“Eyes Closed” and “Eyes Open” Expectations Guide Fixations in Real-World Search

Tom Foulsham (foulsham@essex.ac.uk)
Department of Psychology, University of Essex
Wivenhoe Park, Colchester, CO4 3SQ, UK

Abstract

Investigations of search within realistic scenes have identified both bottom-up and top-down influences on performance. Here, I describe two types of top-down expectations that might guide observers looking for objects. Initially, likely locations can be predicted based only on the target identity but without any visual information from the scene (“Eyes closed”). When a visual preview becomes available, a more refined prediction can be made based on scene layout (“Eyes open”). In two experiments participants guessed the location of a target with or without a brief preview of the scene. Responses were consistent between observers and were used to predict the eye movements of new observers in a third experiment. The results confirm that participants use both types of top-down cues during search, and provide a simple method for estimating these expectations in predictive models.

Keywords: scene perception; visual search; attention; eye movements

Introduction

Imagine walking into a friend’s kitchen to look for a coffee mug when you have never been there before. Even before you open the door you would already have a significant amount of knowledge about where this object might be. For example, you would not expect it to be on the floor or near the ceiling, so you would be unlikely to look in these locations. When entering the room, the first glance tells you that, in this particular kitchen, there is a large window on one side of the room and shelving with cupboards on the opposite side. You refine your expectations about where the object will be and subsequently recognize the mug on a shelf.

As this scenario reveals, searching for something in the real world involves not just the matching of visual information to a stored template but also the use of detailed semantic knowledge about scenes and objects. Although visual search has been very well studied in cognitive psychology, this has mostly been in the context of simple search displays and models that predict performance based on target features (e.g., Wolfe, 1998). More recently, there has been significant interest in exploring the mechanisms involved in directing attention during search within natural scenes, and in testing these mechanisms by measuring eye fixations.

In one approach, computational models of bottom-up visual salience have been proposed that select targets based on the degree to which they stand out from their background (Itti & Koch, 2000). By this account, participants will attend to regions of high salience until they find what they are looking for. However, it is clear from several experiments that, when searchers know what they are looking for, this knowledge, and not simple visual salience, dominates the locations inspected during search (Chen & Zelinsky, 2006; Foulsham & Underwood, 2007).

This has led to more realistic models that combine top-down knowledge of the searcher to prioritize those locations in a scene in which an object is likely to appear. One way to do this is to compare scene locations with a representation of target *appearance*. If one knows that the target is red, locations with this colour should be more likely to be fixated. This principle underlies several models of search guidance (Wolfe, 1994; Navalpakkam & Itti, 2005), and can be successful at predicting fixations in real scenes (Zelinsky, 2008; Kanan et al., 2009).

However, it is clear from the scenario at the beginning of this paper that searchers also have access to detailed expectations about target *location*. There is evidence from several different experiments that these expectations are used to direct attention during search. For example, scrambling an image—so that local visual features remain the same but their configuration is altered—impedes search and alters eye movements (Biederman et al., 1973; Foulsham, Alan & Kingstone, 2011). Objects that are incongruent with their context or out of place may be found more slowly (Henderson, Weeks & Hollingworth, 1999). The contextual guidance model proposed by Torralba et al. (2006) accounts for these effects by combining bottom-up salience with a Bayesian prior for where an object is likely to be, conditioned on the global features of an image. In essence, the model recognizes the gist and layout of a scene (e.g., finding street level in an urban environment), learns the target’s likely location within this representation, and searches accordingly (e.g., by looking for people at street level).

The top-down guidance by context discussed so far emerges early, with the first glance of a scene, but requires visual input in the form of low spatial-frequency features and “gist”. On the other hand, it is likely that the semantic information associated with different objects might include general expectations about position within a scene-centered or person-centered frame-of-reference which could be activated *before* exposure to the to-be-searched scene. The present paper investigates whether these expectations are reliable and whether they effect the distribution of attention in real-world search. If so, they could be incorporated into probabilistic models (e.g., Kanan et al., 2009; Torralba et al., 2006).

I will distinguish between “Eyes closed” expectations, which can be made prior to any perception of the scene, and “Eyes open” expectations, which are affected by a rapid perception of scene gist, as might be available during the first fixation on a scene. I describe two simple experiments to quantify “Eyes closed” and “Eyes open” predictions, and these predictions are then compared to the eye fixations made by independent searchers. If contextual guidance of attention occurs only in response to scene features then ”Eyes closed” expectations will not be a good description of where people look during search.

