An Information Search Model Integrating Visual, Semantic and Memory Processes

Myriam Chanceaux (myriam.chanceaux@imag.fr)
University of Grenoble, Laboratoire TIMC-IMAG, Domaine de la Merci
38700 La Tronche FRANCE

Anne Guérin-Dugué (anne.guerin@gipsa-lab.inpg.fr)
University of Grenoble, Laboratoire Gipsa-Lab, Domaine universitaire
38402 Saint Martin d'Hères FRANCE

Benoît Lemaire (benoit.lemaire@imag.fr)
University of Grenoble, Laboratoire TIMC-IMAG, Domaine de la Merci
38700 La Tronche FRANCE

Thierry Baccino (baccino@lutin-userlab.fr)
Laboratoire LUTIN, Cité des sciences et de l'industrie de la Villette
75930 Paris cedex 19 FRANCE

Abstract

This study aims at presenting a computational model of visual search including a visual, a semantic and a memory process in order to simulate human behavior during information seeking. We implemented the memory process, which is the most important part of the model based on the Variable Memory Model (Arani, Karwan, & G., 1984; Horowitz, 2006). To compare model and humans, we designed two experiments where participants were asked to find the word among forty distributed on the display, which best answers a question. The results showed good fits on different features extracted between empirical and simulated scanpaths.

Keywords: Visual Memory; Information Seeking; Computational Model; Eye Movements; Semantic Similarities

Introduction

Nowadays information seeking, on a Web page for example, is a very common task. That is why for several years research has been conducted to try to answer this question: what guide user's attention in this particular task, especially on the web? The literature contains theoretical models of information seeking activity (Marchionini, 1997), especially in electronic documents, and also computational models which simulate navigation between pages, with cognitive architectures like ACT-R (Pirolli & Fu, 2003). There are also a lot of studies based on the Feature Integration Theory (Treisman & Gelade, 1980) with models which take into account the visual features of stimuli: colors, orientation, contrast (Itti & Koch, 2000) to determine the most salient part of the stimulus. However even if some models take into account the semantic information of the material, in addition to visual information (Navalpakkam & Itti, 2005), experiments and modelling are missing in this field. The purpose of this paper is to describe a cognitively plausible model that takes into account semantic and visual features of stimuli when searching for information. This model is an improvement of a simpler one (Chanceaux, Guérin-Dugué, Lemaire, & Baccino, 2008). It has been implemented and compared to experimental data collected during two experiments where participants had to search for information.

Model architecture

In this section we will describe our model. This model simulates human eye movements during a simple task of information seeking involving text and visual features. We developed an architecture in 3 parts, (see Figure 1), corresponding to 3 main cognitive processes involved in a task of information search. These three processes are respectively related to bottom-up visual information, top-down semantic information and memory. These processes will be detailed thereafter. The principle of the model is to predict from a fixed location and an history of previous fixations where will be the next fixation. It is assumed that from a given point each location of the image has a visual weight (calculated by the visual process) and a semantic weight (calculated by the semantic process). These values are modulated by the process of memory. The basic process of memory is that if a location has already been visited it has less interest than if it never was. At each iteration the location of the fixation is thus determined until the end of the scanpath. The number of fixations is fixed.

Visual part

The visual part of the model is itself divided into 2 parts. The first part is a very simple retinal filter which provide a good fit to acuity limitations in the visual human field (Zelinsky, Zhang, Yu, Chen, & Samaras, 2006; Geisler & Perry, 2002) . The filtering output has a maximal resolution in areas near the fixation point, and the resolution decreases with eccentricity. More practically, locations close to the current fixation will have a strong visual weight, and those further away a low weight. The second part of this process concerns visual stimuli. As in many models of visual attention (Itti & Koch, 2001) we also take into account this information in a pseudo saliency map. In the first experiment, this information is the

