frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 06 February 2014
doi: 10.3389/fpsyg.2014.00054

Guidance of visual attention by semantic information in real-world scenes

Chia-Chien Wu**, Farahnaz Ahmed Wick and Marc Pomplun

Department of Computer Science, University of Massachusetts, Boston, MA, USA

Edited by:
Gregory Zelinsky, Stony Brook University, USA

Reviewed by:
Aude Oliva, Massachusetts Institute of Technology, USA
John M. Henderson, University of South Carolina, USA

****Correspondence:***
Chia-Chien Wu, Department of Computer Science, University of Massachusetts, 100 Morrissey Boulevard, Boston, MA 02125, USA
e-mail: chiachie@cs.umb.edu

Recent research on attentional guidance in real-world scenes has focused on object recognition within the context of a scene. This approach has been valuable for determining some factors that drive the allocation of visual attention and determine visual selection. This article provides a review of experimental work on how different components of context, especially semantic information, affect attentional deployment. We review work from the areas of object recognition, scene perception, and visual search, highlighting recent studies examining semantic structure in real-world scenes. A better understanding on how humans parse scene representations will not only improve current models of visual attention but also advance next-generation computer vision systems and human-computer interfaces.

Keywords: scene understanding, semantics, attention, scene perception, real-world scenes

INTRODUCTION

For the past two decades, research on the deployment of visual attention has shifted its focus from synthetic stimuli to real-world scenes. Unlike simple statistical structures of synthetic stimuli, real-world environments provide complex layers of information that cannot be processed all at once by our visual system. Despite this overwhelming amount of information, people perform daily visual tasks such as visual search or inspection with only a few glances. Therefore, effective vision greatly depends on the information observers acquire to help them decide where to look next. This has drawn a vast amount of research interest in recent years.

Current models of attentional deployment are divided into two camps – focusing on either top-down (or "endogenous") mechanisms or bottom-up (or "exogenous") mechanisms. One of the most influential studies on attentional guidance by bottom-up mechanisms was conducted by Koch and Ullman (1985), proposing the idea of a saliency map. In their model, features of entities in the scene such as edge density, color, intensity, and motion are computed in parallel as by different retinotopic maps in early visual areas. These maps are then combined into a single scalar saliency map representing relative conspicuities across the visual scene. The regions with "high salience" can be used to predict gaze fixation distribution in the scene, which indicates how attention is allocated in a visual scene (Kowler et al., 1995; Findlay, 1997). Therefore, the modeling of bottom-up mechanisms driven by saliency maps has been used extensively to predict the regions where attention is likely to be deployed during natural viewing (e.g., Itti and Koch, 2001; Bruce and Tsotsos, 2009).

Though the saliency map serves an important heuristic function in the study of eye movements, predictions from it begin to falter in real-world scenes and often fail to explain how gaze is directed. For example, Yarbus (1967) showed that the effect of context can direct eye movements to important locations in the scene, such as human faces (**Figure 1**). In addition to scene context, the observer's current task strongly influences allocation of attention. For instance, Hayhoe et al. (2003) asked observers to make a sandwich while their eye and hand movements were recorded. Their results showed that while the participants were performing these tasks in the real environment, they made clusters of fixations only to task-relevant regions. In contrast, task-irrelevant regions, sometimes having higher saliency, were rarely fixated. In order to account for the dominant control of visual attention by top-down mechanisms, many studies have incorporated top-down components into the saliency map to improve the prediction of gaze distribution (see Borji and Itti, 2013, for a review).

Unlike bottom-up mechanisms, which are mainly driven by the physical properties of the scene, top-down mechanisms process visual input in a way that is shaped by the observer's experience. Top-down mechanisms assign meaning to perceived information based on long-term memory content or knowledge that can be generalized from memory (e.g., inferring that the location of an unknown truck is likely near the ground level). For instance, when observers view the picture used in the study by Loftus and MacKworth (1978), they see "an octopus in a farmyard" rather than "an object at the bottom of the picture" (see **Figure 2**). Top-down mechanisms have been extensively investigated, with a focus on two aspects-one is goal directed, task-driven control (Hayhoe et al., 2003; Jovancevic et al., 2006), and the other is understanding of scene content, that is, how the content of a scene is learned and influences visual behavior (Neider and Zelinsky, 2006; Henderson et al., 2009; Tatler et al., 2010).

There has been a growing interest in the role of semantic information in attentional guidance. In order to access semantic information, the visual input has to be processed in a memory-based manner so that it can be assigned an existing meaning or be associated with a known category. This knowledge-based information has been referred to as "semantic" or "contextual" information by many studies, which have examined scene content at various levels. For example, some research focused on the

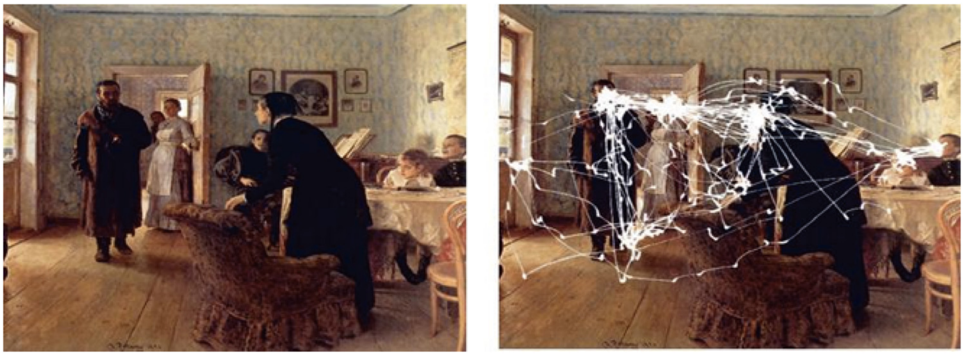

FIGURE 1 | Left: The oil painting "An Unexpected Visitor" painted by Ilya Repin. Right: an example of an eye trace measured by Yarbus during the free viewing condition (1967). This figure was originally referred from http://www.cabinetmagazine.org/issues/30/archibald.php.

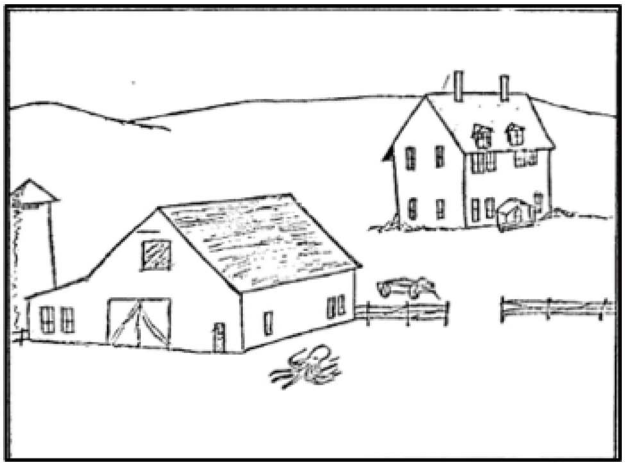

FIGURE 2 | A stimulus used in Loftus and MacKworth (1978). A line drawing of an octopus was placed on a line-drawn farmyard. Figure reproduced from Loftus and MacKworth (1978).

coarse-level category information of scenes, known as "scene gist" (Friedman, 1979; Schyns and Oliva, 1994; Oliva and Torralba, 2001; Torralba et al., 2006). Other work has examined the relations between scene and object or between objects themselves (Hollingworth and Henderson, 1998; Henderson et al., 1999; Joubert et al., 2007; Võ and Henderson, 2009). Therefore, when prior studies investigated whether and how semantic information in the scene can bias attentional deployment, it was sometimes confusing to discern which type of semantic information was actually being referred to. In order to understand how the visual system uses semantic information provided in the scene and incorporate this knowledge into the current state of attentional models, further clarification is needed.

Previous work has also reviewed semantic information in natural scenes and the role it plays in scene perception (Biederman et al., 1982; Rayner and Pollatsek, 1992; Henderson and Hollingworth, 1998, 1999; Henderson, 2003; Oliva, 2005, 2009; Oliva and Torralba, 2007). It did not, however, directly compare all aspects of semantic information. The current review takes a closer look at the different types of semantic information retrieved from natural scenes and how they may associate with each other. Note that the goal of the current review is not to exhaustively discuss all aspects of semantic information but to provide an overview of some important ongoing debates in this field of research.

Our review is structured as follows: in Section "Contextual Information – the Gist of a Scene," we summarize research on scene gist, one of the most studied types of semantic information, and discuss how this factor might influence attentional guidance even without accessing the meaning of objects. Section "Scene-Object Relations" examines how different aspects of semantic information can be extracted when meanings of objects are recognized and how the relation between objects and scene may guide visual attention. In Section "Object–Object Relations: Co-Occurrence and Spatial Dependency of Objects," we further discuss the effects of co-occurrence and spatial dependency between different objects in a scene on the allocation of attention. Section "Conceptual Semantic Associations between Objects in the Scene" discusses how conceptual semantic similarity between objects can be used by the visual system to guide visual attention. In the last section, we summarize the current debates on how different pieces of semantic information are perceived and how they are used to direct attention and facilitate scene understanding. By exploring the different aspects of semantic information, we hope to resolve some inconsistencies in the literature and shed light on how these factors influence visual attention during natural viewing.

