The Role of Scene Gist and Spatial Dependency among Objects in the Semantic Guidance of Attention

Chia-Chien Wu (chiachie@cs.umb.edu)
Hsueh-Cheng Wang (hchengwang@gmail.com)
Marc Pomplun (marc@cs.umb.edu)
Department of Computer Science, University of Massachusetts at Boston
100 Morrissey Boulevard, Boston, MA, 02125-3393, USA

Abstract

A previous study (Hwang et al., 2011) 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, these results are not necessarily indicative of semantic processing of individual objects but may be confounded by knowledge of the scene gist, which does not require object recognition (Torralba et al., 2006), or by known spatial dependency among objects (Oliva & Torralba, 2007). To examine the mechanisms underlying semantic guidance, in the present study, subjects were asked to view a series of displays with the scene gist removed and spatial dependency varied. Our results confirm the previous finding of semantic guidance and show that it is not entirely due to either the effect of scene gist or the spatial dependency among objects. Even without scene gist or spatial dependency, subjects still retrieved semantic information to guide their attention. This strategy may facilitate scene understanding and object memorization.

Keywords: Attention, semantics, eye movements, visual guidance, real-world scenes.

Introduction

Real-world scenes contain rich information, which usually is not thoroughly processed during natural viewing. Therefore, the way in which the visual system deploys the limited attention resources is crucial for effective vision and has drawn huge interest over the last two decades. The guidance of attention based on the features of stimuli in the visual environment has been well investigated in both its bottom-up (Itti & Koch, 2001; Koch & Ullman, 1985) and top-down aspects (Hayhoe et al, 2003; Hwang, Higgins & Pomplun, 2009; Pomplun, 2006).

Visual attention is not only affected by factors based on the overt visual appearance, but also by inherent factors, such as meaning and semantic relations among objects. Hwang, Wang and Pomplun (2011) found that during natural scene viewing, humans tend to bring their gaze to the objects that are semantically similar either to the currently fixated one or to the specified search target. This result, however, may have been confounded by the observers' knowledge of the global scene context. That is, instead of considering the semantic relation between the currently fixated object and the objects located in the extrafoveal visual field, observers may simply use their knowledge about the scene type to decide where to look next. For example, if observers are aware that the viewed image is a kitchen, they may only attend the regions nearby the counter or sink, where most of the kitchenware is likely located.

The ways in which people acquire such global contextual information is not well understood. Torralba, Oliva, Castelhano and Henderson (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.

Even when the global context, which usually comes from visual background information, is missing, it is still possible to learn some context of the scene. Chun (2000) showed that some contextual information can be learned merely by the typical arrangement of elements and affect the deployment of attention. Oliva and Torralba (2007) also found that spatial dependency among objects could provide different contextual information about a scene. 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 the scene gist and the spatial dependency among scene objects may have caused a bias in observers' gaze patterns that could explain the results of Hwang et al. (2011) without the need for semantic analysis of extrafoveal scene objects. If gist, object dependency, or both were entirely responsible for the effect observed in that study, the concept of semantic guidance would not be a new phenomenon but rather a bias introduced by already known factors. The aim of the present study was to discern the contributions of scene gist, object dependency, and semantic object analysis to semantic guidance in order to address this problem.

To study the influence of spatial dependency among scene objects, we employed the LabelMe object annotated image data base (Russell, Torralba, Murphy, & Freeman, 2008) in which scene images were manually segmented into annotated objects by volunteers. In addition, the locations of objects are provided as coordinates of polygon corner and all objects are labeled with English words or phrases. It provides an excellent opportunity for not only segregating each object from its scene, but also shifting the object's coordinates to any desired location in the image.

One way to eliminate potential influence of scene gist on attentional guidance is to remove all background information and only keep segregated objects in the scene. We used the resulting images in an experimental condition

referred to as 'fixed condition'. This procedure effectively removes the relation between scene and objects as defined by Torralba et al. (2006). For example, it is easier to predict where a plate is located in a scene when the plate is shown on a dining table than when it is shown by itself. The spatial dependency among objects, however, is still retained when the background information is excluded. For instance, it is possible to predict the likely location of a glass merely based on the location of a seen plate in a scene, even when no context is provided. To remove the spatial dependency among objects as well, we created another set of stimuli ('scrambled condition') by generating scenes without background as in the fixed condition and then randomly shifting the objects within the scene.

