served as predictors of gaze behavior, as described in the following section.

Fig. 3 shows examples of semantic saliency maps generated for a subject's individual fixations on various objects in the scene

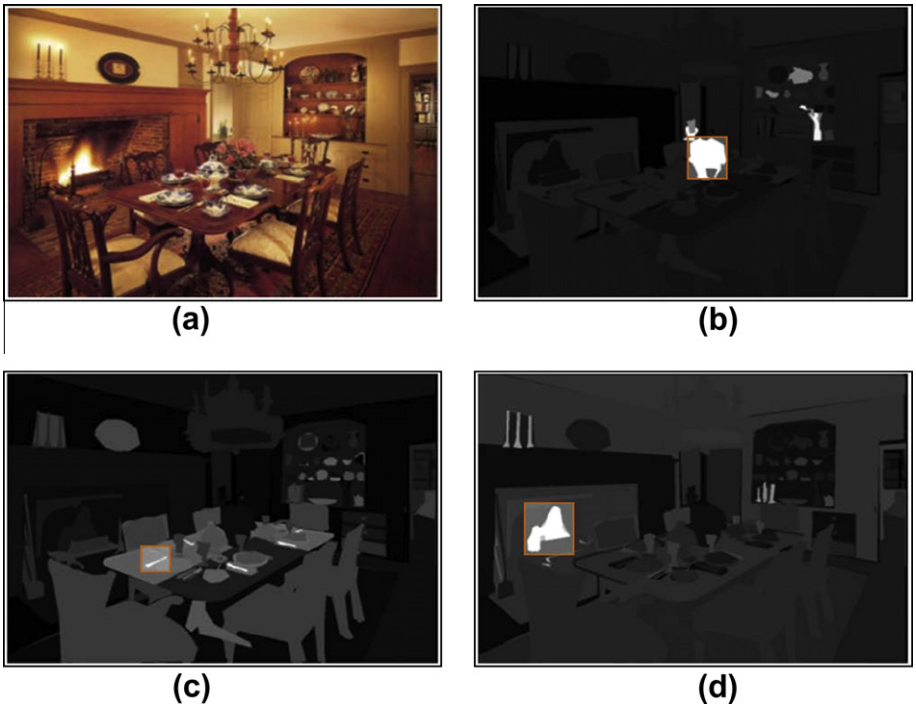

Fig. 3. Examples of semantic saliency maps. The reference object (for instance, the currently fixated one) is marked with an orange square. (a) Original scene image (Dining20). (b) Semantic saliency map during gaze fixation on an object labeled as "PLANT IN POT"; greater brightness indicates higher activation. (c) Semantic saliency map when the observer fixates on an object labeled as "FORK". (d) Semantic saliency map while fixating on an object labeled as "FLAME". As it can clearly be seen, semantically more similar objects receive higher activation; for example, candle sticks in (d) are activated by the reference object labeled "FLAME".

("Dining20"). As expected, when subjects fixate on an object labeled "PLANT IN POT" (highlighted in Fig. 3b), semantically similar objects like "FLOWER IN VASE" and "DECORATIVE POT" receive high semantic activations. The LSA cosine values between "PLANT IN POT" and "FLOWER IN VASE", and between "PLANT IN POT" and "DECORATIVE POT" are 0.53 and 0.37, respectively. In Fig. 3c, where the subject is currently looking at one of the forks on the table, the semantically most similar objects are the other forks on the table (LSA cosine = 1.00), which are maximally activated in the semantic saliency map. Objects that are semantically similar to "FORK", such as "BOWL" (LSA cosine = 0.46) and "PLATE" (LSA cosine = 0.34) still get higher activations compared to rather dissimilar objects like "FIREPLACE" (LSA cosine = 0.19) or "CHANDELIER" (LSA cosine = 0.13). Similar semantic saliency elevation between "FLAME" and "CANDLE STICKS" (LSA cosine = 0.59) can be seen in Fig. 3d.

2.2.3. Measuring semantic guidance

Similar to bottom-up and top-down effects in visual feature guidance, we defined two kinds of hypothetical semantic effects that might guide eye movements in real-world scenes. One is *transitional semantic guidance*, which can be computed for both scene inspection and scene search, and the other is *target-induced semantic guidance*, which can only be computed for scene search and will be discussed in the context of Experiment 2.

Transitional semantic guidance affects immediate gaze transitions from one object to another. In other words, this guidance influences the choice of the next object to be inspected; our hypothesis is that there is a bias toward selecting objects that are semantically similar to the currently fixated object. Since this type of guidance is thought to influence transitions between objects, we measured it by analyzing only those eye movements that transitioned from one object to another. This restriction led to the exclusion of 36.2% of the saccades (23.5% within object saccades and 12.7% saccades starting or landing outside of any marked objects) from the analysis of transitional semantic guidance in Experiment 1 (fixation sequences and durations within objects were examined in a separate study by Wang, Hwang, & Pomplun, 2010). To be clear, this exclusion only affected saccades in the semantic guidance analysis, and no data were excluded from any fixation analyses.

In order to compute the transitional guidance measure, we first translated the sequences of eye fixations into sequences of inspected objects. For each gaze transition in a given scene, a semantic saliency map was generated based on the currently fixated object (see Fig. 4). Subsequently, the ROC value was computed for the semantic saliency map as a predictor of the next object to be fixated by the subject. This calculation was very similar to previous studies using visual saliency maps (Hwang et al., 2009; Tatler, Baddeley, & Gilchrist, 2005). All ROC values computed along scan paths, excluding successive fixations on the same object, were averaged to obtain the extent of transitional semantic guidance during the inspection of a real-world scene. If gaze transitions were exclusively guided by semantic information, making semantic saliency a perfect predictor of gaze transitions, then the average ROC value across all scenes should be close to one. If there were no semantic effects on gaze transitions at all, the average ROC value should be close to 0.5, indicating prediction at chance level.

Similar to the majority of studies using visual saliency maps (e.g., Bruce & Tsotsos, 2006; Hwang et al., 2009; Itti & Koch, 2001; Parkhurst et al., 2002), our semantic saliency maps for both inspection and search were static, i.e., did not account for the observer's gain in scene knowledge over time (see Najemnik & Geisler, 2005). Clearly, as with the visual saliency maps, this characteristic does not imply that, at stimulus onset, observers instantly build a complete, static semantic map that guides all of their subsequent eye movements. Observers certainly do not identify all objects in the scene at once, which would be necessary to instantly build a complete semantic map (see Torralba et al., 2006). Instead, we assume the semantic exploration of the scene to be an iterative process. For example, at the beginning of the inspection or search process, subjects may mostly identify objects that are close to the initial (central) fixation position. From this initial set, in the case of transitional semantic guidance, subjects tend to choose objects that are semantically similar to the initially fixated one. As inspection progresses, subjects identify more objects in the scene. While these dynamics are not reflected in our saccadic similarity maps, for the purpose of the current study, they are a straightforward, initial approach to investigating the existence and extent of semantic guidance.

