Incremental Learning of Target Locations in Visual Search

Michal Dziemianko (m.dziemianko@sms.ed.ac.uk)
Frank Keller (keller@inf.ed.ac.uk)
Moreno I. Coco (m.i.coco@sms.ed.ac.uk)
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, 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 the Contextual Guidance Model can achieve (CGM, Torralba et al. 2006; Ehinger et al. 2009), even though our model does not perform scene recognition or compute global image statistics.

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 their spatial relations. 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., Einhauser et al., 2008) have shown that scan patterns on visual scenes can vary greatly between subjects. On the other hand, the task that participants have to perform has an influence on visual attention, resulting in fixated regions being relatively consistent across participants for the same experimental conditions (e.g., Torralba et al., 2006).

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. A number of 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 in case of the visual input that does not provide any useful high-level information (e.g., Peters et al., 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 et al. (1998), which builds saliency maps based on color, orientation, and scale filters inspired by neurobiological results.

The second group of models assume the existence of top-down supervision of attention which contributes to the selection of fixation targets. Various types of such supervision have been observed experimentally. Humans show the ability to learn general statistics of the appearance, position, size, spatial arrangement of objects, and their relationships (e.g., Zelinsky, 2008). They also exploit visual memory during scene comprehension tasks (e.g., Shore & Klein, 2000). Moreover, studies such as those of Chun & Jiang (1998) show that participants benefit from learning spatial arrangement of the objects in consecutive searches.

A series of studies have also shown the importance of context in scene comprehension. Context not only provides information about scene layout and type (Schyns & Oliva, 1994; Renninger & Malik, 2004), but also about object presence, location, and appearance (e.g., Oliva & Torralba, 2007; Bar, 2004). 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, which combines bottom-up saliency with a prior encoding global scene information. A detailed description of the Contextual Guidance Model can be found in the Background section below.

In this paper, we introduce 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. This avoids not only an expensive training phase, but also enables fast adaptation to different data sets, tasks, and experimental conditions.

Background

The Contextual Guidance Model (Torralba et al. 2006) combines saliency with a model of global scene information (gist) in a Bayesian fashion. The central quantity computed by the CGM is the probability that a target object O is present at

In Laura Carlson, Christoph Hoelscher, and Thomas F. Shipley, eds., *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 1729–1734. Boston, 2011.

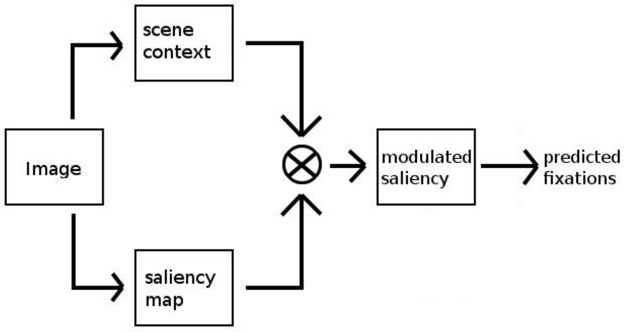

Figure 1: The architecture of the CGM. First, a saliency map is computed for the image. It is then modulated with a contextual prior conditioned on global scene features. The resulting map is thresholded to select the areas most likely to be fixated.

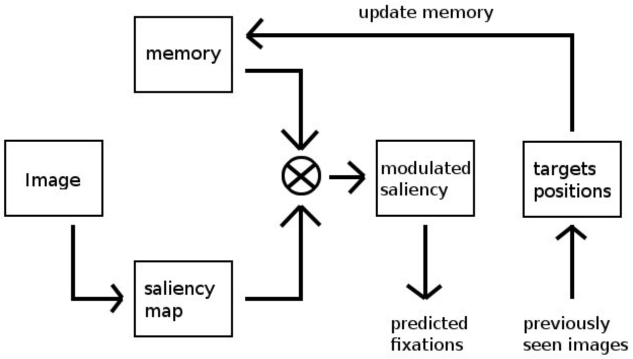

Figure 2: The architecture of the proposed MMS model. First, a saliency map is computed for the image. It is then modulated with a memory map estimated using fixations landing within the targets on previously seen images. The resulting map is thresholded to select the areas most likely to be fixated.

point X in the image:

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

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 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, the second and the forth terms are omitted, yielding:

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

This 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.

