Semantic guidance of eye movements during real-world scene inspection

Alex D. Hwang (ahwang@cs.umb.edu)
Hsueh-Cheng Wang (hcwang@cs.umb.edu)
Marc Pomplun (marc@cs.umb.edu)
Department of Computer Science, University of Massachusetts Boston
100 Morrissey Blvd., Boston, MA 02125-3393, USA

Abstract

This is the first study to measure semantic guidance during scene inspection, based on the efforts by two other research groups, namely the development of the LabelMe object-annotated image database and the LSA@CU text/word latent semantic analysis tool, which computes the conceptual distance between two terms. Our analysis reveals the existence of semantic guidance during scene inspection, that is, eye movements during scene inspection being guided by a semantic factor reflecting the conceptual relation between the currently fixated object and the target object of the following saccade. This guidance may facilitate memorization of the scene for later recall by viewing semantically related objects consecutively.

Keywords: semantic guidance; contextual guidance; eye tracking; eye movement, scene inspection.

Introduction

Real-world scenes are filled with objects representing not only visual information, but also meanings and semantic relations with other objects in the scene. The guidance of eye movements based on visual appearance (low-level visual features) has been well studied in terms of both bottom-up (e.g., Bruce & Tsotsos, 2006; Henderson, 2003; Itti & Koch, 2001; Parkhurst, Law & Niebur, 2002) and top-down control of visual attention (e.g., Henderson, Brockmole, Castelhano & Mack, 2007; Hwang, Higgins & Pomplun, 2009; Peters & Itti, 2007; Pomplun, 2006; Zelinsky, 2008; Zelinsky, Zhang, Yu, Chen & Samaras, 2006) as well as neurological aspects (e.g., Corbetta & Shulman, 2002; Egner, Monti, Trittschuh, Wienecke, Hirsch & Mesulam, 2008).

Although there has been research on high-level contextual effects on visual search using global features (e.g., Neider & Zelinski, 2006; Torralba, Oliva, Castelhano & Henderson, 2006) and primitive semantic effects based on co-occurrence of objects in term of implicit learning (e.g., Chun & Jiang, 1998; Chun & Phelps, 1999; Manginelli & Pollmann, 2008), effects on eye movements by object meaning and object relations, *Semantic guidance*, have not been studied because of a few hurdles that make such study more complicated: (1) Object segmentation is difficult, (2) semantic relations among objects are hard to define, and (3) a quantitative measure of semantic guidance has to be developed.

Automated segmentation of images and labeling is one of the crucial steps for further understanding of image context, and there have been numerous attempts to solve this problem, ranging from global classification of scenes to individual region labeling (Athanasiadis, Mylonas, Avrithis, & Kollias, (2007); Boutell, Luo, Shena & Brown, 2004; Le Saux, & Amato, 2004; Luo & Savakis, 2001), but results were rather disappointing compared to human performance. Thanks to the LabelMe object-annotated image database (Russell, Torralba, Murphy & Freeman, 2008) developed by the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), various scenes with annotated objects are available to the public, which helps to bypass the first hurdle.

In order to convincingly compute semantic or conceptual relations between objects, the co-occurrence of objects has to be analyzed in a large number of scenes images. Unfortunately, collecting and analyzing a sufficient amount of annotated scenes is unfeasible with the currently available data sources. Since semantic relations are formed at the conceptual rather than at the visual level, relations do not have to be derived from image databases. Any database that can generate a collection of contexts or knowledge can be used to represent the semantic meaning of objects.

A useful mathematical method for such representation for computer modeling and simulation is Latent Semantic Analysis (LSA), which is based on the analysis of representative corpora of natural text. It transforms the occurrence matrix from large corpora into a relation between the terms/concepts, and a relation between those concepts and the documents (Landauer, Foltz & Laham, 1998). Since annotated data in LabelMe are text descriptions of objects, their semantic or conceptual relation can be processed with LSA. In this study, the LSA@CU text/word latent semantic analysis tool is used to pass the second hurdle.