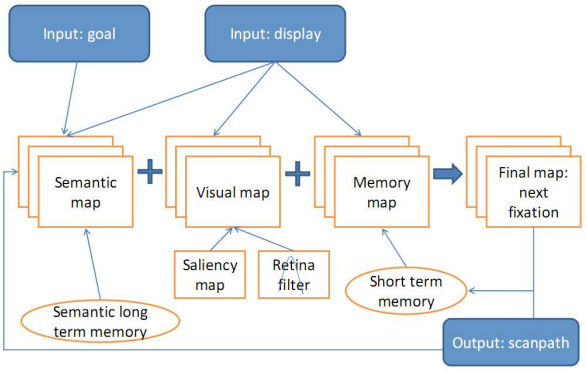

Figure 1: Model's architecture

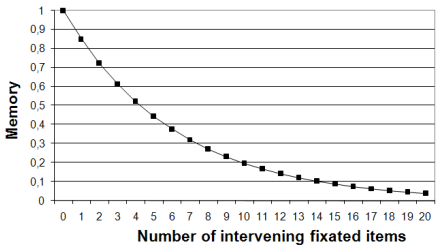

Figure 2: Memory weight as a function of intervening items

size of items, and in the second one, their color. Our visual model is defined as such:

$$VisualWeight(i) = VisualSaliency(i) * VisualAcuity(i,c)$$

i represents each location in the display and *c* the current fixation.

Semantic part

It is well known that the semantic meaning of the search has an impact on eye movement (Yarbus, 1967), but this information is tricky to implement. That is why our semantic model takes into account the semantic similarity between a word and the goal of the search. This similarity is calculated by LSA (Latent Semantic Analysis). This method, popularized by Landauer and Dumais (1997), takes a huge corpus as input and yields a high-dimensional vector representation for each word, usually about 300 dimensions. It is based on a singular value decomposition of a word $\times$ paragraph occurrence matrix, which implements the idea that words occurring in similar contexts are represented by close vectors. Such a vector representation is very convenient to give a representation to sentences that were not in the corpus: the meaning of a sentence is represented as a linear combination of its word vectors. Therefore, we can virtually take any sentence and give it a representation. Once this vector is computed, we can compute the semantic similarity between any word and this sentence, using the cosine function. The higher the cosine value, the more similar the words are. The results of this method have been experimentally tested and are close to the capabilities of human in judging of semantic similarities (Landauer, McNamara, Dennis, & Kintsch, 2007). Specifically in our model, our assumption is that elements which are spatially close are also semantically close, as it is often the case in reality, and it is the case by construction in our stimuli. If a fixated word is semantically close to the goal, elements nearby will receive a high weight. On the contrary if that word is far from the goal, elements close to it will receive weak weight. The model will therefore tend to move away from the areas considered unrelevant. More precisely if the similarity between the current fixated word and the instruction is under 0.2[1], the weights are decreased following a Gaussian around the current fixation. If the similarity is above 0.2, the weights are increased still following a Gaussian around the current fixation.

Memory part

Most computational models of attention have implemented an Inhibition Of Return (IOR) mechanism for driving attention through a scene (Klein & MacInnes, 1999). However, evidence for IOR during scene viewing is inconclusive (Motter & Belky, 1998; Hooge, Over, Wezel, & Frens, 2005). In these studies, results indicate that there is a tendency for saccades to continue the trajectory of the previous saccade, but contrary to the *foraging facilitator* hypothesis of IOR, there is also a distinct population of saccades directed back to the previous fixation location. To capture the pattern of saccadic eye movements during scene viewing we need to model the dynamics of visual encoding. That is why we implemented the Variable Memory Model instead of using the simple mechanism of IOR.

Variable Memory Model The memory part is based on the Variable Memory Model (VMM), developed by Arani et al. (1984) and revisited by Horowitz (2006). VMM is a non-deterministic model originally designed to accomodate eye movements data. It is based on 2 parameters: θ the probability of encoding and φ the probability of recovering information in memory[2]. When an item is attended, the location of that item could be encoded by the model, or not, depending of the encoding probability θ. The second parameter, φ, simulates a forgetting mechanism. In fact when a stimulus is encoded, it will be gradually deteriorated following the curve drawn in Figure 2. When the memory weight is 1 the item is perfectly remembered, when it is 0 there is no trace of this item in memory. When an item is attended, its weight is 1 if it is encoded then progressively goes back to 0: at each step, its memory weight decreases as a function of the number of intervening items.

[1] This value of 0.2 is usually considered in the LSA literature as a threshold under which items are unrelated.

[2] A third parameter is used in the original model, representing the probability of correctly identifying the target.