Model

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

As in the CGM, we approximate saliency as the probability of the local images feature L in a given location based on the global distribution of these features:

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

Here, with 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 computed as a set of Gabor filter responses computed over three color channels for six orientations and four scales, totalling 72 values at each position.

The contextual component of our model is based on memorized information, without access to image statistics or global scene representations. The MMS model learns a distribution of target objects 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 horizontal positions is considered, while vertical position assumed to be uniform. This is similar to an assumption made by Torralba et al. (2006).

Even with these simplifying assumptions, the formulation of the model is still challenging. The main issue is the memory depth, i.e., the number of previous images that are taken into account during the estimation of the target distribution. Moreover, the Gaussian distribution assumed can only capture the mean position of the targets. People are able to capture and exploit more specific information such as position of interesting areas or the spatial arrangement of objects (e.g., De Graef et al., 1990; Chun & Jiang, 1998). Additionally, memory decay effects and limited size of short term memory are not modelled by the MMS, even though they have shown to have an effect on visual tasks (e.g., Davelaar et al., 2005). Although there is ongoing discussion whether memory is present in visual search (see, e.g., Horowitz & Wolfe,

[1] Although the histograms of target positions (see Figure 4) suggest that a simple mixture of Gaussians may be worth investigating.

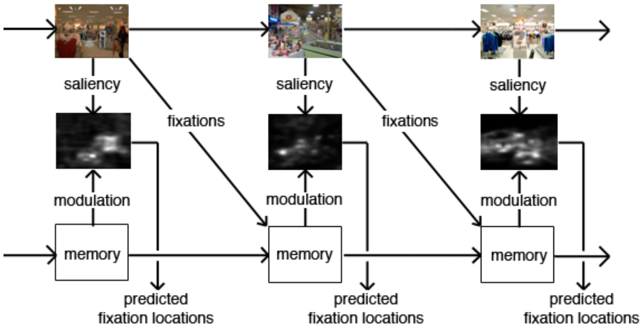

Figure 3: The process performed by the MMS model. The incoming image is converted into a saliency map. The map is then modulated with a bound calculated based on memorized target fixations. The resulting map is thresholded to select likely fixation locations.

1998; Hollingworth, 2006) and our assumptions are not entirely consistent with current theories of memory, we believe they are sufficient, as previous studies have either been conducted on artificial stimuli, or focused on a particular phenomenon rather than investigated memory as the top-down supervision of low level attentional mechanisms.

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 saliency map contains the model prediction for the fixation locations. The distribution of the target objects for the first images in the sequence is assumed to be uniform. In the second step, the positions of target objects found by the participant are estimated. As the position of a target object can be specified in different ways, this step requires more detailed explanation. A naive choice would be to use the center of mass of the object as its position. This however does not capture the fact that objects are often relatively large, non-homogeneous entities, and several unrelated fixations can fall within their area. Moreover, this would not use the information provided by saccades and fixations directly. Hence the position of the object is approximated using following rules:

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

2. If more than one fixation falls into the object area, than only the first one is taken into account, the other ones are discarded.

3. If no fixations fall within the object area, then the fixations within one degree of visual field are considered (with rules 1 and 2 modified appropriately).

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.

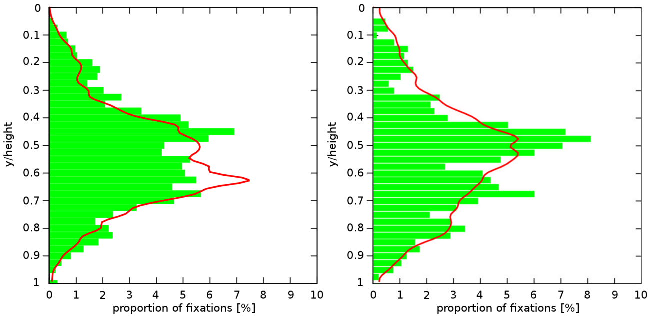

Figure 4: Histograms of vertical coordinates of fixations in visual search (left) and visual counting (right). The green bars depict percentages of fixations on the target objects; the red line shows percentages of all fixations.

The approximated target objects positions are used to update the memorized distribution of the positions, which is then used to modulate saliency for consecutive images. In the evaluation, we will also consider a variation of the model with separate memories for animate and inanimate targets (dual memory MMS), as there is some evidence that animacy affects visual attention (Fletcher-Watson et al., 2008; Coco & Keller, 2009).