Equipped with above tools, we computed a series of semantic salience maps for each labeled object in a subject's visual scan path. These salience maps approximated the transition probabilities for the following saccade to the other labeled objects in the scene, assuming that eye movements were entirely guided by the semantic relations between objects. Under this assumption, the probability of a gaze transition between two objects is proportionate to the strength of their semantic relation. Subsequently, the amount of semantic guidance was measured by the Receiver Operator Characteristic (ROC), which computes the extent

to which the actual gaze transitions followed the ideal semantic salience map.

Method

Participants

Ten participants performed this experiment. All were students at the University of Massachusetts Boston, aged between 19 to 40 years old. Each was entitled to a $10 honorarium.

Apparatus

Eye movements were tracked and recorded using an SR Research EyeLink II system. After calibration, the average error of visual angle in this system is 0.5˚. Its sampling frequency is 500 Hz. Stimuli were presented on a 19-inch Dell P992 monitor. Its refresh rate was set to 85 Hz and its resolution was set to 1280×1024 pixels. Participant responses were entered using a game-pad.

Materials

A total of 200 photographs (1024×768 pixels) of real-world scenes, including landscapes, home interiors, and city scenes, were selected from the LabelMe database (http://labelme.csail.mit.edu/) as stimuli (see Figure 1). Objects in each scene were annotated with polygon coordinates which define the outline of the object shape, and they were labeled with representative English words.

When displayed on the screen, the photographs covered 16˚×12˚ of visual angle. Each scene contained an average of 53.03±38.14 labeled objects (in the present work, '±' always indicates a mean value and its standard deviation), and the median value for the number of objects in each image was 40. On average, labeled objects covered 92.88±10.52% of the scene area.

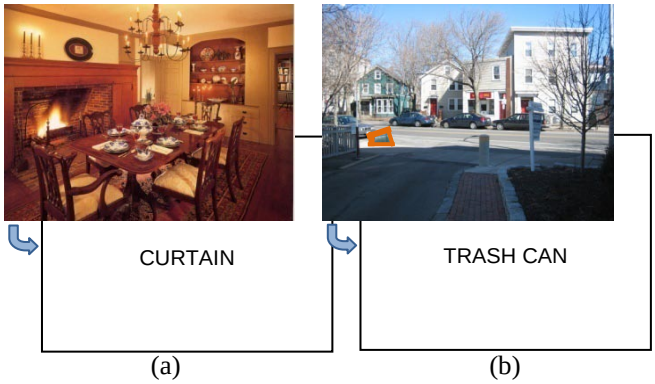

Figure 1: Sample trials. Scene is displayed for five seconds, followed by target object name for two seconds. (a) Target object absent case; (b) Target object present case (target object is marked for illustrative purpose).

Procedure

After five practice trials, subjects viewed 200 randomly ordered scenes. Subjects were instructed to inspect the scenes and memorize them for subsequent object recall tests. After the five-second presentation of each scene, an English word was shown and subjects were asked whether the object indicated by the word had been present in the previously viewed scene. If subjects believed that the object in question had been in that scene, they had to press a button on the game pad within two seconds. If they were unable to make the decision within that period, the trial would time out and the next trial would begin (see Figure 1). Target object present cases and target object absent cases were evenly distributed among the 200 trials.

Data Analysis

Object Z-Order

As noted in the Materials section, scenes were selected from the LabelMe annotated image database. However, there is one problem that prevents using the annotated data directly for the analysis of the eye-tracking data. The LabelMe annotation tool allows users to submit the object boundary as they assume it to be, regardless of whether it is partially or fully overlapped by other objects. As a result, a gaze fixation at the intersection of two or more objects regions could be assigned to multiple objects. In other words, the object membership of the intersected area is ambiguous.

In order to solve this membership ambiguity problem, the *'Z-order'* among intersecting objects must be resolved. Two comparison methods, one based on the number of characteristic polygons in the intersection area and the other based on visual feature similarity were used to solve this problem, mainly following the suggestions by Russell et al (2008). The logic of the first method is the simple fact that visible objects that are on top will contribute more characteristic polygons defining the intersection area than the other intersecting objects. The logic of the second method is that the object whose non-intersecting part is visually most similar to the intersection area will take the on-top position.