At the *i*th fixation the model will remember or not that a particular item was encoding during the *k*th fixation, as following:

$$MemoryWeight(i,k) = \varphi^{i-k}$$

The last point about this part is the importance of the meaning of the stimulus: θ is modulated by semantics. If the stimulus is really close (semantically similar) to the user's goal, the encoding is more difficult, because the treatment of this stimulus takes more time and cognitive resources. In contrary, if the stimulus has anything to do with the information search the encoding will be easier. This point explains why when a stimulus is interesting we sometimes need to go back to it. In fact in the observed data there are more return on words semantically close to the information seeking goal, than others.

Integration

Each part of the model (visual, semantic and memory) generates a map. On this map each location has a weight for this component. The integration of maps to form the main map is a weighted sum of these three components:

$$M_{general} = \alpha_V . M_{visual} + \alpha_S . M_{semantic} + \alpha_M . M_{memory},$$

$$\alpha_V + \alpha_S + \alpha_M = 1$$

The weights α_V, α_S and α_M are not fixed a priori. We are instead looking at the respective role of each component.

Experiments

In order to test the model, we will now present two experiments, which enabled to compare the computational model with experimental data. We formalized the goal the user is pursuing by considering that this user is seeking a particular piece of information. Our goal is to apply our model to complex Web pages, but before that we tested it on a simpler task. The user is asked to find the best answer to a question. 40 words are spread on the display, and the target is defined by the class it belongs to and its specific features within this class. For instance, the instruction could be: Find the biggest animal (class: animal and feature: big), as in Figure 3. Each of the 40 words of each image has a visual feature and a semantic feature. We also organized the 40 words in order to reproduce the fact that in our world objects that are semantically similar are also often close to each other; in supermarket, vegetables are in the same place and they are close to fruits. In each image, seven words, including the target word, belong to the same category (i.e. category of the instruction, in our example *animal*), and all 33 other words are of decreasing semantic similarity with the instruction. To calculate these similarities we used the LSA method, briefly described above.

Methods

In both experiments, eye movements were monitored by an SR Research Eyelink 2 eyetracker. Viewing was binocular, but only one eye was tracked. The images were presented on a 19 inch CRT monitor at a viewing distance of 50 cm. Participants were told to read the instruction, to fix a fixation cross and to view the display in order to find the best answer to the instruction, with no time limitation. Participants were presented 18 1024 x 768 pixels images (subtending 42 horizontal deg. of visual angle). These 2 experiments have the same methods, but differ in one point: the visual features of the stimuli.

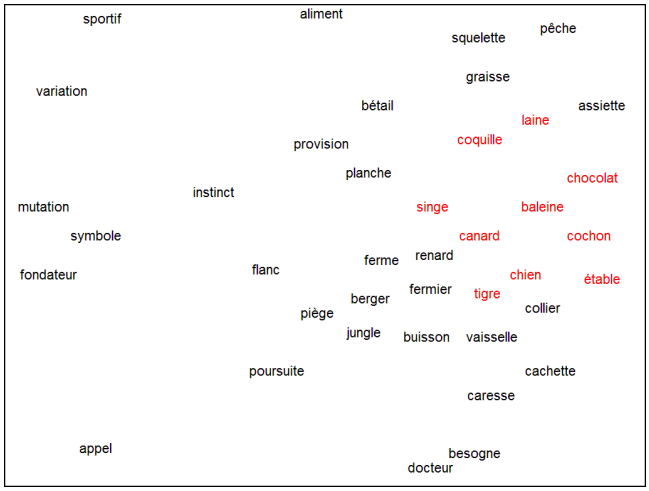

Figure 3: Example of image. The instruction is *Find the biggest animal*. The answer is *whale ("baleine" in french)*

Experiment 1 In the first experiment the visual feature is the font size of the stimuli. Each word has a font size between 13 and 19, allocated in this way (19: 5 words, 18: 5 words, 17: 6 words, 16: 7 words, 15: 6 words, 14: 6 words and 13: 5 words). They can be grouped into two classes: V+ (size 18-19) and V- (size 13-17). There are three visual conditions which are (1) random assignments of visual features to words; (2) no visual features at all; (3) visual features congruent to spatial locations: words that are close to the target have bigger font size. In this experiment the visual feature is multi-varied and gradual. 43 students of Grenoble University participated (30 female; mean age = 20.9 years). All participants had normal or corrected to normal vision and were naïve with respect to the purposes of the study.