CONTEXTUAL INFORMATION – THE GIST OF A SCENE

What kind of information can people perceive in the early stages of visual processing during natural viewing? Imagine that we are watching TV and rapidly flipping channels from one to another. With only a single glance we can identify the channels we want to skip and have no trouble recognizing whether they are showing news, sports, music, or a movie. This phenomenal ability to recognize each picture with a single glimpse has drawn substantial research interest. Potter (1976) found that, during an RSVP task, observers could detect the target picture, which was prespecified by a picture or descriptive title, within an exposure as short as 113 ms per image. Potter and Levy (1969) also found that observers were able to memorize an image within a presentation duration of only 100 ms when the image was presented by itself. These results suggest that, during the early visual processing stages, observers can extract many types of basic-level information (such as spatial configuration) and use them to identify the basic content of the scene (such as its category). This information is referred as the "gist" of a scene (Friedman, 1979).

Many following studies have investigated how the gist of a scene might affect visual attention and facilitate object recognition. The term "gist" has been widely used to refer to scene content ranging from low-level features (e.g., color or luminance) to high-level information (e.g., events occurring in the scene, see Friedman, 1979). Sometimes it has been termed "scene context" as well, although the context of a scene is often used to refer to co-occurrence relations among objects in the scene (see Section "Object–Object Relations: Co-Occurrence and Spatial Dependency of Objects"). Oliva (2005) provided a brief review about the different levels of scene gist, which includes the gist built during perception (referred to as perceptual gist) and a higher level of gist inferred from more complex semantic information, such as the meaning of an object or the relation between scene and objects (referred to as conceptual gist). According to this distinction, conceptual gist can persist after perceptual information is no longer available. Nevertheless, in the literature, the term "gist of a scene" typically refers to the essential level of information that is able to convey the basic meaning of the scene. For example, Oliva and Torralba (2001) found that humans are sensitive to the spatial structure of natural scenes, which can be used to infer the scene category. **Figure 3** shows an example of how a few global features (contour, density, and color in this case) are sufficient to form the spatial envelope and represent the gist of a scene. Torralba et al. (2006) also found that observers could extract some global scene properties without recognizing individual objects and use this information to guide their attention and eye movements.

Although the studies above have demonstrated that observers are able to use scene gist to facilitate scene understanding, little is known about when and where the gist of a scene is learned. Potter (1976) demonstrated that, within approximately 100 ms, observers were able to not only identify the category of an image, but also recognize some objects and their features. Thorpe et al. (1996) discovered that some meaning of a scene could be understood when it was presented for only 20 ms. These results show that the time course of perceiving the gist is clearly shorter than the time required for object recognition or typical saccadic preparation. This implies a minor role of foveal processing in perceiving a scene gist. A comparable result was also found by Larson and Loschky (2009). They compared observers' performance in scene recognition between two conditions: a gaze-contingent window condition in which peripheral vision was blocked and only central vision was permitted, and a "scotoma" condition in which central vision was blocked and only peripheral vision was intact. They found that when central vision was blocked, peripheral vision provided sufficient information for recognizing scene gist and performance was unimpaired even when the central "scotoma" was as large as 5°. A similar result was obtained by Boucart et al. (2013). They showed that observers could categorize scenes (highway vs. forest) even at 70° eccentricity. These findings, along with other studies suggesting that scene gist can be extracted from low spatial frequency information alone (Schyns and Oliva, 1994; Oliva and Torralba, 2006), demonstrate that peripheral vision is sufficient for recognizing scene gist.

Since observers can capture the gist of a scene within 100 ms (Potter, 1976) and central vision is not even necessary for gist recognition, this raises the question whether visual attention is required for recognizing scene gist. Li et al. (2002) found that the ability to classify a natural scene presented in peripheral vision is not impaired when the observer is performing a concurrent task presented in the central visual field (known as the dual-task paradigm). This implies that the perception of scene gist may be a "pop-out" preattentive process and does not require focal attention. The ability to detect scene gist instantly even when attention is directed to another task has become a commonly cited evidence for awareness of scene gist without attention (Rensink et al., 1997; Koch and Tsuchiya, 2007; see also Fabre-Thorpe, 2011, for a review). Furthermore, Serre et al. (2007) proposed a hierarchical feed-forward model that made passable predictions of human performance in a categorization

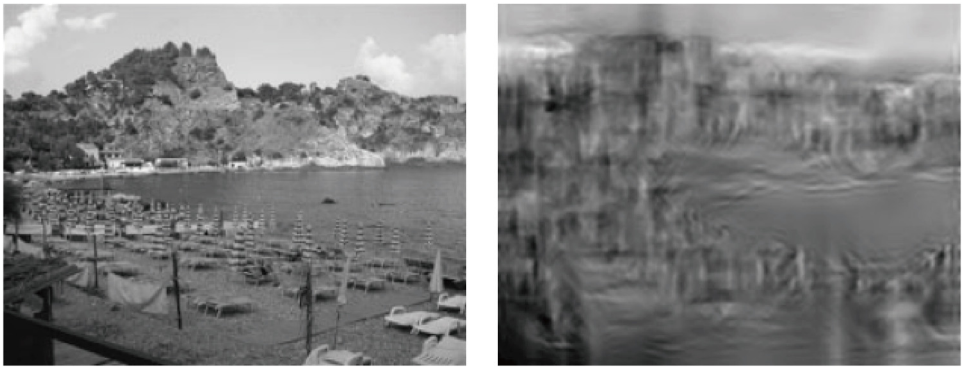

FIGURE 3 | Illustration of a natural scene (left) and its scene gist (right), generated by an image processing algorithm, that conserves sufficient spatial perceptual dimensions to infer the category of the scene (Figure reproduced from Oliva, 2005, Figure 41.3).

task in which each image was only presented for 20 ms. This result shows the neurophysiological plausibility of the rapid categorization task being performed without feedback loops for attentional modulation.

In contrast to these findings, Cohen et al. (2011) argued that the reason prior studies did not find impaired performance for gist detection in the dual-task paradigm was insufficient attentional demand in the central task. They conducted a dual-task experiment with a variety of demanding attention tasks and found that gist detection was impaired when the central task was sufficiently difficult. Their result suggests that awareness of scene gist is not preattentive and attention is essential to the process of perceiving scene gist. In order to reconcile their finding with the demonstrated human ability of extremely fast scene categorization, Cohen et al. (2011) proposed that some components of scene processing may be accomplished preattentively and can bias categorical decisions, whereas actual awareness of scene gist requires at least a small amount of attention. In agreement with this assumption, Kihara and Takeda (2012) found that observers can integrate information from different spatial frequencies without attention.

Regardless of the requirement of attention, the role of object recognition in acquiring the gist is still unclear. As mentioned earlier, some research has claimed that scene gist can be retrieved based on the processing of spatial layout, texture, volume, or other low-level image features and does not depend on recognizing objects (Schyns and Oliva, 1994; Oliva and Torralba, 2001; Torralba et al., 2006). On the other hand, other studies argued that scene gist is not processed independently but can be processed more accurately when the representative or diagnostic object in the scene is recognized. For example, recognizing a reverend in a scene can help in classifying the scene category as a church (Friedman, 1979; Davenport and Potter, 2004; see also Henderson and Hollingworth, 1999, for a detailed review of object recognition and scene context).

To summarize, in the literature on visual attention, scene gist may refer to different types of information provided in the scene. Nevertheless, in most cases the term "gist" indicates the information extracted in early visual processing (20–100 ms) that can convey the meaning of a scene and is sufficient to categorize the scene. Whether object recognition is necessary before perceiving scene gist, or whether visual attention is needed for gist recognition are still open questions.

SCENE-OBJECT RELATIONS

As discussed above, scene gist may be perceived without recognizing any object in the scene. However, to accomplish the arguably most common visual task – visual search-accessing the meanings of task-related items becomes essential. The gist of a scene not only enables us to recognize the category of environment we are looking at, but also facilitates object recognition, enabling us to locate the most informative region without serially inspecting every single position in a scene (Brockmole et al., 2006; Brockmole and Henderson, 2006a). Visual search in natural scenes is currently a prominent research paradigm because the human visual system still outperforms state-of-the-art computer vision systems. How is this high efficiency achieved? One important factor seems to be that, unlike artificial systems, the human visual system can use the context of a scene to guide attention before most of the scene objects are recognized.

When reviewing the literature, we first need to clarify what information is referred to as the "context" of a scene. In addition to scene gist, which may only provide some super ordinate category information about the scene, additional contextual information is conveyed when the meaning of an object is known. Upon recognition of an object, the observer's visual system considers both its semantic relationship (whether this object's identity fits in the scene) and spatial relationship within the scene (whether the object's location is appropriate). Biederman et al. (1982) used the terms "semantic" and "syntactic" to describe the relations between an object and its setting. Semantic relations require access to the object's meaning and involve probability, position, and expected size of objects in a scene. On the other hand, syntactic relations involve support and interposition, that is, it describes the laws of physics (e.g., whether an object should rest on a surface, or occlude the background). **Figure 4** show examples of semantic and syntactic relations and corresponding violations in a scene.

