Vision Research 105 (2014) 10–20

Contents lists available at ScienceDirect

Vision Research

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

The roles of scene gist and spatial dependency among objects in the semantic guidance of attention in real-world scenes

Chia-Chien Wu [a,*], Hsueh-Cheng Wang [b], Marc Pomplun [a]

[a] Department of Computer Science, University of Massachusetts at Boston, USA
[b] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA

ARTICLE INFO

Article history:
Received 3 April 2014
Received in revised form 19 August 2014
Available online 6 September 2014

Keywords:
Attention
Semantics
Eye movements
Visual guidance
Real-world scenes

ABSTRACT

A previous study (Vision Research 51 (2011) 1192–1205) found evidence for semantic guidance of visual attention during the inspection of real-world scenes, i.e., an influence of semantic relationships among scene objects on overt shifts of attention. In particular, the results revealed an observer bias toward gaze transitions between semantically similar objects. However, this effect is not necessarily indicative of semantic processing of individual objects but may be mediated by knowledge of the scene gist, which does not require object recognition, or by known spatial dependency among objects. To examine the mechanisms underlying semantic guidance, in the present study, participants were asked to view a series of displays with the scene gist excluded and spatial dependency varied. Our results show that spatial dependency among objects seems to be sufficient to induce semantic guidance. Scene gist, on the other hand, does not seem to affect how observers use semantic information to guide attention while viewing natural scenes. Extracting semantic information mainly based on spatial dependency may be an efficient strategy of the visual system that only adds little cognitive load to the viewing task.

1. Introduction

One of the main reasons why visual search in human observers is still drawing a vast amount of research interest after many decades of study is that the human visual system still outperforms state-of-the-art computer vision systems. This advantage is particularly significant for the visual analysis of high-level information, as contained in natural scenes. While viewing such scenes, people only fixate on a few regions of interest to infer the type of the scene and where the most important information is located. Modern computer vision systems, however, seem to hardly achieve the same level of efficiency. To understand the mechanisms that enable human vision to gain such an advantage, many previous studies have extensively investigated the guidance of attention in real-world environments either with regard to its bottom-up, stimulus driven (Bruce & Tsotsos, 2009; Itti & Koch, 2000; Koch & Ullman, 1985), or top-down, task driven aspects (Hayhoe et al., 2003; Hwang, Higgins, & Pomplun, 2009; Navalpakkam & Itti, 2007; Pomplun, 2006). These studies show that the observers' attention is biased toward regions with high visual saliency, e.g., high-contrast areas, or toward areas related to the task goal, respectively.

Human visual attention is not only affected by factors based on overt visual appearance, but also by inherent factors such as meaning and semantic relations among objects. Prior studies have found that eye movements and attention were affected by contextual knowledge of the scene. For example, Gordon (2004, 2006) found that observers preferred to attend to semantically inconsistent objects in scenes that were only presented for 150 ms. Furthermore, Davenport and Potter (2004) showed that during early visual processing, the relation between objects and the scene context in which they are located can be established, which could influence observers' cognitive performance, including scanning strategies and scene understanding. The contextual knowledge provided by a typical scene, however, is much richer than the object–scene relationship mentioned above. Torralba et al. (2006) found that observers could extract some global scene properties – referred to as scene gist – without recognizing individual objects and use this information to guide their attention and eye movements. Moreover, even when the global context, which usually comes from visual background information, is missing, it is still possible for observers to learn some context of the scene. Chun (2000) showed that some contextual information could be learned simply by the typical arrangement of display elements and that this learning can affect the deployment of attention. Moreover, Oliva and Torralba (2007) found that spatial dependency among objects, such as object co-occurrence and local spatial layout of the scene,

* Corresponding author. Address: 100 Morrissey Boulevard, Boston, MA 02125-3393, USA.
E-mail address: cchienwu@gmail.com (C.-C. Wu).

http://dx.doi.org/10.1016/j.visres.2014.08.019
0042-6989/

could provide a different type of contextual information about a scene. These results imply that, in addition to the scene gist and the object–scene relationship, the object–object relationship can also provide a wealth of semantic information for attentional guidance during natural viewing (see Wu, Ahmed-Wick, & Pomplun, 2014, for a review of the various aspects of sematic information).

While these previous studies indicate the relevance of semantics to scene inspection and visual search, little is known about how and when semantic relationships are learned and to which extent this conceptualization of contextual information could influence attention. Many investigations of semantic effects on eye movements have been simply based on a single object–scene relation, which rarely occurs in real-world environments (Gordon, 2004, 2006; Hollingworth & Henderson, 1999; Loftus & Mackworth, 1978). During natural viewing, the conceptual relations should be able to continuously impact observers' viewing strategy when needed, integrate with either low-level stimulus features or task goal over time, and ultimately improve scene understanding. Using a single object–scene relation (either semantically or syntactically) to investigate whether the contextual information could bias the deployment of attention may underestimate the utility of semantic information in attentional guidance.

Only few studies have asked how the visual scan path is impacted by semantic information during natural scene viewing, presumably because of the complexity of object segmentation and difficulty of defining semantic relations among objects in the scene. Hwang, Wang, and Pomplun (2011) attempted to investigate how the semantic similarity among scene objects influences attention and eye movements by analyzing gaze transitions between scene objects. They found that during natural scene viewing, observers tend to bring their gaze to those objects that are semantically similar, as measured by Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), either to the currently fixated one or to the specified search target. This result, however, may have been influenced by the observers' knowledge of the global scene context that they obtained during early visual processing. That is, instead of directly analyzing the semantic relation between the currently fixated object and the objects located in the extrafoveal visual field, observers could simply have used their knowledge about the scene type to decide where to look next. For example, if observers had been aware that the viewed image was a kitchen, they may have only attended to the regions nearby the counter or sink, where most of the kitchenware was likely located. This strategy could be executed by merely using the scene gist perceived during the initial glance without further semantic analysis (Oliva & Torralba, 2001, 2006).

In addition to using scene gist, observers could also obtain contextual information by exploiting the spatial dependency among objects and use it to predict the most likely location of the search target or a semantically relevant object to inspect (Oliva & Torralba, 2007). For example, a chair may be expected to be located behind a table, or a fork may be expected to be next to a spoon. In summary, both scene gist and spatial dependency among scene objects may have caused a bias in observers' gaze patterns that could explain the results of Hwang, Wang, and Pomplun (2011) without the need for semantic analysis of extrafoveal scene objects.

The aim of the present study was to discern the contributions of scene gist, spatial object dependency, and semantic object analysis to semantic guidance. In order to investigate the factors contributing to attentional guidance, we measured observers' eye movement while they explored a series of scenes. Prior studies have shown that eye movement and visual attention are tightly coupled and visual scanning in natural scenes or complex images is rarely driven by covert attention (Findlay, 2004; Henderson, 2003; Kowler et al., 1995). Therefore, using gaze transitions as an indication of attention allocation in a scene is a reasonable way to study attentional guidance in natural viewing. To examine whether scene gist is the essential factor in semantic guidance and what role spatial dependency plays in it, we conducted Experiment 1. In this experiment, we removed the background information and only left objects in the scene so that subjects could not extract the scene gist from natural scenes during early visual processing. As an additional factor, the presence of spatial dependency was manipulated. If observers can evaluate the semantic information of objects in the extrafoveal visual field, we should still see, at least partially, the effect of semantic guidance even when scene gist, spatial object dependency, or both are not available.