Experiments 1 and 2

In Experiments 1 and 2, participants guessed where a target object would be located based on very little information.

Method

Participants Eighteen student volunteers (12 females) took part in return for course credit. All participants took part in Experiment 1 first, followed by Experiment 2. The mean age was 19.4 years.

Stimuli and Apparatus The stimuli for all experiments were derived from 72 colour photographs of indoor and outdoor scenes collected from the Internet. Scenes were chosen which contained a single example of an easily nameable target object that was not located directly in the centre of the image. The name of this object was the matching target label for the scene.

Each target label was also matched to another scene from the set in which it could plausibly be found. This led to 144 label-scene pairs, half of which were “target-present” trials, where the scene contained the target, and half of which were “target-absent”. The same target labels were used in both experiments.

Target labels were presented in large black font centred above a grey rectangle representing the scene. In Experiment 2, scene images were presented at a resolution of 1024 x 768 pixels. Stimuli were presented on a 19-inch CRT monitor with a refresh rate of 60Hz. Presentation was controlled by PsychoPy (Pierce, 2007), and responses were entered with the mouse.

Procedure Figure 1 depicts the procedure. In Experiment 1, participants were instructed to “make their best guess” where a target was located in an image. The experiment began with a practice example of a target and scene. In the experimental trials, a target label was presented alongside a grey rectangle representing the image frame, and participants were instructed to click with a mouse cursor where in the frame they thought the target was located. In order to motivate participants, feedback was given after every 12 trials in the form of a percentage score representing how close their mouse clicks had been to the actual target locations.

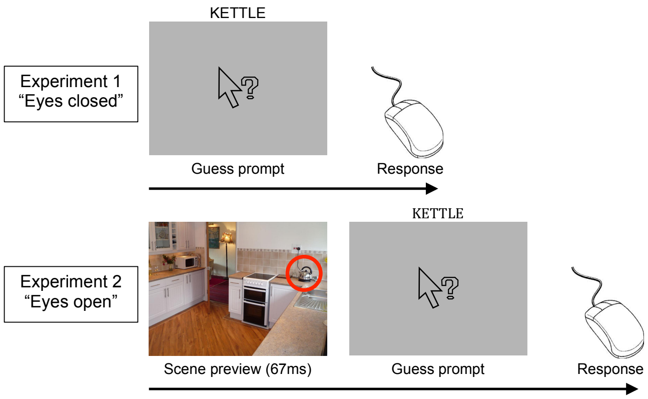

Figure 1: The procedure for one trial in Experiment 1 (top) and Experiment 2 (bottom). The target is highlighted in the scene preview, for display purposes only.

Scores were calculated from the average Euclidian distance between the chosen location and the centre of the target object, in target present trials only, and normalized by the scene diagonal. All 72 target labels were presented in a random order, but the actual scenes were not shown to participants.

In Experiment 2, participants were given a brief preview of the scene in which the target was located before they made their response. In each trial, a text prompt told participants to get ready, and a fixation cross was then presented in the centre of the screen for 1s. The scene was then presented briefly for 67ms, followed by a target label presented alongside a grey rectangle representing the image frame. The brief preview was chosen because it is known that scene gist can be perceived very quickly (Biederman et al., 1973), and also to limit the possibility that targets would be attended during the preview. A pattern mask was not included, and so after-images may have persisted, although the guess prompt had the effect of partly masking the display and drawing attention away from the scene. As in Experiment 1, participants were instructed to guess the location of the target with a mouse click, and feedback was given regarding average performance over the previous 12 trials. In Experiment 2, all 144 label-scene pairs were presented in a random order.

Analysis and Results

The results were analysed in order to estimate the inter-observer agreement, i.e., the degree to which different participants “guessed” in similar locations for each target label. The approach used closely followed that in previous studies of fixations in real-world search (Torralba et al., 2006; Ehinger et al., 2009). Participant-selected locations were first combined to produce a spatial model of target predictions—an “expectation map”—which was then used to predict search behaviour in Experiment 3.