Scene consistency involving semantic and syntactic relations has been studied in the context of object recognition and visual search. Loftus and MacKworth (1978) investigated how fixations were distributed over consistent and inconsistent objects during picture viewing. In their experiments, participants were asked to memorize line drawings of natural scenes which were composed of "consistent" (e.g., a tractor in a farmyard) versus "inconsistent" (e.g., an octopus in a farmyard, see **Figure 2**) objects. The object congruency violations in their experiments were semantic in nature. Their findings showed that an inconsistent object in the scene was fixated on earlier and for a longer duration during free viewing than a consistent object. They suggested that allocating more attention to inconsistent objects might have been a memorization strategy to distinguish the informative regions in the scene. Some researchers have claimed that these categories (semantic vs. syntactic) are terms of linguistics and do not reflect distinct cognitive signals (see Henderson and Ferreira, 2004). However, a recent EEG study by Võ and Wolfe (2013) found differences in evoked potentials between semantic and syntactic violations during scene perception, indicating that they are being processed in categorically different ways.

A number of studies involving visual search and inspection tasks reported that inconsistent objects in scenes not only drew attention immediately but they also affected early eye movements (Loftus and MacKworth, 1978; Biederman et al., 1982; Hollingworth and Henderson, 1999; Gordon, 2004, 2006; Stirk and Underwood, 2007; Underwood et al., 2007; Bonitz and Gordon, 2008). These findings suggest that object-scene inconsistency attracts attention and gaze without the need for full object identification, also known as the pre-attentive pop-out effect (Johnston et al., 1990; Marks et al., 1992; Brockmole and Henderson, 2008).

Alternatively, many studies have found that consistent objects are easier to detect and identify than inconsistent objects (Boyce and Pollatsek, 1992a,b; De Graef, 1992; Henderson, 1992; Rayner and Pollatsek, 1992; Rensink et al., 1997, 2000; Kelley et al., 2003; Davenport and Potter, 2004; Malcolm and Henderson, 2010). Numerous studies have found that inconsistent object detection

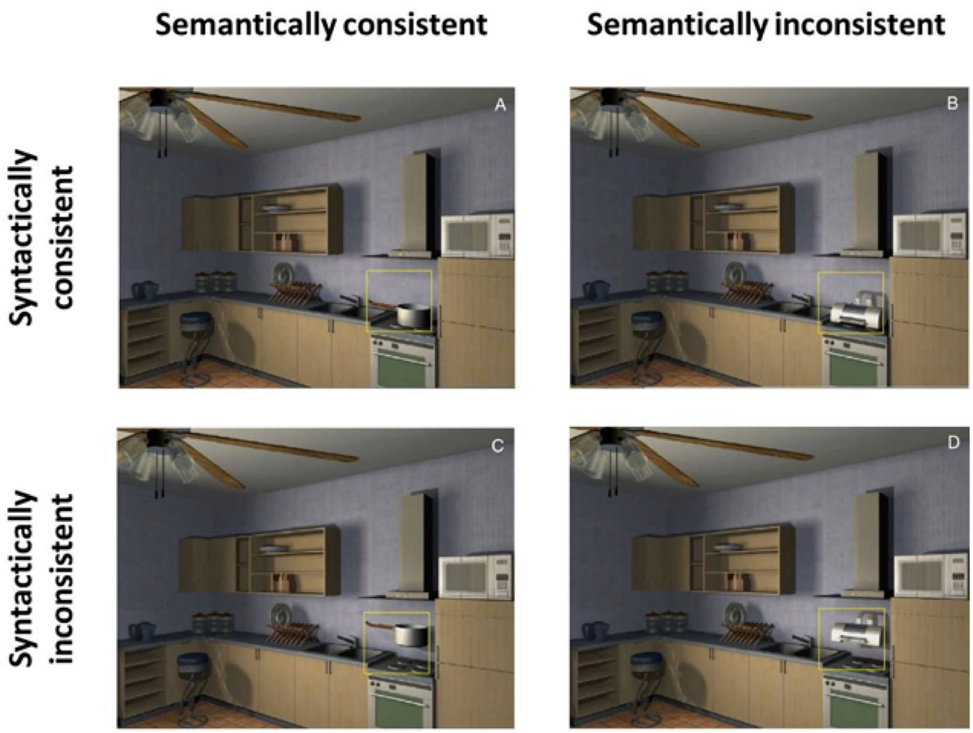

FIGURE 4 | Examples of semantic and syntactic relations in a kitchen scene. (A) Semantically and syntacticallyconsistent. **(B)** Semantically inconsistent (the printer does not belong in the kitchen scene) but syntactically consistent. **(C)** Semantically consistent but syntactically inconsistent (a floating pot violates gravity). **(D)** Semantically and syntactically inconsistent. Figure reproduced from Võ and Henderson (2009).

was slower and less accurate in scenes containing semantic, syntactic or both violations (Biederman et al., 1982; Henderson et al., 1999). Unlike earlier studies that used line drawings or photographs, Võ and Henderson (2009) used 3D-rendered images of real-world scenes to control for low-level cues. In agreement with the studies mentioned above, they did not find early effects of scene inconsistencies either during scene memorization or visual search. Together these results suggest that the pop-out effects found by previous studies may have been due to inconsistencies in bottom-up saliency such as inconsistent lighting, brightness, shading, or transitions between object and background when violations were created in the stimuli. It is thus possible that attention was drawn to these inconsistencies due to low-level features rather than semantic/syntactic violations in the stimuli.

The main criticisms against pre-attentive pop-out effects are: (1) low-level visual conspicuity of inconsistent objects from the rest of the scene; (2) failure to find fixational precedence for inconsistent over consistent objects regardless of their spatial structure in the scene (Friedman and Liebelt, 1981; Henderson and Hollingworth, 1998); (3) better discrimination performance for consistent extrafoveal objects than for inconsistent ones (De Graef et al., 1990; Hollingworth and Henderson, 1998; Võ and Henderson, 2009). There is more support for the claim that foveal processing is necessary before such inconsistent objects can be detected (Võ and Henderson, 2011) or even affect early eye movements (Henderson et al., 1999; Võ and Henderson, 2009).

How does the background of the scene play a role in deploying attention to objects in that scene? Though Henderson and Hollingworth (1998, 1999) found that objects in scenes were processed independently from their background, Davenport and Potter (2004) proposed an interaction between objects and their background during scene processing. They presented color photographs containing a salient or highly distinctive object in a scene, such as a road with a cyclist. The photographs were presented for 80 ms followed by a mask, and a naming task was used in which subjects were instructed to identify either the object or the background. The results showed that objects were more accurately detected in consistent settings, and backgrounds were perceived more accurately with consistent foreground objects. In another study, Davenport (2007) showed that in addition to the background, objects in scenes exert contextual influences on each other, suggesting that objects and their settings are processed together. Brockmole and Henderson (2006a,b) also found that the association between artificial objects and a natural scene could be learned via repeated exposure and used to facilitate search performance.

A question at this point is how the interplay between visual salience of objects in the scene and their semantic properties affects the allocation of attention during scene perception. The traditional view (Henderson et al., 1999) is that early fixations on a scene are determined by low-level visual features. A number of studies have reported that visual salience plays a dominant role in change detection tasks more than in search tasks (Pringle et al., 2001; Carmi and Itti, 2006; Spotorno and Faure, 2011; Spotorno et al., 2013). When target categories or specific items are altered during change detection tasks, this may impact the visual salience of the objects more than the semantic or syntactic congruency of the stimuli. Kollmorgen et al. (2010) examined the guidance of eye movements during a classification task using three measures: low-level features, high-level task dependent components (e.g., expression or gender of human faces) and spatial bias in stimuli composed of small image patches of either human faces or outdoor scenes. Each of these measures had a significant effect on eye movements. Spatial bias had the strongest effect on the guidance of eye movements. This was closely followed by high-level task-dependent components, and low-level features had less of an impact. The authors also found that task-dependent components had an especially strong effect when categorizing facial expressions. Other findings suggest an interaction between salience and semantic content of the scene (Nyström and Holmqvist, 2008) with semantic content causing more influence over time than visual salience. It is still unclear what proportions of these factors contribute to gaze guidance and how these proportions might vary over viewing time.

Regardless of visual salience or semantic relations between an object and its setting, there are particular classes of objects that can immediately capture our attention in an image. Texts were found to attract more attention than regions with similar size and position in real-world scenes (Cerf et al., 2009; Wang and Pomplun, 2012). Faces also belong to this special category of objects. Yarbus (1967) showed that during free viewing of a painting without any other instruction, fixations of observers were not evenly distributed but clustered around faces of the individuals in the painted scene (see **Figure 1**). In the first study of this kind, Buswell (1935) noted that human figures were disproportionately likely to be fixated on. Recently, Fletcher-Watson et al. (2008) found a similar bias toward human bodies and faces. This attentional preference is not limited to humans but applies to animals as well. In a series of studies, Kirchner and his colleagues (Kirchner et al., 2003; Kirchner and Thorpe, 2006) reported that participants were rapidly, within 120 ms of stimulus onset, able to saccade toward a natural scene with an animal when presented with two such natural images simultaneously.