2.2.4. Excluding possible confounds with control data sets and analyses

In order to control for possible confounds in the measurement of semantic guidance, ROC values were computed for three control data sets, namely (1) random fixations, (2) dissociated fixations, and (3) Greedy Model fixations. The random case, consisting of randomly positioned fixations, served as a test for correct and unbiased computation of ROC values. For example, if the normalized semantic saliency maps were biased toward greater saliency for larger objects, we may receive above-chance ROC values even for random fixations, because larger objects are likely to receive more fixations than small objects. The random data for each subject and trial were sequences of randomly selected fixation positions in the scene, simulating unbiased and unguided fixations. We used a homogeneous pseudo-random function to place fixations at random pixel coordinates (x, y) on the image. For each simulated trial, the number of gaze transitions during the inspection period was kept identical to the empirical number in a given

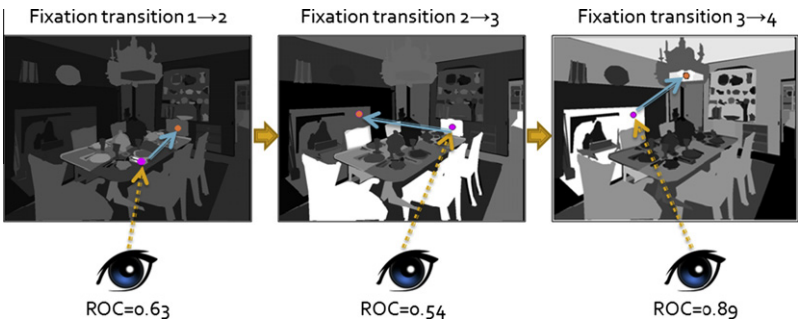

Fig. 4. Examples of *transitional semantic guidance* computation. For each fixation transition, a semantic saliency map of the scene is generated based on the currently fixated object (dashed arrows). The semantic saliency of the next fixation target determines the guidance score (ROC value) of that gaze transition (solid arrows). The average of these scores across all gaze transitions during a trial is computed as the transitional semantic guidance for that trial.

subject's data (see Fig. 5a and b). Any ROC values for the random case that deviate substantially from the chance level of 0.5 would indicate a bias in our ROC measure.

The dissociated case was introduced to control for likely confounds in the guidance measures: It is possible that semantically more similar objects tend to be spatially closer to each other in real-world images (proximity effect). Since amplitudes of empirical saccades during both scene inspection ($5.81 \pm 4.30°$ of visual angle) and scene search ($6.43 \pm 5.27°$) are significantly shorter (both $ts(9) > 24.125$, $ps < 0.001$) than those of random saccades ($12.59 \pm 6.08°$ and $13.17 \pm 6.50°$, respectively), we might overestimate the extent of semantic guidance of actual eye movements simply because they favor transitions between spatially close objects. Furthermore, it is known that our eye fixations are biased toward the center of a presented image during experiments under laboratory conditions (Tatler, 2007), and real-world images are often biased by a tendency of photographers to put interesting objects in the center. Therefore, the empirical eye fixation distribution is unlikely to resemble the artificial, homogeneous distribution created in the random control case.

To measure the potential proximity effect on our guidance measure, we computed the ROC value for dissociated fixations and scenes, that is, we analyzed the eye fixation data measured in scene n against the object data from scene $(n + 1)$, and the eye fixation data in scene 200 against the object data from scene 1, in the randomized sequence of scenes. This technique conserved the spatial distribution statistics of the empirical eye movements while eliminating semantic guidance effects (see Fig. 5a and c). Consequently, an ROC elevation above 0.5 in the dissociated case would indicate distribution (e.g., proximity) effects, and the ROC difference between empirical and dissociated fixations measures the actual strength of semantic guidance.

However, even the dissociated case may not provide sufficient assurance against artifacts entering the data, because it may be distorted by breaking the mapping between fixations and objects. While inspecting or searching through a scene, subjects presumably tend to fixate on objects. However, in the dissociated case, these fixations are superimposed on a different scene and do not necessarily land on objects anymore. Furthermore, successive fixations that transitioned between objects in the original scene may, in the dissociated case, land on the same object and would then be excluded from analysis.

In order to ensure that this characteristic of the dissociated case did not lead to a misinterpretation of the guidance data, we implemented the Greedy Model of gaze transitions. Following the idea of greedy, i.e., locally optimizing algorithms, this model always transitions from its current fixation location to the center of the display object with the shortest Euclidean distance from it, excluding the currently fixated object. If the semantic similarity maps predict the empirical transitions better than they predict the Greedy Model's transitions, this would further support the existence of transitional semantic guidance. For this evaluation, due to possible proximity effects (see above), it is important to only compare saccades of similar amplitudes.

When testing the Greedy Model, we found that if we simply started it at the screen center and let it perform a series of transitions that matched the number of transitional saccades in empirical scan paths, in most cases the model remained substantially closer to the screen center than the human gaze trajectories would. This was the case even when the model was prevented from visiting

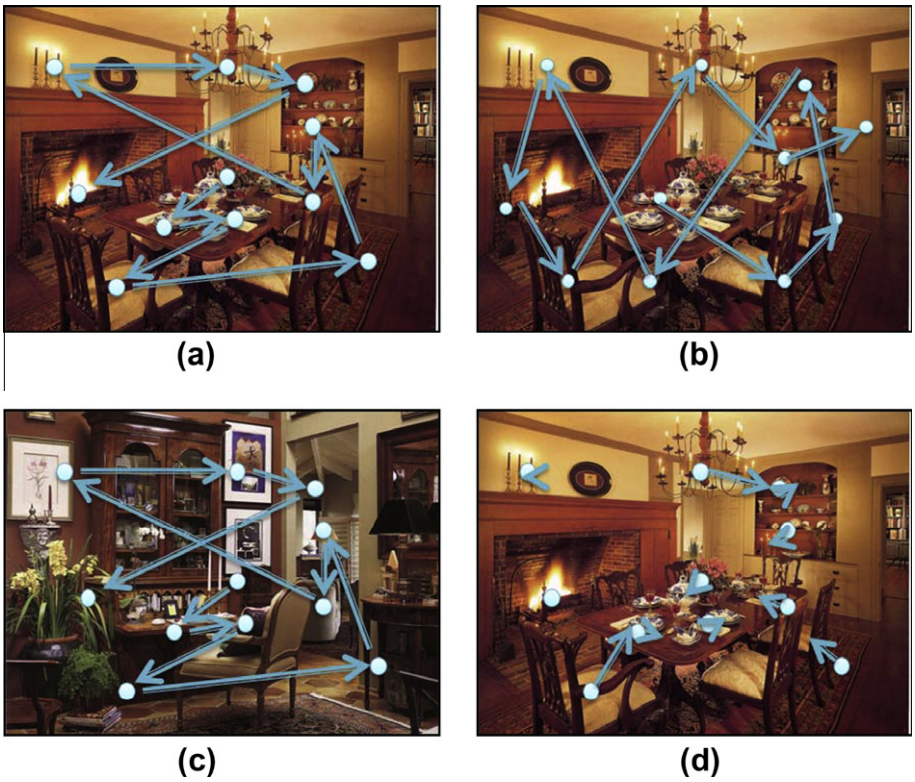

Fig. 5. Examples of the three control cases. (a) Empirical eye movements in one of the scenes. (b) Random case, in which we computed the ROC value of simulated random fixations in the same scene. (c) Dissociated case, in which empirical eye fixation data were analyzed for different scenes than those in which they actually occurred. We computed the ROC value of eye movements made in scene 1 as predicted by the saliency map for scene 2, eye movements made in scene 2 as predicted by saliency in scene 3, and so on. (d) Gaze transitions produced by the Greedy Model, which uses empirical fixations as starting points and the spatially closest display objects as endpoints of its transitions.

any object more than once. However, as discussed above, it is problematic to compare the guidance characteristics of scan paths with clearly distinct spatial distributions, and therefore we decided to use the empirical fixations as the starting points for all of the model's gaze transitions. To be precise, we took every object fixation from every subject and scene, recorded a transition from the current fixation to the closest object, and continued with the subject's next fixation, and so on (see Fig. 5a and d). This approach allowed us to compute transitions that were not guided by semantic information but only by proximity, while preserving the spatial distribution of eye movements and their targeting of scene objects. Since each model transition was associated with a given subject's fixation, we compared empirical and model ROC values within subjects.

As a final control measure, we computed transitional guidance by visual similarity, in order to rule out that transitional guidance is caused by low-level visual similarity rather than semantic similarity. Conceivably, visual and semantic similarities are positively correlated – semantically similar objects may be more likely to share visual features than do semantically dissimilar ones. For example, intuitively, different kinds of plants are semantically similar, and they are also visually similar, as green is their predominant color. Therefore, gaze guidance by both visual and semantic similarity of objects and the correlation between the two similarities have to be considered in order to get conclusive results.