Expectation maps were formed by representing each participant’s guessed location as a 2-dimensional Gaussian and summing across all participants. The highest points on this map indicate locations that, according to the participants, are most likely to contain the target. The dispersion of the map will reflect between-participant agreement: maps that cluster into a few small areas signify that participants agreed on where a target was likely to appear. Maps were computed separately for each target label in Experiment 1, and for target-present and target-absent trials in Experiment 2.

The receiver operating characteristic (ROC) curve was used to evaluate expectation maps. The ROC curve is a non-parametric measure of sensitivity originating from signal detection theory. This measure has become common in machine learning, and also in studies of spatial attention and eye movements, as it allows spatial distributions (e.g., salience maps) to be compared to specific locations (e.g., eye fixations). Full coverage of this method can be found elsewhere (e.g., Ehinger et al., 2009). The area under the curve (AUC) was computed as a summary statistic. AUC values indicate the probability that the map will rank a selected location more highly than a non-selected location and range from 0 to 1, with a score of 0.5 indicating chance performance.

Inter-Observer Agreement An all-except-one method was used to compute the between-participant agreement for each expectation map. In this analysis, a map was computed based on the responses of all participants except one, and the ROC curve was used to evaluate how well this model predicted the location chosen by the remaining participant. This process was repeated for all participants, and the mean AUC value obtained is an indicator of the inter-observer agreement in guessed locations.

Figure 2 shows a target expectation map for two example target labels from Experiment 1. The first is from a target on which participants showed considerable agreement, while the mouse clicks in the second are more distributed. Table 1 summarizes the between-participant AUC scores across all targets in Experiments 1 and 2. Critically, all the scores are much greater than 0.5, confirming that participants were indeed consistent in the points that they chose. This was true for Experiment 2, where participants saw a brief preview of the search scene, and also for Experiment 1, when participants guessed ("eyes closed") with only the target identity to go on.

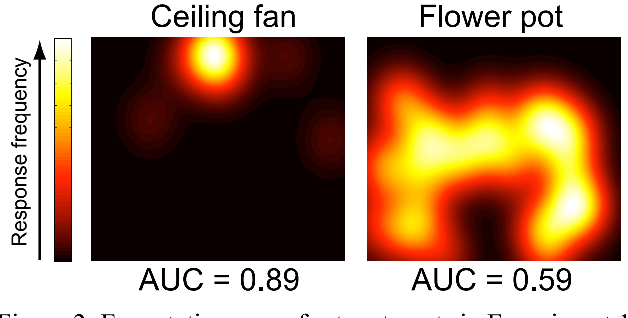

Figure 2: Expectation maps for two targets in Experiment 1. AUC scores represent the inter-observer agreement, which is high for one target (left) and much lower for the other (right).

The targets used were distributed throughout the scene, and could appear anywhere. However, some of the agreement may have originated because, across all trials, objects and mouse clicks were more likely to be in some locations (such as the centre of the image) than others. To control for this, an additional "between-target" analysis was performed using the method described above but with the responses associated with each object used to predict those from a *different* target (e.g., how well do guesses for the location of a flower pot predict those for a ceiling fan or a TV?). This control analysis will therefore quantify convergence that is independent of the particular target. This between-target control was higher than 0.5, probably because some objects were in a similar position. Importantly, inter-observer agreement was significantly higher than the between-target control AUC, in both Experiments (all *t*s(71)>3.8, *p*s<.001).

Table 1: Inter-observer agreement in target guesses in Experiments 1 and 2. AUC values give the mean and standard deviation across all targets.

Trial type	AUC	
	Mean	SD
Experiment 1: "Eyes closed"		
All trials	0.79	0.07
Experiment 2: "Eyes open"		
Target-present	0.88	0.08
Target-absent	0.83	0.08
Between-target control	0.71	0.1

Inter-observer agreement was significantly higher in the preview Experiment 2 than in Experiment 1. Moreover, this was the case in both target-present scenes (where participants could have, in theory, perceived the target object during the preview; *t*(71)=5.6, *p*<.0001) and in target-absent scenes (where there was no target to find; *t*(71)=3.5, *p*<.001). In other words, exposure to a brief glimpse of a scene made participants more likely to predict the same location for an object, even when that object was not present. Figure 3 shows the expectation map for the target label "TV", from responses in Experiment 1 (where participants made a blind guess) and for a target absent trial in Experiment 2 (where participants saw the depicted preview scene which did not contain a TV). Participants responding in Experiment 2 changed their guesses considerably and focused on a spot where a TV might appear.

