Memory Modulated Saliency: A Computational Model of the Incremental Learning of Target Locations in Visual Search

Michal Dziemianko and Frank Keller
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
m.dziemianko@sms.ed.ac.uk, keller@inf.ed.ac.uk

Abstract

The top-down guidance of visual attention is one of the main factors allowing humans to effectively process vast amounts of incoming visual information. Nevertheless we still lack a full understanding of the visual, semantic, and memory processes governing visual attention. In this paper, we present a computational model of visual search capable of predicting the most likely positions of target objects. The model does not require a separate training phase, but learns likely target positions in an incremental fashion based on a memory of previous fixations. We evaluate the model on two search tasks and show that it outperforms saliency alone and comes close to the maximal performance of the Contextual Guidance Model (CGM, Torralba, Oliva, Castelhano, & Henderson, 2006; Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009), even though our model does not perform scene recognition or compute global image statistics. The search performance of our model can be further improved by combining it with the CGM.

Keywords: visual search; contextual guidance; eye-tracking; incremental learning.

Introduction

Virtually every human activity occurs within a visual context and requires visual attention in order to be successfully accomplished (Land & Hayhoe, 2001). When processing a visual scene, humans have to localize objects, identify them, and establish the relations that hold between them.

The support of the European Research Council under award number 203427 “Synchronous Linguistic and Visual Processing” is gratefully acknowledged.

We would like to thank Moreno I. Coco for sharing his data, and for numerous comments and suggetions regarding this work. We are also grateful to the authors of Ehinger et al. (2009) and Torralba et al. (2006) for sharing the image corpora and eye-tracking data used in their studies.

A preliminary version of study reported in this paper has been published as Dziemianko, Keller, and Coco (2011).

Visual Cognition, 21: 3, 277–305, 2013.

The eye-movements involved in these processes provide important information about the cognitive processes that unfold during scene comprehension (Henderson, 2003).

Studies of free viewing (e.g., Yarbus, 1967; Einhauser, Spain, & Perona, 2008) have shown that scan patterns on visual scenes can vary greatly between participants. On the other hand, the task that participants have to perform drives visual attention, resulting in fixated regions that are relatively consistent across participants both in search tasks (Torralba et al., 2006; Henderson, Malcolm, & Schandl, 2009) and in everyday activities (Pelz & Canosa, 2001; Hayhoe, Shrivastava, Mruczek, & Pelz, 2003).

A number of models have been proposed to predict eye-movements during scene comprehension; they can be broadly divided into two categories. The first one consists of bottom-up models exploiting low-level visual features to predict areas likely to be fixated. Several studies have shown that certain features and their statistical unexpectedness attract human attention (e.g., Bruce & Tsotsos, 2006). Moreover, low-level features are believed to contribute to the selection of fixated areas, especially for visual input that does not provide any useful high-level information (e.g., Peters, Iyer, Itti, & Koch, 2005). These experimental results are captured by models that detect salient areas of visual input and predict attention in a bottom-up fashion. The best-known example is the model of Itti, Koch, and Niebur (1998), which builds saliency maps based on color, orientation, and scale filters inspired by neurobiological studies of human vision. While there is evidence that saliency is predictive of eye-movement behavior (Itti, 2005), other authors have argued that this is merely a consequence of the fact that saliency is correlated with high-level properties that guide attention, such as objecthood (Castelhano & Henderson, 2007; Nuthmann & Henderson, 2010). Similarly Koostra, Nedereen, and De Boer (2008) and Zhang, Tong, Marks, Shan, and Cottrell (2008) have shown that a range of others factors, including symmetry and Bayesian surprise, need to be taken into account when predicting fixation locations.

The second group of models assume that top-down supervision of attention contributes to the selection of fixation targets. Various types of top-down supervision have been observed experimentally. Humans show the ability to learn general statistics pertaining to the appearance, position, size, spatial arrangement of objects, and their semantic relationships. Chun and Jiang (1999) show that observers are able to temporarily learn contingencies between objects. Similarly, Green and Hummel (2006) show that perception is sensitive to the relative pose of pairs of objects. Hwang, Wang, and Pomplun (2011) demonstrate that observers tend to fixate objects that are semantically related in sequence.

A series of studies have also shown the importance of context in scene comprehension. Context not only provides information about scene layout scene and type (Schyns & Oliva, 1994; Renninger & Malik, 2004), but also about object presence, location, and appearance (see, e.g., Bar, 2004; also Oliva & Torralba, 2007, discuss the effects of context on object recognition in detail). Another important manifestation of context in scene understanding is contextual cueing: observers are able to associate the locations of target objects with arbitrary scene contexts, and use this information to speed up visual search when exposed to the same scene again (Brockmole & Henderson, 2006a, 2006b; Brockmole, Castelhano, & Henderson, 2006). Furthermore, it has been shown that observers are able to extract low-level contextual information (scene gist) at very short exposures, without need for high-level visual processing (Castelhano & Henderson, 2007). This has inspired models that condition visual search on scene gist, such as the Contextual Guidance Model (Torralba et al., 2006), to which we will return below.

Whether visual memory is used during scene comprehension, as well as the exact form of such memory, is the subject of an ongoing debate in the literature. Several studies have indicated that visual search is memory-free (e.g., Horowitz & Wolfe, 1998; Wolfe, Klempen, & Dahlen, 2000). Wolfe (1999) explains this result by proposing that vision produces loose groupings of simple visual features such as the pre-attentive object files of Wolfe and Bennett (1997) or the proto-objects of Rensink (2000), which dissolve upon the withdrawal of attention, meaning that visual search is memory-free.

But there is also a considerable amount of evidence for the opposite effect, i.e., the influence of visual memory in a range of search paradigms. For instance Gibson, Li, Skow, Brown, and Cooke (2000), Klein (1988), Klein and MacInnes (1999), and Takeda and Yagi (2000) all show that vision exploits information about which objects have been accessed within the same trial. Chun and Jiang (1998, 1999) show that memory can also be used across trials to guide attention. Also the contextual cueing effect discussed above is an example of visual search making use of information retained in memory across trials. In the context of the present paper, the study by McPeek, Maljkovic, and Nakayama (1999) is particularly relevant. Using a visual search paradigm, the authors show that targets that match previously fixated targets are re-fixated more accurately and quickly than mismatching targets, indicating that attention is guided by short-term memory of visual features. Along the same lines, Maljkovic and Martini (2005) show that short-term memory can be used to explain effects of target frequency in visual search. In addition to this, memory effects have also been observed in other experimental paradigms (e.g., change blindness); for a more detailed discussion, refer to Hollingworth (2006), Shore and Klein (2000), or Woodman and Chun (2006).

The aim of the present paper is to explore the relationship between scene context and visual memory. Existing experimental and modeling studies dealing with context effects rely on an implicit form of memory, by assuming that participants remember, e.g., where objects are typically located in a scene (Torralba et al., 2006), or which object typically co-occur together (Hwang et al., 2011). In this paper, we postulate a more direct link between visual memory and context. We test the hypothesis that the locations of the fixation that participant makes on a given scene can be predicted based on their fixations on directly preceding scenes. We present a model that stores fixation locations in memory, and compare its accuracy in predicting fixation locations to a model that relies on object context (Torralba et al., 2006). Our evaluation uses two data sets: an existing visual search data set from the literature (Ehinger et al., 2009), and a novel visual counting data set that we collected.