Our visual similarity measure considered the following four important object characteristics: color, size, compactness, and orientation. Color similarity between two objects was measured by a simple, robust histogram matching method called Histogram Intersection Similarity Method (HISM; Swain & Ballard, 1991); Chan (2008) demonstrated the accuracy of the HISM method for estimating perceptual color similarity. Following these studies, our current color similarity measure included the three components of the DKL color model which is based on the human eye's cone receptor sensitivity regarding three wavelengths (short, medium and long) and double opponent (red–green, blue–yellow and luminance[1]) cell responses (see Krauskopf, Lennie, & Sclar, 1990; Lennie, Derrington, & Krauskopf, 1984). The computation of these features and their similarity is described in detail in Hwang et al. (2009).

Object size was measured as the number of pixels covered by the polygon that outlined the object. The compactness of an object was defined as the square of its perimeter, measured as the sum of the length of the enclosing polygon's edges, divided by the object's area. Compactness tells us whether an object is rather disc-shaped (low value) or elongated like a stick (high value). Finally, the orientation of an object was determined by computing a linear regression on all pixels belonging to that object, and taking the angle between the horizontal axis and the resulting regression line as the orientation measure. Size, compactness and orientation values were scaled to vary between 0 and 1.

The overall similarity between two objects was then computed as the product of similarity values along the four feature dimensions (color, size, compactness, and orientation), where color similarity was measured by the HISM method, size and compactness similarity were defined as one minus the absolute distance between feature values, and angular similarity was defined as the minimum angle difference between two line orientations. This type of multiplicative feature similarity was found to yield more robust results than additive techniques (e.g., Hwang et al., 2009).

It is clear that adding more visual feature dimensions to our similarity measure could still, at least slightly, improve that measure. In order to estimate the extent of such improvements, we also computed the measure with only the DKL color component, making it insensitive to size, compactness, and orientation. To assess the quality of visual similarity measurement, we computed the correlation between visual and semantic similarity across all object pairs in our study. This correlation is assumed to yield a positive coefficient for sound visual similarity measures (see below). We found a correlation of $r = 0.15$ for the full visual similarity measure and only a very small decrease, $r = 0.147$, for the color-only measure. This finding is in line with our previous studies (e.g., Hwang et al., 2009), showing that among commonly computed low-level visual features, color features exert by far the strongest guidance of eye movements. Moreover, this finding suggests that adding even more visual features is unlikely to drastically increase the correlation between visual and semantic similarity. It thus seems appropriate to use the full version of the current measure for estimating visual similarity guidance in the present context.

3.2. Results and discussion

3.2.1. Basic performance measures

Subjects produced an average of 15.5 ± 4.0 fixations per trial. Among those fixations, subjects made 11.1 ± 1.7 gaze transitions between distinct objects, with average fixation duration of 248 ± 29 ms. The average saccade amplitude was 5.81 ± 4.30°. Response accuracy, measured as the percentage of correctly identified target-present and target-absent cases, was 72.0%.

3.2.2. Transitional semantic guidance and control cases

As described above, we computed four ROC values to study transitional semantic guidance, which were based on (1) empirical data, (2) random fixations, (3) dissociated fixations, and (4) the Greedy Model's gaze transitions. As shown in Fig. 6a, the transitional semantic guidance value of simulated random fixations during scene inspection (0.508 ± 0.123) was close to 0.5, i.e., chance level. This result indicates that the ROC computation was applied correctly and that the normalized saliency maps used for our analysis were unbiased. Moreover, the ROC value was significantly greater for the dissociated gaze-scene pairs (0.583 ± 0.143) than for the random fixations, $t(9) = 17.16$, $p < 0.001$, evidencing the hypothesized proximity effect. Finally, the empirical eye movements had a significantly higher ROC value (0.646 ± 0.127) than the random fixations, $t(9) = 23.28$, $p < 0.001$, and disassociated ones, $t(9) = 12.46$, $p < 0.001$. Consequently, we can conclude that although there is a significant proximity effect, actual transitional guidance still plays a significant role independently of proximity.

As discussed above, due to proximity effects, the ROC analysis for the Greedy Model had to be performed separately for different saccade amplitude intervals. At the same time, this analysis had the added benefit of providing some insight into both the nature of the proximity effect and semantic guidance as a function of saccade amplitude. Fig. 7 shows an ROC comparison of the empirical gaze transitions and those generated by the Greedy Model. Since the transitions produced by the Greedy Model tended to be shorter (3.43 ± 2.75°) than the subjects' transitions (5.81 ± 4.30°), $t(9) = 7.69$, $p < 0.001$, there were not many transitions larger than 10° (3.2% of all transitions) to allow interval-based analysis. The cut-off point for empirical transitions was set to 18°, with 1.8% of the transitions being longer. It can clearly be seen that the ROC values for the empirical transitions were consistently greater than those for the modeled ones. Comparing the average ROC values for saccade amplitudes below 10° within subjects revealed a significant difference, $t(9) = 26.13$, $p < 0.001$, between empirical (0.667) and model data (0.586). Furthermore, the data for the Greedy Model show strong proximity effects, as evidenced by ROC values above 0.6 for object-to-object transitions shorter than 3°. The ROC values decrease with longer transitions and seems to virtually disappear for transitions longer than 9°. This pattern contrasts with the

[1] For interpretation of color in Fig. 5, the reader is referred to the web version of this article.

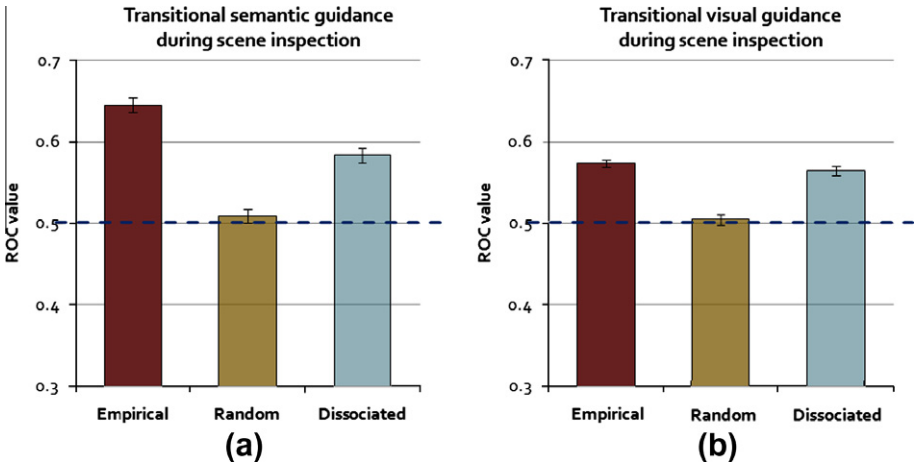

Fig. 6. (a) Transitional semantic guidance and (b) transitional visual guidance during scene inspection (Experiment 1) as measured by the ROC method, with dashed lines indicating chance level and error bars representing standard error of the mean. The difference between the empirical and dissociated cases indicates the existence of semantic guidance.

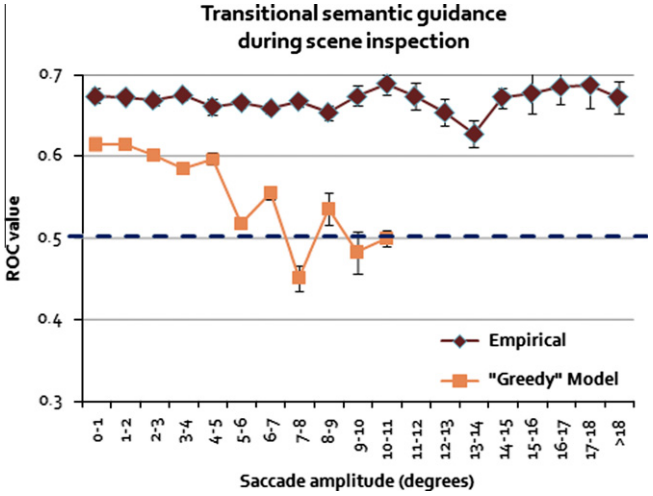

Fig. 7. Comparison of transitional semantic guidance during scene inspection (Experiment 1) between empirical gaze transitions and transitions generated by the Greedy Model. Results are shown separately for different saccade amplitude (distance between transition starting point and endpoint) intervals. Note that all ROC values for saccades longer than 18° and 10° for the empirical and model data, respectively, were collapsed into one data point for each series. The dashed line indicates ROC chance level, and error bars show the standard error of the mean.