Evaluation Experiments

Method

We evaluate the performance of the MMS model on eye-tracking data collected during a visual counting task. In this task, 25 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 in 72 photo-realistic scenes (both indoor and outdoor), containing zero to three instances of the target object. 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 34 × 30 degrees. The data set consists of 54,029 fixations. Moreover, in order to directly compare the performance of our model with the CGM, we used the data set collected by Ehinger et al. (2009) in a visual search task, where 14 participants were asked to locate an animate target object, i.e., a pedestrian, in 912 naturalistic urban scenes, half of which containing 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.

Figure 4 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 datasets. This finding confirms the hypothesis that visual attention is efficiently allocated to regions which are contextually relevant for the task at hand.

Analysis

We evaluate the performance of the MMS model against a simple saliency model and a context oracle, which Ehinger et al. (2009) suggest to be the upper bound of what can be achieved with a context-based model such as the CGM. A context oracle is created by using manually annotated ground-truth maps. Human participants are asked to mark on the y-axis the regions where the target object is likely to be found. Then, these regions are blurred using a Gaussian filter, and aggregated over the different participants to obtain a single map for each image.[2]

In the Results and Discussion section below, we show how the different models perform by using **receiver operating characteristic** (ROC) plots, which indicate the sensitivity (i.e., true positive rate vs. false positive rate) of a classifier as its discrimination threshold varies. Moreover, 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.[3] 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 et al., 2008), we expect models to perform better on animate targets, as they are more quickly and efficiently identified than inanimate targets. The identification of inanimate targets is further complicated by their larger contextual variability (all animate object were of type person).

Results and Discussion

Figures 5 and 6 show the ROC curves obtained by the different models for the two data sets. Overall, we find that 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 MMS models with small memory perform better than saliency, especially in the visual search data set, where we find that both MMS3 ($F(1,13) = 27.8, p < 0.0001$) and MMS10 ($F(1,13) = 192.8, p < 0.0001$) perform better than saliency. We obtain similar results for the visual counting data, where MMS3 is not significantly different from saliency ($F(1,24) = 2.0, p > 0.1$), but MMS with a memory of 10 fixations outperforms 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, which is introduced by both the animacy of the targets and the variable number of target occurrences per scene. Animate objects are often located at the bottom of the image, e.g., a pedestrian on the cross-walk, whereas inanimate objects can be found at a wide range of locations. Moreover, the possibility of having more than a single target causes participants to inspect the scene longer, which increases the variability of visual responses.

Table 1: The performance of the proposed models with respect to the animacy of the target objects for visual counting task expressed as mean percentage and standard deviation of the area under the ROC curve for each experimental subject.

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
MMSunrestricted	85.13±1.44	82.43±1.98	83.78±1.52
MMS10	84.61±1.51	81.84±1.90	83.22±1.47

When comparing the MMS models with the context oracle (i.e., the 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 context oracle, 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 separate memories for animate and inanimate objects, i.e., MMSdual, however the difference with context oracle fails to reach significance ($F(1,24) = 2.9, p > 0.09$). Any model with smaller memory performs worse than the context oracle on both data sets.

As argued above, the difference in model performance observed for the two data sets is due to the nature of the task, as well as the variability of targets, both in terms of their animacy and the number of occurrences in the scene. In the visual search task, only a few fixations are needed to ascertain the presence or absence of the single animate target. In visual counting, however, up to three targets can be present, which results in longer and more widely distributed fixations. Furthermore, the variability of visual responses in the visual counting task is increased by the use of both animate and inanimate targets. Animate targets are more quickly identified than inanimate ones (Fletcher-Watson et al., 2008). Moreover, inanimate targets have larger contextual variability, as all animate objects belonged to the same object class (i.e., person), which was not the case for inanimate objects. Members of different object classes vary more in their contextual associations and hence their likely locations.