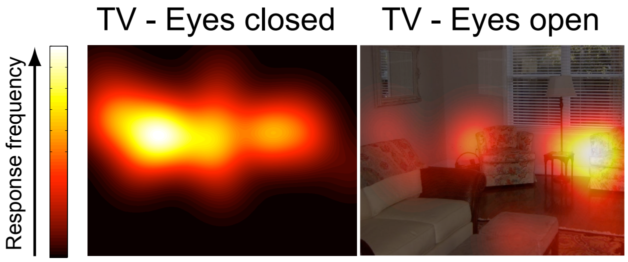

Figure 3: Expectation maps for the target label "TV" when guessing in Experiment 1 (left) and in Experiment 2 (right, superimposed over the preview scene that was shown).

Experiment 3

Experiments 1 and 2 confirmed that people were consistent in their expectations about where a named target would be likely to appear in a real world scene. The target maps provide a simple way of representing these expectations. Experiment 3 tested whether the eye movements of a new group of observers could be predicted from the target guesses.

Method

Participants Eighteen new participants (12 females), who had not taken part in Experiments 1 and 2, were recruited for payment. All participants had normal or corrected-to-normal vision, and their mean age was 22.5 years.

Stimuli and Apparatus The same set of target labels and scenes was used as in the previous experiments. To avoid trial-to-trial learning, each scene was presented once only, with half of the scenes containing the target and half without (i.e. matched with a different target label not present in the scene as in Experiment 2). Across participants, each scene appeared in both target-present and target-absent conditions.

Stimuli were presented on a 19-inch monitor positioned 60cm from the observers. Participants rested on a chin-rest, which ensured a constant viewing distance and restricted head movements. Scene images filled the screen, subtending 33 x 26 degrees of visual angle at this viewing distance.

Eye movements were recorded during search using the EyeLink 1000 system (SR Research), which used a desk-mounted camera to record monocular eye position from a video image of the pupil and the corneal reflection. This eye-tracker has a high spatial resolution (error of less than 0.5 degrees on average) and captured eye position at 1000Hz. Samples were parsed into oculomotor events using the EyeLink system's default algorithm, which identifies saccades and fixations based on velocity and acceleration thresholds. Search responses were entered via a button box.

Procedure The experiment began with an eye-tracker calibration (using a nine-dot grid), followed by instructions and 8 practice trials. The experimental trials followed a standard visual search procedure. First, a target label was presented, written in black font in the centre of the screen for 1s. This was replaced by a fixation point presented in the centre of the screen and participants pressed a button to proceed with the search. At this time the eyetracker checked that fixation was on the centre. The search scene then appeared, and participants were told to press one of two buttons as quickly and accurately as possible to identify whether or not the target was present in the scene. The search response terminated the trial, which ended with a blank screen for 500ms. All 72 trials were presented in the same way, in a random order, and the eye-tracker was recalibrated at the halfway point.

Results

Search Performance Participants responded accurately on a mean of 89% of all trials. In correct, target-present trials, the mean reaction time was 1350ms (standard error of the mean, SEM = 134) and participants made 5.5 fixations on average, per trial (SEM = 0.4). As with most visual search tasks, target absent trials were responded to more slowly (M = 2063ms, SEM = 237) and with more fixations per trial (M = 7.9, SEM = 0.7). Figure 4 gives an example of the locations fixated during a trial.

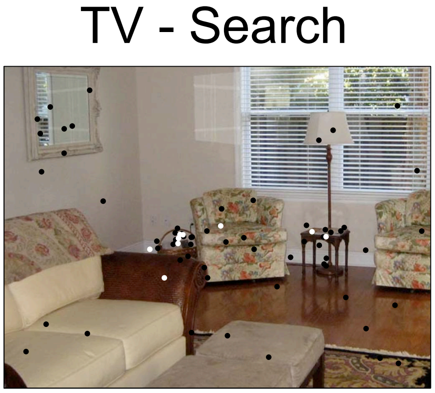

Figure 4: The locations of fixations made by all participants searching for the target "TV" in a target-absent trial. White markers indicate the first fixation in the trial.