Models of Context in Visual Attention

A number of models have been proposed to capture context effects on visual attention. A prominent example is Torralba et al.'s (2006) Contextual Guidance Model (CGM), which combines bottom-up saliency with a prior probability distribution encoding global scene information (gist). The central quantity computed by the CGM is the probability that a target object O is present at point X in the image:

$$p(O=1,X|L,G) \quad = \quad \frac{1}{p(L|G)} p(L|O=1,X,G)p(X|O=1,G)p(O=1|G) \tag{1}$$

Here, L is a set of local image features at X and G is a set of global features representing scene gist. The first term $\frac{1}{p(L|G)}$ is the saliency model. The second term $p(L|O=1,X,G)$ has the effect of enhancing the features of X that belong to the target object. The third term $p(X|O=1,G)$ is the

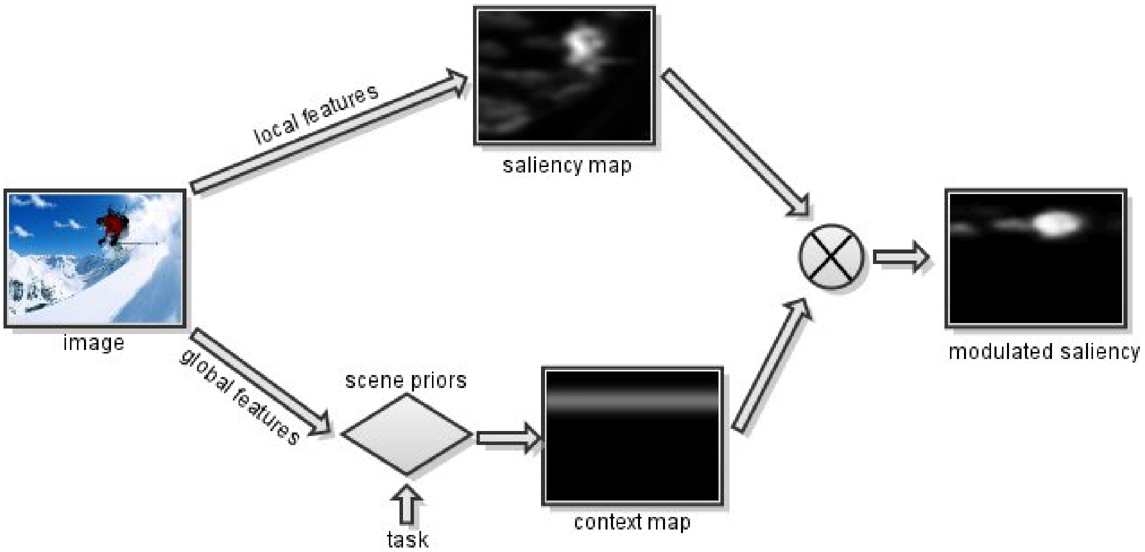

Figure 1. The architecture of the CGM. First, a saliency map is computed for the image. It is then modulated with a contextual prior conditioned on global scene features. The resulting map is thresholded to select the areas most likely to be fixated.

contextual prior, which provides information about likely target locations. The fourth term $p(O = 1|G)$ is the probability that O is present in the scene. The model is illustrated schematically in Figure 1. In Torralba et al.'s (2006) implementation of the CGM, the second and the forth terms are omitted, yielding:

$$S(X) = \frac{1}{p(L|G)} p(X|O = 1, G) \tag{2}$$

This equation describes contextually modulated saliency $S(X)$ as the combination of bottom-up saliency and a prior on the likely location of the target, both conditioned on global features representing scene gist. These global features are computed by pooling local features over 4×4 non-overlapping windows; the resulting vectors are reduced using principal component analysis.

In following sections, we describe a model of visual attention that predicts fixation locations in visual search tasks. Our proposal is conceptually similar to the CGM, but the top-down modulation of saliency in our model is based on the memory of previously found targets, rather than on global scene properties. Moreover, we show that the knowledge of expected object locations can be learned incrementally, and that no prior is needed to achieve satisfactory results in predicting fixation positions. Additionally we show that combining both sources of knowledge (context and memory) enhances search performance.

Methods

Model Architecture

We propose the Memory Modulated Saliency (MMS) model of eye-movements in scene comprehension. Like the CGM, our model combines bottom-up saliency with a top-down estimate of

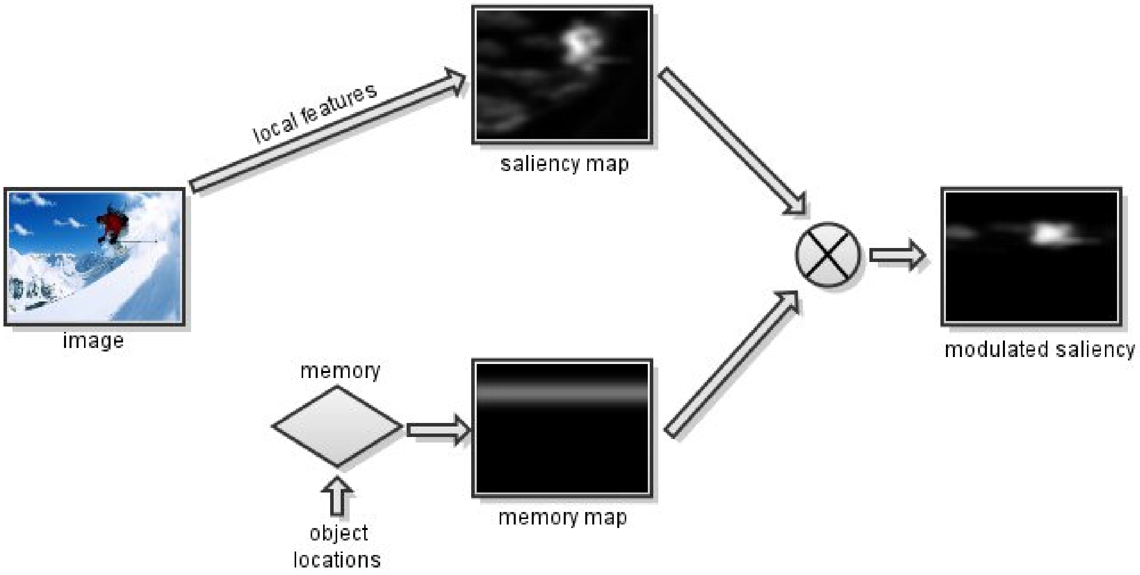

Figure 2. The architecture of the proposed MMS model. First, a saliency map is computed for the image. It is then modulated with a memory map estimated using fixations landing within the target objects or their center of mass on previously seen images. The resulting map is thresholded to select the areas most likely to be fixated.

likely target positions. In contrast to the CGM, our model does not assume any correspondence between global representations such as scene gist and human behavior. Instead, we assume that to estimate likely target positions, viewers rely on their memory of targets encountered in previous scenes. This information is then used to modulate a standard saliency map. The schematic architecture of the MMS model is shown in Figure 2.

Figure 3 presents an example of the computations performed by the model when fed a series of images. In the first step of each cycle, the saliency map of the image is calculated and modulated with the learned target position distribution. The resulting modulated map contains the model prediction for the fixation locations for this image. In the next cycle, the distribution of target object locations is updated based on the fixations the participant made on the targets in the previous image. The resulting updated memory map is then used to modulate the saliency map for the current image, resulting in fixation predictions for this image. The actual fixations are then again used to update the memory map in the next cycle, and so forth.

Salience Map We approximate saliency as the probability of the local images feature L in a given location based on the global distribution of these features (similar to Torralba et al., 2006):

$$p(L) \propto e^{-\frac{1}{2}[(L-\mu)^T\Sigma^{-1}(L-\mu)]} \tag{3}$$

Here μ is the mean vector and Σ the covariance matrix of the Gaussian distribution of local features estimated over the currently processed image. The local features are a set of Gabor filter responses computed over three color channels for six orientations and four scales, totaling 72 values at each position.

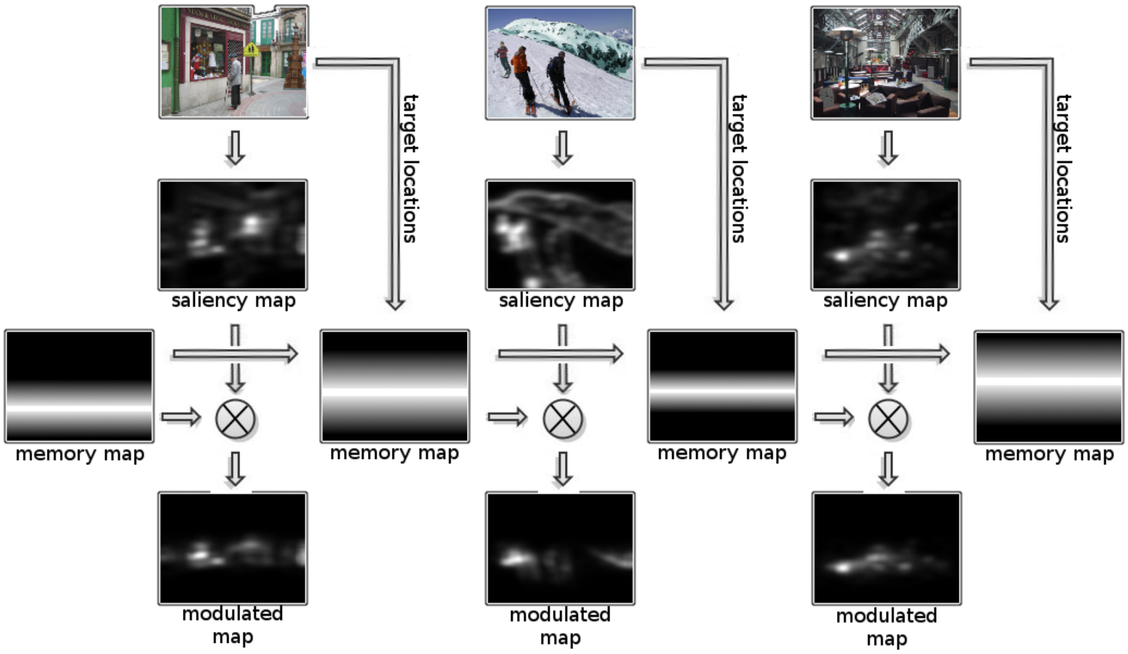

Figure 3. The computations performed by the MMS model. The incoming image is converted into a saliency map. The map is then modulated with a memory map computed based on target positions on previous images. resulting map is thresholded to select likely fixation locations.