These intuitions are confirmed when comparing at the performance of the different models for animate and inanimate targets in the visual counting task; see Table 1 for AUC values. We observe that all models have a better performance on animate targets than on inanimate ones ($F(1,24) = 40.8, p < 0.0001$). The introduction of a dual memory improves the performance when compared to a model with memory of 10 fixations ($F(1,24) = 3.9, p = 0.05$), but is not sufficient to outperform MMS with unrestricted memory ($F(1,24) = 0.7, p = 0.39$). Further investigation with less types of inanimate ob-

[2] The context oracle information of Ehinger et al. (2009) was obtained from seven participants; for our data set we used five participants.

[3] The AUC is equivalent to a Wilcoxon test of ranks, and closely related to the Mann-Whitney U-test.

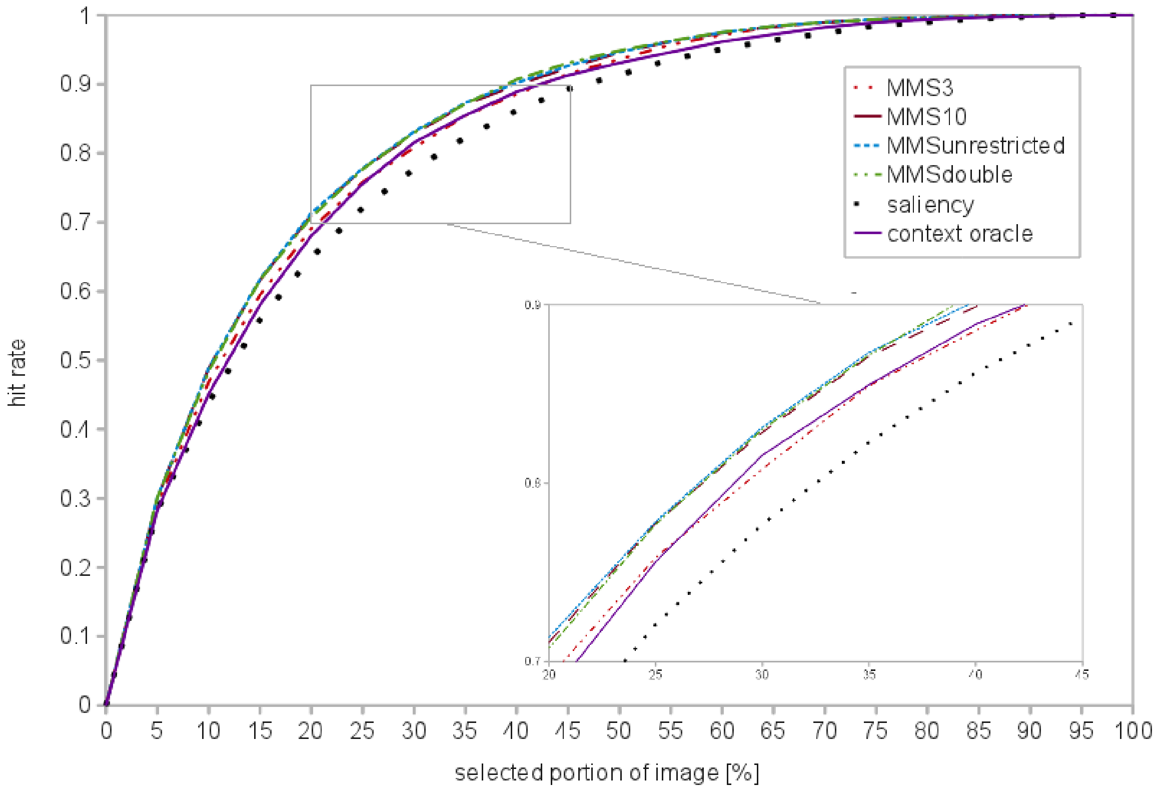

Figure 5: Prediction performance for the visual counting task for MMS with memory of 3, 10 and an unrestricted number of fixations (MMS3, MMS10 and MMSunrestricted), MMS with a separate memory for animate and inanimate objects (MMSdual), the approximation of a contextual upper bound (context oracle), and the saliency baseline.

jects and a larger number of images is needed to test whether the dual memory model is able to improve performance above the level of the other models presented.

Overall, our results demonstrate that a simple model of visual search based on the memory of previous fixations can perform equally good, 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 the MMS model performance does not degrade on a visual count dataset consisting of different scenes with radically different context. Instead the model still performs better than saliency and comparable to the context oracle.