In Experiments 2 and 3, we manipulated the presence of scene gist in two different ways while either keeping the spatial dependency among objects in a scene or eliminating it to further investigate the role and function of scene gist in semantic guidance. If the semantic guidance found in the previous study (Hwang, Wang, & Pomplun, 2011) was contributed by scene gist at any level, presenting scene gist should induce stronger semantic guidance than excluding scene gist. By discerning different aspects of semantic information that likely induce guidance of attention, we may obtain a more refined understanding of the way people perceive natural scenes. Such knowledge may, among other applications, have an impact on future computer vision and human–computer interaction approaches.

2. Experiment 1

To examine whether the semantic guidance of attention found by Hwang, Wang, and Pomplun (2011) was at least partially due to the effects of scene gist or spatial dependency among objects, we removed the gist from the natural scenes and manipulated the presence of spatial dependency among objects. The gist was removed by excluding any background of the scene and only showing a small number of foreground items, while eliminating the spatial dependency by randomly displacing the foreground objects within the scene. Similar to Torralba et al. (2006), this study refers to the term "scene gist" as indicating the global statistics of the scene, which can be extracted during early visual processing and provide some superordinate information about the scene. Note that scene gist is generally defined as some coarse global scene characteristic and rarely specified as a quantitative property. Furthermore, how and when the scene gist is learned is still unknown. Therefore, it is impossible to effectively remove the gist from a scene. Nevertheless, taking away the background information from a scene can impair observers' ability of extracting the gist information. We use the term "removing the gist" to mean that the scene background is removed and the remaining information is insufficient to allow subjects to extract the gist information during early visual processing in the same way as when the whole scene is provided. If the semantic guidance found in Hwang, Wang, and Pomplun (2011) was entirely due to either the scene gist or the spatial dependency among objects, excluding the respective factor should diminish the effect, and thus gaze transitions would no longer depend on the semantic similarity of objects.

2.1. Method

2.1.1. Participants

Twenty observers, aged between 19 and 40 years old, were tested. All had normal or corrected to normal vision and were naïve as to the purpose of the study. Each subject received a $10 honorarium. Experimental procedures were approved by the University of Massachusetts Boston IRB in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki) for exper-

iments involving humans. Each subject signed an informed consent form before data collection.

2.1.2. Apparatus

Eye movements were tracked and recorded using an SR Research EyeLink-1k system. Its sampling frequency is 1000 Hz. Stimuli were presented on a 22-in. ViewSonic LCD monitor. Its refresh rate was set to 75 Hz and its resolution was set to 1024×768 pixels. Participant responses were entered using a keyboard.

2.1.3. Stimulus display

To study the influence of spatial dependency among scene objects without interference by the scene gist, we employed the LabelMe object annotated image database (http://labelme.csail.mit.edu; see Russell et al., 2008) in which scene images are manually segmented into annotated objects by volunteers. In addition, the locations of objects are provided as coordinates of polygon corners, and all objects are labeled with English words or phrases. This database thus provides an opportunity for not only segregating each object from its scene, but also shifting the object's coordinates to any desired location in the image.

A total of 60 stimulus images (1024×768 pixels) were generated. Each image was composed of 13–15 objects selected from a real-world scene from the LabelMe database. The selected scenes included home interiors, landscapes and city scenes. Objects with extreme size (occupying more than ~2% of the scene or less than ~0.2%) were not chosen as scene objects. To remove the scene gist or other global regularity from the scene, all objects were segregated from the image and were pasted on a grey canvas. Each object was placed at either the same coordinates as in the original scene, which was referred to as 'fixed condition', or at randomly selected locations on the canvas, referred to as 'scrambled condition'. In the scrambled condition, different objects were placed manually to avoid overlap and clutter (see Fig. 1 for an example). Thus, only spatial dependency, but not scene gist was retained in the fixed condition. On the other hand, both scene gist and spatial dependency were removed in the scrambled condition.

2.1.4. Preliminary psychophysical testing

As explained above, the term "removing the scene gist" referred to in this study does not mean that all the global statistics in the scene were removed. Instead, most of the background information of the scene was removed and the remaining information was insufficient to allow subjects to extract the same gist information as when the whole image was provided. To verify this methodology and show that excluding the background information can prevent subjects from retrieving the gist information, a preliminary psychophysical test was conducted on Amazon Mechanical Turk (MTurk).

MTurk has been used in the past few years to get data inexpensively and rapidly, and the quality of the resulting data has been verified by many studies (Buhrmester, Kwang, & Gosling, 2011; Callison-Burch, 2009; Crump, McDonnell, & Gureckis, 2013; Mason & Suri, 2012). In our study, to investigate whether observers are able to extract the gist information when the scene background was removed, we asked human observers from MTurk to conduct a scene classification task. The images used in both the fixed condition and the scrambled condition, as well as the original images used to create the stimulus displays were tested separately. In each trial, observers fixated a fixation cross for 1000 ms. Once the cross disappeared, an image was shown for 80 ms. Observers were instructed to choose one out of 8 given categories which could describe the scene best. In each condition (whole scene, fixed and scrambled), 30 observers were tested. Each observer only conducted one condition for all 120 images used in the current study so that they could not learn any image cues from the other conditions. Table 1 shows the observers' performance.

This result shows that, in line with findings from prior studies (Potter & Levy, 1969; Thorpe, Fize, & Marlot, 1996), observers were able to extract some information (referred as the scene gist) to efficiently categorize the whole scenes in early visual processing. When we removed the background information, this ability was significantly impaired (one-way ANOVA for all three conditions: $F(2,87) = 176.5$, $p < .01$). In addition, the performance in the fixed condition was better than in the scrambled condition ($t(29) = 2.76$, $p < .01$). Since the same amount of global statistics of the scene was eliminated in both conditions and the only difference between them was the spatial layout of objects, observers' slightly better performance in the fixed condition suggests that the remaining spatial dependency provided more semantic information to help them categorize the scenes. Interestingly, when both background information and spatial dependency were removed in the scrambled condition, observers still performed above chance level ($t(29) = 9.86$, $p < .01$). This finding suggests that some objects were representative enough to allow observers to infer the category of a scene (for example, a bed would signify that the current scene is a bedroom).

2.2. Procedure

Subjects were instructed to inspect the scenes and memorize them for the subsequent object recall test (see Fig. 1, bottom panel). Each stimulus image was presented for 5 s. After the image had disappeared, an English word was shown and subjects were asked whether the object indicated by the word had been shown in the previous scene. Subjects responded by key press. The next trial would begin once subjects made a response. Subjects performed a total of 60 trials (30 trials each in the fixed and scrambled conditions). Each scene was only presented once to each subject, either in the fixed condition or in the scrambled condition, and the order of both conditions was counterbalanced.

2.3. Data analysis

2.3.1. Assigning fixations to objects

Since all images excluded the global contextual information by only leaving the selected objects on a grey canvas, some fixations landed on the blank area rather than on any object in the image. When this happened, we assumed that this fixation was aimed at the nearest object, i.e., the one whose center had the shortest Euclidean distance to the current fixation location. In addition, only the transitions between different objects were counted so that the result would not be affected by reconfirming fixations on the same object.

2.3.2. Computing semantic similarity between two objects based on Latent Semantic Analysis (LSA)

Similar to the original semantic guidance study (Hwang, Wang, & Pomplun, 2011), we used Latent Semantic Analysis (referred to as LSA; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998) to serve as a quantitative measure of semantic similarity between objects. LSA is able to extract and represent the contextual usage-meaning of words by statistical computations applied to a large corpus of text. It is often used in linguistic studies to compute semantic similarity between two texts or phrases from a given large text corpus. Here the concept was applied to assess the semantic similarity between two text labels in the same scene.