Since the first method is computationally inexpensive, it is applied first. If the first method cannot make a definitive decision above some certainty threshold, the second method is applied, which involves more complex visual feature comparison between intersection and object regions using the Histogram Intersection Similarity Method (Swan & Ballard, 1991) over pixel intensity.

After applying the z-order computation to all object pairs in the scene, we can derive their final z-order. For each object, this value represents the number of other objects on top of it. Therefore, whenever an eye fixation falls onto a location where multiple annotated objects intersect, the lowest z-ordered object is assigned to the eye fixation (see Figure 2).

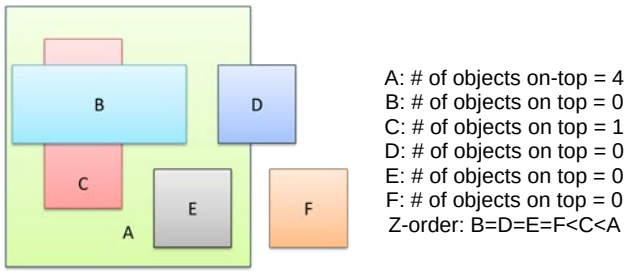

Figure 2: Sample of z-ordered objects. All objects in the scene have characteristic polygons with four corners. The Z-order among object A and other objects except object F is resolved by the number of characteristic polygons in the intersection area. Z-order between object B and object C is resolved by visual feature similarity.

Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the contextual usage-meaning of words by statistical computations applied to a large corpus of text (Landauer & Dumais, 1997). The basic premise in LSA is that the aggregate contexts in which a word does and does not appear provide a set of mutual constraints to induce the word's meaning.

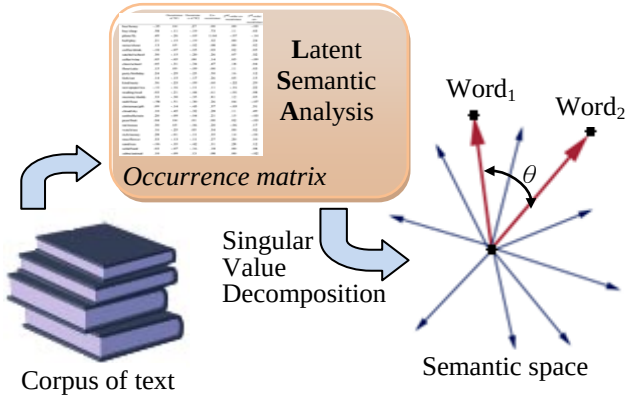

Figure 3: Semantic similarity between two words using LSA tools. Cosine of the angle θ between the vectors representing two words, $Word_1$ and $Word_2$, in semantic space is the semantic similarity between two words.

In a nutshell, LSA similarity computation can be described as follows: At first, a word occurrence matrix is constructed from the large corpus of text, where each row stands for a unique word, each column stands for a text passage or context and each cell contains the frequency with which the word appeared in the context.

Next, each cell frequency is normalized by its information-theoretic measure, in other words, by its importance in the context, which is computed by the usual Shannon entropy, -p log p over all entries in its row.

Then a form of factor analysis called Singular Value Decomposition (SVD) is applied to reduce this huge-dimensional matrix to a smaller-dimensional vector space called '*semantic space*'. Since this is a vector space generated from individual word appearance frequency, it has the nice property that even if two words have never co-occurred in the same document it can easily estimate their latent semantic relationship (Jones & Mewhort, 2007; Landauer, et al, 1997).

Every term, every text, and every novel combination of terms has a high-dimensional vector representation in the space. So the semantic similarities between two terms and two or more novel combinations of terms can be calculated by the cosine value of the angle between two vectors representing them in semantic space. The greater the cosine value, the closer is the semantic relationship of the objects.