If the semantic guidance found in the previous study (Hwang et al., 2011) were due to the spatial dependency among objects, this effect should be eliminated once the background information and spatial arrangement are removed. On the other hand, if observers are able to use conceptual semantic information between objects to guide attention, their gaze transitions should still show an above-chance semantic relevance.

Method

Subjects

Ten subjects, aged between 19-40 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.

Apparatus

Eye movements were tracked and recorded using an SR Research EyeLink-2k system. Its sampling frequency was set to 1000 Hz. Stimuli were presented on a 22-inch ViewSonic LCD monitor. Its refresh rate was set to 75 Hz and its resolution was set to 1024 x 768 pixels. Participant responses were entered using a keyboard.

Stimulus display

A total of 60 images (1024 x 768 pixels) were generated. Each image was composed of 13 to 15 objects selected from a real-world scene from the LabelMe database (http://labelme.csail.mit.edu). The selected scenes included home interiors, landscapes and city scenes. Objects of extreme size (small or large) 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 Figure 1 for an example).

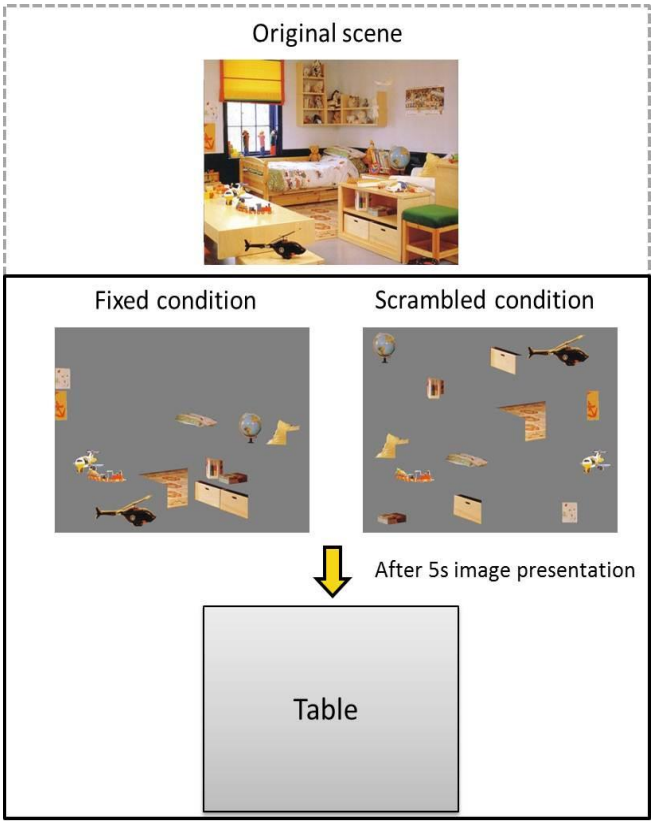

Figure 1: Original scene (top) and a sample trial (bottom). The upper panel shows the original scene used to generate stimulus displays. The scene would be used to generate an image with objects at same coordinates (fixed condition) and an image with objects at randomly selected locations (scrambled condition). During each trial, the created image was presented for 5 seconds. After the stimulus image disappeared, a word was presented and subjects had to report whether the indicated object had been shown in the previous display.

Procedure

Subjects were instructed to inspect the scenes and memorize them for the subsequent object recall test (see Figure 1, bottom panel). Each image was presented for 5 seconds. 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 pressing one of two possible keys on a keyboard. If they believed the indicated object was shown in the previous image, they would press the left arrow key. Otherwise, they would press the right arrow key. 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.

Data Analysis

Assigning fixations to objects

Since all images excluded the global contextual information by only leaving the selected objects on a grey canvas, some fixations may land on the blank area rather than on any object in the image. When this happened, we assumed this fixation was aimed at the nearest object, i.e., the one whose center had the shortest Euclidean distance to the current fixation location.