ROC values for the empirical transitions, which not only exceed the model's ROC for short transitions but remain at a constantly high level, even for transitions longer than 18°. These results further support the view that the elevated ROC for the empirical eye movements is not an artifact caused by the arrangement of objects and their local contexts.

As discussed in the previous section, the final step in the guidance data analysis was to rule out the possibility that the observed guidance effects were purely caused by low-level visual similarity of objects instead of their semantic similarity. In order to quantify the influence of visual similarity on gaze movements, the correlation between visual and semantic similarity within objects was computed. For this computation, all possible pairings of objects across all 200 scenes used for the current experiments were analyzed (879,998 object pairs, average visual similarity of 0.161 ± 0.152). As expected, the result shows a slight positive correlation between the two similarity measures, $r = 0.15$, $p < 0.001$. This finding suggests the possibility that the semantic guidance measured above could be an artifact resulting from strong guidance of eye movements by visual similarity and its correlation with semantic similarity. To examine this possibility, we computed ROC values based on the visual similarity of objects for the empirical, random, and dissociated cases in a manner analogous to our semantic guidance calculation. In this computation, the saliency maps that were generated for each gaze transition between distinct objects represented visual similarity, instead of semantic similarity, between the currently fixated object and all other objects in the scene.

As illustrated in Fig. 6b, the random control case showed a near-chance ROC level of 0.510 ± 0.092, and the dissociated case revealed an elevated ROC value (0.564 ± 0.088) as compared to the random case, $t(9) = 7.09$, $p < 0.001$, demonstrating a visual proximity effect. However, the difference between visual similarity guidance for the empirical data (0.573 ± 0.059) and the dissociated data did not reach statistical significance, $t(9) = 1.29$, $p > 0.1$. This finding indicates that semantic similarity, and not visual similarity, is the main factor underlying the current results.

3.2.3. Time course of transitional semantic guidance

Since we found significant semantic guidance effects, it is sensible to ask when this guidance starts and whether it is sustained throughout the trial. These temporal changes of semantic guidance during scene perception might help to understand the underlying mechanisms. We decided to examine transitional semantic guidance for each of the first nine gaze transitions after stimulus onset, which include 68.6% of all fixations, and an average value over 10th or later fixations.

As shown in Fig. 8, transitional semantic guidance influences gaze movements throughout the trial, starting from the first gaze transition. However, since only saccades transitioning between different objects were included in the analysis, the present data cannot conclusively show whether this guidance is already fully present at the first saccade in a trial. Nevertheless, the data suggest that semantic saliency guides the attentional selection of visual objects in a continuous and constant manner.

To analyze in more detail the timing of guidance, the gaze transitions for each scene and each subject were separated into those visiting an object for the first time and those revisiting an object. We found no significant difference in ROC values, $t(9) = 0.49$, $p > 0.6$, between the first-time visit (0.647 ± 0.020) and re-visit groups (0.645 ± 0.012). This finding suggests that transitional semantic guidance was not limited to revisiting of objects but also occurred, to the same extent, before an object was fixated for the first time.

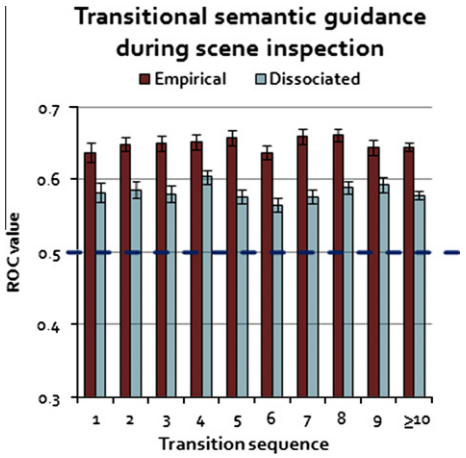

Fig. 8. Temporal variation of semantic guidance during scene inspection indicated by the difference between the ROC values for the empirical data and the dissociated control case. The dashed line represents chance level, and error bars indicate standard error of the mean. Note that the rightmost column, labeled "$\geqslant$10", includes the data for not only the tenth transition, but also for all subsequent ones.

While the results of Experiment 1 provide evidence for transitional semantic guidance during scene inspection, it also raises some questions. As discussed above, we can assume some level of object recognition to be necessary for accessing an object's semantic information (e.g., Torralba et al., 2006). Consequently, in order for transitional semantic guidance to take effect, it seems that such level of recognition must have been achieved for a set of potential target objects prior to the programming of a saccade. The guidance mechanism could then semantically relate these objects with the currently attended one and bias the selection of the next saccade target toward the most similar object.

However, as the Greedy Model (Fig. 7) and fixation re-visit analyses show, transitional guidance does not substantially decrease with greater eccentricity of saccade targets, even for large angles and for targets that have not previously been fixated. This finding seems to imply that recognition performance does not differ between objects at, for example, eccentricities of 1° and 18°, which is clearly implausible. To explain this pattern of results, it should be noted that long saccades are rather infrequent; for example, only 19.4% of all saccades are longer than 8°, and only 1.8% of them are longer than 18°. For most of these long saccades, it is conceivable to assume that a particularly salient peripheral object attracted the observer's attention prior to the saccade. This allocation of attention likely enabled at least some rudimentary processing of the object's visual information, possibly including some semantic analysis, before the saccade was programmed. In such situations, the semantic guidance mechanism may bias saccade-target selection toward either the peripheral object or one of the more central objects whose semantic information has already been accessed. Such a bias could prevent a large proportion of long saccades to peripheral objects that are likely unrelated to the currently fixated object based on the semantic information available. Thus, transitional semantic guidance could still exert a significant influence even on long saccades.

4. Experiment 2

4.1. Method

4.1.1. Participants

Ten subjects, who did not participate in Experiment 1, participated in Experiment 2, all of them were students at the University of Massachusetts Boston, aged between 19 and 40 years old, with normal or corrected-to-normal vision. Each of them received a $10 honorarium.

4.1.2. Apparatus and materials

The apparatus and materials used in Experiment 2 were identical to those in Experiment 1.

4.1.3. Procedure

Subjects were instructed to search for objects whose name or description was shown prior to the scene. After a two-second presentation of the object name, the search scene was shown for five seconds. During each scene presentation, whenever subjects thought they had found a target object, they were to press a button while fixating on that object. Since there could be multiple target objects in the same scene, subjects were asked to continue searching for target objects until the trial ended (see Fig. 2b). This task design allowed a fixed five-second duration of scene presentation as in the scene inspection experiment (Experiment 1) and thereby enabled a useful between-experiments comparison of results. After scene presentation, the correct location of targets would be indicated or the text "Object does not exist" would be shown, in the target-present or the target-absent case, respectively. Search targets were randomly selected, then inspected, and possibly newly selected to avoid target objects that can be detected very easily. Subjects performed 200 randomly ordered trials preceded by five practice trials. Target-present and target-absent cases were evenly distributed among the 200 trials.