Table 1 shows examples of LSA cosine values for various objects' labels used in scene image Dining20 in terms of the object label "FORK". The semantic similarity between "FORK" and "TABLE TOP" is higher than the semantic similarity between "FORK" and "SHELVES". There can be lots of reasons why conceptual similarity between fork and table top is higher than similarity between fork and shelves, for example, because forks are usually put on table tops but rarely on shelves, or because forks are typically used for eating food on table tops rather than on shelves. However, the important part is that LSA can quantify higher-level conceptual semantic similarity, regardless of whether it came from geometrical proximity, functional similarity or even shape similarity.

Table 1: Sample LSA cosine values.

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

In this study, to compute semantic similarity among object labels, a web-based LSA tool, LSA@CU (http://lsa.colorado.edu), developed by the University of Colorado at Boulder was used. This tool was set to create a semantic space from the general readings up to 1st year college with 300 factors. The system will compute a similarity score between 0 and 1 for each submitted text compared to all other submitted texts for the same image. For the annotated scenes used in our experiment, the average score is 0.245±0.061.

Semantic Guidance

In this study, the semantic guidance effect is 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 consecutively inspected object.

In order to compute this effect quantitatively, the computation has to follow each subject's eye movements. Since we are interested in the effect of semantic similarity on gaze transitions, i.e., the selection of the next object to inspect, only eye movements that transition 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 4. The semantic landscapes, excluding the area occupied by the currently observed 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, submitted; Tatler, Baddeley & Gilchrist, 2005) for the semantic landscape as a predictor of the target point of the 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.

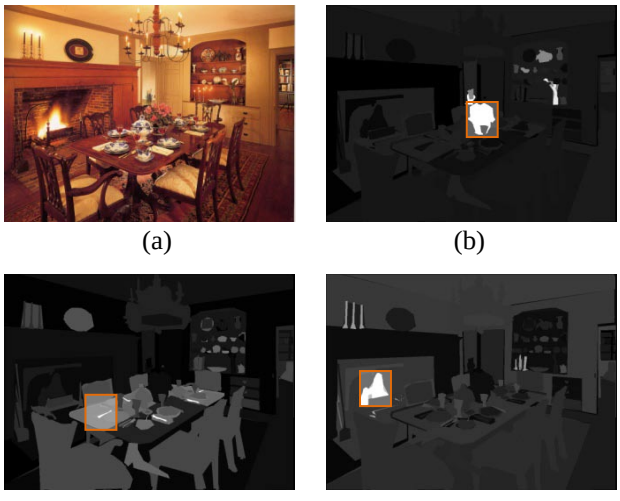

Figure 4: Examples of semantic landscapes. Currently fixated object is marked with orange square. (a) Original scene image (Dining20). (b) Semantic landscape during gaze fixation on an object labeled as "PLANT IN POT". (c) Semantic landscape when eye is currently fixated on an object labeled as "FORK". (d) Semantic landscape when eye is currently fixated on an object labeled as "FLAME". As it can be seen from above, conceptually closer objects receive higher activation, for example, candle sticks in (d) are activated by the object label "FLAME".

Experimental Results

As described in the previous section, semantic guidance effects in terms of the ROC measure were computed for all subjects' empirical data for all scenes, considering only transitions between distinct objects during scene inspection. In order to get controlled comparative results, hypothetical eye movements that always choose a random object regardless of the current object's semantic relation with other objects in the scene were used to simulate the random guidance case. For each simulated trial, the number of fixations during the inspection period was kept identical to the empirical data for fair comparison.

As it is shown in Figure 5, the semantic guidance of the actual, empirical transitions between distinct objects is clearly greater (0.649±0.080) than that of random simulation, which is approximately at chance level (0.508±0.073), and the difference between them is statistically significant, $t(199) = 29.676$, $p < 0.001$.