Predicting Fixation Locations From Expectation Maps The remaining analyses aimed to assess whether fixation locations in the visual search task could be predicted based on the expectation maps derived from each target in Experiments 1 and 2. As previously, an ROC approach was followed. For each target-scene pair, the analysis asked how well the expectation maps formed from guesses could discriminate between fixated and non-fixated locations. Because it was anticipated from theory and previous experiments that attentional priorities might change over time, separate ROC curves were computed from each participant's first saccade target (i.e., the location of the first fixation away from the central starting point) and from all fixations in the trial. It is also essential to take into account the general tendencies for fixations (and probably mouse clicks) to be located near to the centre of the image and away from the scene edges.

Table 2: Predicting fixation locations from the guessed locations in Experiments 1 and 2.

Prediction	Trial type	AUC (all saccades)		AUC (first saccade)	
		M	SD	M	SD
Experiment 1: "Eyes closed"	Target-present	0.67	0.12	0.71	0.16
	Target-absent	0.68	0.13	0.73	0.17
	Between-trial control	0.62	0.05	0.67	0.05
Experiment 2: "Eyes open"	Target-present	0.82	0.12	0.83	0.13
	Target-absent	0.76	0.11	0.79	0.10
	Between-trial control	0.64	0.07	0.72	0.11

Therefore, following Ehinger et al. (2009), I computed a between-trial control comparison where the expectation map for one target/scene was used to predict the fixations made while searching for a different object.

Table 2 displays AUC summary statistics for the comparison between expectation maps and fixated locations. There were several noteworthy results. First, all the AUC values are greater than 0.5, showing that fixation locations could be predicted on the basis of the mouse responses made in Experiments 1 and 2. Moreover, in all trial types, the observed AUC values are greater than the between-trial control estimate. This was statistically reliable across the different target/scene pairs (all *t*(71)>2.6, *p*s<.01) and confirms that the results cannot be attributed to general spatial biases.

In addition, both expectation models were better predictors of the first saccade in a trial than they were of all saccades. This may be because the initial saccade was most likely to move toward the expected location, whereas later saccades might be exploring different areas of the picture. However, the between-trial control also led to higher AUC values when only the first saccade was evaluated, so it seems the first saccade is more predictable *in general*. This might be because of a strong central bias in scene viewing which tends to decrease over time, particularly when viewing starts in the centre (as it did here).

Most importantly, the guesses made by participants who saw a brief preview of the scene ("Eyes open") were a significantly better predictor of fixation locations than those who guessed blindly without seeing the scene. The best performance came in target-present trials, which indicates that participants in Experiment 2 had seen the target at least some of the time when guessing. Searchers in Experiment 3 were obviously highly likely to look at this correct target location, whereas there was more variability in target-absent images. However, it is important to note that, even when there was no target, the "Eyes closed" guesses of an independent group of participants were a significant predictor of fixation.

Predicting Between-target Variation An additional question concerns the relationship between expectations and search performance. If target objects are strongly associated with a particular location then we would expect a considerable amount of inter-observer agreement in the expectation maps (e.g., compare the two maps in Figure 2). If these expectations are an important factor in real world search, then the inter-observer agreement should correlate with reaction time in Experiment 3.

The mean reaction time was calculated across all participants for each correct, target-present trial, and then correlated with the AUC values representing inter-observer agreement from Experiments 1 and 2. In both cases there was a negative correlation (see Figure 5). When participants were more consistent in their guesses about where an object would appear, this object was found more quickly. The correlation with "Eyes closed" guesses approached significance (*r*=-.21, *p*=.08), while the correlation with "Eyes open" guesses was larger and statistically reliable (*r*=.50, *p*<.001).