Computational approaches predicting attentional allocation via fixational distribution performed much better than the traditional saliency map model when detection of special objects such as faces, texts, or both were incorporated. Cerf et al. (2008) integrated the Viola-Jones face detection algorithm (Viola and Jones, 2001) into their saliency model and demonstrated that, with this simple addition, the new model was better at predicting gaze allocation than the original one. Even superior performance was achieved by models using a nonlinear combination of several top-down *cognitive* features which affect eye-movements, such as human bodies and interesting objects such as cars, dogs, or computer monitors (see Elazary and Itti, 2008), along with bottom-up features (Borji, 2012; Zhao and Koch, 2012).

To summarize, the majority of findings from the literature suggest that the visual system utilizes knowledge of semantic coherence of a scene during search. This makes detection of inconsistent objects difficult unless these objects violate extreme semantic or syntactic rules. The spatial configuration of objects seems to define the basic structure of a scene and perhaps contributes the most in forming a scene schema.

OBJECT-OBJECT RELATIONS: CO-OCCURRENCE AND SPATIAL DEPENDENCY OF OBJECTS

The contextual information discussed so far involves either the scene gist or the relation between an object and the background of the scene. However, in the natural environment, objects rarely appear in isolation. The co-occurrence of objects and local spatial layout of the scene can be an alternate conceptualization of scene context (Bar, 2004; Mack and Eckstein, 2011). Object co-occurrence provides information about the likelihood of an object appearing in a scene when a reference object is recognized. For instance, if a scene contains a keyboard, it is likely that there is a mouse next to it. Typically, the concept of co-occurrence is associated with spatial proximity, that is, the tendency of certain objects to be located near each other.

Only a few studies have investigated the role of object co-occurrence in the deployment of attention in real-world scenes. Mack and Eckstein (2011) investigated the effects of object co-occurrence in natural environments on visual search. They found that viewers searched for targets (e.g., a headphone) at expected locations (e.g., next to an iPod) more efficiently than for targets at unexpected locations (e.g., next to a cup). Furthermore, a disproportionate amount of fixations landed on the relevant cue item (e.g., an iPod when headphones were the search target) as compared to the other scene objects. Similarly, using a flash-preview-moving-window paradigm, Castelhano and Heaven (2011) found that the presence of an object that co-occurs with the target can guide attention and facilitate search even when these pairings are shown on an inconsistent background (e.g., a can of paint in a bedroom).

While our visual environment is rich and complex, it also has regular and redundant spatial structure that reduces some of this complexity. In addition to object co-occurrence, the local layout of objects can also guide attention. The local layout constrains the probability of finding an object at a certain location relative to a reference object. For example, when searching for a keyhole, observers tend to first inspect the areas below any doorknobs. Such regularities are commonly referred to as spatial dependency between objects. Spatial dependency between objects in a scene arises because: (1) some objects fulfill a function together that requires a specific spatial arrangement, for instance, a computer mouse is likely located to the right of a keyboard; (2) some objects are physically supported by other objects, such as a computer monitor on a desk; and (3) real-world scenes have syntactic structure; for example, the sky usually appears in the upper half of an image, and pedestrians often appear in the middle of an image on a sidewalk or crosswalk.

Oliva and Torralba (2007) reviewed the effect of scene context extracted from spatial dependencies among objects. When an object was recognized in the scene, the identities and locations of other objects became highly constrained. **Figure 5** shows some examples of inferring objects based on a reference object. Before the reference object (red object) is recognized, other objects

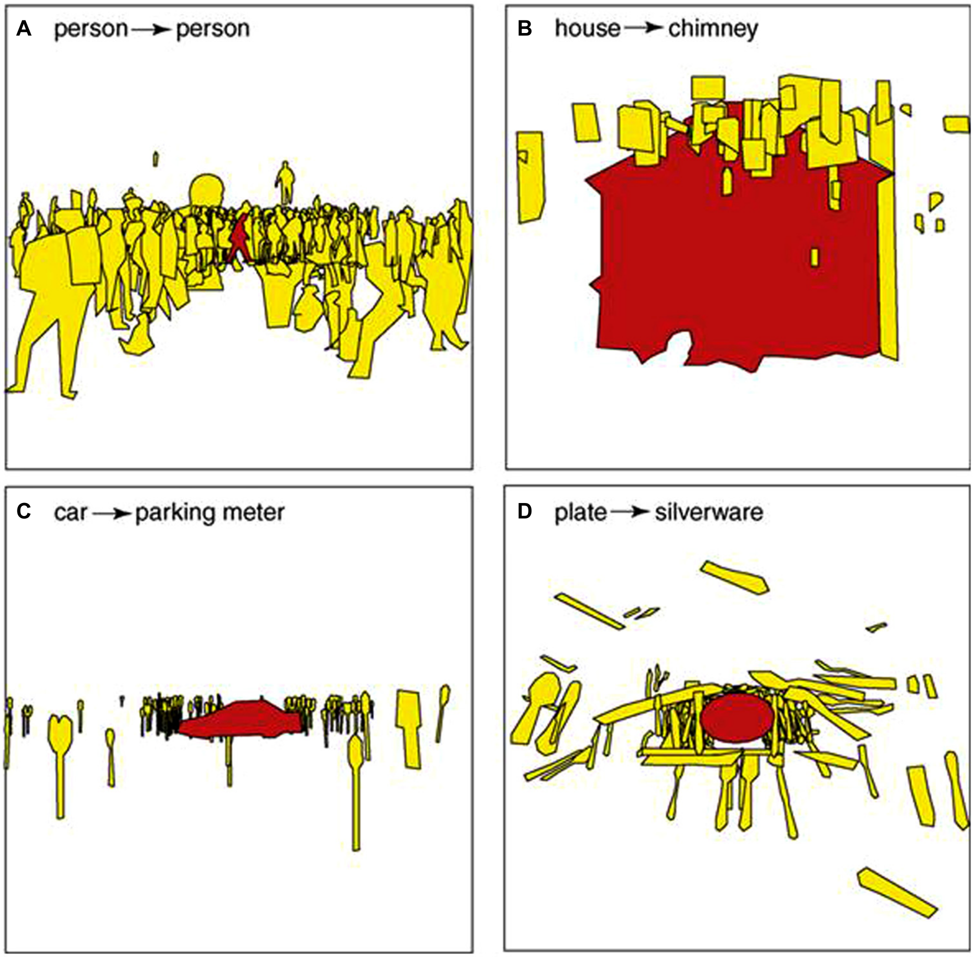

FIGURE 5 | Illustration of spatial dependency among objects. Without being directly fixated, the identities and locations of target objects (shown in yellow) are conditional to the recognition of a reference object (shown in red). The reference objects are **(A)** a person, **(B)** a house, **(C)** a car and **(D)** a plate. The inferred objects are **(A)** other people, **(B)** a chimney, **(C)** parking meters and **(D)** a silverware. Figure reproduced from Oliva and Torralba (2007).

around it can only be inferred based on their visual features. For example, the yellow objects in **Figure 5C** could be any items with elongated shape. Once the reference object has been recognized as a car, the other objects are most likely to be identified as parking meters.

In their pioneering work, Chun and Jiang (1998) coined the term *contextual cueing* for a paradigm where synthetic stimuli composed of L's and T's in certain spatial configurations were repeated throughout the experiment. They demonstrated that search time to locate targets in repeated configurations were significantly shorter than in novel arrangements of elements (Chun and Jiang, 1998, 1999, 2003; Olson and Chun, 2002; Jiang and Wagner, 2004; Brady and Chun, 2007). Later Brockmole and Henderson (2006a,b) found a similar contextual cueing effect in real-world scenes. In real-world scenes, it was shown that the first saccade during a search process would reliably go toward the predictive location of the target embedded in a real-world scene even when the target was absent, e.g., toward the top of a house when the target was a chimney (Beutter et al., 2003; Hidalgo-Sotelo et al., 2005; Eckstein et al., 2006; Droll and Eckstein, 2008; Ehinger et al., 2009; Castelhano and Heaven, 2010). The implication of this line of research is that knowledge of spatial dependency among objects is acquired through experience and is a top-down mechanism affecting visual processes (Chun, 2000).

There is an ongoing debate on how knowledge of spatial dependency among objects and knowledge of scene gist is utilized. In particular, there is some controversy regarding whether scene gist and spatial dependency are separate sources of information or hierarchically organized. The possibility most supported in the literature is that retrieving scene gist leads to knowledge of spatial dependency of objects in the scene (see Tatler, 2009). However, recent studies (Kanan et al., 2009; Castelhano and Heaven, 2011) question this assumption by showing that search can be guided by learned spatial dependency of objects and object appearance even without consistent gist information. For example, Brockmole et al. (2006) showed that the association between a target and its local context can be learned and bias attention when the global context in unpredictable.

Desimone and Duncan's (1995) Biased Competition Model proposes that attentional guidance and motor behavior in a scene emerges from competition among objects, moderated by both top-down and bottom-up processes. It is plausible to assume that associated objects can be processed together more easily than unrelated objects. Moores et al. (2003) found such top-down associative effects on the deployment of attention during visual search. They showed participants a display of four objects in which some objects were semantically related to each other (e.g., a motorbike and a motorbike helmet). The participants were asked to search for a target (e.g., a motorbike) before such a display was flashed briefly (presentation duration ranged from 47 to 97 ms). Their results showed that participants were able to recognize and recall more often those objects that were semantically related with the target than unrelated distractors. The authors speculated that the template of a target causes our visual system to activate templates of semantically related items. Belke et al. (2008), using a similar paradigm, also found that attention was preferentially attracted to those objects that were semantically related to targets and that perceptual load did not affect this bias.