Latent Semantic Analysis

Similar to the original semantic guidance study (Hwang et al., 2011), we used Latent Semantic Analysis (referred to as LSA; Landauer & Dumais, 1997) 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. The basic premise in LSA is that the aggregate contexts in which a word does or does not appear provide a set of mutual constraints to deduce the word's meaning (Landauer, Foltz, & Laham, 1998). The greater the cosine value, the higher is the semantic similarity. Since annotated objects in LabelMe have descriptive text labels, their semantic similarity can be estimated by calculating cosine values for the labels of object pairs.

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 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, a form of factor analysis called Singular Value Decomposition (SVD; see Berry, Dumais, & Obrien, 1995) is applied to reduce the matrix to a lower-dimensional vector space called 'semantic space'. 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 1 shows examples of LSA cosine values for various object labels used in the LabelMe scene image ''Child4'' (see Figure 1) in terms of the reference object label ''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 ''AIRPLANE'' and ''PILLOW''. One of the nice features of LSA is that it can quantify higher-level conceptual semantic similarity, regardless of any geometrical relation, functional relation or visual relation.

Table 1: *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
-	-	-

To compute semantic similarity for each pair of object labels in our experiment, 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.

Measuring semantic guidance

In this study, the semantic guidance effect was defined as the extent to which the semantic relation/similarity between the currently fixated object and the other objects in the scene influences the choice of the next fixated object. In order to compute this effect quantitatively, the computation had to follow 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 landscape 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 Figure 2. The semantic landscapes, excluding the area occupied by the currently fixated object, were normalized so that the sum of all activation was one. With the normalized semantic landscape, the Receiver Operating Characteristic (ROC) value was computed in a similar way as it was done in previous studies (Hwang et al., 2009; Tatler, Baddeley & Gilchrist, 2005). Overall, each fixation would build its own semantic landscape as a predictor of the target point of the next transition. All ROC values computed along scan paths were averaged across scenes to obtain the extent of semantic guidance during the inspection of a scene. If eye movements were exclusively guided by semantic information, this average ROC value should be close to one. If there were no semantic effect on eye movements at all, the average ROC value should be close to 0.5, indicating prediction at chance level.

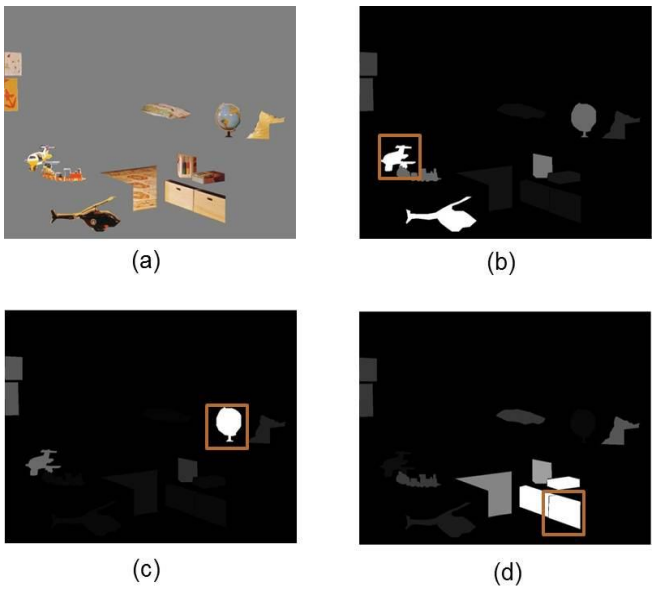

Figure 2: Example of semantic landscapes. The currently fixated object is marked with an orange square. (a) The original image that subjects inspected. (b) Semantic landscape during gaze fixation on the object labeled as "AIRPLANE". (c) Semantic landscape during gaze fixation on the object labeled as "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 sematic similarity to the currently fixated object - receive higher activation (brightness), for example, the helicopter in (a) shows a higher activation due to the fixated object labeled as 'AIRPLANE'.

Excluding potential confounds by computing control analyses

Following Hwang et al. (2011), to control for possible confounds in the measurements of semantic guidance, subjects' ROC values computed from their empirical gaze transition data were compared with two control data sets: (1) random fixations and (2) dissociated fixations. The random fixations were generated by replacing subjects' fixation positions with randomly positioned coordinates in the scene. This data set served as an unbiased test of ROC values. That is, since gaze transitions of the random data set were not affected by any other factor, we should always receive a chance level ROC value (ROC = 0.5).