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, and each column stands for a document, which is typically a collection of

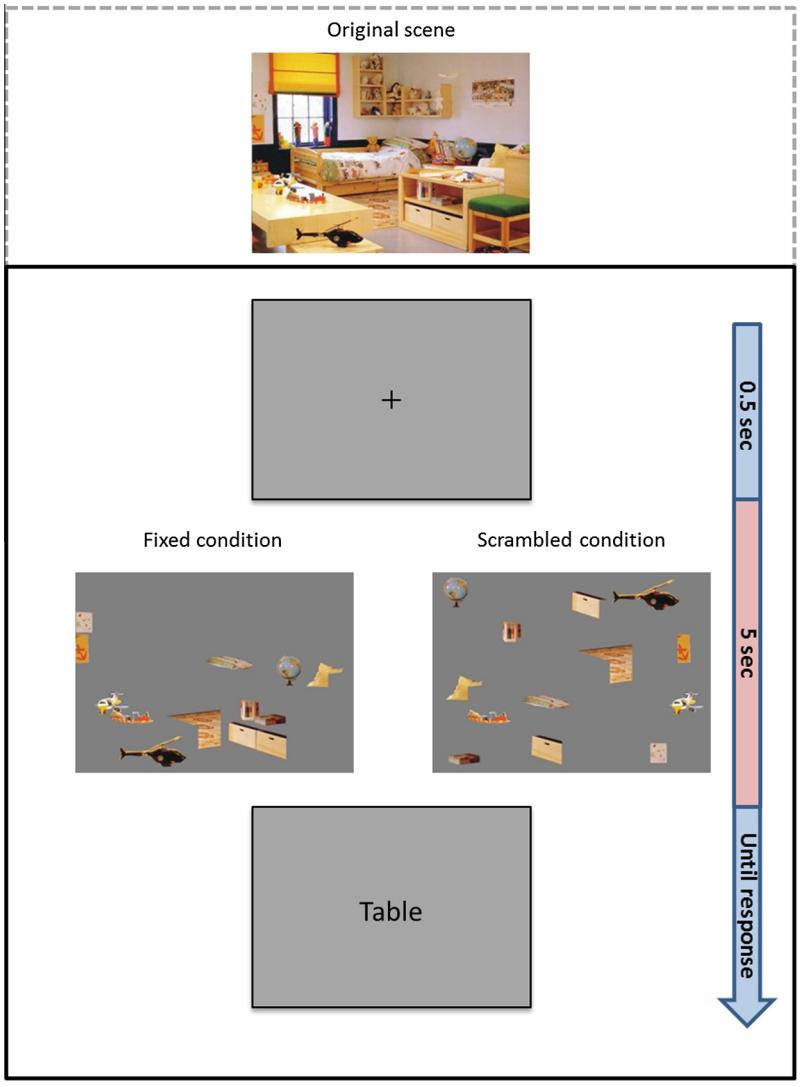

Fig. 1. Original scene (top) and a sample trial (bottom). The upper panel shows the original scene used to create stimulus displays. The scene would be used to generate an image showing only a subset of the objects, which were either located at their original coordinates (fixed condition) or at randomly selected locations (scrambled condition). During each trial, the created image was presented for 5 s. After the image disappeared, a word was presented and subjects were asked to report whether the object indicated by it had been shown in the previous display.

Table 1
Accuracy of performing the classification task (proportion of correct classifications).

	Amazon Mechanical Turk human rating
Whole scene	0.683
Fixed	0.326
Scrambled	0.253
Chance level (1/8)	0.125

Table 2
Sample LSA cosine values.

Label 1	Label 2	Cosine
–	–	–
AIRPLANE	HELICOPTER	0.62
AIRPLANE	TOY TRAIN	0.28
AIRPLANE	PICTURE	0.14
AIRPLANE	PILLOW	0.03
–	–	–

words. 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, Singular Value Decomposition (SVD; see Berry, Dumais, & O'Brien, 1995) is applied to reduce the matrix to a lower-dimensional vector space, referred to as 'semantic space'. LSA can still estimate the semantic similarity of two words even when they never co-occur in the same document (Jones & Mewhort, 2007; Landauer & Dumais, 1997).

Every term, every document, and every novel collection of terms 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 2 shows examples of LSA cosine values for various object labels used in scene image "Child4" (see Fig. 1) in terms of the reference object labeled as "AIRPLANE". This label has, for instance, a higher cosine value (greater semantic similarity) with "HELICOPTER" (0.62) than with "PILLOW" (0.03). This difference indicates that in the text corpus, "AIRPLANE" and "HELICOPTER" occur in more similar contexts than do "AIRPLANE" and "PILLOW". One of the nice features of LSA is that it can quantify higher-level

conceptual semantic similarity, regardless of any geometrical or functional relation. 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.

To compute semantic similarity for each pair of object labels in the current study, 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 with 300 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.

2.3.3. Measuring semantic guidance

In this study, the semantic guidance effect was defined as the extent to which the semantic relation (i.e., similarity) between the currently fixated object and the other objects in the scene influences the choice of the next fixated object. Semantic guidance would be indicated by a tendency of saccades to land on objects with above-chance level semantic similarity to the previously fixated item. In order to quantify this effect, its computation followed each subject's eye movements. Since we were interested in the effect of semantic similarity on gaze transitions, i.e., which object would be inspected next, only eye movements that transitioned between distinct objects were analyzed. For the starting point of each of these transitions, a semantic map was generated based on the LSA cosine value between the labels of the currently fixated object and each other object in the scene, as shown in Fig. 2.

The semantic maps, excluding the area occupied by the currently fixated object and the areas not containing any object, were normalized by linear scaling so that the mean of all activation was zero and its standard deviation was one. With this normalized semantic map, we computed the average activation along scan paths. That is, each fixation would build its own semantic map as a predictor of the target point of the next transition. All normalized values computed along scan paths were averaged across all transitions to obtain the extent of semantic guidance during the inspection of a scene. If eye movements were exclusively guided by semantic information, this average Normalized Scanpath Saliency (NSS) value should be significantly greater than zero. On the other hand, if there were no semantic effect on eye movements at all, the average NSS value should be close to zero, indicating prediction at chance level (Peters & Itti, 2007; Peters et al., 2005).

2.3.4. Evaluating semantic guidance and the NSS measure

To evaluate whether subjects show semantic guidance and to test the NSS measure, subjects' NSS values computed from their empirical gaze transition data were compared with a control data set of random fixations that were generated by replacing subjects' fixation positions with the center positions of randomly selected objects in the scene. This data set served as an unbiased test of NSS values. That is, since gaze transitions of the random data set were not affected by any object properties, we should always receive a chance level NSS value (NSS = 0).

It is important to note, however, that an average above-zero NSS value for the empirical gaze transitions does not necessarily indicate semantic guidance. This is due to the fact that objects in close proximity to each other in a real-world scene tend to have greater semantic similarity than those with long distances between them. Since most saccades made by human observers are short compared to the display size, most of the resulting transitions occur between objects that are close to each other. Therefore, we could find an elevated average NSS value even if a subject's attention were not guided by semantic similarity at all; this was termed the *proximity effect* by Hwang, Wang, and Pomplun (2011).