The landmarks of the environment provide cues as to where attention should be deployed. Chun and Jiang (1998, 1999) demonstrated that participants could learn arbitrary configurations of targets and distractors that were repeated over epochs at a global level, that is, the arrangement of items or relative positions of targets and distractors. They also showed that such learning could take place at a local level, i.e., through the co-occurrence of novel objects in the display and even motion trajectories of items in the display. In fact, the literature in this domain supports the view that the local layout of objects within a global configuration plays a dominant role for selection and maybe sufficient to demonstrate many of the major properties of contextual cueing (Peterson and Kramer, 2001; Olson and Chun, 2002; Jiang and Wagner, 2004; Brady and Chun, 2007; but see Brooks et al., 2010). This is an intuitively credible claim because observers usually cannot perceive the bird's eye view of a scene immediately and must either search through their visual environment serially or use associative knowledge of items in their immediate view.

Unlike scene gist or object-scene relations, which can be acquired almost instantly in early visual processing and require little attentional resources, we need to identify at least one object in order to perceive relative associations among objects. Therefore, co-occurrence and spatial dependency of objects are perceived later than either scene gist or object-scene relations. Even though using object-object relations for attentional guidance may require more time and cognitive resources than relying on scene gist or scene-object relations, it still seems to be a commonly used strategy that observers adapt during natural scene viewing. It is likely that scene gist and scene-object consistency may affect only the initial stage of viewing (Brockmole and Henderson, 2006a). Zoest et al. (2004) suggested that top-down influences require at least 100 ms after scene onset to affect saccadic guidance. In real-world situations, however, natural viewing behavior usually involves inspecting the same scene for several seconds. Once an object is recognized, humans can form the spatial structure of a scene from their memory to infer the likely location of other objects (see Hollingworth, 2012, for a review). Thus, using spatial dependency between different objects may be a more efficient way to continuously decrease uncertainty about uninspected scene objects and continuously update the current search strategy.

CONCEPTUAL SEMANTIC ASSOCIATIONS BETWEEN OBJECTS IN THE SCENE

As discussed above, many studies on the effect of semantic information on visual attention were based mainly on a single object-scene relation, which can be considered to be a simplistic approach. During natural scene inspection, semantic information may be continuously impacting observers' viewing strategy, integrated with either low-level stimulus features or task goals. Therefore, any conclusions from studies using a single object-scene relation (either semantic or syntactic) might underestimate the use of semantic information in attentional guidance. Hwang et al. (2011) investigated how conceptual semantic similarity among scene objects influences attention and eye movements in real-world scenes. They asked observers either to view a natural scene and memorize its content or to search for a pre-specified target in a scene. In their experiments, each scene was selected from Label Me, an object annotated image data base (Russell et al., 2008) in which scene images were manually segmented into annotated objects by volunteers. They applied Latent Semantic Analysis (referred to as LSA; Landauer and Dumais, 1997) to measure semantic similarity between objects. Since annotated objects in Label Me have descriptive text labels, their semantic similarity can be estimated by computing the vector representations of object labels in a semantic space. Semantic similarity is then calculated as the cosine of the vector angle between object pairs, with larger values indicating greater similarity. Hwang et al. (2011) used this method to generate a semantic saliency map for each scene based on the semantic similarity of objects to the currently fixated object in an inspection task or the search target in a search task (see **Figure 6**).

Hwang et al. (2011) found that, during scene inspection, observers tended to shift their gaze toward those objects that were semantically similar to the previously fixated one. **Figure 6** illustrates this tendency: when the currently fixated object was a dishwasher, the next fixation was more likely to land on a bowl than on a sink, because the LSA cosine value was greater for the labels "dishwasher" and "bowl" than for the labels "dishwasher" and "sink." Surprisingly, the use of semantic relevance between objects to guide visual attention, which was referred to as "semantic guidance," still existed for transitions with long saccades of amplitudes exceeding 10° of visual angle. The authors showed that while the visual similarity between objects (semantically similar objects may share similar visual features) and their spatial proximity (semantically similar objects tend to be located close to each other) did contribute to the observed semantic guidance, the effect of semantic guidance did not disappear when both factors were ruled out. This finding implies that the role of peripheral vision in scene viewing is not limited to perceiving gist. Peripheral vision may also help in object recognition. This interpretation is supported by the results obtained by Kotowicz et al. (2010), who found that in a simple conjunction search task, the target

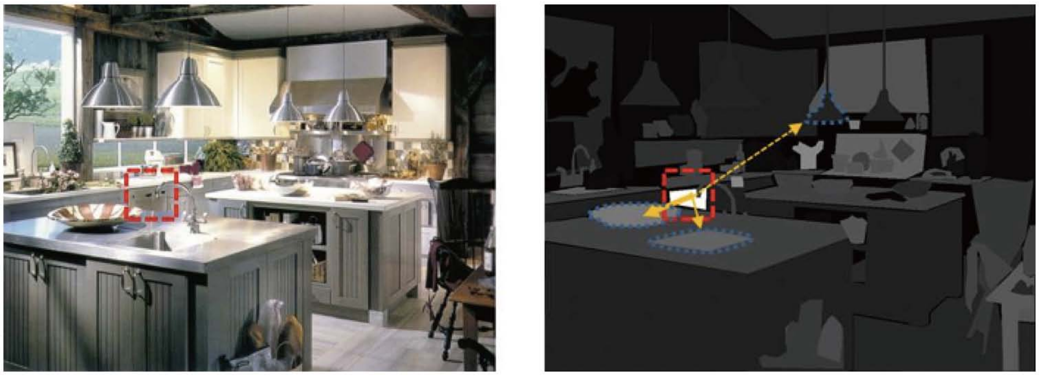

FIGURE 6 | An illustration of a semantic saliency map in an inspection task, as proposed by Hwang et al. (2011). Left: the original scene. Right: The semantic saliency map based on the currently fixated object labeled "dishwasher." The luminance of each object indicates how semantically similar it is to the dishwasher, as quantified by the corresponding LSA cosine value. The weight of the arrows indicates the likelihood of subsequent gaze transitions based on semantic guidance. The LSA cosine values for "bowl," "sink," and "hanging lamp" are 0.47, 0.39, and 0.14, respectively. Note that the values only indicate the relatively tendency for the subsequent gaze transition. They are not the probabilities.

was recognized before it was fixated upon. The function of the final saccade to the target was to simply increase the confidence of judgment. What is the function of semantic guidance? Observers may inspect semantically similar objects consecutively in order to quickly construct the concept for a given scene. For example, if the first few fixations are located on a pan, a stove, and a microwave oven, an observer may quickly develop the concept of a kitchen scene and also infer the likely appearance and location of other objects which are often found in a kitchen. The tendency of using conceptual semantic information may be an attempt to decrease memory load by grouping semantically similar objects so that the content of a scene can be encoded efficiently.

While performing a visual search task, participants in the Hwang et al. (2011) study tended to fixate on objects that were semantically similar to the verbally specified search target. This bias became more pronounced over the course of the search. It is possible that observers exploit semantic information in a scene for efficient search performance, leading to increased semantic similarity between fixated objects and the target as search progresses (Hwang et al., 2011). This finding is corroborated by Moores et al. (2003) and Belke et al. (2008), who found that, attention was attracted to an object which was semantically similar to the verbally specified target.

Interestingly, explicit assessment of semantic similarity between objects requires prior knowledge of their meanings. Consequently, attentional deployment would have to wait for the process of object recognition to completed, which seems to be an inefficient strategy of visual exploration. Thus, it is possible that the semantic guidance observed by Hwang et al. (2011) was facilitated by the use of observers' knowledge about scene gist that they obtained in early visual processing. That is, instead of considering the semantic relation between the currently fixated object and the objects located in the extrafoveal visual field, observers can use their knowledge about the scene type to decide where to look next. For example, if observers were aware that the image was a kitchen, they may only attend to the regions nearby the counter or sink, where most of the – semantically related – kitchenware is likely located. This strategy could be executed by using the scene gist perceived during the initial glance without assessing semantic associations between objects (Oliva and Torralba, 2001, 2006). In addition to 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 a semantically related object or the search target (Oliva and Torralba, 2007). For example, a fork may be expected to be next to a spoon on top of a table. Therefore, if there was frequent gaze shifting from a table to a chair in a natural viewing task, it is possible that the visual system used the semantic similarity as a cue to make this decision, leading to the observation of semantic guidance. On the other hand, knowing the meaning of the "chair" object may not be necessary for making this transition. It is possible that when the table was fixated, the identities of other objects near the table were highly constrained. That is, the presence of the table limits the probabilities of other objects to appear next to it; they are very likely to be chairs or other furniture that is typically located near a table. Consequently, the decision of fixating on the chair was not necessarily made as a result of identifying it beforehand. The visual system may instead have chosen to fixate on a location where the possible occurrence of objects was highly constrained and thus contained the least uncertainty. Fixating on a predictive location may help recognize the object faster than fixating on a location with greater uncertainty. Other studies found a similar strategy in visual search tasks in which visual attention can be reliably directed to the predictive location of the target embedded in real world scenes even when the target was absent, e.g., toward the top of a house when the target was a chimney (Hidalgo-Sotelo et al., 2005; Eckstein et al., 2006; Droll and Eckstein, 2008; Ehinger et al., 2009; Castelhano and Heaven, 2010). Altogether, both scene gist and the spatial dependency among scene objects could contribute to the observed effect of semantic guidance without the need to identify extrafoveal scene objects. Further studies are needed to clarify whether the semantic guidance effect is due to the actual evaluation of semantic relevance between objects.

