Predicting Visual Search Targets via Eye-Tracking Data

Manisha Gupta
University of Texas at Austin

Fawad Ahmed
University of Central Florida

Dr. Ali Borji
University of Central Florida

The purpose of this work is to train a binary support vector machine to develop a learning mechanism to predict visual search targets and eventually over an unknown set of targets. We used a dataset that contains fixation points of 18 participants searching for a particular image within three types of collages. After plotting the visual patterns of each of the users, we compare the features of each of the images the participants fixated on to the features of the designated search target using a Siamese Convolutional Neural Network. These cumulation of these features indicate attributes of the target, which can be used to train the support vector machine accurately predict it.

I. Introduction

Half of the human brain is involved in the visual cortex - either directly or indirectly, which makes image processing and computer vision and integral part of machine learning (Glickstein, 1988) []. By teaching a Support Vector Machine (SVM) how to interpret features from a specific image, we can teach a camera to look for specific attributes (such as small, metal) to scan the visible environment for a particular object such as a lost set of keys or the computer to search for a specific face within a crowd when it only knows certain attributes. The purpose of this project is to be able to predict the target of a visual search in both an open and closed-world setting. A closed-world setting is that in which the SVM is trained and tested on the given dataset. In an open-world setting we no longer assume the SVM has been trained on the test set, so we develop a learning mechanism over an unknown set of images.

Approach

With the combination of computer vision and human behaviour originates the idea behind this project - teaching the computer to interpret visual behaviour which include fixations and saccades from the gaze information to better understand the participants' behaviour and then predict their search target within a collage of images. Because there has been research conducted on the importance of predicting visual search targets in closed-world settings (Haji-Abolhassani, 2014) [], this project aimed to address the same problem in an open-world setting. We used a pre-trained Siamese model This dataset for this problem was acquired from the Max Planck Institute in Saarbucken, Germany (Sattar, 2015) []. 18 participants wore an eye-tracker while searching for five targets within three sets of collages (one of faces, the other two of book covers) of approximately 80 images.

II. Related Works



Figure 1. 3x3 image patches around each fixation point

Predicting Search Targets

Many works in the past have addressed the task of analysing gaze information to better understand user behaviour and predict search targets using fixations. Zelinsky's et al. (Zelinsky, 2010) [] and Alexander's et al. work (Alexander, 2011) [] are particularly similar to ours because they set up a categorical task. The users had to find two search targets within a set of distractors. Then they made predictions about the search task from the fixation data collected. This sets up the idea of our project, however ours uses a much more challenging data set. Boisvert et al. (Boisvert, 2016) [] also worked on predicting search targets from recorded eye movements but focused on spatial dynamics and image features. They developed statistical information while trying to predict the designated task from the gaze information collected. At the Max Planck Institute (Sattar, 2015) [], they addressed the same task that we approached. However, there they took 3x3 image patches

(Figure 1) around each fixation point and trained the SVM from that. While this compensates for eye-tracker inaccuracy, the SVM does not learn features from the entire book cover. They also use a bag-of-words model with the fixations as their input.

Our Modifications

This project intended on improving results from the Max Planck Institute. Here, the images sampled was the entire book cover that had a fixation point on it. To compare the book covers against the search targets, we needed a model that learned directly from the images instead of using hand-crafted features. Zagoruyko et al. (Zagoruyko, 2015) []trained a CNN model to compare raw image patches from scratch even with wide disparity.

III. Dataset

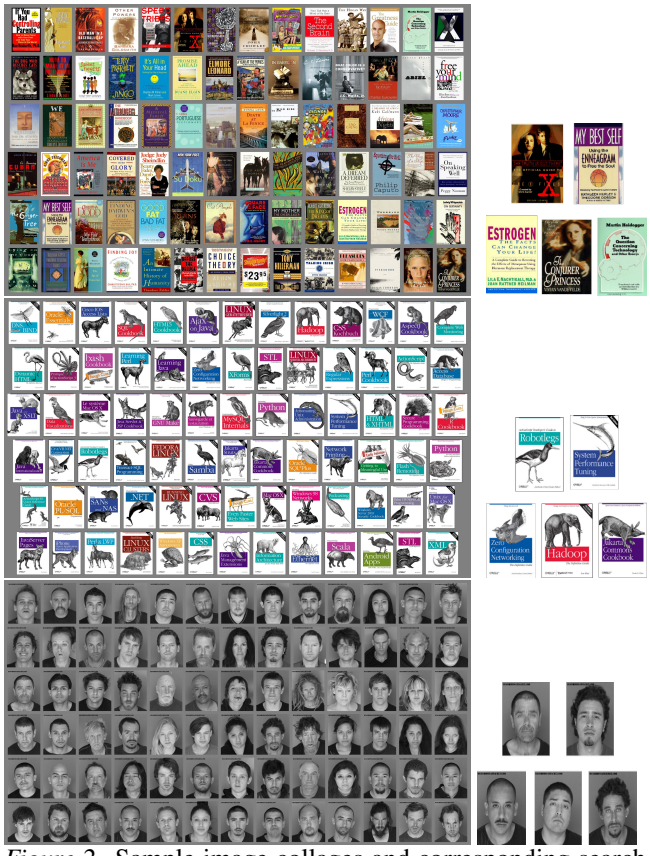


Figure 2. Sample image collages and corresponding search targets from the data set. Amazon book covers (top), O'Reilly book covers (middle), mugshots (end).