Memory Map The top-down component of our model is implemented using memorized information, without access to image statistics or global scene representations. The MMS model learns a distribution over target object positions, and uses this distribution to modulate saliency. We make the simplifying assumption that this distribution is Gaussian.

[1] An additional simplification is that only the distribution of vertical positions is considered, while horizontal position assumed to be uniform. This is similar to an assumption made by Torralba et al. (2006).

For some images, no memory map can be estimated, because there are not enough target objects present in past scenes within allowed memory depth. This usually happens when the first few images in the experimental sequence are processed. A uniform distribution of target positions is assumed in this case.

Object Positions The position of a fixated object can be stored in memory in a number of ways. A naive choice would be to use the center of mass of the fixated object as its position. This however does not capture the fact that objects are can be large, non-homogeneous entities, with fixations not always landing on the center of mass, or several unrelated fixations falling within an object's area. Moreover, this approach would not use the information provided by saccades and fixations directly. Hence the position of an object is approximated using following rules:

[1]

1. If a fixation falls within the object area, then the object position is approximated by the fixation coordinates.

2.

3. [2]

4. If no position can be calculated using rules 1–3 then the object is assumed not to have been noticed by the participant, and thus discarded.

Once the object position has been approximated in this way, it is used to update the memorized distribution over target objects, as detailed above. After each update, the saliency map is modulated with the memory map to obtain the overall attention map.

Updates happen once per image based on the fixations on the target object in that image. If an image does not contain a target (which is the case for half of the images in the visual search data set), then no update is performed. If an image contains multiple targets, then all of them are used for the update. This situation occurs in the visual count data set.

Figure 3 shows an example of how the various maps evolve over time in our model.

Memory Depth An important questions regarding the computation of the memory map is what memory depth to use, i.e., how many previous fixations should be taken into account when estimating the distribution over target positions. In the experiments reported below, we manipulated memory depth by computing the memory map based on the three most recent fixations (MMS3), the ten most recent fixations (MMS10), or all previous fixations (MMSunrestricted). The memory depth of three was chosen as it provides a lower bound on what can be achieved by memorizing fixations locations: at least three fixation points are needed to estimate a Gaussian distribution over target locations. The memory depth of ten is based on the assumption that ten is the maximum number of fixations that can plausibly be held in human short term memory. MMSunrestricted is included to provide an upper bound on what a memory-based model can achieve. (Note that we will later also add MMSdual as a way of simulating category-specific memory.)

Note that assuming a Gaussian distribution over targets has potential limitation. People are able to capture and exploit more specific information such as the position of interesting areas or the spatial arrangement of objects (e.g., De Graef, Christiaens, & d'Ydewalle, 1990; Chun & Jiang, 1998). Additionally, memory decay effects and the distinction between long and short term memory are not modeled by the MMS, even though they have been shown to have an effect on visual tasks (e.g., Davelaar, Goshen-Gottstein, Haarmann, & Usher, 2005). As mentioned in Introduction, there is an ongoing discussion whether memory plays a role in visual search. However, it is important to note that previous studies have either been conducted on artificial stimuli (e.g., visual arrays), or focused on a particular phenomenon. Our aim, in contrast, is the more general one of investigating the role of memory as top-down supervision of low-level attentional mechanisms. For this, we believe, a simplified implementation of visual memory is sufficient.

Combined Model We also investigate an extended version of the MMS model which combines the memory map with a contextual map representing prior knowledge as it used by the CGM. The modulation map M is constructed as a simple weighted mean of the memory map MMS and a context map CO derived from the context oracle (see below for details on the context oracle). The

[2] For the visual search data, the mean size of an object is 0.93° visual angle horizontally and 1.92° visual angle vertically. For the visual count data, the mean size is 1.77° horizontally and 3.90° vertically.

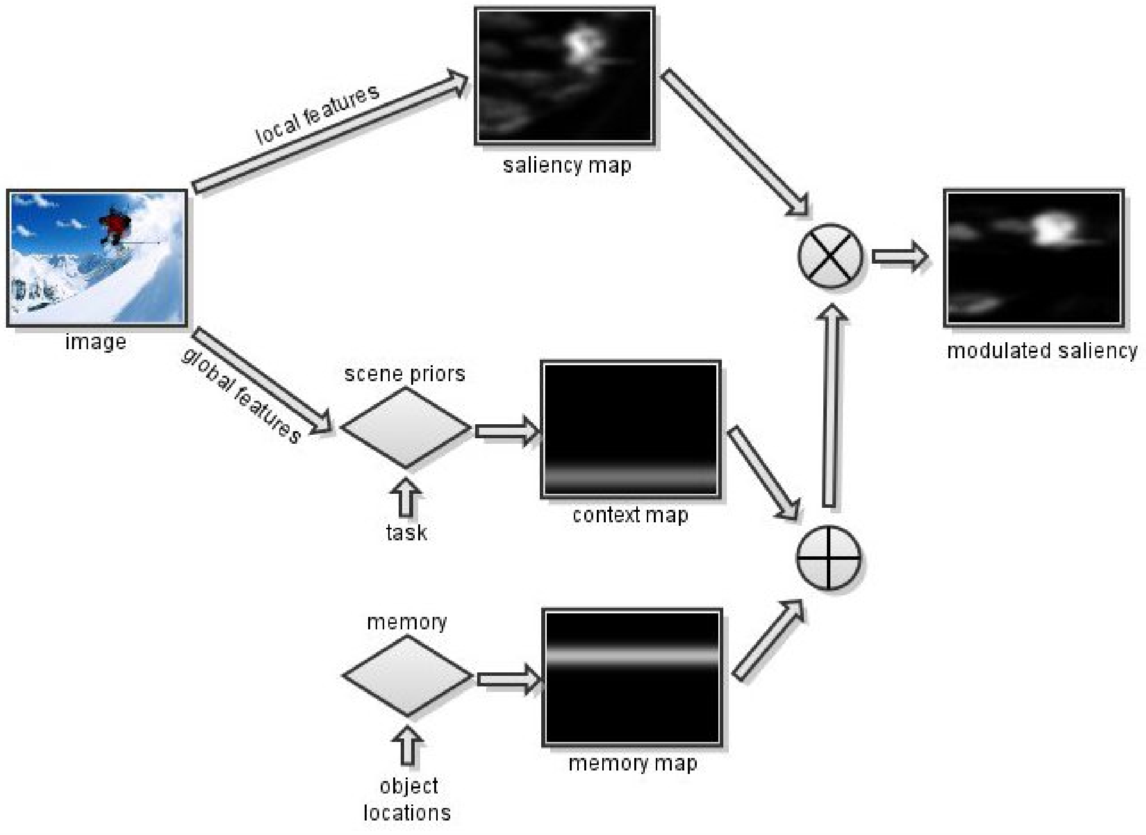

Figure 4. The architecture of the proposed joint model. First, a saliency map is computed for the image. It is then modulated with a map computed as weighted sum of the memory and context maps. The resulting map is thresholded to select the areas most likely to be fixated.

value of the resulting map at position x, y is computed as:

$$M(x,y) = \omega \cdot CO(x,y) + (1-\omega) \cdot MMS(x,y) \tag{4}$$

Here, ω is a weight parameter determining the proportions at which the maps are combined. The architecture of this model is depicted in Figure 4.