To avoid this possible confound, for any NSS measure that we conducted below, we also analyzed NSS as a function of saccade size. In addition, "ground truth" NSS values were computed to serve as a baseline that exists in the spatial arrangement of each pair of objects. In other words, the baseline NSS values are the NSS values we would expect if no actual semantic guidance occurred, i.e., transitions were randomly selected. To compute this baseline data, we first computed the average semantic similarity

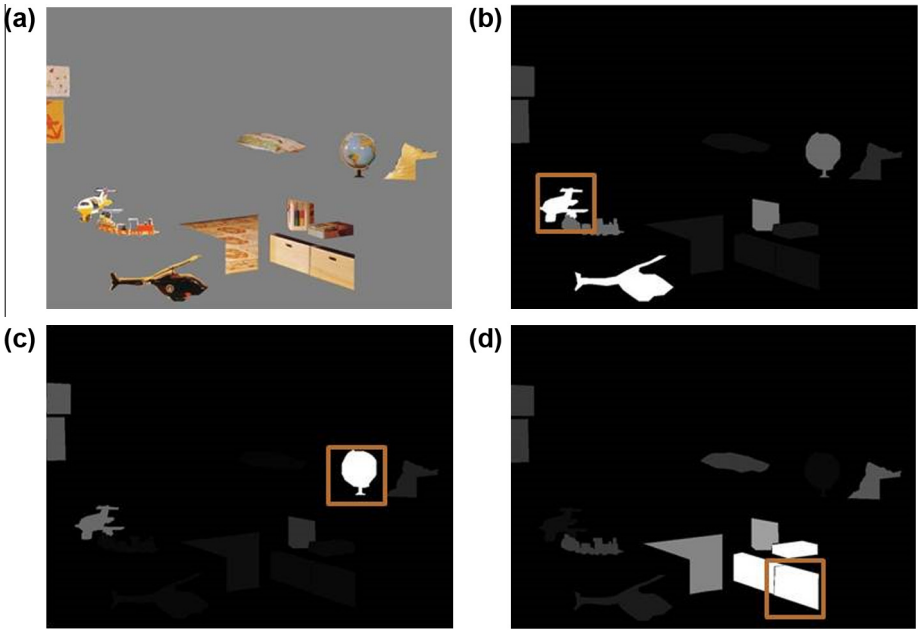

Fig. 2. Examples of semantic landscapes. The currently fixated object is marked with an orange square. (a) The original image that subjects inspected. (b) Semantic landscape during gaze fixation on the object labeled as "AIRPLANE". (c) Semantic landscape during gaze fixation on the object labeled as "GLOBE". (d) Semantic landscape during gaze fixation on the object labeled as "STORAGE BOX". As shown above, objects with conceptually higher relevance – measured as greater semantic similarity to the currently fixated object – receive higher activation (illustrated by greater brightness); for example, the helicopter in (b) shows high activation due to the fixated object labeled as 'AIRPLANE'. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

between objects whose distance from each other fell into a given saccade size interval. All pairwise distances between objects in all stimulus images were included in this computation. This ground truth reflects the influence of the proximity effect on the data. The NSS values measured from the actual saccades were averaged separately for the same saccade size intervals, and the corresponding ground truth values were subtracted from them. Plotting the resulting NSS difference for each saccade size interval allowed us to determine whether for any given interval the semantic guidance was above chance level, i.e., ground truth. This would be indicated by an NSS difference above zero.

Another confound that needs to be considered is the possibility that objects that are semantically similar according to the LSA measure may also tend to be visually similar. If that were the case, it could be visual similarity, rather than semantic similarity, that causes the observed scanning behavior. Indeed, Hwang, Wang, and Pomplun (2011) used four feature dimensions (size, color, compactness, orientation) to measure the visual similarity between objects and found a direct correlation between visual and semantic similarity. However, it was weak, $r = 0.15$, and computing guidance by visual similarity yielded a much smaller effect than guidance by semantic similarity. It thus seems justified to rule out the possibility of visual similarity substantially influencing the NSS measure.

2.4. Results

First, we investigated whether observers were able to recognize the selected objects and perform the recall test. Note that the selected objects used in the current study were polygonal regions cropped from a natural scene. Therefore, some of the regions may only have shown an incomplete object if it was partially occluded by other objects in the original scene. The recall performance might have been impaired if observers could not recognize the selected objects. To test this possibility, we analyzed the observers' performance in the recall test. Their recall performance was clearly above chance level in both the fixed and scrambled conditions (fixed condition: 76%, $t(19) = 22.78$, $p < .01$; scrambled condition: 70%, $t(19) = 8.15$, $p < .01$). This implies that observers were able to reliably recognize the selected objects in the stimulus display when the contextual information was removed, even when the spatial dependency was excluded as well (scrambled condition).

The aim of the current study was to investigate whether observers would use the semantic relations among objects to guide their attention. To examine semantic guidance, we computed NSS values for the two experimental conditions (fixed vs. scrambled) and both data sets (empirical and random). Fig. 3 shows that the semantic guidance values of random fixations were close to 0 in both the fixed and scrambled conditions. This result shows that the NSS computation was applied properly and the normalized semantic landscapes used in our analysis were unbiased.

As expected, observers' NSS values were small in both conditions since the contextual information was removed from the scene so that the remaining semantic information in the scene was very limited. Another reason for this finding may be that the small number of visible objects increased the noise in the NSS measurement. Nevertheless, the NSS value in the fixed condition (NSS = 0.21) was still significantly higher than that in the scrambled condition (NSS = −0.002), $t(19) = 11.08$, $p < 0.01$. This result suggests that, although the effect was small, the spatial dependency among objects preserved in the fixed condition provided additional semantic information and facilitated semantic guidance. When the spatial dependency was eliminated by shuffling the locations of objects in the scrambled condition, observers' NSS values decreased and showed no difference with the random condition ($t(19) = 1.46$, $p = 0.162$). As noted earlier, however, any above-

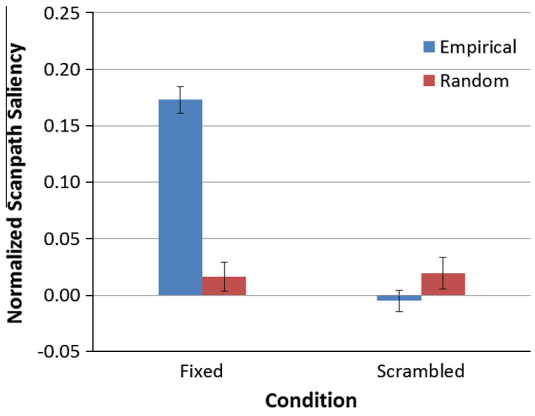

Fig. 3. Semantic guidance as measured by the NSS method in the fixed condition and the scrambled condition in Experiment 1. The errors represent ±1 standard error of the mean.

chance effect may not necessarily indicate semantic guidance due to the proximity effect (see Section 2.1). To clarify whether the observed effect was simply due to the proximity effect or due to the semantic guidance induced by the remaining spatial dependency, we subtracted the NSS values of the ground truth from that of the empirical gaze transitions and analyzed NSS as a function of saccade size. Since the ground truth of NSS values represents the average semantic similarity among objects in our stimulus displays, any above zero value after subtracting the ground truth would indicate the use of semantic guidance.