Experiment 2 In the second experiment the visual feature is the color of the stimuli. Each word is black or red (black: 30 words, red: 10 words). There are then two classes: V+ (red) and V- (black). There are two visual conditions which are (1) random assignments of color; (2) visual features congruent to spatial locations: words that are close to the target are red. In this experiment the visual feature is bimodal and dichotomic. 29 students of Grenoble University participated (14 female; mean age = 21.2 years). All participants had normal or corrected to normal vision and were naïve with respect to the purposes of the study.

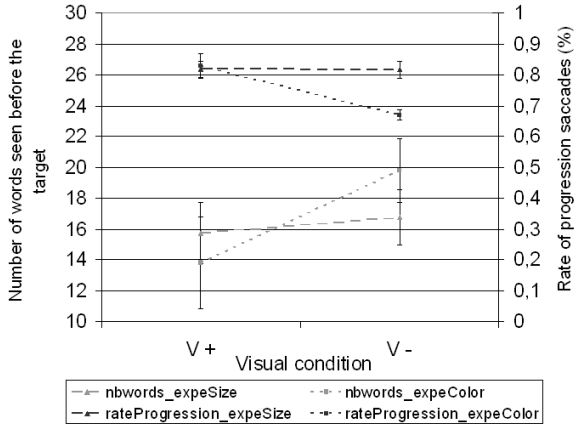

Figure 4: *Number of words* and *rate of progression saccades* according to visual conditions for both experiments

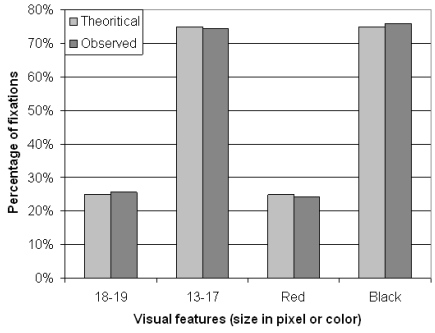

Figure 5: Theoritical and observed percentage of fixations depending on visual features

Results

We have here two different experiments, with different results and we will see if the model can account for both data. In both experiments we identified three features that allow us to characterize the performance of participants and to compare them afterwards with model data. The selected features are the number of words seen before reaching the target, the rate of progression saccades, i.e. the number of fixations closer to the target than the previous one divided by the number of all fixations and the angles in the scanpath. These features were chosen for their discriminating power between the different conditions of the experiment. We are interested in the values of all these features for the last 6 maps seen by the participants, once they were well aware of the task and the semantic organization of words on the display.

The difference between the 2 experiments was the visual factor. The results show (see Figure 4) gains in performance for the experiment *color* that are not in experiment *size* between the two visual conditions (colored words help, but big words do not). For the feature *number of words*: $T(17)=2.06$, $p=0.055$ against $T(17)=0.96$, $p=0.349$ and for the feature *rate of progression saccades* $T(17)=4.07$, $p\leq 0.001$ against $T(17)=0.73$, $p=0.474$.

To see why there are differences in performance between the 2 experiments, and especially, why the factor *size* does not affect performance unlike the factor *color*, we looked at whether the words in color were over fixated compared to the words in black and similarly if the biggest words were more fixated than smaller words (in random visual condition) Results show (Figure 5) that the words with more salient visual features are not more fixated than others. There is no significant difference between observed and theoretical values (factor *size*: $\chi^2 = 1{,}26$, dl = 1, p = 0.26 and factor *color*: $\chi^2 = 2.36$, dl = 1, p = 0.12). The gradualness of factor *size* is certainly the reason for the difference previously observed.

We will now see what is the behavior of the model previously described on these quite different tasks.

Comparisons between model and human scanpaths

Weights combination

We have done 50 iterations of the model for each experiment, like a simulation of 50 participants. The total number of fixations was fixed for each image and equals to the average number of fixations made by the participants who saw this picture. We made these simulations for all combinations of α_V, α_S and α_M from 0 to 1 with 0.05 steps. For the experiment with the factor *color* we also varied the weight of the red words for the visual map (representing the visual saliency of the words). We also varied the parameters of the memory model φ and θ, to compare them thereafter with data from the literature.

To select the best parameters of the model we compared the average relative errors between the model and participants for the two features described above, the number of words seen before the target and the rate of progression saccades. The average angle of the scanpath gives us similar results that we do not present here. We took into account for each feature the 10 best combinations, to average the parameters. In fact the difference in the relative errors are too weak to only take the best one. These results are shown in Table 1.