4.1.4. Data analysis

Besides the analysis of transitional guidance that was introduced in Experiment 1, the search task used in Experiment 2 motivated the additional study of a hypothesized second kind of semantic guidance, termed *target-induced semantic guidance*, influencing gaze distribution during scene search. This guidance reflects the extent to which semantic similarity between the target object and the objects in the search image determines the choice of fixated objects. Its computation for a given search scene-target pair only requires a single semantic saliency map, which represents the spatial configuration of semantic similarity between the target object and all non-target objects in the scene. As shown in Fig. 9, the ROC value was measured for this saliency map as a predictor of all eye fixations made during a trial.

4.2. Results and discussion

4.2.1. Basic performance measures

Subjects made an average of 16.2 ± 1.6 fixations per trial in Experiment 2, with no statistical difference to Experiment 1 (15.5 ± 4.0 fixations), $t(18) = 0.57$, $p > 0.5$. Among those fixations were 9.3 ± 3.6 gaze transitions per trial between distinct objects. The average fixation duration was 301 ± 87 ms, which was significantly greater than that measured in Experiment 1 (248 ± 29 ms), $t(18) = 2.67$, $p < 0.05$. Even if we exclude all fixations on targets in Experiment 2, which may have been prolonged by verification processes and by executing button presses, the resulting fixation duration (286 ± 69 ms) was still significantly greater than that in Experiment 1, $t(18) = 3.27$, $p < 0.05$. This difference in fixation duration between scene inspection and scene search clearly differs from previous studies (e.g., Castelhano, Mack, & Henderson, 2009; Võ & Henderson, 2009). A possible reason for the current pattern of results is that in the current scene inspection task, subjects were asked to memorize the objects in the scene. Given the large average amount of objects in the stimulus images that needed to be memorized in a short amount of time, subjects may have produced more saccades than they would have without any explicit task instruction.

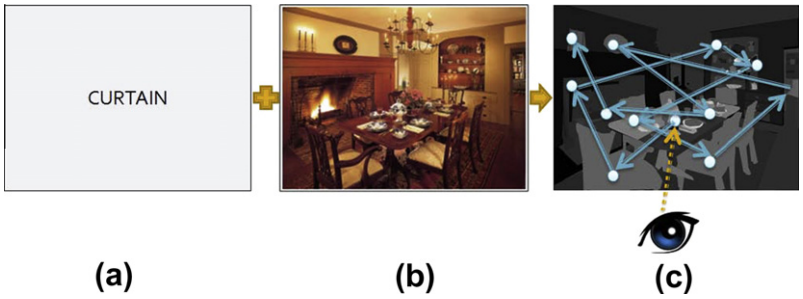

Fig. 9. Example of *target-induced semantic guidance* computation in the scene search experiment. For each trial, a single semantic saliency map is generated based on the search target (a) and all objects in the scene (b). The ROC value for this map as a predictor of all fixated objects during that trial is taken as the guidance measure (c).

The average saccade amplitude in Experiment 2 was 6.43 ± 5.27°, which was significantly larger than the one measured in Experiment 1 (5.81 ± 4.30°), $t(18) = 2.33$, $p < 0.05$. The subjects' response accuracy in Experiment 2, measured as the percentage of correctly identified target-present and target-absent cases, was 70.1%, which was very similar to Experiment 1 (72.0%). In target-present trials, subjects manually reported the detection of the first target after an average of 6.2 ± 3.5 fixations.

4.2.2. Transitional semantic guidance and control cases

Analogous to Experiment 1, in Experiment 2 we excluded all saccades that did not transition from one object to a different one, which amounted to an elimination of 36.8% of the saccades, from all further analysis (25.9% within-object saccades and 10.9% saccades starting or landing outside of any labeled objects). As in Experiment 1, subsequent fixation analyses were not affected by this exclusion. Once again, we examined transitional semantic guidance through four ROC analyses based on (1) empirical data, (2) the random control case, (3) the dissociated control case and (4) the Greedy Model. As shown in Fig. 10a, the ROC value for simulated random fixations, 0.504 ± 0.104, was close to 0.5, indicating unbiased saliency maps. The ROC value for the dissociated gaze-scene pairs was significantly elevated (0.566 ± 0.127) above the random-fixation ROC value, $t(9) = 17.10$, $p < 0.001$, revealing a proximity effect similar to the one observed in Experiment 1. Moreover, the ROC value for the empirical eye movements was slightly greater (0.583 ± 0.134) than that for the dissociated case, $t(9) = 4.71$, $p < 0.001$. Even though this difference (0.017 ± 0.012) was statistically significant, it was substantially smaller than the corresponding difference for scene inspection (0.063 ± 0.016), $t(18) = 7.27$, $p < 0.001$.

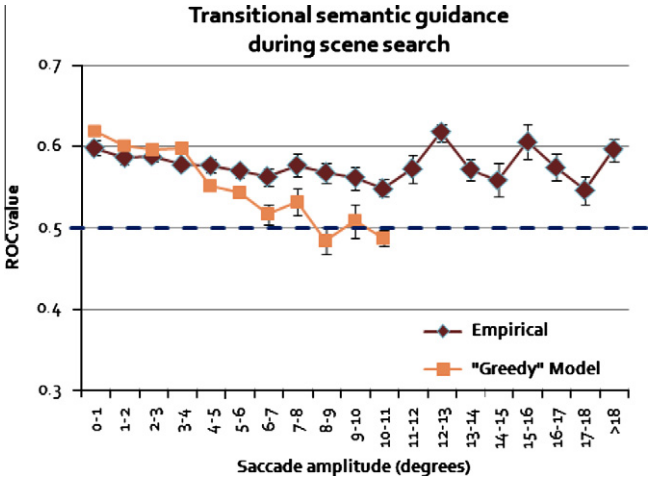

Fig. 11. Comparison of transitional semantic guidance during scene search (Experiment 2) between empirical gaze transitions and transitions generated by the Greedy Model. Results are shown separately for different saccade amplitude (distance between transition starting point and endpoint) intervals. Note that all ROC values for saccades longer than 18° and 10° for the empirical and model data, respectively, were collapsed into one data point for each series. The dashed line indicates ROC chance level, and error bars show the standard error of the mean.

In order to verify that the dissociated fixations did not bias the results by breaking the fixation-to-object mapping, we also applied the Greedy Model to the data of Experiment 2. Fig. 11 illustrates the ROC values for the empirical and the modeled gaze transitions for different saccade amplitude intervals. Empirical transitions

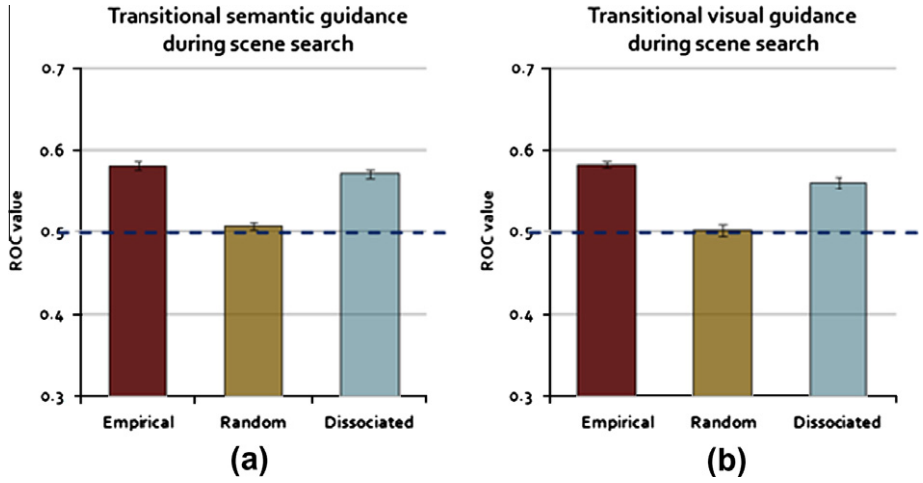

Fig. 10. (a) Transitional semantic guidance and (b) transitional visual guidance in Experiment 2 (scene search) as measured by the ROC method, with dashed lines indicating chance level and error bars representing standard error of the mean.

longer than 18° (4.7% of the data) were pooled into a single data point, and for the model this was done with all transitions above 10°, which also affected 4.7% of the respective data. The model's average saccade amplitude was 3.58 ± 3.01°. In contrast to Experiment 1, the ROC comparison between empirical and model data in Experiment 2 did not show a clear distinction. Comparing the mean values for saccade amplitudes below 10° did not reveal a significant difference between the empirical (0.582) and model data (0.578), $t(9) = 1.18$, $p > 0.25$. Given this result and the small difference between empirical and dissociated ROC, the present data do not provide any conclusive evidence of transitional guidance during search. However, Fig. 11 suggests that with greater amplitude, the model ROC decreases faster than the empirical ROC, allowing the speculation that there may be weak, long-range transitional semantic guidance effects during search.