Visual Counting Experiment

Method We evaluate the performance of the MMS model on eye-tracking data collected during a visual counting task. In this task, 24 participants were asked to count the number of occurrences of a cued target object, which was either animate (e.g., man, woman) or inanimate (e.g., bin). The data set consisted of 72 photo-realistic scenes (both indoor and outdoor scenes), each containing one to three instances of the target object. The animate targets were all people, the inanimate targets were drawn from a wider range of categories; Figure 7 shows the frequency with which each target category occurred in the experiment.

A random order of the 72 scenes was generated for each participant (no blocking was used). Participants viewed each scene for as long as they liked, and then pressed one of three response

Table 1: Breakdown of missed targets by target animacy (rows) and number of targets in a scene (columns)

Cue	One target	Two targets	Three targets	Sum
Animate	4.08	0.63	0.50	5.21
Inanimate	7.19	7.07	4.11	18.37
Sum	11.27	7.70	4.61	23.58

buttons to indicate whether one, two, or three targets were present in the scene. Then the next scenes appeared; no feedback was provided.

The data was collected using a head-mounted eye-tracker with a sampling rate of 500 Hz. The images were displayed with a resolution of 1024 × 768 pixels, subtending a visual field of approximately 20 degrees.

Results and Discussion The data set consists of 54,029 fixations collected over total of 1,738 trials. The average trial length was 4.84 seconds, with a standard deviation of 3.96.

Model Evaluation

Visual Search Data In addition to the visual counting data described in the previous section, we also evaluated our model against the visual search data of Ehinger et al. (2009). In their experiment, 14 participants were asked to locate an animate target object, i.e., a pedestrian, in 912 naturalistic urban scenes, half of which contained the target. The data was collected using an eye-tracker with a sampling rate of 240 Hz, the images were displayed with a resolution of 800 × 600 pixels, subtending a visual field of about 24 × 18 degrees. This data set consists of 38,334 fixations.

CGM with Context Oracle The context oracle is based on manually annotated ground-truth maps, which were generated as follows. Participants are asked to mark on the y-axis the regions where the target object is likely to be found. Then these regions are then blurred using a Gaussian filter and aggregated over the different participants to obtain a single map for each image. We use the context oracle maps collected by Ehinger et al. (2009) for their data set, which are based on the context judgments of seven participants. For visual counting data, we generated our own context oracle maps, based on the judgments of five participants collected using the same procedure as used by Ehinger et al. (2009).

It is important to note that the can only serve as a approximate upper bound of CGM performance. It is not meant to estimate how much contextual guidance is possible in general. The context oracle is limited by the fact that each participant had to select a single y-axis location per target, even if there were multiple possible target locations. Effectively, the probability of not selected locations is estimated at zero; while this may be acceptable for some objects, it is unlikely to work for targets that can occur in a wide range of possible locations. Scene complexity is also potentially important: the assumption of a single location per scene is likely to work less well for complex scenes.

Performance Measures In the Results and Discussion section below, we compare how the different models using *receiver operating characteristic* (ROC) curves. These curves plot the true positive rate of a model (also called hit rate) against its false positive rate. Our ROC curves are computed over all fixations a participant makes on a given image, as we are interested in how well the model predicts fixations in general, not just fixations on target objects. A true positive therefore is a fixation location correctly predicted by the model, a false positive is a fixation location incorrectly predicted by the model.

The models under investigation do not assume a fixed number of fixations per image; how many fixations a model predicts for an image depends on a threshold that determines what percentage of the image is selected for evaluation.[3] As the threshold is proportional to the false positive rate of the model, we will simply plot the threshold values on the x-axis of our ROC curves. In order to statistically compare model performance, we calculate the area under the ROC curve (AUC) of each participant. The AUC measures the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one, and is equivalent to a Wilcoxon test of ranks, and closely related to the Mann-Whitney U-test (see e.g., Fawcett, 2006). We submit AUC means to an ANOVA analysis, where we compare the performance of the different models pairwise, e.g., Saliency against MMSunrestricted.

In the visual counting data set, we also test the impact of target animacy on model performance. In line with the visual cognition literature (Fletcher-Watson, Findlay, Leekam, & Benson, 2008), we expect our models to perform better on animate targets, as they are more quickly and accurately identified than inanimate targets, therefore exhibiting less variance in fixation behavior. Note that the identification of inanimate objects is also complicated by the fact that they are more variable than animate objects, both in the terms of the range of object categories they belong to, and in terms of the positions at which they can appear in the image. We will return to this point in the section *Varying Memory Depth* below (see also Figure 8).

Results and Discussion

Distribution of Fixations

Figure 5 gives histograms of the vertical coordinates of the fixations in the two data sets. The histograms show percentages of all fixations (red lines) and percentages of fixations on the target objects (green bars). We find that these distributions are similar for both of the data sets. This finding is consistent with the hypothesis that visual attention is efficiently allocated to regions which are contextually relevant.

Alternatively, Figure 5 could also be explained by a central bias for both fixations and object locations, which has been reported in the literature (Tatler, 2007). This is a point to which we will return below, when we test a baseline model which remembers a random set of fixations, rather than storing the *n* most recent fixations. The random model matches the distribution of the fixations in the data set it is trained on; it therefore has an inherent central bias and should also pick up occulomotor biases that are present in human search behavior (Tatler & Vincent, 2009). Crucially, the random model does not use information about the order of fixations and therefore can serve as realistic baseline against which to compare the MMS model, which makes use of order information (see section Random Baseline below).

[3] Thresholding works by selecting the points with the highest model values until the threshold is reached. For example, a threshold of 10% on a saliency map means that we select the points with the highest saliency until we have selected 10% of the image. We then count how many of the fixations fall within these 10%. If we select 100% of image, we trivially predict all fixations correctly.

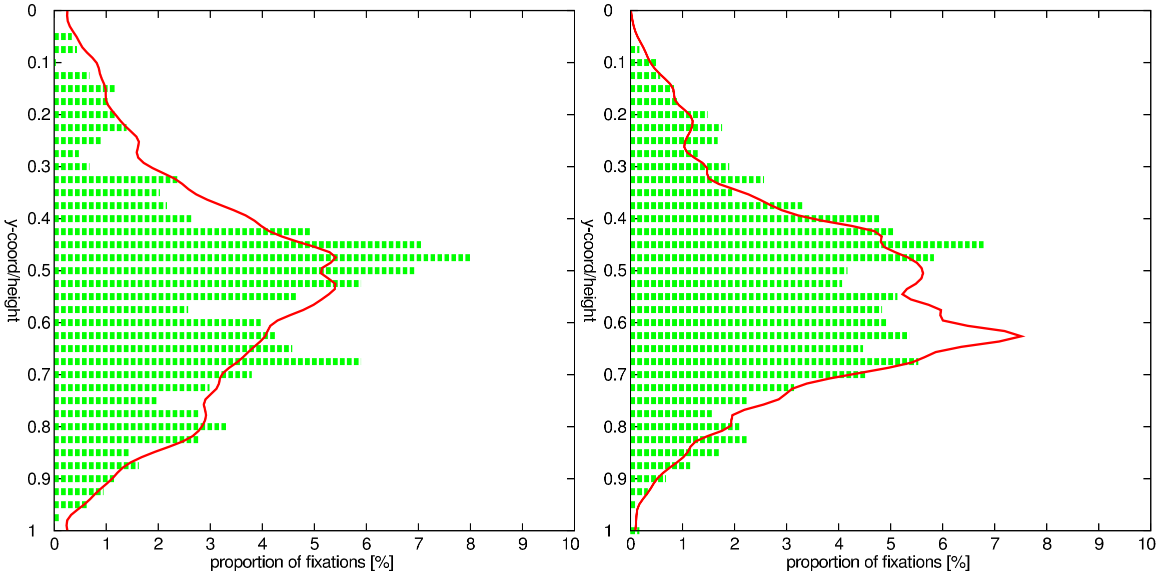

Figure 5. Histograms of vertical coordinates of fixations in visual counting (left) and visual search (right). The green bars depict percentages of fixations on the target objects; the red line shows percentages of all fixations.