Furthermore, it is likely that any above-chance ROC value was simply caused by the proximity effect. This effect is due to the previous finding (Hwang et al., 2011) that semantically similar objects tend to be located closer to each other and subjects' saccades tend to be shorter than gaze transitions in the random data set. To examine this possible confound, subjects' data were also compared with a "dissociated" data set. The dissociated data were analyzed using the eye movement data recorded in scene n against object data from scene n+1, and the eye movement data recorded from the last scene against the object data from the first scene. This mismatch conserved the spatial distribution of both the scene objects and the observers' fixations and therefore the proximity effect (at least in the fixed condition in which the coordinates of selected objects were not changed). This method allowed us to examine whether any observed above chance level ROC value for the empirical data was simply caused by proximity, which would be indicated by ROC values in the dissociated case being similar to the actual ROC values).

Experimental Results

Results showed that subjects recall performances were above chance level in both the fixed and scrambled conditions (Recall performance in the fixed condition, 79%, $t(9) = 21.50$, $p < 0.05$; recall performance in the scrambled condition: 70 %, $t(9) = 4.36$, $p < 0.05$).

As mentioned earlier, in order to examine semantic guidance, we computed ROC values for the two experimental conditions (fixed vs. scrambled) for all three data sets (empirical, random and dissociated). Figure 3 shows that the transitional semantic guidance values of random fixations were close to 0.5 in both the fixed and scrambled conditions. This result shows that the ROC computation was applied properly and the normalized semantic landscapes used in our analysis were unbiased.

The ROC value in the fixed condition (ROC = 0.704 ± 0.14) was significantly higher than that in the scrambled condition (ROC = 0.65 ± 0.19), $t(9) = 4.76$, $p < 0.05$. '±' here indicates a mean value and its standard error.

This result suggests that the spatial dependency preserved in the fixed condition provided additional semantic information and facilitated semantic guidance. The ROC value decreased when this spatial dependency was eliminated by shuffling the locations of objects.

Interestingly, in the scrambled condition, where the spatial dependency among objects was destroyed, the ROC value of empirical transition between distinct objects was still substantially greater than both ROC values for the other two control cases. A one-way ANOVA showed that the effect was significant, $F(2,27) = 51.61$, $p < 0.05$. A post hoc Tukey test indicated that there was no difference between the dissociated and random cases, $p = 0.53$. This finding shows that the proximity effect, at least in our experiment, had no impact on semantic guidance.

Overall, the results indicate that, even without scene gist and the spatial dependency among objects, subjects were still able to extract the semantic relevance between objects to guide their attention.

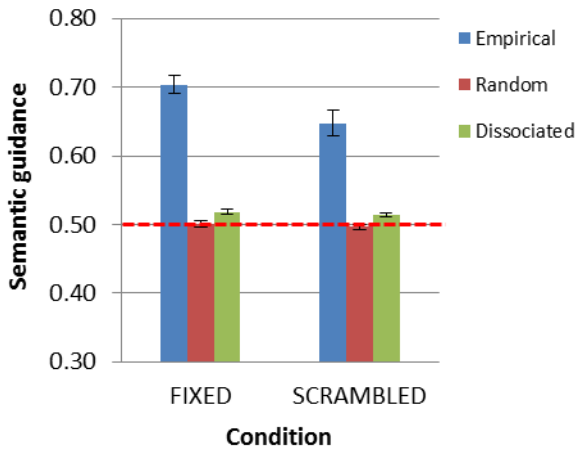

Figure 3: Transitional semantic guidance as measured by the ROC method in the fixed condition and the scrambled condition. The red dashed line represents the chance level (ROC = 0.5) and errors represent +/-1 standard error of the mean.

Conclusions

Hwang et al. (2011) found that, during scene inspection, observers tend to bring the line of sight to objects that are semantically relevant to the currently fixated object. Based on these previous data alone, it cannot be ruled out that the high semantic relevance of gaze transitions was contributed by scene gist information or by subjects' prediction of local scene context based on the spatial layout of objects. In other words, observers may not actually evaluate the semantics of peripheral objects for saccade target selection, and consequently, semantic guidance could not be considered a new phenomenon but rather an effect caused by other known mechanisms.