Table 1: Best combination of weights

	Experiment *size*		Experiment *color*	
	Rate prog	Nb words	Rate prog	Nb words
α_V	0,4	0,34	0,38	0,31
α_M	0,50	0,62	0,45	0,58
α_S	0,11	0,05	0,18	0,12
θ	0,83	0,84	0,83	0,85
φ	0,84	0,86	0,84	0,86

We can first notice that the best parameters are very close for the 2 experiments, i.e. the model is suitable for 2 types of

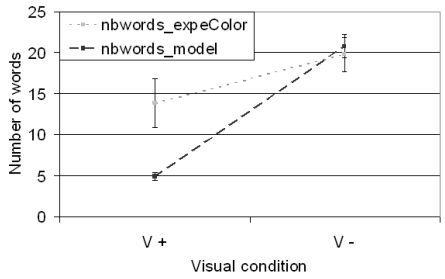

Figure 6: Number of words seen before the target for model and observed data

visual feature, bimodal (*color) or gradual (*size). If we now look at the meaning of these values the most important weight is for the memory component, which enables the model to remember where it has already been. The visual component also plays an important role, because it takes into account the visual acuity and guide the attention on the closest words, as human do. Finally the semantic part is less important, especially in the first experiment. Second the parameters θ and φ are also close to those found in literature which are: θ=0,82 et φ=0,86 (Horowitz, 2006).

Finally looking at the distinction between images with red words around the target and images with red words randomly displayed both models and participants have differences in performances, as shown for example in Figure 6 for the feature *number of words*.

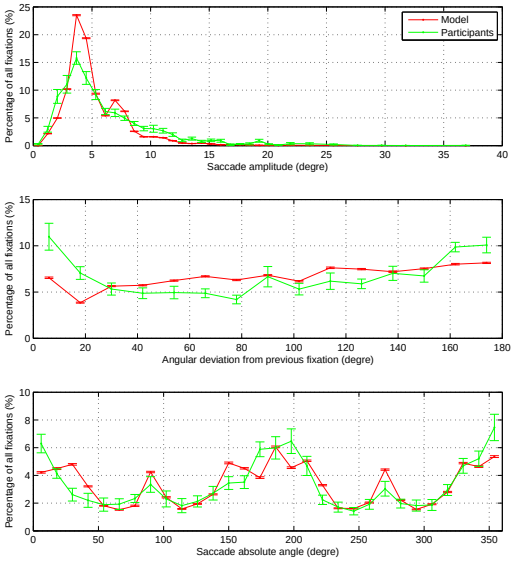

Figure 7: Comparisons between model and data on the experiment *size*

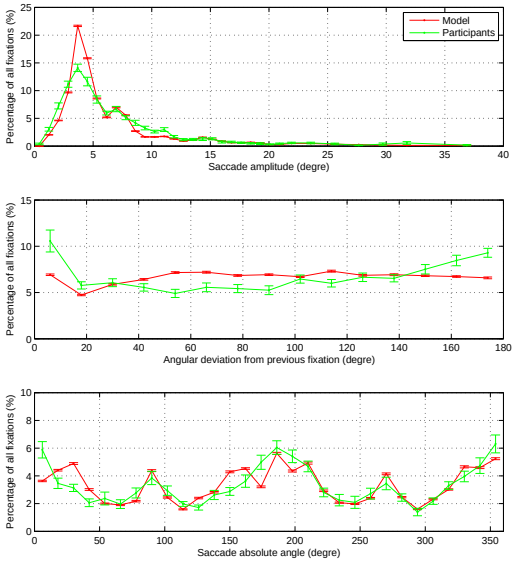

Figure 8: Comparisons between model and data on the experiment *color*

Angle and saccade distributions

To have a more complete comparison between humans and model, we studied two distributions which are typical of human scanpath: saccade length distribution, and angle distributions (both relative and absolute angles). A relative angle is an angle between two saccades, an absolute angle is between a saccade and the horizontal line.

Figure 7 describes the first experiment comparisons and Figure 8 the second ones. There is a good fit of the data for the saccade amplitude distribution curve (upper panel of both figures) with a peak at about 4 degrees of saccade amplitude in all cases.