When we plot horizontal fixation positions for the visual counting data set (see Figure 6, left panel), we find a uniform distribution, which means that there is no general central bias for horizontal positions in this data set. For the visual search data, we find a bimodal distributions of horizontal positions, rather than a central bias (see Figure 6, right panel). The bimodality is an artifact of the experimental design which underlies this data set.[4]

Varying Memory Depth

Figures 9 and 10 show the ROC curves obtained by the different models for the two data sets. Overall, we find that the MMS models have a higher *hit rate*, i.e., proportion of fixations on target areas, than saliency in both data sets. This finding confirms that top-down knowledge is fundamental for model performance in goal-directed tasks, such as search. Crucially, we observe that even MMS models with small memory perform better than saliency.

In order to confirm this visual impression, we performed an ANOVA comparing the area under the RUC curve of the saliency-only model with the area under the curve of the MMS models (the AUC values are averaged over participants for both data sets, so the degrees of freedom are derived from the number of participants). The AUC values are summarized in Table 2.

[4] Ehinger et al. (2009) designed their stimuli as follows:

> For the target-present images, targets were spatially distributed across the image periphery (target locations ranged from 2.7° to 13° from the screen centre; median eccentricity was 8.6°), and were located in each quadrant of the screen with approximately equal frequency. (Ehinger et al., 2009, p. 950)

The fact that the authors placed target deliberately at the screen periphery explains the bimodality of horizontal positions in Figure 6 (right panel). There is only a weak bimodality in vertical positions in Figure 5 (right panel), which is probably due the fact that their target objects (which were always pedestrians) show a central bias vertically, which presumably counteracts the peripheral bias in the stimulus design.

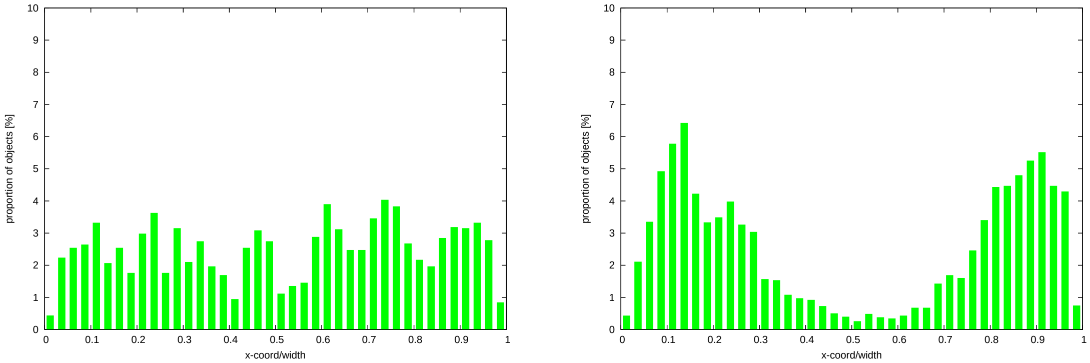

Figure 6. Histograms of horizontal coordinates of fixations in visual counting (left) and visual search (right).

Table 2: Performance of the models on the visual counting and visual search data sets. Given is the area under the ROC curve, averaged over participants (the table lists means and standard deviations).

Model	Visual search	Visual count
Saliency	75.33±1.10	80.91±1.68
MMS3	77.18±0.72	81.55±1.50
MMS10	79.60±0.89	83.22±1.47
MMSunrestricted	82.89±0.82	83.78±1.52
CGM	82.67±1.01	83.19±1.64
Random3	70.61±0.83	78.54±2.00
Random10	75.21±1.10	82.65±1.61

For the visual search data set, we found a significantly larger area under curve for MMS3, the MMS model with a memory depth of three fixations, when compared to saliency ($F(1,13) = 27.8$, $p < 0.0001$). Also MMS10, with a memory depth of ten fixations, outperformed saliency ($F(1,13) = 192.8$, $p < 0.0001$). We obtained similar results for the visual counting data, where the area under the curve was not significantly different between saliency and the MMS3 model ($F(1,24) = 2.0$, $p > 0.1$), but it was larger for MMS10 compared to saliency ($F(1,24) = 26.6$, $p < 0.0001$).

The difference observed between the two data sets is due to the larger variability in the visual counting task. The counting task used both animate and inanimate targets, while the search task only used one specific type of animate target (i.e., pedestrians). Furthermore, in the counting task, most animate objects belong to three frequent object categories, while inanimate objects belong to a larger number of categories, each of which only occurs once or twice (see Figure 7). It is also the case that animate objects are often located at the center and bottom part of the image, e.g., a pedestrian on a cross-walk, whereas inanimate objects can be found at a wide range of locations, see Figure 8. This source of variation is not present in the search data (compare Figure 8 to Figure 5). Moreover, the possibility of having multiple target causes participants to inspect the scene longer than during the

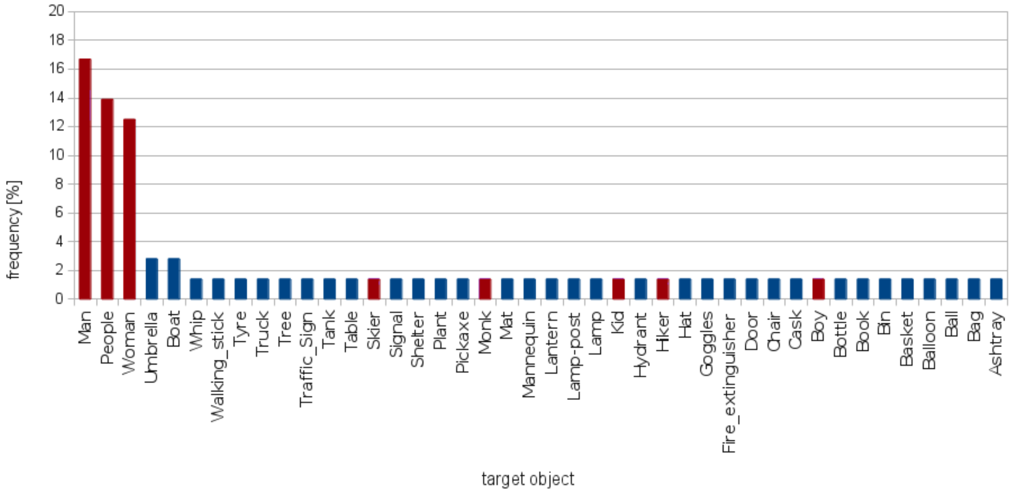

Figure 7. Frequency of different targets in the visual counting task. Marked red are animate objects while inanimate objects are blue. Note that most animate objects belong to just three categories, which while for inanimate objects are distributed over a larger number of infrequent categories.

search task, which again increases the variability of visual responses.

When comparing the MMS models with the (i.e., the approximation of an upper bound of the performance of the CGM), we find that only MMSunrestricted, i.e., the memory model using all available fixations, is better than the , and only on the visual search data set ($F(1,13) = 5.4$, $p = 0.02$). We observe an improvement on the visual counting data set when we assume separate memories for animate and inanimate objects, i.e., MMSdual (to be discussed in more detail below). The performance of MMSdual is not statistically different from that of the ($F(1,24) = 2.9$, $p > 0.09$). Any model with a smaller memory performs worse than the on both data sets.

We repeated the evaluation using the position of the center of the mass of the target objects. In this analysis, the MMS did not memorize the fixation positions directly, but instead we computed the center of mass of the fixated object, and used this as the target position for the MMS to memorize. This analysis was meant to simulate a situation in which no fixation data is available to the model, and it instead has to rely on object positions, just as the CGM does during training time.

This analysis revealed no difference in performance for visual count data. In the case of the visual search data, a difference was only observed for the smallest memory size, where using center of mass led to significantly improved performance (about 1.5% increase in AUC, $F(1,13) = 32.84$, $p < 0.0001$). Presumably, fixation data at a memory depth of three is fairly noisy, and this noise is smoothed out by using the center of mass of objects, rather than the fixation data directly.

More generally, the lack of a significant difference in most conditions between models using real fixation locations and the centers of mass means that the MMS does not have to rely on fixation data in order to update the memory. On more theoretical level, this result supports the finding of Nuthmann and Henderson (2010), who show that the *preferred viewing location* of an object is close to its center of mass in naturalistic scenes; this is in turn predicted by the *cognitive relevance hypothesis* of Henderson, Brockmole, and Castelhano (2007) and Henderson et al. (2009).

