Prediction of Search Targets From Fixations in Open-World Settings

Hosnieh Sattar[1,2], Sabine Müller[1], Mario Fritz[2], Andreas Bulling[1]
[1]Perceptual User Interfaces Group, [2]Scalable Learning and Perception Group, Max Planck Institute for Informatics, Saarbrücken, Germany
{sattar,smueller,mfritz,bulling}@mpi-inf.mpg.de

Previous work on predicting the target of visual search from human fixations ([2], [1]) only considered closed-world settings. In this work we go beyond the state of the art by studying search target prediction in an open-world setting in which we no longer assume that we have fixation data to train for the search targets. We present a dataset containing fixation data of 18 users searching for natural images from three image categories within synthesised image collages of about 80 images.

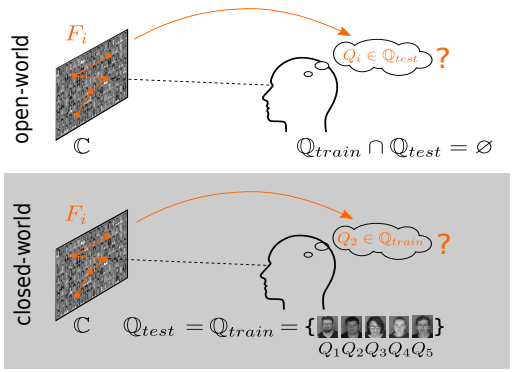

Figure 1: Experiments conducted in this work. In the *closed-world* experiment we aim to predict which target image (here Q_2) out of a candidate set of five images $\mathbb{Q}_{train} = \mathbb{Q}_{test}$ the user is searching for by analysing fixations F_i on an image collage $\mathbb{C}$. In the *open-world* experiments we aim to predict Q_i on the whole $\mathbb{Q}_{test}$.

Given a query image $Q \in \mathbb{Q}$ and a stimulus collage $C \in \mathbb{C}$, during a search task participants $P \in \mathbb{P}$ perform fixations $F(C,Q,P) = \{(x_i, y_i, a_i), i = 1,\ldots,N\}$, where each fixation is a triplet of positions x_i, y_i in screen coordinates and appearance a_i at the fixated location. To recognise search targets we aim to find a mapping from fixations to query images (Figure 1). We use a bag of visual world featurisation ϕ of the fixations. We interpret fixations as key points around which we extract local image patches. These are clustered into a visual vocabulary V and accumulated in a count histogram. This leads to a fixed-length vector representation of dimension $|V|$ commonly known as a bag of words. Therefore, our recognition problem can more specifically be expressed as:

$$\phi(F(C,Q,P),V) \mapsto Q \in \mathbb{Q} \tag{1}$$

In our new open-world setting, we no longer assume that we observe fixations to train for test queries. Therefore $\mathbb{Q}_{\text{test}} \cap \mathbb{Q}_{\text{train}} = \emptyset$. The main challenge that arises from this setting is to develop a learning mechanism that can predict over a set of classes that is unknown at training time. To circumvent the problem of training for a fixed number of search targets, we propose to encode the search target into the feature vector, rather than considering it a class that is to be recognised. This leads to a formulation where we learn compatibilities between observed fixations and query images:

$$(F(C,Q_i,P),Q_j) \mapsto Y \in \{0,1\} \tag{2}$$

Training is performed by generating data points of all pairs of Q_i and Q_j in $\mathbb{Q}_{\text{train}}$ and assigning a compatibility label Y accordingly:

$$Y = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{3}$$

We do not have fixations for the query images. Therefore, we introduce a sampling strategy S which still allows us to generate a bag-of-words representation for a given query. We stack the representation of the fixation and the query images. This leads to the following learning problem:

$$\begin{pmatrix} \phi(F(C,Q_i,P),V) \\ \phi(S(Q_j)) \end{pmatrix} \mapsto Y \in \{0,1\} \tag{4}$$

We learn a model for the problem by training a single binary SVM $\mathcal{B}$ classifier according to the labelling as described above. At test time we find the query image describing the search target by

$$Q = \underset{Q_j \in \mathbb{Q}_{\text{test}}}{\arg\max}\, \mathcal{B} \begin{pmatrix} \phi(F_{\text{test}},V) \\ \phi(S(Q_j)) \end{pmatrix} \tag{5}$$

In both closed-world and open-world evaluation we distinguish between within-participant and cross-participant predictions. In the "within participant" condition we predict the search target for each participant individually using their own training data. In the closed-world setting accuracies were well above chance for all participants for the Amazon book covers (average accuracy 75%) and the O'Reilly book covers (average accuracy 69%). Accuracies were lower for mugshots but still above chance level (average accuracy 30%, chance level 20%). In the open-world setting the average performance of all participants in each group was for Amazon: 70.33%, O'Reilly: 59.66%, mugshots: 50.83% (chance level 20%). In contrast, for

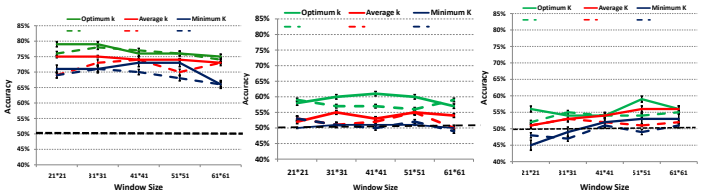

Figure 2: Open-World evaluation results showing mean and standard deviation of cross-participant prediction accuracy for Amazon book covers (left), O'Reilly book covers (middle), and mugshots (right). Results are shown with (straight lines) and without (dashed lines) using the proposed sampling approach around fixation locations. The chance level is indicated with the dashed line.

the "cross participant" condition, we predict the search target across participants. The "cross participant" condition is more challenging as the algorithm has to generalise across users. In the closed-world the prediction accuracies for Amazon book covers was best, followed by O'Reilly book covers and mugshots. Accuracies were between $61\% \pm 2\%$ and $78\% \pm 2\%$ for Amazon and O'Reilly book covers but only around chance level for mugshots. In the open-world setting the model achieves an accuracy of 75% for Amazon book covers, which is significantly higher than chance at 50%. For O'Reilly book covers accuracy reaches 55% and for mugshots we reach 56%. Figure 2 summarises the cross-participant prediction accuracies.

In this paper we demonstrated how to predict the search target during visual search from human fixations in an open-world setting. We showed that this formulation is effective for search target prediction from human fixations.

References

[1] Ali Borji, Andreas Lennartz, and Marc Pomplun. What do eyes reveal about the mind?: Algorithmic inference of search targets from fixations. *Neurocomputing*, 2014.

[2] Gregory J Zelinsky, Yifan Peng, and Dimitris Samaras. Eye can read your mind: Decoding gaze fixations to reveal categorical search targets. *Journal of Vision*, 13(14):10, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.