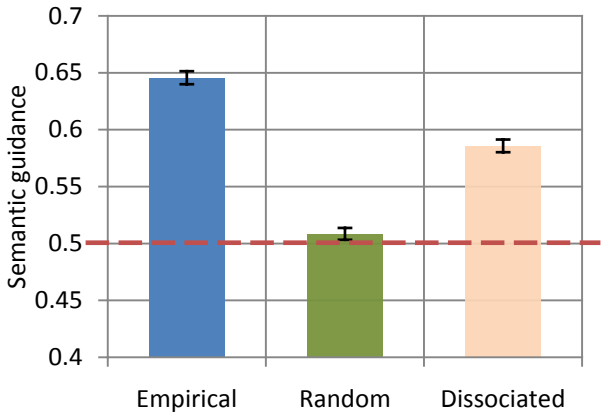

Figure 5: Semantic guidance as measured by the ROC method. Dotted line indicates chance level. The average guidance of empirical human data is 0.649±0.080 and it is significantly higher, all ts(199) > 13.099, ps < 0.001, than for the hypothetical random case (0.508±0.073), in which consecutive objects for inspection are chosen randomly, and the hypothetically dissociated eye fixation data-scene pair case (0.576±0.065). The difference between the empirical and dissociated cases confirms the existence of semantic guidance.

Is this difference possibly created by a proximity effect, since semantically closer items may tend to be located more closely together in the visual scenes than other objects, and saccadic eye movements tend to be shorter than transitions between randomly chosen points? To test for such a confound, an additional hypothetical case is generated by dissociating eye movement data and scene images. For example, for scene 1, we analyze the eye movements made in scene 2, for scene 2, we analyze the eye movements made in scene 3, and so on. This manipulation preserves the saccade amplitude statistics, i.e., objects closer to each other

will still receive more transitions than others, but it eliminates semantic guidance entirely. If the resulting ROC value is close to 0.5, we can conclude that it is entirely the semantics that guide eye movements. If, on the other hand, the value is closer to the empirical semantic guidance value of 0.65, then we can conclude that semantic guidance does not play a role in guiding eye movements, and what we have measured is purely a proximity effect. Since the result of dissociated-pair simulation (0.576±0.065) is significantly lower than that for the empirical data, $t(199) = 13.099$, $p < 0.001$, and is also significantly bigger than the random transition case, $t(199) = 12.618$, $p < 0.001$, we can conclude that although there is a significant proximity effect during scene inspection, semantic guidance also plays a significant role independently of proximity.

Conclusions

Unlike other contextual guidance models that focus on defining the context of the scene from its gist and create a model of eye fixation distribution or study the context in terms of implicit learning by measuring co-occurrence or contextual cueing of objects, our study focused on how this context is constructed by following the observers' eye movements.

Based on empirical data and a linguistic measure on objects labels, we have shown that there is a significant semantic effect that guides eye movement during scene viewing. In other words, our eyes tend to move to objects that are conceptually similar/close to the currently inspected object.

The function of such a guidance mechanism could be to optimize the encoding of scenes for later object recall. Object information may be encoded more efficiently if semantically related objects are inspected and memorized in close temporal succession.

From the current results it cannot be concluded whether this scan path planning is happening at the beginning of the scene viewing by analyzing the scene gist or it is incrementally constructed during the inspection period (similar to finding the optimal path using a "greedy" algorithm). Nevertheless, the present study has introduced a new interdisciplinary approach combining visual context research and linguistic research and pointed out a new way of looking at the semantics of the visual world.

Acknowledgments

This research was supported by Grant Number R15EY017988 from the National Eye Institute to M.P.

References

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

Boutell, M., Luo, J., Shena, X. & Brown, C., (2004). *Learning multilabel scene classification,* Pattern Recognition, vol. 37, no. 9, pp. 1757–1771

Chun, M., M. & Jiang, Y. (1998). *Contextual cueing: Implicit learning and memory of visual context guides spatial attention*. Cognitive Psychology, 36, 28-71

Chun, M. M. & Phelps, E. A. (1999). *Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage*. Nature Neuroscience. 2 (9):775-6.

Corbetta, M. & Shulman, G. L., (2002). *Control of goal-directed and stimulus-driven attention in the brain*, Nature Review 3, 3, 201-215.