CONCLUSIONS

It is well known that semantic information in natural scenes can influence attentional guidance. Research in this field has addressed a variety of different aspects of information retrieved or even inferred from the scene. These aspects are often generalized as a single concept such as semantic or contextual information from the scene. The current review attempted to disentangle the major semantic factors that guide visual attention. In summary, semantic information can be contributed from scene gist, scene-object relations, spatial associations between objects, or the semantic similarity between objects. Though most of these factors have been extensively investigated in the context of attention deployment, the issues listed below are still not well understood:

WHAT IS THE ROLE OF ATTENTION DURING THE INITIAL LEVEL OF SCENE PERCEPTION?

It is clear that attention is necessary when semantic information involves recognizing the meaning of an object. However, many prior studies have shown that some semantic information such as scene gist could be perceived within less than 100 ms, which is too fast for focal attention and saccadic planning. Therefore, instant pop-out scene perception may be achieved without attention. In contrast, other studies found that scene perception was impaired when attention was fully engaged to an unrelated, concurrent task. This suggests that perception of natural scenes indeed requires visual attention. Whether attention can be exempted at any level of natural scene awareness, or if it is even essential for coarse-level scene perception is still an ongoing debate.

WHAT IS THE ROLE OF OBJECT RECOGNITION WHEN PERCEIVING SCENE GIST AND OTHER SEMANTIC INFORMATION?

As discussed in Section "Contextual Information – The Gist of a Scene," previous studies have found that humans can perceive scene gist without recognizing any individual object. Nevertheless, this does not necessarily imply that object recognition has no impact on the perception of gist. To determine the category of a scene, recognizing a representative object may be more useful than evaluating some global properties of a scene or its spatial layout. For example, recognizing a bed in a scene is more informative than evaluating the coarse spatial layout of the visual information for inferring a bedroom scene. Moreover, object recognition is needed for accessing the scene-object, object–object, and conceptual semantic relations. That is, observers need to recognize the currently attended object to infer the locations and identities of other objects in the scene. However, it is currently unknown whether the object that is going to be fixated next is recognized even before a saccade is initiated. Although some studies found that observers used semantic similarity as a cue to guide their attention (Moores et al., 2003; Belke et al., 2008; Hwang et al., 2011), this does not imply that the identities of other objects have been confirmed before they were fixated. It is possible that the meaning of objects cannot be confirmed before they are fixated, but the likelihood of their occurrence in a given location can be inferred based on the information retrieved from the currently fixated object. This strategy may induce conceptual semantic guidance without knowing the identity or meaning of any objects before attending to them. Whether observers adopt this strategy instead of evaluating the meaning of objects located in the periphery needs to be further investigated.

HOW DOES THE USE OF SEMANTIC INFORMATION CHANGE OVER TIME WHEN VIEWING A NATURAL SCENE, AND HOW DOES IT INTERACT WITH THE TASK GOAL?

Previous literature has demonstrated that we perceive scene gist faster than other semantic information or the identity of objects. In spite of this, the time course of perceiving different pieces of semantic information does not have to align with the order in which these types first affect attention deployment. In other words, perceiving scene gist first does not necessarily indicate that attention is influenced by scene gist earlier than by other semantic information. Since scene gist only provides some coarse information about a scene such as its category, it is likely that the visual system may not deploy attention to other locations until the first object is recognized in order to avoid shifting gaze too early. In addition, the use of different aspects of semantic information must be sensitive to the goal of observers' behavior. For a given scene, the use of different types of semantic information can be prioritized based on different task goals. How the task goal influences perception and use of semantic information is still not well understood.

Note that we are not claiming these different aspects are isolated independent processes. In fact, they may be tightly coupled. For example, the functions of scene gist and spatial dependency among objects not only help in understanding the content of a scene but also facilitate the process of object recognition. By accessing scene gist and spatial dependency among objects, people are able to infer the existence and location of other objects in a scene without directly fixating them. Therefore, when attention is guided to a new target, it may be difficult to determine whether this decision is made because the target has been recognized, or because the visual system was trying to decrease target uncertainty by using its knowledge about scene gist or spatial associations among objects.

The current state of attention models often considers each property extracted from a scene – such as faces, texts, or object luminance – as an isolated factor that can attract attention by itself. This approach seems incomplete, as it does not take into account the semantic associations between objects, arguably the main factor allowing human search performance in real-world scenes to still surpass modern computer vision approaches. Perhaps the relation between scene perception and different aspects of semantics can be described by the famous Gestalt notion: the whole is greater than the sum of its parts. Each component in a scene may contribute a different piece of semantic information, and these pieces can be treated as different variables that affect attention deployment. Nevertheless, the relations between these pieces may convey different types of semantic information (e.g., semantic consistency) which can be regarded as the interactions between different variables. In order to understand the whole perception of a scene and improve the search algorithms in current computer vision systems, knowing each part of its semantics and their associations is indispensable.

ACKNOWLEDGMENTS

The preparation of this article was supported by Grant R01EY021802 from the National Eye Institute to Marc Pomplun. The authors would like to thank Hsueh-Cheng Wang and Lucia Saura for their comments on earlier versions of this text.

REFERENCES

Bar, M. (2004). Visual objects in context. *Nat. Rev. Neurosci.* 5, 617–629. doi: 10.1038/nrn1476

Belke, E., Humphreys, G. W., Watson, D. G., Meyer, A. S., and Telling, A. L. (2008). Top-down effects of semantic knowledge in visual search are modulated by cognitive but not perceptual load. *Percept. Psychophys.* 70, 1444–1458. doi: 10.3758/PP.70.8.1444

Beutter, B. R., Eckstein, M. P., and Stone, L. S. (2003). Saccadic and perceptual performance in visual search tasks. I. Contrast detection and discrimination. *J. Optic. Soc. Am. A* 20, 1341–1355. doi: 10.1364/JOSAA.20.001341

Biederman, I., Mezzanotte, R. J., and Rabinowitz, J. C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cogn. Psychol.* 14, 143–177. doi: 10.1016/0010-0285(82)90007-X

Bonitz, V. S., and Gordon, R. D. (2008). Attention to smoking-related and incongruous objects during scene viewing. *Acta Psychol.* 129, 255–263. doi: 10.1016/j.actpsy.2008.08.006

Borji, A. (2012). "Boosting bottom-up and top-down visual features for saliency estimation," in *Conference on Computer Vision and Pattern Recognition (CVPR), 2012 IEEE (IEEE)*, (Providence, RI), 438–445.

Borji, A., and Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Analysis Mach. Intell.* 35, 185–207. doi: 10.1109/TPAMI.2012.89

Boucart, M., Moroni, C., Thibaut, M., Szaffarczyk, S., and Greene, M. (2013). Scene categorization at large visual eccentricities. *Vision Res.* 86, 35–42. doi: 10.1016/j.visres.2013.04.006

Boyce, S. J., and Pollatsek, A. (1992a). "An exploration of the effects of scene context on object identification," in *Eye Movements and Visual Cognition: Scene Perception and Reading*, ed. K. Rayner (New York: Springer Verlag), 227–242. doi: 10.1007/978-1-4612-2852-3_13

Boyce, S. J., and Pollatsek, A. (1992b). Identification of objects in scenes: the role of scene background in object naming. *J. Exp. Psychol. Learn. Mem. Cogn.* 18, 531–543. doi: 10.1037/0278-7393.18.3.531

Brady, T. F., and Chun, M. M. (2007). Spatial constraints on learning in visual search: modeling contextual cuing. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 798–815. doi: 10.1037/0096-1523.33.4.798

Brockmole, J. R., Castelhano, M. S., and Henderson, J. M. (2006). Contextual cuing in naturalistic scenes: global and local contexts. *J. Exp. Psychol. Learn. Mem. Cogn.* 32, 699–706. doi: 10.1037/0278-7393.32.4.699

Brockmole, J. R., and Henderson, J. M. (2006a). Recognition and attention guidance during contextual cueing in real-world scenes: evidence from eye movements. *Q. J. Exp. Psychol.* 59, 1177–1187. doi: 10.1080/17470210600665996

Brockmole, J. R., and Henderson, J. M. (2006b). Using real-world scenes as contextual cues for search. *Visual Cogn.* 13, 99–108. doi: 10.1080/13506280500165188

Brockmole, J. R., and Henderson, J. M. (2008). Prioritizing new objects for eye fixation in real-world scenes: effects of object-scene consistency. *Visual Cogn.* 16, 375–390. doi: 10.1080/13506280701453623

Brooks, D. I., Rasmussen, I. P., and Hollingworth, A. (2010). The nesting of search contexts within natural scenes: evidence from contextual cuing. *J. Exp. Psychol. Hum. Percept. Perform.* 36, 1406–1418. doi: 10.1037/a0019257

Bruce, N. D., and Tsotsos, J. K. (2009). Salience, attention and visual search: an information theoretic approach. *J. Vis.* 9, 1–24. doi: 10.1167/9.3.5

Buswell, G. T. (1935). *How People Look at Pictures: A Study of the Psychology of Perception in Art.* Chicago: University of Chicago Press.