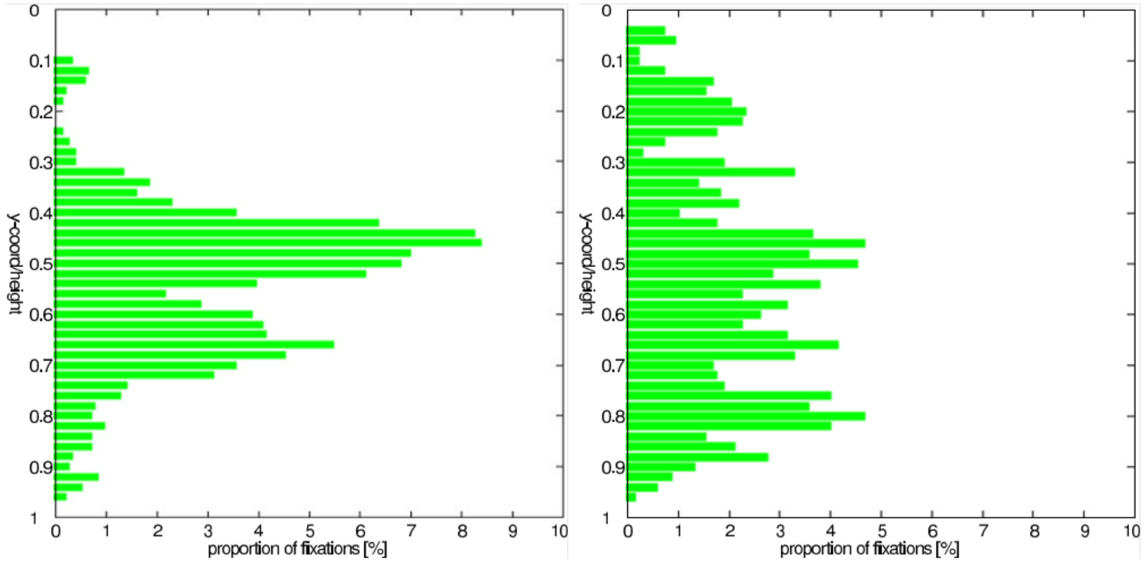

Figure 8. Distribution of vertical locations for animate (left) and inanimate (right) targets on the visual counting data. Animate targets are usually located at between half and two thirds of the image height, while inanimate objects are distributed more evenly across the image height.

Random Baseline

In order to provide a baseline against which to compare the MMS model, we also tested a version of the model that does not remember the n previous fixations, but a set of n randomly chosen fixations. This means that the baseline model does not have access information about the order in which the fixations occurred, but should capture general biases in fixation behavior and target locations, such as the central bias observed in the data (see Figure 5).

The baseline model was implemented by randomly scrambling the order of the fixations on the target objects. In doing so, we preserved the number of fixations per image, and fixations were randomized only within participants. The scrambled data set therefore has the same overall distribution of fixation locations as the original data set, and the same number of fixations is used to compute the memory map for each person/image combination.

We computed two version of the random baseline model, Random3, which with a memory depth of three, and Random10, which a memory depth of ten (if we were to assume unrestricted memory then the random baseline would be identical to MMSunrestricted). The AUC values for these models for both the search and the counting data sets are displayed in Table 2. For the visual search data, we find that Random3 performs significantly worse than saliency alone, the worst predictor of fixation locations ($F(1,13) = 97.33$, $p < 0.001$), while Random10 is not significantly different from saliency ($F(1,13) = 0.1$, $p = 0.754$). On the visual count data, we again find that Random3 is significantly worse than saliency ($F(1,24) = 21.86$, $p < 0.001$), while Random10 is significantly better than saliency ($F(1,24) = 13.65$, $p < 0.001$) but significantly worse than the corresponding memory-based model MMS10 ($F(1,24) = 6.43$, $p = 0.0146$).

This set of results indicates that the performance of the MMS can not be attributed to general biases in fixation behavior and target locations, but is driven by the fact that the MMS uses the locations of the most recent target fixations to predict the location of the current fixation.

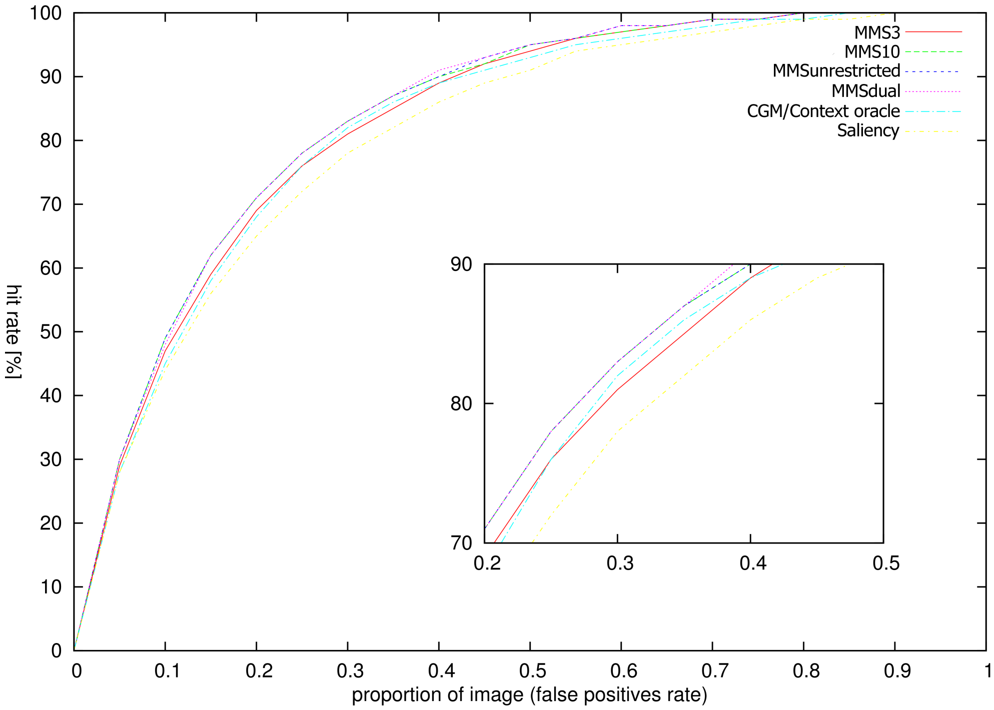

Figure 9. Prediction performance for the visual counting task for MMS with memory of three, ten, and an unrestricted number of fixations (MMS3, MMS10 and MMSunrestricted), MMS with a separate memory for animate and inanimate objects (MMSdual), the approximation of a CGM performance upper bound (CGM with context oracle), and the Saliency baseline. The curve is an ROC curve which plots true positives (hit rate) against false positives (proportion of image selected by the model). The red line marks the 20% threshold used by Torralba et al. (2006) in their evaluation.

Dual Model and Combined Model

Given the differences between animate and inanimate objects in terms of their typical location and their distribution over categories, it makes sense to consider evaluate model performance on animate and inanimate targets separately. Table 3 provides the relevant AUC values. We observe that all models have a better performance on animate targets than on inanimate ones ($F(1,24) = 40.8$, $p < 0.0001$). This motivates the introduction of a dual memory version of the MMS model, which maintains separate memory maps for animate and inanimate objects,

This model (MMSdual) improves performance compared to ($F(1,24) = 3.9$, $p = 0.05$). MMSdual fails to outperform an MMS with unrestricted memory ($F(1,24) = 0.7$, $p = 0.39$). While this result is encouraging, it also raises questions regarding the level of granularity that is appropriate for category specific memories. It is possible that the animate/inanimate distinction needs to be refined further, for example suing subdivisions such as human/animal for animate and artifact/natural object for inanimate. It seems plausible to assume that the MMS tracks a small number of object types and keeps separate memory maps for each of them. Furthermore, the granularity of the object types may be task-dependent. This is an issue that should be addressed in future research.

An analysis of the fixations generated by the MMS model and by the CGM reveals that both models tend to predict fixations at different locations, in spite of their similar overall performance.