Fig. 4 shows the difference of NSS values between empirical gaze transitions and ground truth over the different saccade size intervals. Note that there were not enough empirical transitions smaller than 2° (2.9% of all transitions in the fixed condition and 0.3% of all transitions in the scrambled condition), thus we collapsed all transitions smaller than 4° to allow interval-based analysis. The result shows that only for the short saccades (<4°), NSS values were smaller than zero. This is due to the fact that the objects near each other indeed have higher semantic similarity so that the difference was small or even negative when the ground truth of NSS values was subtracted from the NSS values of empirical gaze transitions. For the transitions with larger saccade size (>4°), however, the difference in NSS values between empirical transitions and ground truth was consistently above zero for the fixed condition ($t(7) = 9.1$, $p < 0.01$) but not in the scrambled condition ($t(7) = 0.67$, $p = 0.53$). This implies that the observed semantic

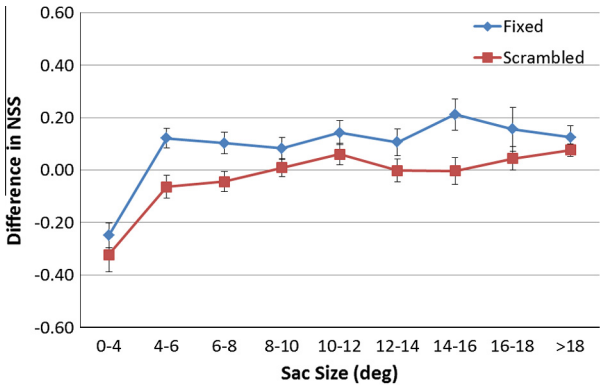

Fig. 4. The difference of NSS values between empirical gaze transitions and ground truth in Experiment 1. Results are shown separately for different saccade size intervals. The error bars represent ±1 standard error of the mean in the interval. Note that, to ensure that each data point has a sufficient number of transition samples, NSS values for saccades shorter than 4° were collapsed into one data point and those for saccades longer than 18° were collapsed into another data point.

guidance during scene viewing cannot be simply due to the proximity effect among objects in the scene. Therefore, it seems that semantic guidance in Hwang, Wang, and Pomplun (2011) was contributed, at least partially, by the spatial dependency among objects in the scene.

The results so far only show that spatial dependency does affect semantic guidance of visual attention. Nevertheless, they cannot rule out the possibility that semantic guidance also, to some extent, depends on scene gist, since Experiment 1 removed this factor completely. Many prior studies had found that, in early visual processing, observers are able to extract scene gist based on spatial layout, texture, volume or other low-level image features without the need of recognizing objects in a scene, and use this information to guide their attention and eye movements (Oliva & Torralba, 2001; Schyns & Oliva, 1994; Torralba et al., 2006). Therefore, it is possible that the semantic guidance found in the previous study may depend on scene gist as well.

To investigate the role of scene gist on semantic guidance, in Experiment 2 we conducted a scene inspection task that was similar to Experiment 1 and manipulated the presence of scene gist using a preview paradigm. If scene gist can induce semantic guidance as spatial dependency does, providing extra semantic information along with spatial dependency should lead to even greater semantic guidance than providing spatial dependency alone.

3. Experiment 2

To investigate the role of the scene gist on semantic guidance without interference by other factors, the spatial dependency was preserved in all stimuli as in the fixed condition in Experiment 1. In one condition, we provided scene gist by letting subjects preview the complete scene before the inspection stimulus appeared (*with-preview condition*) so that subjects could use both scene gist and spatial dependency to guide attention. In the other condition, no preview was provided (*without-preview condition*). If scene gist can contribute semantic guidance, providing it to the observers should facilitate the use of semantic information and produce stronger semantic guidance.

3.1. Participants

Twenty new subjects, aged between 19 and 40 years old, were tested. All had normal or corrected to normal vision, and were naïve as to the experimental design and hypothesis.

3.2. Apparatus

The apparatus was identical to that used in Experiment 1.

3.3. Stimulus display

Stimulus displays were generated in the same way as in Experiment 1. Each image was composed of 13–15 objects which were manually selected from a natural scene. Subsequently, all selected objects were pasted onto a grey canvas, and each of them was placed at the same coordinates as in the original image to retain the spatial dependency as in the fixed condition in Experiment 1. In addition, 120 distinct scenes were used in Experiment 2 to increase the variety of scenes and thereby minimize any bias by image selection.

3.4. Procedure

In one condition (with-preview condition), we used a preview of the entire original image from which the objects had been extracted. The preview was shown for 80 ms, followed by a mask for 500 ms so that observers could retrieve scene gist information during early visual processing while being unable to obtain and memorize a significant amount of semantic information for individual scene objects. The mask was made by randomly selecting an RGB color code in each pixel. In the other condition (without-preview condition), there was no preview shown so that the materials were identical to the fixed condition in Experiment 1. Both conditions were run separately (with-preview & without-preview). All images were only shown once in the experiment and both conditions were administered in a counterbalanced order.

In the with-preview condition, observers were told that there would be a flash preview followed by a mask before the search display. They were also told that the preview was the original image that was used to generate the following search image and they should attend to these previews since they may provide additional information to facilitate the scene inspection. In both conditions, observers were instructed to inspect the search image and memorize it for the subsequent recall test. After the five-second presentation of each scene, an English word was shown and observers were asked whether the object indicated by the word had been shown in the previous scene image. Fig. 5 shows an example trial sequence in the with-preview condition.

3.5. Results

Observers' recall performance revealed no statistically relevant difference between the with-preview condition (71%) and the without-preview condition (73%), $t(19) = 0.68$, $p = 0.508$.

To investigate whether scene gist could induce semantic guidance, we conducted the same NSS analysis as in Experiment 1, including "random fixation" control data, and compared the NSS values between both experimental conditions. Fig. 6 shows the result, which is similar to the one in Experiment 1. In the without-preview condition, observers had a stronger semantic effect (NSS = 0.27) than in the control condition (NSS = 0.04), $t(19) = 9.91$, $p < 0.001$. This finding basically replicates the result from Experiment 1, while using a larger image set. In Experiment 1, observers were still able to use the remaining spatial dependency among objects to obtain semantic guidance even when scene gist was not available. Surprisingly, in Experiment 2, adding contextual information by providing the preview of the complete scene did not seem to enhance the semantic guidance induced by the spatial dependency. Fig. 6 shows that the NSS values did not significantly differ between the with-preview condition (NSS = 0.25) and the without-preview condition (NSS = 0.27), $t(19) = 0.97$, $p = 0.34$. This finding suggests that knowing the scene category in advance does not help observers use the semantic information to guide their attention.

To exclude the effect of proximity which may contribute to the NSS values observed in Experiment 2, we subtracted the NSS values of the ground truth from that of the empirical gaze transitions as we did in Experiment 1. Fig. 7 shows that the difference between empirical transitions and ground truth was significantly larger than zero in both the with-preview and without-preview conditions ($t(8) = 4.56$, $p < 0.01$ for the with-preview condition; $t(8) = 8.31$, $p < 0.01$ for the without-preview condition). This suggests that both conditions induced actual semantic guidance. In addition, there was no difference in NSS values between the with-preview and without-preview conditions across different saccade size intervals ($t(8) = 0.17$, $p = 0.87$) and the overall saccade sizes were similar between the two conditions (averaging 9.8° in the with-preview condition and 9.5° in the without-preview condition). This implies that providing the extra scene gist information did not change the subjects' viewing strategy or help subjects access the semantic information better.