To investigate the contribution of low-level visual similarity to transitional guidance of eye movements during search, we computed ROC values based on visual similarity in the same manner as for Experiment 1. As illustrated in Fig. 10b, an elevated ROC measure of visual guidance in the dissociated case (0.560 ± 0.095) as compared to the random case (0.501 ± 0.069), $t(9) = 7.00$, $p < 0.001$, demonstrated a proximity effect. There was also a slight effect of visual similarity guidance, as indicated by significant differences between the empirical case (0.583 ± 0.058) and the dissociated case, $t(9) = 3.12$, $p < 0.005$.

Comparing Fig. 10a and b, we find that, in contrast to Experiment 1, the ROC difference between the empirical and dissociated cases is greater for visual similarity than for semantic similarity. However, since both effects are very modest, we can only speculate about their behavioral relevance and underlying mechanisms. It is possible that both types of guidance slightly influence transitional eye movements during scene search, or that the transitional semantic guidance may, at least in part, be due to both transitional visual guidance and the correlation between the two similarity measures.

In summary, while there were statistically significant effects of both semantic and visual guidance on transitional eye movements in both scene inspection and scene search, the pattern of results differed noticeably between the two tasks. During scene inspection, subjects were guided much more strongly by semantic similarity as compared to low-level visual similarity. This finding suggests that during scene inspection tasks, subjects may inspect semantically similar objects consecutively to enhance scene memorization for later recall. In the scene search task, on the other hand, visual guidance is stronger than semantic guidance, but both influences are clearly weaker than that of semantic similarity in the scene inspection task. It seems that when subjects are assigned a specific task, 'search for the target object', this task takes precedence over scene inspection. As a result, the strategy of gaze control may be shifted to a target-focused mode that is not aimed at object memorization. The slight transitional visual guidance found during search could be a result of subjects forming a visual template whose low-level features guide their search (cf. Schmidt & Zelinsky, 2009; Yang & Zelinsky, 2009).

4.2.3. Target-induced semantic guidance

The analysis of target-induced semantic guidance during scene search was similar to the analysis of transitional semantic guidance. Target-induced semantic guidance represents the influence of semantic similarity between the search target and all non-target scene objects on fixation distribution. As shown in Fig. 12, target-induced semantic guidance for the random control case (0.497 ± 0.076) was close to 0.5, confirming that the target-based semantic saliency maps were unbiased. For the dissociated control case, target-induced semantic guidance was 0.506 ± 0.145, which was very slightly, but significantly, greater than the value for the random case, $t(9) = 4.35$, $p < 0.005$, indicating a very small proximity effect in the scene search task. The empirical ROC value was significantly higher (0.637 ± 0.159) than both the random, $t(9) = 18.79$, $p < 0.001$, and dissociated ones, $t(9) = 17.04$, $p < 0.001$. We can thus conclude that target-induced semantic guidance plays a significant role in scene search, independently of proximity effects.

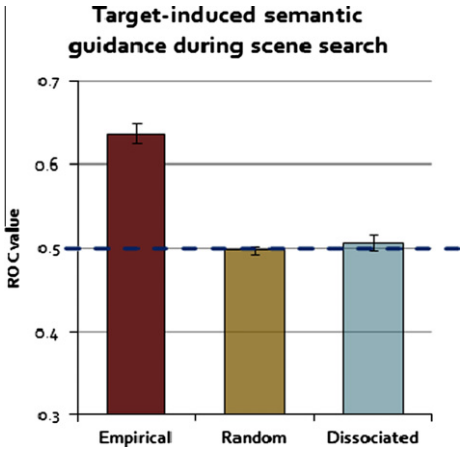

Fig. 12. Target-induced semantic guidance during scene search, with a dashed line indicating chance level and error bars representing standard error of the mean.

Note that, in the search task, the target object was specified only by its verbal description, not by its visual features. Due to the large visual variation among those objects that match a given description, determining dependable and representative visual features of the target that could define visual similarity between the target and other objects in the scene is computationally infeasible. As a consequence, we did not attempt to compute target-induced visual similarity guidance in Experiment 2. Nevertheless, it is still possible that the ROC values for target-induced semantic guidance were partially due to visual guidance. For example, an observer searching for a fork may look at a knife not because the two are semantically similar, but because they look alike. While such an effect cannot be ruled out, the weak correlation ($r = 0.15$) between visual and semantic similarity makes it seem unlikely that visual similarity, rather than semantic similarity, plays a major role in producing the current results.

Comparing the results for transitional and target-induced semantic guidance during scene search (Figs. 10a and 12, respectively), it is noticeable that while transitional semantic guidance is hardly detectable (0.017 ± 0.012), target-induced guidance is very pronounced (0.131 ± 0.023). This finding further supports our interpretation that the specific priorities in the scene search task are responsible for reduced transitional semantic guidance as compared to the scene inspection task. More insight into this issue may be obtained by analyzing the time course of target-induced semantic guidance, which is reported in the following section.

4.2.4. Time course of target-induced semantic guidance

Following the same grouping used in Experiment 1, we examined the time course of target-induced semantic guidance for each of the first nine gaze transitions after stimulus onset, covering 59.8% of all fixations, and an average value over 10th or later fixations.

As shown in Fig. 13a, target-induced semantic guidance increased gradually during search in a given scene, followed by a decrease after approximately the sixth fixation. This pattern may be due to interference between visual saliency and semantic saliency.

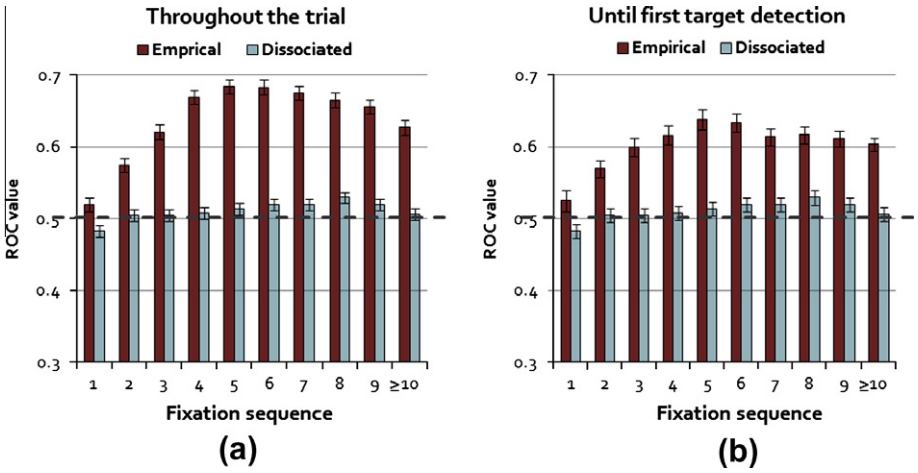

Fig. 13. Temporal change of target-induced semantic guidance during scene search indicated by the difference between the ROC values for the empirical data and the dissociated control case, including fixations (a) throughout the trial and (b) from stimulus onset until first target detection. Dashed lines represent chance level, and error bars indicate standard error of the mean. The rightmost column, labeled "$\geqslant 10$", includes not only the data for the tenth transition or fixation, but also all subsequent ones.