In the case of relative angles the results show a curve for participants with more return saccades (0 ˚) and forward saccades (180 ˚), same results that those found by Tatler and Vincent (2008). For the model, the curve is more horizontal (middle panel of both figures).

The values that we get in the case of absolute angles are very interesting. They show very clearly an horizontal trend (peaks at 0 ˚ and 180 ˚) rather than vertical (small peaks at 90 ˚ and 270 ˚) for both humans and model.

χ^2 tests give us no significant differences for each comparison (all $p \geq 0.9$), meaning the model distributions are similar to human's ones for experiments 1 and 2.

Conclusion

We have considered two experiments in order to validate a model. This model takes into account both semantic and visual information, associated with a model of visual memory. There are many parameters we had to determined, and

to do that we tried first of all to make cognitively plausible choices. In fact the visual part takes into account both visual human acuity and the saliency of the stimuli. This saliency is here really simple because of the simplicity of the stimuli, but could be more complex if necessary. For the semantic part we used a well known method and theory which provides us a good measure of similarity between the aim of the information search and the items. Finally the memory component, which is the most important according to the results enables the model to remember items previously seen and to forgetting them without being too strict, unlike a classical inhibition of return, which is generally used in such a model. In fact it is more realistic, because humans often go back to the previously fixated item, as shown in the relative angles distribution.

We chose two different scales for visual saliency of stimuli: bimodal for color and linear for size. The color is not linear, the reverse would have been difficult to control on effects of saliency. This choice allowed us to see differences in performance, and even if the two experiments are not similar, the best parameters for the model are almost the same, meaning that this model is robust to, at least, such a change.

In further experiments with text paragraphs instead of single words, more similar to Web pages, we will test this model again, to see if in a new task the weight of these components are always the same or not. This model, able to explain simple stimuli, tends to be more complex, integrating for example a reading model.

Acknowledgements

We would like to thank Gelu Ionescu for providing us LisEyeLink software; and Nicolas Betton for his work during the experiment. We also thank participants who accepted to pass the experiment.

References

Arani, T., Karwan, M. H., & G., D. C. (1984). A variable-memory model of visual search. *Human factors*, *26*, 631-639.

Chanceaux, M., Guérin-Dugué, A., Lemaire, B., & Baccino, T. (2008). Towards a model of information seeking by integrating visual, semantic and memory maps. In *Proceedings of the 4th international cognitive vision workshop* (pp. 65–78). Lecture Notes in Computer Science 5329, Berlin: Springer Verlag.

Geisler, W. S., & Perry, J. S. (2002). Real-time simulation of arbitrary visual fields. In *Proceedings of the 2002 symposium on eye tracking research & applications* (pp. 83–87). ACM New York, NY, USA.

Hooge, I. T., Over, E. A., Wezel, R. J. van, & Frens, M. A. (2005). Inhibition of return is not a foraging facilitator in saccadic search and free viewing. *Vision Research*, *45*, 1901–1908.

Horowitz, T. (2006). Revisiting the variable memory model of visual search. *Visual Cognition*, *14*, 668–684.

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

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience*, *2*, 194-203.

Klein, R. M., & MacInnes, J. W. (1999). Inhibition of return is a foraging facilitator in visual search. *Psychological Science*, *10*, 346–352.

Landauer, T., & Dumais, S. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, *104*, 211–240.

Landauer, T., McNamara, D., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of latent semantic analysis.* Lawrence Erlbaum Associates.

Marchionini, G. (1997). *Information seeking in electronic environments (Cambridge series on Human-Computer interaction).* Cambridge University Press.

Motter, B. C., & Belky, E. J. (1998). The guidance of eye movements during active visual search. *Vision Research*, *38(12)*, 1805-1815.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, *45*, 205-31.

Pirolli, P., & Fu, W. (2003). SNIF-ACT: a model of information foraging on the world wide web. In P. Brusilovsky, A. Corbett, & F. de Rosis (Eds.), *User modeling 2003, 9th International Conference, UM 2003* (Vol. 2702, p. 45-54). Johnstown, PA: Springer-Verlag.

Tatler, B. W., & Vincent, B. T. (2008). Systematic tendencies in scene viewing. *Journal of Eye Movement Research*, *2(2):5*, 1-18.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97-136.

Yarbus, A. L. (1967). *Eye movements and vision.* New York: Plenum Press.

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, p. 1609-1616). Cambridge, MA: MIT Press.