Conclusions

We presented a 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 on a large set of images with manually provided object labels. The Memory Modulated Saliency model that we propose, on the other hand, does not require offline training 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 the experimental data sets, demonstrating the benefit of memory for the prediction of fixation locations. An MMS model with unrestricted memory outperforms the theoretically possible upper bound of the CGM on the visual search data (but not on the visual counting data). Unlike the CGM, the MMS does not require training data, but incrementally learns likely target positions. This means that the model can adapt easily to new data sets, tasks, and experimental conditions (while the CGM is sensitive to the nature of the training data).

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

Acknowledgments

The support of the European Research Council under award number 203427 ”Synchronous Linguistic and Visual Process-

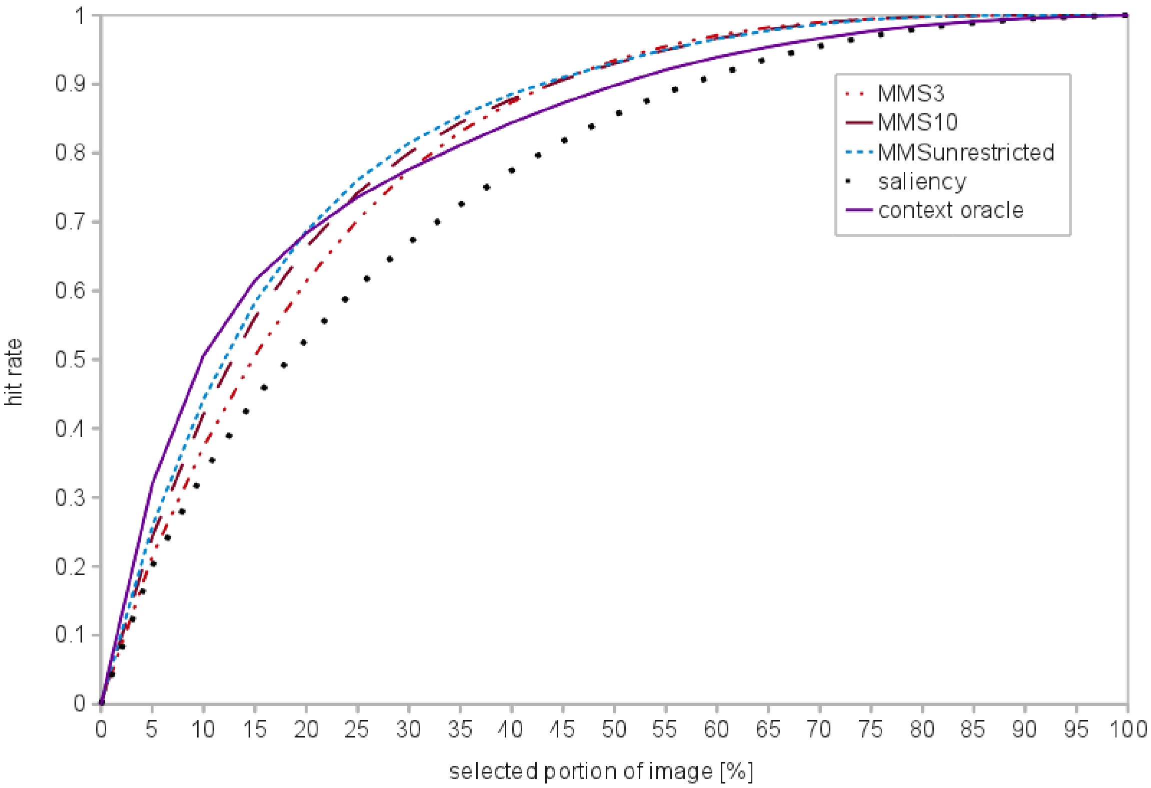

Figure 6: Prediction performance for MMS for the visual search task.

ing" is gratefully acknowledged.

We would also like to thank 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.

References

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

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.

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

Coco, M. I., & Keller, F. (2009). The impact of visual information on reference assignment in sentence production. In N. Taatgen, & H. van Rijn (eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, (pp. 274–279), Amsterdam. Cognitive Science Society.

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*, 3–42.

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

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*, 945–978.

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

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

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

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

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

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*, 1254–1259.

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

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

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.

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.

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.

Zelinsky, G. (2008). A theory of eye movements during target acquisition. *Psychological Review*, *115*, 419–433.