At the beginning of the search, although the target is specified by words, a visual representation of the target object, a 'search-template', may be generated and maintained in working memory in order to continuously match it with objects in the scene (e.g., Desimone & Duncan, 1995; Houtkamp & Roelfsema, 2009; Schmidt & Zelinsky, 2009; Wolfe, 1994; Yang & Zelinsky, 2009). Therefore, initially eye movements may be strongly governed by bottom-up and top-down processing of low-level visual saliency, with only little influence by semantic information. As search progresses, a better semantic understanding of the scene may develop, and if visual feature saliency by itself fails to detect the target, target-induced semantic guidance may start to dominate the control of eye movements. Conceivably, after the first target object in the scene has been found (as reported above, it occurs after an average number of 6.2 ± 3.5 fixations), subjects may maintain strong semantic guidance to detect further targets. When no more targets can be found, guidance may slowly decrease.

We should consider the possibility, however, that the specific search task - requiring subjects to continue search for further targets after detecting the first one – may have artificially induced the steady increase in semantic guidance over time. The reason for such a data artifact could be that after the first target detection, subjects might transition back-and-forth between the first target and further target candidates. Due to the possibly high semantic similarity between these objects with the target label, such gaze behavior would likely increase target-induced guidance in the later stages of the search.

Our data revealed that subjects did in fact frequently re-visit the first target object they detected; on average, this occurred 1.02 ± 0.44 times per trial. However, this number does not appear large enough to suggest a significant impact of back-and-forth scanning behavior on guidance measurements. In order to rule out such potential bias from the data, we recomputed our analysis, but this time excluded all eye-movement data that were recorded in any given trial after the first target detection was reported. The resulting time course of target-induced guidance (Fig. 13b) shows slightly reduced ROC values but does not differ qualitatively from the initial one and thus supports the notion of a steady increase of guidance over the course of the search.

5. General discussion

Previous studies on semantic effects on visual processing have focused on global contextual effects based on scene gist and eye fixation distribution, semantic effects in simple, artificial visual search tasks, or context effects based on co-occurrence or contextual cueing of objects. In contrast, the present work investigated semantic guidance of eye movements in real-world scenes, induced by the semantic similarity of scene objects to each other or to a search target.

We conducted two experiments to demonstrate semantic guidance of gaze transitions during scene inspection and semantic guidance of gaze distribution during scene search. To accomplish this, we introduced a novel interdisciplinary approach combining visual context research and linguistic research. Using eye-movement recording and linguistics-based LSA on object labels, we demonstrated that our visual scan paths in the inspection of everyday scenes are significantly controlled by the semantic similarity of objects. Our gaze tends to transition to objects that are semantically similar to the currently fixated one, basically unaffected by the time course of the inspection or whether an object is fixated for the first time or is re-visited.

When interpreting the current data, we have to consider the possibility that, besides semantic guidance, contextual guidance may also have influenced the subjects' gaze transitions. While the dissociated control case allowed us to account for proximity effects, it did not fully control for the fact that semantically similar objects are often also located in contextually restrained parts of a scene in similar ways. For example, a spoon and a plate are often placed on a horizontal surface within the scene, such as a table. Eye movements during search in real-world scenes can be guided by such contextual factors (Castelhano & Henderson, 2007), relying on global scene statistics rather than the identification of individual objects and their semantics (Torralba et al., 2006). Since 'scene context' is ultimately built on the spatial layout of semantically related objects, it is difficult to rule out the possibility of contextual guidance in the current study. However, our data show that during scene inspection, transitional guidance by semantic similarity does not decrease with greater distance between the currently fixated object and the saccade target object. This finding is important because longer saccades should be more likely to move the observer's gaze beyond a contextually constrained part of the scene. Although such scene parts are sometimes large, e.g., the sky, we would expect at least a small reduction in average empirical ROC values for longer saccades if contextual guidance, and not semantic guidance, were the main factor driving the observed effects. This is clearly not supported by our data. Nevertheless, the contribution of contextual guidance

to the effects observed in this study needs to be examined in future experiments.

While the current study demonstrates the existence of semantic guidance, it only allows a rough characterization of its underlying mechanisms. Clearly, it is impossible for observers to semantically analyze each individual scene object prior to their first eye movement in a scene. Instead, there has to be an iterative semantic exploration of the scene. The present data suggests that it involves parafoveal and even peripheral semantic analysis, since even long saccades tend to land on objects that are semantically similar to the previously fixated one. This finding is in line with several of the semantic inconsistence studies such as Underwood et al. (2007), Becker et al. (2007), and Bonitz and Gordon (2008), but it conflicts with others such as Võ and Henderson (2009). The last study used well-controlled, computer-generated displays to control for some confounds in earlier work, and it did not find an effect of peripheral analysis. There are two possible reasons for the discrepancy between Võ and Henderson's (2009) and the present data: First, Võ and Henderson (2009) had subjects inspect a large number of scenes without presenting intermittent questions about the scene content. In contrast, our study required subjects to memorize a potentially large number of objects within five seconds of scene presentation, which may have induced a strategy of peripheral semantic analysis. Second, it is likely that the detection of semantic inconsistency differs from semantic analysis of individual visual objects. Guidance toward semantic information that is related to the currently attended information is a plausibly useful mechanism allowing us in everyday life to efficiently explore the semantic content of a visual scene. Detection of semantic inconsistencies, however, is a rather unusual task as such inconsistencies rarely occur in the real world. It is thus possible that the human visual system has developed an ability for at least a rough semantic analysis of peripheral objects, whereas the detection of semantic inconsistencies requires focal attention.

In a larger context, the transitional semantic guidance data may reveal a general mechanism of high-level attentional guidance by semantic association. As put, most prominently, by James (1890), there are different "varieties of attention", among them visual attention and internal attention to our thoughts, with the latter variety producing trains of thought by association, i.e., transitions between semantically related concepts. The current study demonstrates that the former variety, visual attention, also proceeds by semantic association when exploring a visual scene, and this is the first time that any such general attentional mechanism has been studied quantitatively.

Once a visual search task is involved, search seems to take precedence over scene inspection. In a search task, the fixation order is prioritized by similarity to the target, and as a result, guidance of individual gaze transitions by semantic factors almost disappears. However, the overall distribution of fixations during the search task shows strong target-induced semantic guidance – observers tend to inspect objects that are semantically similar to the search target. This result demonstrates that semantic bias of attention, as previously shown for artificial search displays (Huettig & Altmann, 2006; Yee & Sedivy, 2006), also exists during search in real-world scenes. Unlike transitional semantic guidance during scene inspection, target-induced guidance increases gradually during the time course of the search task. This increase in guidance is similar to, but slower than, the one observed in guidance of visual search by low-level visual features in real-world scenes (Hwang, Higgins, & Pomplun, 2007). For such low-level guidance, it is assumed that observers first examine the overall composition of the scene in terms of its low-level visual features before using those features to guide their search (Pomplun, 2006). With regard to semantic guidance, a similar assumption seems plausible: It is possible that the progressively developing semantic understanding of the scene during the course of the search task is accompanied by an increased influence of target-induced semantic guidance on eye movements. Furthermore, it is likely that observers start their search under the guidance of a generic visual template that they create based on the verbal description of the target. This visual feature guidance may initially dominate the search process, at the cost of semantic guidance, until it either fails to detect the target, or more semantic context information becomes cognitively available, or both.