Castelhano, M. S., and Heaven, C. (2010). The relative contribution of scene context and target features to visual search in scenes. *Attent. Percept. Psychophys.* 72, 1283–1297. doi: 10.3758/APP.72.5.1283

Castelhano, M. S., and Heaven, C. (2011). Scene context influences without scene gist: eye movements guided by spatial associations in visual search. *Psychon. Bull. Rev.* 18, 890–896. doi: 10.3758/s13423-011-0107-8

Carmi, R., and Itti, L. (2006). Visual causes versus correlates of attentional selection in dynamicscenes. *Vision Res.* 46, 4333–4345. doi: 10.1016/j.visres.2006.08.019

Cerf, M., Frady, E. P., and Koch, C. (2009). Faces and text attract gaze independent of the task: experimental data and computer model. *J. Vis.* 9, 1–15. doi: 10.1167/9.12.10

Cerf, M., Harel, J., Einhäuser, W., and Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. *Adv. Neural Inf. Process. Syst.* 20, 1–7.

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends Cogn. Sci.* 4, 170–178. doi: 10.1016/S1364-6613(00)01476-5

Chun, M. M., and Jiang, Y. (1998). Contextual cueing: implicit learning and memory of visual context guides spatial attention. *Cogn. Psychol.* 36, 28–71. doi: 10.1006/cogp.1998.0681

Chun, M. M., and Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychol. Sci.* 10, 360–365. doi: 10.1111/1467-9280.00168

Chun, M. M., and Jiang, Y. (2003). Implicit, long-term spatial contextual memory. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 224–234. doi: 10.1037/0278-7393.29.2.224

Cohen, M. A., Alvarez, G. A., and Nakayama, K. (2011). Natural-scene perception requires attention. *Psychol. Sci.* 22, 1165–1172. doi: 10.1177/0956797611419168

Davenport, J. L. (2007). Consistency effects between objects in scenes. *Mem. Cogn.* 35, 393–401. doi: 10.3758/BF03193280

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

De Graef, P. (1992). "Scene-context effects and models of real-world perception," in *Eye Movements and Visual Cognition: Scene Perception and Reading*, ed. K. Rayner (New York: Springer-Verlag), 243–259. doi: 10.1007/978-1-4612-2852-3_14

De Graef, P., Christiaens, D., and d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychol. Res.* 52, 317–329. doi: 10.1007/BF00868064

Desimone, R., and Duncan J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18:193–222. doi: 10.1146/annurev.ne.18.030195.001205

Droll, J. A., and Eckstein, M. P. (2008). Expected object position of two hundred fifty observers predicts first fixations of seventy seven separate observers during search. *J. Vis.* 8:320. doi: 10.1167/8.6.320

Eckstein, M. P., Drescher, B. A., and Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. *Psychol. Sci.* 17, 973–980. doi: 10.1111/j.1467-9280.2006.01815.x

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., and Oliva, A. (2009). Modeling search for people in 900 scenes: a combined source model of eye guidance. *Vis. Cogn.* 17, 945–978. doi: 10.1080/13506280902834720

Elazary, L., and Itti, L. (2008). Interesting objects are visually salient. *J. Vis.* 8:3, 1–15. doi: 10.1167/8.3.3

Fabre-Thorpe, M. (2011). The characteristics and limits of rapid visual categorization. *Front. Psychol.* 2:243. doi: 10.3389/fpsyg.2011.00243

Findlay, J. M. (1997). Saccade target selection during visual search. *Vision Res.* 37, 617–631. doi: 10.1016/S0042-6989(96)00218-0

Fletcher-Watson, S., Findlay, J. M., Leekam, S. R., and Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception* 37, 571–583. doi: 10.1068/p5705

Friedman, A. (1979). Framing pictures: the role of knowledge in automatized encoding and memory for gist. *J. Exp. Psychol. Gen.* 108, 316–355. doi: 10.1037/0096-3445.108.3.316

Friedman, A., and Liebelt, L. S. (1981). "On the time course of viewing pictures with a view towards remembering," in *Eye Movements: Cognition and Visual Perception*, eds D. F. Fischer, R. A. Monty, and J. W. Senders (Hillsdale: Erlbaum), 137–155.

Gordon, R. (2004). Attentional allocation during the perception of scenes. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 760–777. doi: 10.1037/0096-1523.30.4.760

Gordon, R. (2006). Selective attention during scene perception: evidence from negative priming. *Mem. Cognit.* 34, 1484–1494. doi: 10.3758/BF03195913

Hayhoe, M. M., Shrivastava, A., Mruczek, R., and Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *J. Vis.* 3:49–63. doi: 10.1167/3.1.6

Henderson, J. M. (1992). Object identification in context: the visual processing of natural scenes. *Can. J. Psychol.* 46, 319–341. doi: 10.1037/h0084325

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends Cogn. Sci.* 7, 498–504. doi: 10.1016/j.tics.2003.09.006

Henderson, J. M., and Ferreira, F. (2004). "Scene perception for psycholinguists," in *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*, eds J. M. Henderson and F. Ferreira (New York, NY: Psychology Press), 1–58.

Henderson, J. M., and Hollingworth, A. (1998). "Eye movements during scene viewing: an overview. eye guidance," in *Reading and Scene Perception*, ed. G. Underwood (Amsterdam), 269–293. doi: 10.1016/B978-008043361-5/50013-4

Henderson, J. M., and Hollingworth, A. (1999). High level scene perception. *Annu. Rev. Psychol.* 50, 243–271. doi: 10.1146/annurev.psych.50.1.243

Henderson, J. M., Malcolm, G. L., and Schandl, C. (2009). Searching in the dark: cognitive relevance drives attention in real-world scenes. *Psychon. Bull. Rev.* 16, 850–856. doi: 10.3758/PBR.16.5.850

Henderson, J. M., Weeks, P. A. Jr., and Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *J. Exp. Psychol. Hum. Percept. Perform.* 25, 210. doi: 10.1037/0096-1523.25.1.210

Hidalgo-Sotelo, B., Oliva, A., and Torralba, A. (2005). "Human learning of contextual priors for object search: where does the time go?," in *Computer Vision and Pattern Recognition Workshop* (Los Alamitos, CA: IEEE Computer Society), 86.

Hollingworth, A. (2012). "Guidance of visual search by memory and knowledge," in *The Influence of Attention, Learning, and Motivation on Visual Search, Nebraska Symposium on Motivation*, eds M. D. Dodd and J. H. Flowers (New York: Springer), 63–89. doi: 10.1007/978-1-4614-4794-8_4

Hollingworth, A., and Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *J. Exp. Psychol. Gen.* 127, 398–415. doi: 10.1037/0096-3445.127.4.398

Hollingworth, A., and Henderson, J. M. (1999). Object identification is isolated from scene semantic constraint: evidence from object type and token discrimination. *Acta Psychol.* 102, 319–343. doi: 10.1016/S0001-6918(98)00053-5

Hwang, A. D., Wang, H.-C., and Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Res.* 51, 1192–1205. doi: 10.1016/j.visres.2011.03.010

Itti, L., and Koch, C. (2001). Computational modelling of visual attention. *Nat. Rev. Neurosci.* 2, 194–203

Jiang, Y., and Wagner, L. C. (2004). What is learned in spatial contextual cuing? Configuration or individual locations? *Percept. Psychophys.* 66, 454–463. doi: 10.3758/BF03194893

Johnston, W. A., Hawley, K. J., Plewe, S. H., Elliott, J. M. G., and DeWitt, M. J. (1990). Attention capture by novel stimuli. *J. Exp. Psychol. Gen.* 119, 397–411. doi: 10.1037/0096-3445.119.4.397

Joubert, O., Rousselet, G., Fize, D., and Fabre-Thorpe, M. (2007). Processing scene context: fast categorization and object interference. *Vision Res.* 47, 3286–3297. doi: 10.1016/j.visres.2007.09.013

Jovancevic, J., Sullivan, B., and Hayhoe, M. (2006). Control of attention and gaze in complex environments. *J. Vis.* 6. doi: 10.1167/6.12.9

Kanan, C., Tong, M. H., Zhang, L., and Cottrell, G. W. (2009). SUN: top-down saliency using natural statistics. *Vis. Cogn.* 17, 979–1003. doi: 10.1080/13506280902771138

Kelley, T. A., Chun, M. M., and Chua, K. P. (2003). Effects of scene inversion on change detection of targets matched for visual salience. *J. Vis.* 2, 1–5. doi: 10.1167/3.1.1

Kihara, K., and Takeda, Y. (2012). Attention-free integration of spatial frequency-based information in natural scenes. *Vision Res.* 65, 38–44. doi: 10.1016/j.visres.2012.06.008

Kirchner, H., Bacon, N., and Thorpe, S. J. (2003). In which of two scenes is the animal? Ultra-rapid visual processing demonstrated with saccadic eye movements. *Perception*, 32(Suppl.), 170. doi: 10.1068/v031005

Kirchner, H., and Thorpe, S J. (2006). Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. *Vision Res.* 46, 1762–1776. doi: 10.1016/j.visres.2005.10.002

Koch, C., and Tsuchiya, N. (2007). Attention and consciousness: two distinct brain processes. *Trends Cogn. Sci.* 11, 16–22. doi: 10.1016/j.tics.2006.10.012

Koch, C., and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* 4, 219–227.