Our present results clarify the influence of these possible confounds in the previous findings. In the fixed condition in which the scene gist was removed, observers still showed strong semantic guidance. This result demonstrates that semantic guidance of visual attention in scene inspection is not entirely due to the scene gist. In fact, semantic guidance in the present study was even higher than that measured in Hwang et al. (2011), suggesting that scene gist only plays a marginal, if any, role in semantic guidance.

In the scrambled condition, in which both scene gist and possible spatial dependency among objects were removed, the effect of semantic guidance was slightly decreased but remained substantially higher than chance level. This finding shows that the spatial arrangement of objects only makes a small contribution to semantic guidance. Moreover, these data reveal that even when the scene gist was excluded and the spatial dependency was removed, subjects could still retrieve semantic information to guide their attention.

Moreover, Hwang et al. (2011) also found an even greater effect of semantic guidance in a visual search task. That is, observers tend to fixate on the objects which are semantically similar to the specified target. Instead of using any verbal probe and search paradigm as they did, the present study used a natural viewing and memory task which was less constrained by cognitive goal and we still found a substantial effect of semantic guidance.

Consequently, the question becomes how observers obtained this semantic information and how it influenced the guidance of attention. It is likely that extrafoveal visual processing may play a crucial role in enabling the semantic effect since observers had to recognize, at least partially, the objects in peripheral vision and processed the semantic relevance in the context of the currently fixated object. Kotowicz, Rutishauser and Koch (2010) found that, during visual search, observers already identified the extrafoveal target before fixating on it. We do not claim that in our task, observers were able to recognize the objects in the extrafoveal field. At the very least, extrafoveal perception may be used to increase the belief of what this object could be. Therefore, when contextual information was removed, people could still learn semantic information to help them determine where to fixate next by using the immediately acquired information from the current fixation and the information accumulated from extrafoveal vision. Such a strategy may facilitate scene understanding and memorization. It is also possible that, instead of using extrafoveal information, observers may construct their own scene representation merely based on the currently fixated object and update it during later fixations. This strategy may become useful when the extrafoveal information is not available or when the cost of processing it is too high.

Overall, the current study showed that semantic guidance of visual attention during the inspection of real-world scenes, as reported by Hwang et al. (2011), is a novel phenomenon that cannot be explained by effects of scene gist (e.g., Torralba et al., 2006) or spatial dependency among scene objects (e.g., Oliva & Torralba, 2007) alone. It has been known that, in addition to the visual saliency from low-level features, attention could be also driven by other particular classes of objects, such as faces (Judd, Ehinger, Durand, & Torralba, 2009) and texts (Wang & Pomplun, 2012). Our result provides a new alternative class of features and suggests that the conceptual semantic effects may need to be considered in the present model of attentional guidance. Further research on semantic guidance, its underlying mechanisms, and its function is necessary before this concept can be integrated into existing models of visual attention.

Acknowledgments

This research was supported by Grant Number R01 EY021802 from NIH 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.

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

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

Hwang, A. D., Higgins, E. C., & Pomplun, M. (2009). A model of top-down attentional control during visual search in complex scenes. *J Vis, 9*(5), 25.1--2518.

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

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res, 40*(10-12), 1489-1506.

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

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look, *IEEE International Conference on Computer Vision (ICCV)*, Kyoto, Japan, 2106 - 2113.

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

Kotowicz, A., Rutishauser, U., & Koch, C. (2010). Time course of target recognition in visual search. *Frontiers in Human Neuroscience, 4*, 12.

Landauer, T. K., Laham, D., & Derr, M. (2004). From paragraph to graph: latent semantic analysis for information visualization. *Proc Natl Acad Sci U S A, 101 Suppl 1*, 5214-5219.

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; 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.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Prog Brain Res, 155*, 23-36.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends Cogn Sci, 11*(12), 520-527.

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

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.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Res, 45*(5), 643-659.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network, 14*(3), 391-412.

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. *Psychol Rev, 113*(4), 766-786.

Wang, H.C. & Pomplun, M. (2012). The attraction of visual attention to texts in real-world scenes. *Journal of Vision*, 12(6):26, 1–17,