The current findings can be considered a first glimpse at the high-level, semantic mechanisms of attentional control in real-world situations. Further experiments are necessary to corroborate the current findings of semantic guidance by using explicit manipulations of semantic scene content and other semantic similarity measures than LSA. Future research should also address the dynamics of semantic guidance in more detail. It would be desirable to develop a dynamic model of semantic guidance that accounts for the iterative semantic exploration of real-world scenes and might be able to predict scanning behavior more accurately. Moreover, for a deeper understanding of these mechanisms, further research needs to address, in particular, the function of the observed semantic guidance. For instance, a crucial question to investigate is whether semantically ordered sequences of object inspections lead to better scene understanding or memorization as compared to random sequences. Furthermore, the processes underlying the gradual increase of target-induced guidance during search have to be examined. Will guidance increase even more slowly or stay at a marginal level in scenes showing unnatural arrangements of objects with no attainable global semantic understanding? Answering such questions promises to reveal the cognitive processes underlying semantic guidance and build a comprehensive, multi-level model of the control of visual attention. As argued above, such a model may generalize, at least in part, toward other "varieties" of attention.

Acknowledgments

This research was supported by Grant Number R15EY017988 from the National Eye Institute to Marc Pomplun.

References

Athanasiadis, T., Mylonas, P., Avrithis, Y., & Kollias, S. (2007). Semantic image segmentation and object labeling. *IEEE Transactions on Circuits and Systems for Video Technology, 17*, 298–312.

Becker, M. W., Pashler, H., & Lubin, J. (2007). Object-intrinsic oddities draw early saccades. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 20–30.

Belke, E., Humphreys, G. W., Watson, D. G., Meyer, A. S., & Telling, A. (2008). Top-down effects of semantic knowledge in visual search are modulated by cognitive but not perceptual load. *Perception and Psychophysics, 70*, 1444–1458.

Berry, M. W., Drmac, Z., & Jessup, E. (1999). Matrices, vector spaces, and information retrieval. *SIAM Review, 41*, 335–362.

Berry, M. W., Dumais, S. T., & Obrien, G. W. (1995). Using linear algebra for intelligent information-retrieval. *SIAM Review, 37*, 573–595.

Biederman, I., Mezzanote, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*, 143–177.

Bonitz, V. S., & Gordon, R. D. (2008). Attention to smoking-related and incongruous objects during scene viewing. *Acta Psychologica, 129*, 255–263.

Bosch, A., Munoz, X., & Marti, R. (2007). Review: Which is the best way to organize/classify images by content? *Image and Vision Computing, 25*, 778–791.

Bruce, N. D. B., & Tsotsos, J. K. (2006). Saliency based on information maximization. *Advances in Neural Information Processing Systems, 18*, 155–162.

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.

Castelhano, M., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision, 9*(3), 1–15 (6).

Chan, H. C. (2008). Empirical comparison of image retrieval color similarity methods with human judgment. *Displays, 29*, 260–267.

Chen, A. Y. C., Corso, J. J., & Wang, L. (2008). HOPS: Efficient region labeling using higher order proxy neighborhoods. In The 18th International Conference on Pattern Recognition (ICPR) (pp. 1–4).

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology, 36*, 28–71.
Chun, M. M., & Phelps, E. A. (1999). Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage. *Nature Neuroscience, 2*(9), 775–776.
Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193–222.
Fellbaum, C. (1998). Wordnet: an electronic lexical database. Bradford Books.
Findlay, J. M. (2004). Eye scanning and visual search. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 135–159). Psychology Press.
Gareze, L., & Findlay, J. M. (2007). Absence of scene context effects in object detection and eye gaze capture. In R. van Gompel, M. Fischer, W. Murray, & R. W. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537–562). Amsterdam: Elsevier.
Grossberg, S., & Huang, T.-R. (2009). ARTSCENE: A neural system for natural scene classification. *Journal of Vision, 9*(4), 1–19 (6).
Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 210–228.
Houtkamp, R., & Roelfsema, P. R. (2009). Matching of visual input to only one item at any one time. *Psychological Research, 73*, 317–326.
Huettig, F., & Altmann, G. T. M. (2006). Word meaning and the control of eye fixation: Semantic competitor effects and visual world paradigm. *Cognition, 96*, B23–B32.
Hwang, A. D., Higgins, E. C., & Pomplun, M. (2007). How chromaticity guides visual search in real-world scenes. In *Proceedings of the 29th Annual Cognitive Science Society* (pp. 371–378). Austin, TX: Cognitive Science Society.
Hwang, A. D., Higgins, E. C., & Pomplun, M. (2009). A model of top-down attentional control during visual search in complex scenes. *Journal of Vision, 9*(5), 1–18 (25).
Itti, L., & Koch, C. (2001). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*(10–12), 1489–1506.
James, W. (1890). *The principles of psychology*. New York: Henry Holt.
Jessup, E., & Martin, J. (2001). Taking a new look at the latent semantic analysis approach to information retrieval. In M. W. Berry (Ed.), *Computational information retrieval* (pp. 121–144). Philadelphia: SIAM.
Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic Lexicon. *Psychological Review, 114*, 1–37.
Joubert, O. R., Fize, D., Rousselet, G. A., & Fabre-Thorpe, M. (2008). Early interference of context congruence on object processing in rapid visual categorization of natural scenes. *Journal of Vision, 8*(13), 1–18 (11).
Krauskopf, J., Lennie, P., & Sclar, G. (1990). Chromatic mechanisms in striate cortex of macaque. *Journal of Neuroscience, 10*, 646–669.
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.
Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259–284.
Le Saux, B., & Amato, G. (2004). Image classifiers for scene analysis. In *Proceedings of the International Conference of Computer Vision & Graphics (ICCVG), Warsaw, Poland*.
Lennie, P., Derrington, A. M., & Krauskopf, J. (1984). Chromatic mechanisms in lateral geniculate nucleus of macaque. *Journal of Physiology, 357*, 241–265.
Li, L., Socher, R., & Li, F. (2009). Towards total scene understanding:classification, annotation and segmentation in an automatic framework. Computer Vision and Pattern Recognition (CVPR).
Lizza, M., & Sartoretto, F. (2001). A comparative analysis of LSI strategies. In M. W. Berry (Ed.), *Computational information retrieval* (pp. 171–181). Philadelphia: SIAM.
Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 565–572.
Manginelli, A. A., & Pollmann, S. (2009). Misleading contextual cues: How do they affect visual search? *Psychological Research, 73*, 212–221.
Moores, E., Laiti, L., & Chelazzi, L. (2003). Associative knowledge controls deployment of visual selective attention. *Nature Neuroscience, 6*, 182–189.
Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature, 434*, 387–391.
Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research, 46*, 614–621.
Parkhurst, D. J., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual selective attention. *Vision Research, 42*, 107–123.
Peters, R. J., & Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)* (pp. 1–8).
Pomplun, M. (2006). Saccadic selectivity in complex visual search displays. *Vision Research, 46*, 1886–1900.
Rasiwasia, N., & Vasconcelos, N. (2008). scene classification with low-dimensional semantic spaces and weak supervision. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage*.
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.
Schmidt, J., & Zelinsky, G. J. (2009). Search guidance is proportional to the categorical specificity of a target cue. *Quarterly Journal of Experimental Psychology, 62*(10), 1904–1914.
Stirk, J. A., & Underwood, G. (2007). Low-level visual saliency does not predict change detection in natural scenes. *Journal of Vision, 7*(10), 1–10 (3).
Swain, M. J., & Ballard, D. H. (1991). Color indexing. *Journal of Computer Vision, 7*(1), 11–32.
Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14), 1–17.
Tatler, B., Baddeley, R., & Gilchrist, L. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45*, 643–659.
Torralba, A., Oliva, A., Castelhano, M., & 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*, 766–786.
Underwood, G., Humphreys, L., & Cross, E. (2007). Congruency, saliency, and gist in the inspection of objects in natural scenes. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 564–579).
Võ, M. 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), 1–15.
Wang, H. C., Hwang, A. D., & Pomplun, M. (2010). Object frequency and predictability effects on eye fixation durations in real-world scene viewing. *Journal of Eye Movement Research, 3*(3), 1–10 (3).
Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review, 1*, 202–238.
Yang, H., & Zelinsky, G. J. (2009). Visual search is guided to categorically-defined targets. *Vision Research, 49*, 2095–2103.
Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32*, 1–14.
Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review, 115*(4), 787–835.