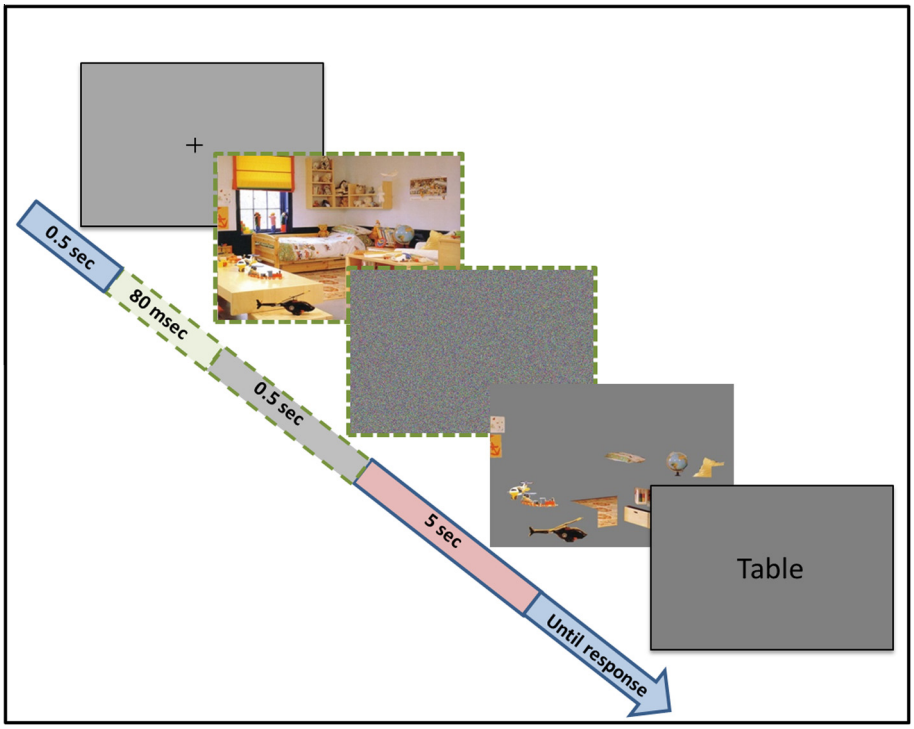

Fig. 5. A sample trial in Experiment 2. The green dashed outline indicates the preview and the mask for the trials in the with-preview condition. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

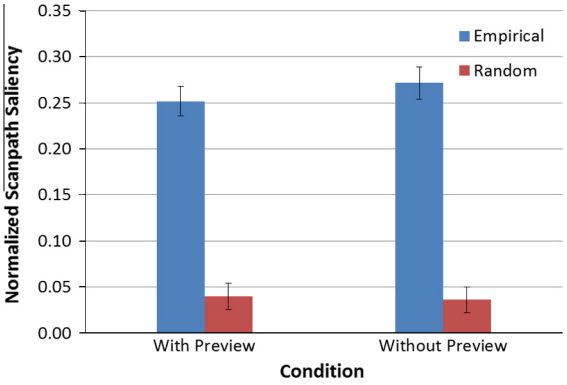

Fig. 6. Semantic guidance as measured by the NSS method in the with-preview condition and the without-preview condition in Experiment 2. The error bars represent ±1 standard error of the mean.

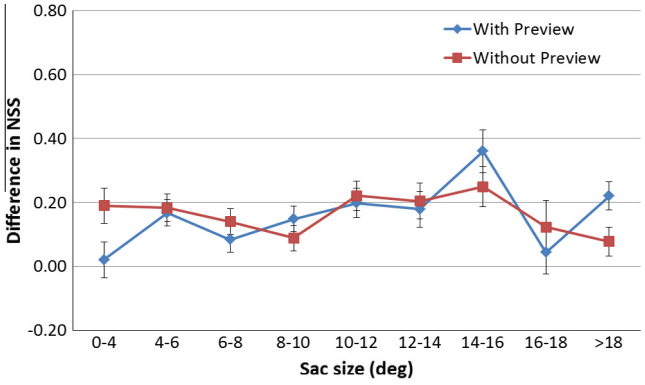

Fig. 7. The difference of NSS values between empirical gaze transitions and ground truth in Experiment 2. Results are shown separately for different saccade size intervals. The error bars represent ±1 standard error of the mean in the interval. Note that, to ensure that each data point has sufficient samples of transitions, NSS values for saccades shorter than 4° were collapsed into one data point and those for saccades longer than 18° were collapsed into another data point.

Many previous studies found that the information of scene gist could be perceived within less than 100 ms without the need for recognizing any particular object in the scene, and that this information could be used for later attentional guidance (Potter, 1976; Torralba et al., 2006). This ability to instantly capture gist information does not imply that scene gist and spatial dependency must be retrieved and used hierarchically. While the above studies have shown that scene gist can be perceived in early visual processing, it is important to note that, under natural viewing conditions, scene gist is available at any given time during scene inspection and may be processed whenever needed. However, Experiment 2 confined this process to occur only in the beginning of the inspection process, which may underestimate the effect of gist on semantic guidance. To resolve this possible limitation, we conducted an additional Experiment 3 in which the scene gist information was provided throughout the inspection process as under natural viewing conditions (*with-gist* condition). Though observers were shown a complete natural image, they were asked to inspect only a marked set of scene objects, which were identical to the objects selected in the fixed condition of the previous two experiments. Therefore, if scene gist can facilitate semantic guidance in any way during scene viewing, we should be able to see stronger semantic guidance when it is provided.

4. Experiment 3

To further examine whether scene gist can strengthen semantic guidance, scene gist was provided by showing the original natural scene without removing the background (*with-gist* condition). In two other conditions, scene gist was removed as in Experiment 1 by placing a subset of objects on a gray background. In the *fixed* condition, the selected objects were placed at their original loca-

tion. If scene gist contributes to semantic guidance, showing the entire image should facilitate the use of semantic information and produce stronger semantic guidance. In addition, to ensure that the significant effect of spatial dependency observed in Experiment 1 was not caused by the bias from the small image set (only 60 images were used), we also included the *scrambled* condition in which both scene gist and spatial dependency were removed.

4.1. Participants

Nineteen subjects, who did not participate in either of the previous two experiments, participated in Experiment 3. 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, and were naïve to the experimental design and hypothesis.

4.2. Apparatus

The apparatus was identical to the one used in Experiments 1 and 2.

4.3. Stimulus display

The 120 distinct scenes from Experiment 2 were also used in Experiment 3. Each scene was used to generate three images: (1) a complete scene with the specified objects indicated by red marks; (2) the same marked objects placed on a gray background at the same coordinates; (3) the same marked objects placed on a gray background at different, randomly chosen coordinates. The latter two types of stimuli were identical to those used in the fixed and scrambled conditions of the previous two experiments except that the location of each object was indicated by a red mark ($\sim$20 minarc $\times$ 20 minarc) at its center in order to ensure comparability with the images in the *with-gist* condition. Fig. 8 shows an example of each type of stimulus (top row) and their semantic saliency maps that were used to measure semantic guidance (bottom row).

4.4. Procedure

The three conditions (with-gist, fixed, scrambled) were run in separate blocks. In the with-gist condition, subjects were told to only focus on the marked objects and answer the recall question based on only those objects. In the other two conditions, the procedures were identical to the previous experiments. Each scene was only presented once to each subject in the form of the with-gist condition, the fixed condition or the scrambled condition. The association between images and conditions as well as the presentation order of all conditions were counterbalanced across subjects.