Dennis, S., Landauer, T., Kintsch, W. & Quesada, J. (2003). *Introduction to Latent Semantic Analysis: LSA@CU Boulder (http://lsa.colorado.edu/)*. Slides from the tutorial given at the 25th Annual Meeting of the Cognitive Science Society, Boston.

Egner, T., Monti, J.M.P., Trittschuh, E.H., Wienecke, C.A., Hirsch, J., & Mesulam, M. M. (2008). *Neural integration of top-down spatial and feature-based information in visual search.* The Journal of Neuroscience, 28, 6141-6151.

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

Henderson, J. M., & Ferreira, F. (2004). *Scene perception for psycholinguists*. In J. M. Henderson and F. Ferreira (Eds.), The interface of language, vision, and action: Eye movements and the visual world (pp. 1-58). New York: Psychology Press.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. L. (2007). *Visual saliency does not account for eye movements during visual search in real-world scenes*. In R. van Gompel, M. Fischer, W. Murray, & R. W. Hill (Eds.), Eye movements: A window on mind and brain (pp. 537–562). Amsterdam: Elsevier.

Hwang, A. D., Higgins, E. C. & Pomplun, M. (under review). *A model of top-down attentional control during visual search in complex scenes.* Under review for Journal of Vision.

Irwin, D. E., (1991). *Information integration across saccadic eye movement*. Cognitive Psychology, 23, 420-458.

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

Jones, M. N. & Mewhort, D. J. (2007). *Representing Word Meaning and Order Information in a Composite Holographic Lexicon.* Psychological Review, Vol. 114, No. 1. (January 2007), pp. 1-37.

Landauer, T. K., & Dumais, S. T. (1997). *A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.* Psychological Review, *104*, 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction *to Latent Semantic Analysis*. Discourse Processes, 25, 259-284.

Le Saux, B. & Amato, G. (2004). *Image classifiers for scene analysis*, Proc. in Int. Conf. of Computer Vision & Graphics (ICCVG), Warsaw, Poland

Luo, J & Savakis, A. (2001). *Indoor versus outdoor classification of consumer photographs using low-level and semantic features*, Proc. IEEE Int. Conf. of Image Process (ICIP01), vol. 2, pp. 745–748.

Manginelli A. A. & Pollmann, S. (2008). *Misleading contextual cues: How do they affect visual search?* Psychology Research, 10.1007/s00426-008-0211-1

Neider, M. B. & Zelinski, G. J. (2006) *Scene context guides eye movements during visual search*, Vision Research, 46, 614-621.

Palmer, S. E. (1999). *Photons to phenomenology*, The MIT press.

Parkhurst, D. J., Law, K., & Niebur, E. (2002). *Modeling the role of salience in the allocation of overt visual selective attention*. Vision Research, 42, 107–123.

Peters, R. J. & Itti, L. (2007). *Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention.* Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2007), 1--8.

Pomplun, M. (2006). *Saccadic selectivity in complex visual search displays.* Vision Research, 46, 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, volume 77, issue1-3, 157-173.

Swain, M. J. & Ballard, D. H. (1991). *Color Indexing*, Journal of Computer Vision, 7(1), 11-32.

Tatler, B., Baddeley, R. & Gilchrist, l. (2005). *Visual correlates of fixation selection: effects of scale and time.* Vision Research 45, 643-659.

Athanasiadis, T., Mylonas, P, Avrithis, Y. & Kollias, S. (2007). *Semantic Image Segmentation and Object labeling*, IEEE Transaction and circuits and systems for video technology, vol. 17, No. 3

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J.M. (2006). *Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search*. Psychological Review, 113, 766-786.

Zelinsky, G. J., Zhang, W., Yu, B., Chen, X., & Samaras, D. (2006). *The role of top-down and bottom-up processes in guiding eye movements during visual search*. In Y. Weiss, B. Scholkopf, & J. Platt (Eds.), Advances in neural information processing systems (Vol. 18, pp. 1569–1576). Cambridge, MA: MIT Press.

Zelinsky, G. J. (2008). *A theory of eye movements during target acquisition.* Psychological Review, 115 (4), 787–835.