Prediction of Search Targets From Fixations in Open-World Settings

Hosnieh Sattar1,2, Sabine Müller1, Mario Fritz2, Andreas Bulling1
1Perceptual User Interfaces Group, 2Scalable 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], [11]) 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 \(Q_t\)) out of a candidate set of five images \(Q_{train} = Q_{test}\) the user is searching for by analysing fixations \(F_i\) on an image collage \(C\). In the open-world experiments we aim to predict \(Q_t\) on the whole \(Q_{test}\).

Given a query image \(Q \in \mathcal{Q}\) and a stimulus collage \(C \in \mathcal{C}\), during a search task participants 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 \(\phi\) 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:

\[
\phi(F(C,Q,P),V) \rightarrow Q \in \mathcal{Q}
\]  

(1)

In our new open-world setting, we no longer assume that we observe fixations to train for test queries. Therefore \(Q_{test} \cap Q_{train} = \emptyset\). 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:

\[
(F(C,Q_i,P),Q_j) \rightarrow Y \in \{0,1\}
\]  

(2)

Training is performed by generating data points of all pairs of \(Q_i\) and \(Q_j\) in \(Q_{train}\) and assigning a compatibility label \(Y\) accordingly:

\[
Y = \begin{cases} 
1 & \text{if } i = j \\
0 & \text{if } i \neq j 
\end{cases}
\]  

(3)

We do not have fixations for the query images. Therefore, we introduce a sampling strategy \(\delta\) 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:

\[
\begin{aligned}
\phi(F(C,Q_t,P),V) & \rightarrow Y \in \{0,1\} \\
\phi(S(Q_i)) & \rightarrow Y \in \{0,1\}
\end{aligned}
\]  

(4)

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

\[
Q = \arg \max_{Q \in \mathcal{Q}_{test}} \left( \phi(F_{test},V) \right)
\]  

(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 for all participants for the Amazon book covers but only around chance level for mugshots. 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 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%±2% and 78%±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.