4.5. Data analysis

The data in the fixed and scrambled conditions were analyzed in the same way as in the previous experiments. In the with-gist condition, since subjects were instructed to only focus on the marked objects, all recorded fixations were analyzed as in the fixed condition. That is, all fixations were assumed to land only on the pre-specified objects. If any fixation landed on another region in the scene, it would be treated in the same way as a fixation landing on a blank area in the fixed condition. Consequently, this fixation would be assumed to be aimed at the pre-specified object with the shortest Euclidean distance to it. Furthermore, the semantic maps of the images in the with-gist condition were generated based on only those objects that were selected in the other two conditions so that we could directly compare semantic guidance across all three conditions.

4.6. Results

Observers' recall performance was 80%, 81% and 78% for the with-gist, fixed, and scrambled conditions, respectively. A one-way ANOVA indicated that the performance did not differ across conditions ($F(2,54) = 0.781$, $p = 0.463$).

Similar to the previous two experiments, we compared subjects' NSS values with the control conditions ("Random") in which fixations were randomly assigned to the center of one marked object.

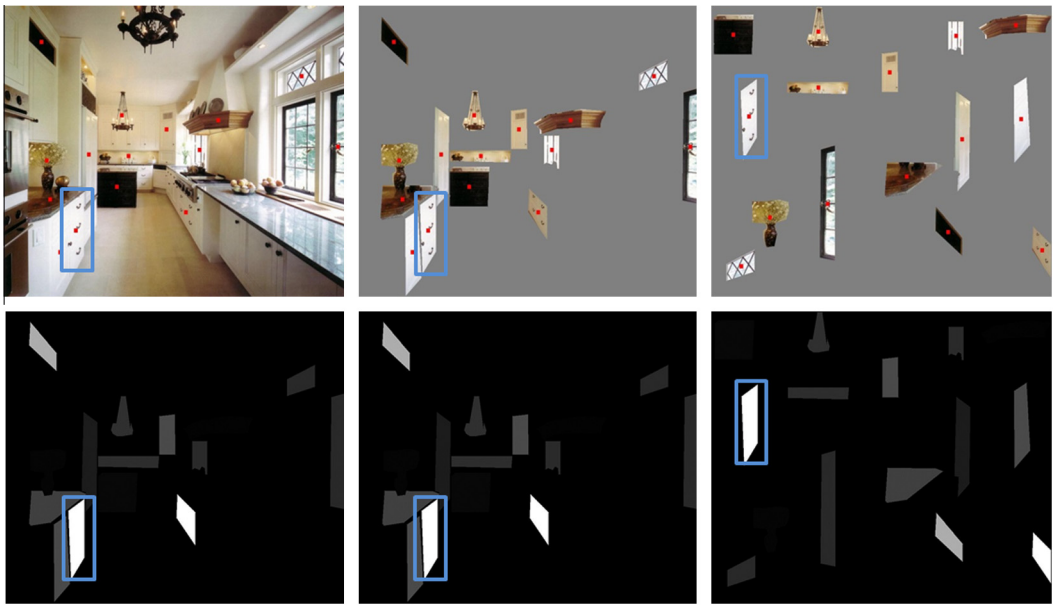

Fig. 8. Examples of stimuli in Experiment 3 (top row). Left: The stimulus in the with-gist condition. The locations of selected objects were indicated by red marks. Middle: The stimulus in the fixed condition. Right: The stimulus in the scrambled condition. The bottom row shows the corresponding semantic saliency map for each condition. The blue box in each image indicates the currently fixated object ("drawers"). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

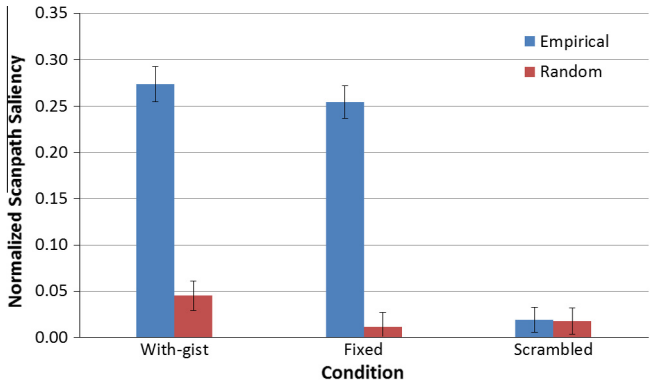

Fig. 9. Semantic guidance as measured by the NSS method in the with-gist condition, the fixed condition, and the scrambled condition in Experiment 3. The errors represent ±1 standard error of the mean.

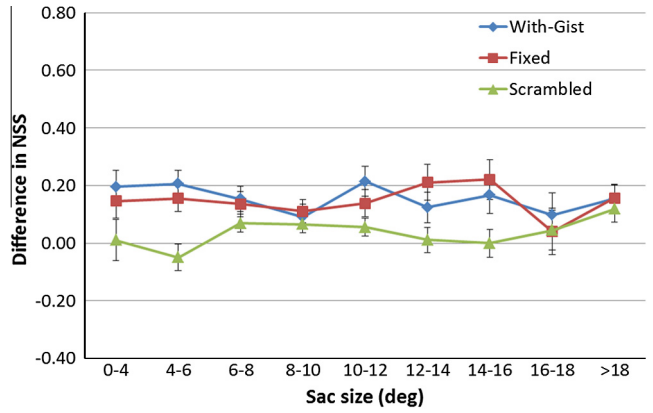

Fig. 10. Semantic guidance as measured by the NSS method across different saccade size intervals in all three conditions of Experiment 3. The error bars represent ±1 standard error of the mean in the interval. Note that, to ensure each data point has sufficient samples of transitions, NSS values for saccades shorter than 4° were collapsed into one data point and those for saccades longer than 18° were collapsed into another data point.

Fig. 9 shows that, as we found in the previous experiments, the effect of semantic guidance was higher in the empirical data than in the control condition even when the scene gist was removed ($t(18) = 7.21$, $p < 0.01$ in the with-gist condition and $t(18) = 11.3$, $p < 0.01$ in the fixed condition). This effect disappeared when the spatial dependency was also eliminated from the scene, and the empirical eye movements had similar NSS values as the random fixations ($t(18) = 0.07$, $p = 0.945$). However, when the information of scene gist was provided throughout the entire inspection process (with-gist condition), the effect (NSS = 0.27) was similar to the semantic guidance in the fixed condition (NSS = 0.25), $t(18) = 0.614$, $p = 0.55$. To verify whether providing the scene gist had no effect on semantic guidance and to exclude the possible proximity effect from the NSS measure, we conducted the same saccade-size based analysis as in the previous two experiments.

Fig. 10 shows that the semantic guidance in the with-gist condition was similar to the guidance in the fixed condition, in which scene gist was removed ($t(8) = 0.43$, $p = 0.68$), but was stronger than the guidance in the scrambled condition, in which both scene gist and spatial dependency were removed ($t(8) = 4.36$, $p < 0.01$). In addition, the average saccade sizes in the with-gist and fixed conditions were similar (9.4° in the with-gist condition and 9.5° in the fixed condition). These results imply that regardless of the availability of scene gist, subjects seemed to always use a consistent strategy of evaluating spatial dependency to guide attention. Moreover, Experiment 3 yields further evidence for the observation in Experiments 1 and 2 that semantic guidance was mainly contributed by spatial dependency among objects in the scene.