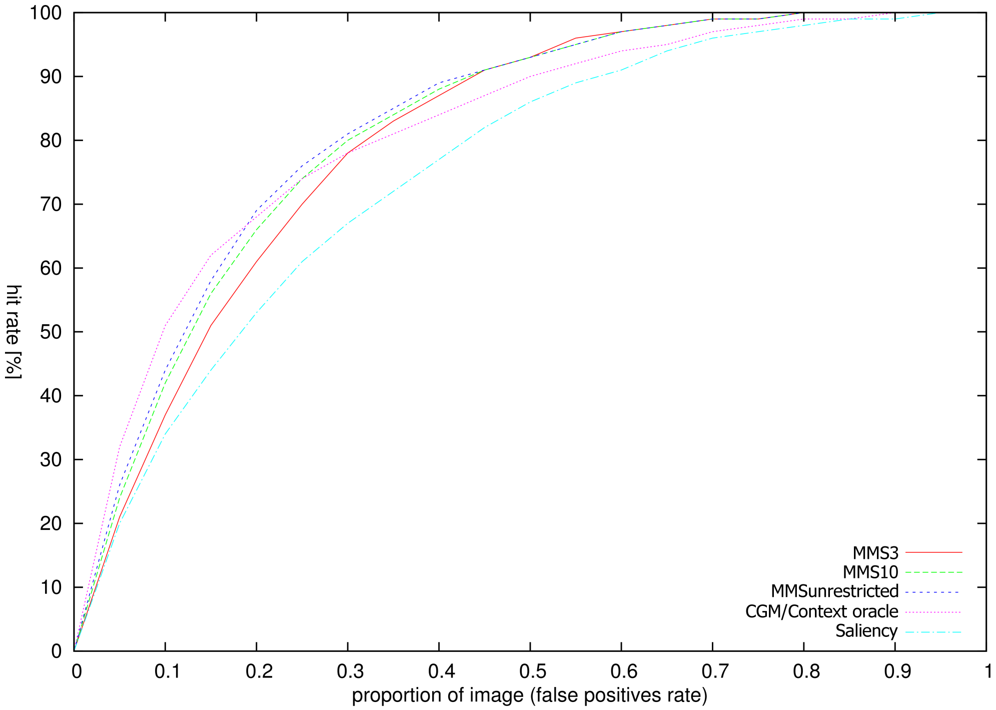

Figure 10. Prediction performance for the visual search task for MMS with memory of three, ten, and an unrestricted number of fixations (MMS3, MMS10 and MMSunrestricted), MMS with a separate memory for animate and inanimate objects (MMSdual), the approximation of a CGM performance upper bound (CGM with context oracle), and the Saliency baseline. The curve is an ROC curve which plots true positives (hit rate) against false positives (proportion of image selected by the model). The red line marks the 20% threshold used by Torralba et al. (2006) in their evaluation.

Figure 11 presents the overlap o of predictions calculated as fraction of fixations found by both models for various threshold sizes:

$$o = \frac{|predicted(MMS) \cap predicted(CO)|}{|predicted(MMS) \cup predicted(CO)|} \tag{5}$$

where $predicted(\cdot)$ denotes set of fixations correctly predicted for a given model (MMS or CGM), and $|\cdot|$ denotes set cardinality.

The relatively low overlap of the predictions for both models for smaller threshold values suggests that combining them should be beneficial. Indeed, we found that the simple joint model described earlier (see Figure 4 and equation (4)) section improves AUC values. The benefit of using a combined model is clear in the case of the visual search data, on which it achieves an AUC value of 86.01 (SD = 1.33) for a weight of $\omega = 0.6$. This AUC value is significantly better than that of the MMS model alone ($F(1,13) = 55.20$, $p < 0.0001$). In the case of the visual count data, the combined model achieves an AUC value of 84.26 (SD = 1.48) for $\omega = 0.4$, which however is not significantly different from the AUC value of the MMS model alone ($F(1,24) = 1.18$, $p = 0.28$). This can be explained by the fact that the overlap ratio for the two models is higher for the visual count data.

Table 3: The performance of the proposed models split by animacy of the target objects for the visual counting task. Given is the area under the ROC curve, averaged over participants (the table lists means and standard deviations).

Model	Animate	Inanimate	All
Saliency	81.16±1.58	80.67±2.23	80.91±1.68
MMSdual	84.74±1.23	82.92±1.95	83.83±1.38
MMS10	84.61±1.51	81.84±1.90	83.22±1.47
MMSunrestricted	85.13±1.44	82.43±1.98	83.78±1.52

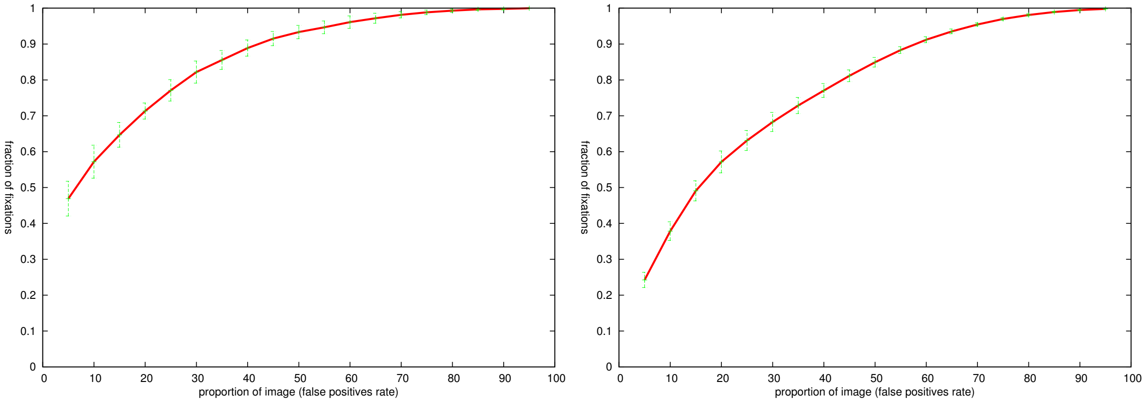

Figure 11. Overlap of fixations locations generated by MMS and CGM with context oracle calculated as the number of fixations found by both models over the total number of fixations predicted. Visual count data on the left, visual search data on the right. Note that this is not an ROC curve; rather, we plot the overlap of the models against the false positive rate.

Overall, our results demonstrate that a simple model of visual search based on the memory of previous fixations can perform as well as, if not better, than a more complex model such as the CGM, which integrates bottom-up saliency with context information conditioned on global scene features.

It is also important to note that MMS model performance does not degrade on a visual count data set consisting of different scenes with a wide range of visual contexts. Instead, the MMS model still performs better than saliency and comparable to the CGM on this data set. Moreover, we have shown that it is beneficial to combine both sources of knowledge: a model that includes prior contextual knowledge and memory of fixation locations showed improved performance, at least for the visual search data.

Conclusions

We presented a computational model that predicts fixation locations in visual search. Our approach is conceptually similar to the Contextual Guidance Model of Torralba et al. (2006), which combines saliency with scene gist and top-down context information about likely target positions. To obtain the context information, the CGM is trained offline on a large set of images with manually provided object labels. The Memory Modulated Saliency model that we propose, on the other hand,

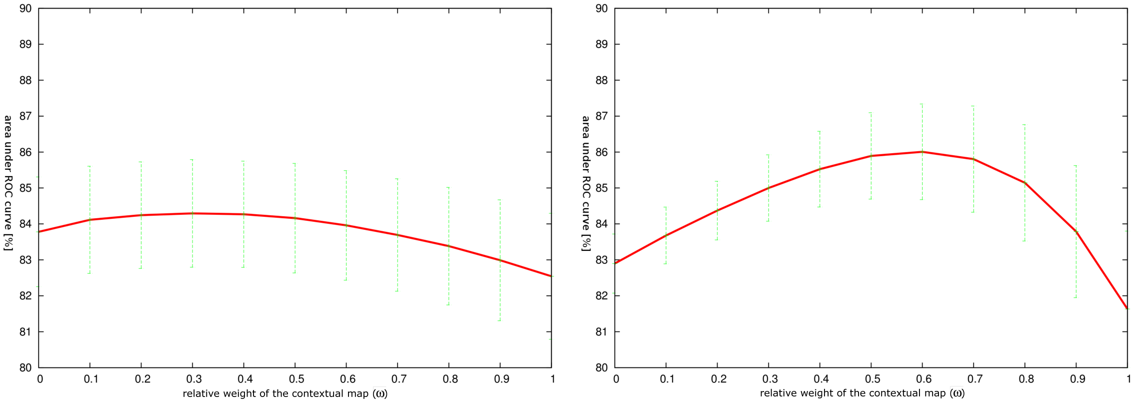

Figure 12. Performance of joint MMS/CGM model on visual count (left) and visual search (right) data. Note that this is not an ROC curve; rather, we plot the AUC achieved by the joint model against the relative weight ω of the contextual maps predicted by the two models used to modulate saliency.

does not require a separate training phase and does not involve the calculation of image or scene statistics. Instead, the MMS model keeps the last few fixations the participant made in memory, and uses them to predict likely positions of target objects.