Kollmorgen, S., Nortmann, N., Schröder, S., and König, P. (2010). Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention. *PLoS Comput. Biol.* 6:1–20. doi: 10.1371/journal.pcbi.1000791

Kotowicz, A., Rutishauser, U., and Koch, C. (2010). Time course of target recognition in visual search. *Front. Hum. Neurosci.* 4:31. doi: 10.3389/fnhum.2010.00031

Kowler, E., Anderson, E., Dosher, B., and Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Res.* 35, 1897–1916. doi: 10.1016/0042-6989(94)00279-U

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the latentsemantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211–240. doi: 10.1037/0033-295X.104.2.211

Larson, A. M., and Loschky, L. C. (2009). The contributions of central versus peripheral vision to scene gist recognition. *J. Vis.* 9, 1–16.

Li, F. F., VanRullen, R., Koch, C., and Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proc. Natl. Acad. Sci. U.S.A.* 99, 9596–9601. doi: 10.1073/pnas.092277599

Loftus, G. R., and MacKworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *J. Exp. Psychol. Hum. Percept. Perform.* 4, 565–572. doi: 10.1037/0096-1523.4.4.565

Mack, S. C., and Eckstein, M. P. (2011). Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *J. Vis.* 11, 1–16. doi: 10.1167/11.9.9

Malcolm, G. L., and Henderson, J. M. (2010). Combining top-down processes to guide eye movements during real-world scene search. *J. Vis.* 10, 1–11. doi: 10.1167/10.2.4

Marks, W., McFalls, E. L., and Hopkinson, P. (1992). Encoding pictures in scene context: does task demand influence effects of encoding congruity? *J. Exp. Psychol. Learn. Mem. Cogn.* 18, 192–198. doi: 10.1037/0278-7393.18.1.192

Moores, E., Laiti, L., and Chelazzi, L. (2003). Associative knowledge controls deployment of visual selective attention. *Nat. Neurosci.* 6, 182–189. doi: 10.1038/nn996

Neider, M. B., and Zelinsky, G. J. (2006). Scene context guides eye movements during search. *Vision Res.* 46, 614–621. doi: 10.1016/j.visres.2005.08.025

Nyström, M., and Holmqvist, K. (2008). Semantic override of low-level features in image viewing – Both initially and overall. *J. Eye Movement Res.* 2, 1–11.

Oliva, A. (2005). "Gist of the scene," in *Neurobiology of Attention*, eds L. Itti, G. Rees, and J. K. Tsotsos (San Diego, CA: Elsevier), 251–256.

Oliva, A. (2009). "Visual scene perception," in *Encyclopedia of Perception*, ed. E. Goldstein (Thousand Oaks, CA: SAGE Publications, Inc.), 1112–1117. doi: 10.4135/9781412972000.n355

Oliva, A., and Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vision* 42, 145–175. doi: 10.1023/A:1011139631724

Oliva, A., and Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Progr. Brain Res. Visual Percept.* 155, 23–36. doi: 10.1016/S0079-6123(06)55002-2

Oliva, A., and Torralba, A. (2007). The role of context in object recognition. *Trends Cogn. Sci.* 11, 520–527. doi: 10.1016/j.tics.2007.09.009

Olson, I. R., and Chun, M. M. (2002). Perceptual constraints on implicit learning of spatial context. *Visual Cogn.* 9, 273–302. doi: 10.1080/13506280042000162

Peterson, M. S., and Kramer, A. F. (2001). Attentional guidance of the eyes by contextual information and abrupt onsets. *Percept. Psychophys.* 63, 1239–1249. doi: 10.3758/BF03194537

Potter, M. C. (1976). Short-term conceptual memory for pictures. *J. Exp. Psychol. Hum. Learn. Mem.* 2, 509–522. doi: 10.1037/0278-7393.2.5.509

Potter, M. C., and Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *J. Exp. Psychol.* 81, 10–15. doi: 10.1037/h0027470

Pringle, H. L., Irwin, D. E., Kramer, A. F., and Atchley, P. (2001). The role of attentional breadth in perceptual change detection. *Psychon. Bull. Rev.* 8, 89–95. doi: 10.3758/BF03196143

Rayner, K., and Pollatsek, A. (1992). Eye movements and scene perception. *Can. J. Psychol.* 46, 342–376. doi: 10.1037/h0084328

Rensink, R. A., O'Regan, J. K., and Clark, J. J. (1997). To see or not to see: the need for attention to perceive changes in scenes. *Psychol. Sci.* 8, 268–373. doi: 10.1111/j.1467-9280.1997.tb00427.x

Rensink, R. A., O'Regan, J. K., and Clark, J. J. (2000). On the failure to detect changes in scenes across brief interruptions. *Visual Cogn.* 7, 127–145. doi: 10.1080/135062800394720

Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. (2008). LabelMe: a database and web-based tool for image annotation. *Int. J. Comput. Vis.* 77, 157–173. doi: 10.1007/s11263-007-0090-8

Schyns, P. G., and Oliva, A. (1994). From blobs to boundary edges: evidence for time-and spatial-scale-dependent scene recognition. *Psychol. Sci.* 5, 195–200. doi: 10.1111/j.1467-9280.1994.tb00500.x

Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6424–6429. doi: 10.1073/pnas.0700622104

Spotorno, S., and Faure, S. (2011). Change detection in complex scenes: hemispheric contribution and the role of perceptual and semantic factors. *Perception* 40, 5–22. doi: 10.1068/p6524

Spotorno, S., Tatler, B. W., and Faure, S. (2013). Semantic consistency versus perceptual salience in visual scenes: findings from change detection. *Acta Psychol.* 142, 168–176. doi: 10.1016/j.actpsy.2012.12.009

Stirk, J. A., and Underwood, G. (2007). Low-level visual saliency does not predict change detection in natural scenes. *J. Vis.* 7, 1–10. doi: 10.1167/7.10.3

Tatler, B. W. (2009). Current understanding of eye guidance. *Visual Cogn.* 17, 777–789. doi: 10.1080/13506280902869213

Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M., and Velichkovsky, B. M. (2010). Yarbus, eye movements, and vision. *i-Perception* 1, 7–27.

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

Torralba, A., Oliva, A., Castelhano, M., and Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychol. Rev.* 113, 766–786. doi: 10.1037/0033-295X.113.4.766

Underwood, G., Humpherys, L., and Cross, E. (2007). "Congruency, saliency and gist in the inspection of objects in natural scenes," in *Eye Movements: A Window on Mind and Brain Elsevier Science Ltd*, eds R. P. G. Van Gompel, M. H. Fischer, W. S. Murray, and R. L. Hill (Oxford), 561–577.

Viola, P., and Jones, M. J. (2001). "Robust object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Kauai, HI), 511–518.

Võ, M. H., and Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *J. Vision* 9, 1–15. doi: 10.1167/9.3.24

Võ, M. L. H., and Henderson, J. M. (2011). Object–scene inconsistencies do not capture gaze: evidence from the flash-preview moving-window paradigm. *Atten. Percept. Psychophys.* 73, 1742–1753. doi: 10.3758/s13414-011-0150-6

Võ, M. L. H., and Wolfe, J. M. (2013). Different electrophysiological signatures of semantic and syntactic scene processing. *Psychol. Sci.* 24, 1816–1823. doi: 10.1177/0956797613476955

Wang, H.-C., and Pomplun, M. (2012). The attraction of visual attention to texts in real-world scenes. *J. Vis.* 12, 1–17. doi: 10.1167/12.6.26

Yarbus, A. L. (1967). *Eye Movements and Vision.* New York: Plenum Press. doi: 10.1007/978-1-4899-5379-7

Zhao, Q., and Koch, C. (2012). Learning visual saliency by combining feature maps in a nonlinear manner using Ada Boost. *J. Vis.* 12, 1–15. doi: 10.1167/12.6.22

Zoest, W. V., Donk, M., and Theeuwes, J. (2004). The role of stimulusdrivenand goal-driven control in saccadic visual attention. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 746–759.doi: 10.1037/0096-1523.30.4.749

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 22 September 2013; paper pending published: 24 December 2013; accepted: 16 January 2014; published online: 06 February 2014.

Citation: Wu C-C, Wick FA and Pomplun M (2014) Guidance of visual attention by semantic information in real-world scenes. Front. Psychol. ***5****:54. doi: 10.3389/fpsyg.2014.00054*

This article was submitted to Perception Science, a section of the journal Frontiers in Psychology.