5. Conclusions

Hwang, Wang, and Pomplun (2011) found that, during scene inspection, observers tend to bring their line of sight to objects that are semantically relevant to the currently fixated object. It was not clear, however, how this semantic guidance of gaze transitions was induced. It may be contributed by three possible factors: (1) scene gist information; (2) local scene context based on the spatial layout of objects or (3) semantic evaluation through extrafoveal vision. The aim of the current study was to investigate the individual contribution of each of these factors to semantic guidance by independently varying the availability of the first two factors in the visual stimuli.

Our results show that, when the information of scene gist was removed, observers could still use the spatial dependency among objects to obtain semantic guidance. When both scene gist and spatial dependency among objects were removed in the scrambled conditions, the effect of semantic guidance completely disappeared and the NSS values were similar to the chance level in the control condition. It is important to note that we are not claiming that observers prefer to use semantic information over other visual cues such as proximity or feature similarity to guide attention. Furthermore, it is possible that the gist information cannot be completely eliminated by removing the scene background. Nevertheless, the global statistics of the scene provided from their background were equally removed in both the fixed and scrambled conditions and the only difference between them was the spatial layout of objects. Thus, any difference performance between these conditions would be only due to the spatial dependency among objects. The current findings demonstrate that the semantic guidance observed by Hwang, Wang, and Pomplun (2011) could not be simply due to the use of scene gist or the effect of proximity. The spatial arrangement of objects, on the other hand, seems to be sufficient to induce semantic guidance.

Experiments 2 and 3 also demonstrate that providing information of scene gist along with the spatial dependency did not facilitate the use of semantic information during natural scene viewing. This suggests that scene gist, at least in our experiment, only played a marginal role in providing semantic guidance. Subjects seemed to infer semantic similarity mainly from spatial dependency to guide their attention.

Note that we are not claiming that the three possible factors mentioned above are separate sources of semantic information in the scene since they may be tightly coupled. For example, knowing the category of the scene from scene gist may facilitate the processing of spatial dependency since the evaluation of spatial dependency could take relatively long due to the need of object recognition, but capturing scene gist is a nearly instant process (Oliva, 2005; Torralba et al., 2006). Furthermore, many studies suggested that retrieving scene gist leads to knowledge of spatial dependency of objects in the scene (see Tatler, 2009). The result of our preliminary psychophysical testing also shows that the impact of spatial dependency could take place at a glance (at least in a categorization task), since the performance at categorizing scenes was better in the fixed condition than in the scrambled condition. This implies that spatial dependency and scene gist, which is often defined as the information extracted during early visual processing, are not always clearly dissociated.

The main contribution from spatial dependency on semantic guidance may simply suggest that when both sources of informa-

tion are available at the same time, spatial dependency plays a more dominant role and observers seem to favor it over scene gist to infer other semantic information and help them decide where to look next. Using spatial dependency to guide attention may take more time than extracting the scene gist to accomplish this since the former involves the process of object recognition. Nonetheless, it may be a more reliable strategy to ensure that the most informative locations in a scene are being fixated.

Even though the current study did not find semantic guidance when both scene gist and spatial dependency were removed, it did not directly evaluate the possible effect of extrafoveal analysis on semantic guidance. Each selected object in the stimulus was directly segregated from a natural scene so that the image of an object was incomplete if that object was partially occluded by other objects in the original scene. It is possible that the ability to evaluate semantic information from extrafoveal vision may be underestimated if the objects cannot be recognized even if subjects fixated them. Nevertheless, the recall performance shows that observers were still capable to recognize the objects when the scene gist and spatial dependency among objects were removed. This suggests that the partial occlusion of objects did not really prevent observers from accessing the semantic information. This may imply that even though extrafoveal vision can be used to perceive scene gist (Larson & Loschky, 2009), it may only play a marginal role of extrafoveal analysis in semantic guidance.

Moreover, it is possible that spatial dependency could not only help understand the content of the scene, but also facilitate the process of recognition. By using spatial dependency among objects, observers may already recognize, at least partially, extrafoveal objects before they are fixated. A comparable result had been shown by Kotowicz, Rutishauser, and Koch (2010), who found that in a simple conjunction search task, the target was recognized before it was fixated. The function of the final saccade to the target was simply to increase the confidence of judgment.

Overall, the current study shows an irreplaceable role of spatial dependency among objects in semantic guidance of visual attention during natural scene inspection. The emphasis on spatial dependency over other types of semantic information provided in the scene (see Wu, Ahmed-Wick, & Pomplun, 2014, for a review) may shed light on the cognitive mechanisms underlying scene perception and memorization. Further research on semantic guidance, its neural basis, and its function is necessary before this concept can be integrated into current models of visual attention.

Acknowledgment

This research was supported by Grant Number R01 EY021802 from NIH - United States to M.P.

References

Berry, M., Dumais, S., & O'Brien, G. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review, 37*(4), 573–595.

Bruce, N. D., & Tsotsos, J. K. (2009). Salience, attention and visual search: An information theoretic approach. *Journal of Vision, 9*(3), 1–24.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3–5.

Callison-Burch, C., 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (Vol. 1, pp. 286–295). Association for Computational Linguistics.

Chun, M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences, 4*(5), 170–178.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE, 8*(3), e57410.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15*, 559–564.

Findlay, J. M. (2004). Eye scanning and visual search. *The Interface of Language, Vision, and Action: Eye movements and the Visual World*, 135–159.

Gordon, R. (2004). Attentional allocation during the perception of scenes. *Journal of Experimental Psychology: Human Perception and Performance, 30*(4), 760–777.

Gordon, R. (2006). Selective attention during scene perception: Evidence from negative priming. *Memory and Cognition, 34*, 1484–1494.

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision, 3*(1), 49–63.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences, 7*(11), 498–504.

Hollingworth, A., & Henderson, J. M. (1999). Object identification is isolated from scene semantic constraint: Evidence from object type and token discrimination. *Acta Psychologica: Special Issue on Visual Object Perception, 102*, 319–343.

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.

Hwang, A. D., Wang, H.-C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research, 51*(10), 1192–1205.

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.

Jones, M., & Mewhort, D. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*(1), 1–37.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*(4), 219–227.

Kotowicz, A., Rutishauser, U., & Koch, C. (2010). Time course of target recognition in visual search. *Frontiers in Human Neuroscience, 4*, 1–11.

Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research, 35*(13), 1897–1916.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211.

Landauer, T., Foltz, P., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2–3), 259–284.

Larson, A. M., & Loschky, L. C. (2009). The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision, 9*(10), 1–16.

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.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44*(1), 1–23.

Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron, 53*(4), 605–617.

Oliva, A. (2005). Gist of the scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (pp. 251–256). San Diego, CA: Elsevier.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*, 145–175.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research, 155*, 23–36.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences, 11*(12), 520–527.

Peters, R. J., & Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *IEEE conference on computer vision and pattern recognition, 2007. CVPR'07* (pp. 1–8). IEEE.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research, 45*(18), 2397–2416.

Pomplun, M. (2006). Saccadic selectivity in complex visual search displays. *Vision Research, 46*(12), 1886–1900.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 509–522.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology, 81*, 10–15.

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.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science, 5*, 195–200.

Tatler, B. W. (2009). Current understanding of eye guidance. *Visual Cognition, 17*, 777–789.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*, 520–522.

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.

Wu, C. C., Ahmed-Wick, F., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology, 5*, 54.