The MMS model performs significantly better than saliency on two experimental data sets, demonstrating the benefit of memory for the prediction of fixation locations. An MMS model with unrestricted memory outperforms the (an approximation of an upper bound on CGM performance) on visual search data, and achieves equal performance on visual count data.

We also investigated whether a memory-based model needs to have access to fixation coordinates. We tested this by replacing the fixation coordinates with the coordinates of the center of mass of the fixated objects. We found that the performance of the resulting model is the same as that of the original version of the MMS that uses fixation coordinates. This means that fixations are not central to the model, they can be eliminated from it without degrading performance. All that the model requires is knowledge of fixated objects and their positions. It is therefore conceivable that the MMS can be trained in an offline fashion using images with object annotations, similar to the CGM, though the details of such a training scheme remain to be worked out.

We also demonstrated that a combined model that uses a weighted sum of the MMS memory map and the CGM context map outperforms the both individual models on the visual search data. This result indicates that a complete model of attentional guidance needs to combine features of both models. One important aspect of the CGM that is not present in the MMS is scene gist. In the CGM, the salience of a location in an image is conditioned on the scene gist (see equation (2)). It seems likely that integrating gist would also be beneficial to the MMS: fixation locations (or alternatively, center of mass points) could be conditioned on gist in the same way. We leave this as an issue for future research.

Another potential limitation of the MMS model is that it requires a predictable, serial trial structure. It seems likely that memorizing fixation locations will only work if all the trials in a experiment are similar to each other (e.g., they are all search trials, or all counting trials, as in the two experimental data sets we tested). In an experiment in which different types of trials alternate, perhaps in an unpredictable fashion, having a memory of fixation locations in immediately preceding trials is likely to be less useful. However, it is conceivable that fixations memory could be conditioned on trial type, or that separate memories for different trial types could be maintained to solve this problem. (This would be similar to the dual memory model that stores animate and inanimate targets separately.)

We also found that the dual memory model, which stores the locations of animate and inanimate objects separately, outperforms a model with just one type of memory. If we assume that animate and inanimate objects differ in the their typical location in the scene, then storing their locations separately provides a restricted form of contextual guidance. It in conceivable to extend this approach and introduce separate memories for a larger number of categories. How many object categories are required is likely to be task specific, so a category-aware version of the MMS potentially should also include a way of learning which (and how many) categories need to be distinguished in a given task. Perhaps this learning could happen in an offline training phase just as in the CGM. This would then provide a less ad-hoc way of integrating the two models, a clear improvement over our combined model, which simply computes the weighted sum of the memory map and the context map.

An important conceptual difference between the two models is the type of learning that they capture. The CGM, by assuming an offline learning phase for target locations based on a large training set, effectively models how humans learn the typical positions of objects during childhood (or even beyond that for novel objects). The MMS, on the other hand, models short-term learning as it happens while human perform a specific task, and learn target locations based on where the targets where a few fixations ago. It seems that human behavior is driven by both types of learning; the two models should therefore be seen as complementary, pointing again towards the need for an model that integrates components from both the CGM and the MMS.

On a more theoretical level, our results provide an alternative explanation for the tendency of experimental participants to only fixate contextually appropriate regions. In addition to using prior context information, participants seem to memorize likely target locations from previous trials, and use this information to guide their search on the current trial.

References

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*, 617–629.

Brockmole, J., Castelhano, M., & Henderson, J. (2006). Contextual cueing in naturalistic scenes: Global and local contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 699-706.

Brockmole, J., & Henderson, J. (2006a). Recognition and attention guidance during contextual cueing in real-world scenes: Evidence from eye movements. *Journal of Experimental Psychology*, *59*, 1177-1187.

Brockmole, J., & Henderson, J. (2006b). Using real-world scenes as contextual cues for search. *Visual Cognition*, *13*, 99–108.

Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. In *Advances in neural information processing systems 18* (pp. 155–162). Cambridge, MA: MIT Press.

Castelhano, M., & Henderson, J. (2007). Initial scene representation facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 753–763.

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

Chun, M., & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science*, *10*, 360–365.

Davelaar, E. J., Goshen-Gottstein, Y., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: empirical and computational investigation of recency effects. *Psychological Review*, *112*(1), 3–42.

De Graef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research*, *52*, 317–329.

Dziemianko, M., Keller, F., & Coco, M. (2011). Incremental learning of target locations in visual search. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Ehinger, K., 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*(6/7), 945–978.

Einhauser, W., Spain, M., & Perona, P. (2008, 11). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14), 1-26.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, *22*, 861-874.

Fletcher-Watson, S., Findlay, J., Leekam, S., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception*, *37*(4), 571–583.

Gibson, B., Li, L., Skow, E., Brown, K., & Cooke, L. (2000). Searching for one versus two identical targets: When visual search has memory. *Psychological Science*, *11*, 324–327.

Green, C., & Hummel, J. (2006). Familiar interacting object pairs are perceptually grouped. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 1107–1119.

Hayhoe, M., Shrivastava, A., Mruczek, R., & Pelz, J. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, *3*, 49–63.

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

Henderson, J., Brockmole, J., & Castelhano, M. (2007). Visual saliency does not account for eye-movements during visual search in real-world scenes. *Eye movements research: insights into mind and brain*.

Henderson, J., Malcolm, G., & Schandl, C. (2009). Searching in dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*, 850–856.

Hollingworth, A. (2006). Visual memory for natural scenes: Evidence from change detection and visual search. *Visual Cognition*(14), 781-807.

Horowitz, T., & Wolfe, J. (1998). Visual search has no memory. *Nature*, *394*(6), 575–577.

Hwang, A. D., Wang, H.-C., & Pomplun, M. (2011). Semantic guidance of eye movements during real-world scene inspection. *Vision Research*, *51*(10), 1192–1205.

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, *12*, 1093–1123.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Klein, R. (1988). Inhibitory tagging system facilitates visual search. *Nature*, *334*, 430–431.

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

Koostra, G., Nedereen, A., & De Boer, B. (2008). Paying attention to symmetry. In *Proceedings of british machine vision conference.* Leeds, UK: British Machine Vision Association.

Land, M., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*, 3559–3565.

Maljkovic, V., & Martini, P. (2005). Implicit short-term memory and event frequency effects in visual search. *Vision Research*, *45*(21), 2831–2846.

McPeek, R. M., Maljkovic, V., & Nakayama, K. (1999). Saccades require focal attention and are facilitated by a short-term memory system. *Vision Research*, *39*(8), 1555–1566.

Nuthmann, A., & Henderson, J. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *10*(8).

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*(12), 520–527.

Pelz, J., & Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research*, *41*, 3587–3596.

Peters, R., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, *45*, 2397–2416.

Renninger, L., & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research*, *44*, 2301-2311.

Rensink, R. (2000). The dynamic representation of scenes. *Vision Cognition*, *1/2/3*(7), 17–42.

Schyns, P., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, *5*, 195–200.

Shore, D., & Klein, R. (2000). On the manifestations of memory in visual search. *Spatial Vision*, *14*, 59–75.

Takeda, Y., & Yagi, A. (2000). Inhibitory tagging in visual search can be found if search stimuli remain visible. *Perception and Psychophysics*, *62*, 927–934.

Tatler, B. W. (2007). The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14), 1–17.

Tatler, B. W., & Vincent, B. T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition*, *17*(6–7), 1029–1054.

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

Wolfe, J. M. (1999). Inattentional amnesia. In V. Coltheart (Ed.), *Fleeting memories* (pp. 71–94). Cambridge, MA: MIT Press.

Wolfe, J. M., & Bennett, S. C. (1997). Preattentive object files: Shapeless bundles of basic features. *Vision Research*, *37*, 25–44.

Wolfe, J. M., Klempen, N., & Dahlen, K. (2000). Postattentive vision. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 693–716.

Woodman, G. F., & Chun, M. M. (2006). The role of working memory and long-term memory in visual search. *Visual Cognition*, *14*, 808–830.

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

Zhang, L., Tong, M., Marks, T., Shan, H., & Cottrell, G. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, *32*(8(7)), 1-20.