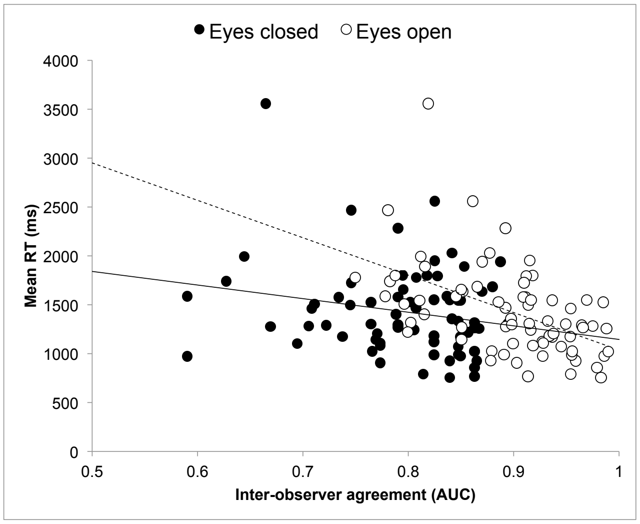

Figure 5: The correlation between inter-observer agreement and search RT. Each data point represents a target/scene pair from Experiments 1 or 2, with least-squares regression lines.

Discussion

In this paper I have proposed a distinction between the different types of top-down information available in guiding search in real-world scenes. Unlike previous descriptions of contextual effects (Biederman et al., 1973; Torralba et al., 2006), I specifically emphasized the fact that some predictions based on semantic knowledge can be made prior to the onset of the search scene. There were several interesting findings, which point to promising future directions for this approach.

First, participants showed a reliable amount of agreement when asked to blindly guess the target location. Although participants initially found this task unusual they were able to do so quickly and often chose the same locations for an object. There was some variation between different objects, with objects showing the largest amount of agreement those which are strongly constrained to spatial locations (such as light fittings). The method described here could be used in further research to characterise different search objects and their effects on performance. It should be noted that, because the present studies were limited to a fixed image frame on a monitor it mainly measured knowledge about picture composition (e.g., where the horizon is likely to be in a scene). How participants use such information in real life, where frames of reference change with head and body position, remains an open question and could be explored by looking at attention in active, real-world environments (see Foulsham, Walker & Kingstone, 2011).

Second, "Eyes closed" predictions were at least partly separable from those made in response to a preview of the scene. A brief preview of the scene gist, prior to seeing the target label, was enough to increase agreement between observers, even when there was no target to find. In other words, additional information about the scene was used by participants in a consistent way. It would be interesting to determine some of the cues that participants are responding to in this situation, as they could potentially be both appearance-based (selecting something which looked like the target) and location-based (selecting a region where the target might reasonably occur).

Third, the guesses of the participants in Experiments 1 and 2 were reliable predictors of fixation locations in independent searchers in Experiment 3. This was true when participants guessed based on a brief preview of the scene, which confirms that searchers look towards the parts of the scene which are contextually relevant given the gist. This finding, in both target-present and target-absent scenes, is similar to that reported by Ehinger et al. (2009), who used a "context oracle" defined by the responses of independent observers predicting the y-coordinate where pedestrians should occur in street scenes. However, what is surprising in the current experiments is that, even without the scene, participants are able to predict target locations, and these predictions are reflected in fixation behavior. In future work this could be modeled by positing a "blind" statistical prior which could then be refined according to global features such as those in the contextual guidance model of Torralba et al., (2006).

References

Biederman, I., Glass, A. L., & Stacy, E. W. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology*, 97, 22–27.

Chen, X., & Zelinsky, G. J. (2006). Real-world visual search is dominated by top-down guidance. *Vision Research, 46*(24), 4118-4133.

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modeling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition, 17*, 945-978.

Foulsham, T., & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual saliency in scene perception? *Perception, 36*, 1123-1138.

Foulsham, T., Alan, R. & Kingstone, A. (2011). Scrambled eyes? Disrupting scene structure impedes focal processing and increases bottom-up guidance. *Attention, Perception and Psychophysics*, *73* (7), 2008-2025.

Foulsham, T., Walker, E. & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, *51* (17), 1920-1931.

Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. Journal Of Experimental Psychology: Human Perception & Performance, 25, 210–228.

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.

Kanan, C., Tong, M. H., Zhang, L., & Cottrell, G. W. (2009). SUN: Top-down saliency using natural statistics. *Visual Cognition, 17*, 979-1003.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research, 45*(2), 205-231.

Peirce, JW (2007) PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1-2), 8-13.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*(4), 766-786.

Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. Psychonomic Bulletin and Review, 1, 202-228.

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science, 9*(1), 33-39.

Zelinsky, G. J. (2008). A Theory of Eye Movements During Target Acquisition. *Psychological Review, 115*(4), 787-835.