The dataset (Figure 2) was acquired online from (Max Planck Institute). The three types of collages were the Amazon book covers, O'Reilly book covers and mugshots. In the Amazon book covers, the six participants relied on both the distinct structure and color to find the designated search target within the collages. In the O'Reilly book covers, color was the most discriminative feature that participants relied on. The mugshots had very similar structure and no color, giving participants no distinguishable feature to rely on while searching for the particular search target. All 18 users were asked to search for a particular search target within a collage whilst wearing a stationary Tobii TX300 eye tracker which recorded their gaze information. The participants were also put on a time restraint (10 seconds to memorize the target and 20 seconds to find it) to prevent them from lingering on a particular image. The Amazon book covers had the highest percentage of lingering (2.45%), then O'Reilly (1.2%) and finally mugshots (0.35%). There were twenty variations of the collage associated with that search target to prevent the participant from learning the collage. With the five search tasks, this yields in 100 search tasks per user. With six users per collage type, we had a total of 1800 search tasks.

IV. Method



Figure 3. Gaze information for user 1 layered on top of the collage (shown left). Extracted book covers are shown on the right.

Figure 3 demonstrates the gaze information for the first participant while they searched for the search target. The size of the fixation points corresponds to the duration and the saccades indicate the path. On the right of Figure 3 are each of the book covers that had a fixation point. These extracted book covers are then compared against the search target because they have some features that attracted the users' attention and gaze. So, we modified the pre-trained Siamese convolutional neural network (CNN) model (Zagoruyko, 2015) []. The model takes in two images as input - the search target and each of the extracted book covers one by one (Figure 4), which then pass through the convolution neural network and output a 2x512 feature vector corresponding to each image. The feature vector goes through the decision network and outputs a euclidean distance, which compares the distance between the features. The higher the number, the more dissimilar the two images are. Figure 5 outlines the inner workings of the Siamese CNN model. A series of convolutional, ReLU and max-pooling layers are applied to each of the two input images. Then the images are concatenated and given

to the decision network, which contains two linear fully connected layers of 512 units (i.e. yields the 2x12 feature vector). In the pseudo-siamese aspect of the CNN model, there are no shared branches - the weights of the two image inputs are not shared. This allows for greater flexibility but still maintains efficiency.

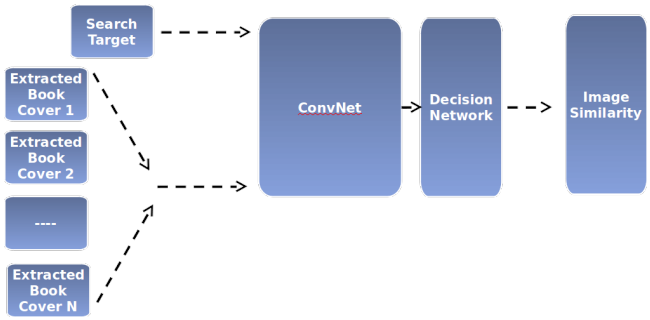


Figure 4. Siamese CNN Model pipeline

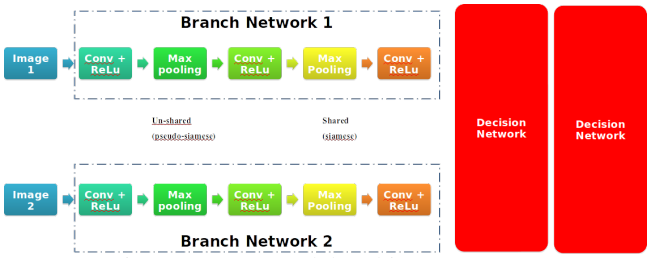


Figure 5. Siamese CNN Model architecture

V. Results

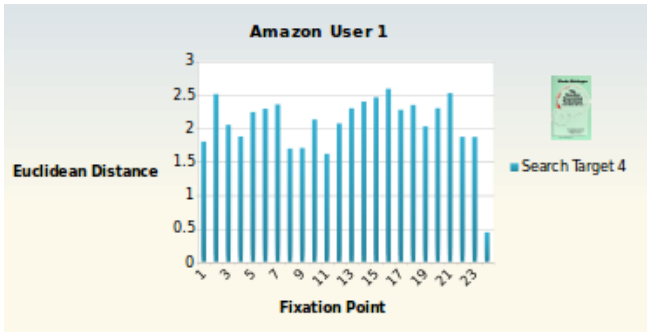


Figure 6. Euclidean distances between each of the fixations points and search target for the first Amazon collage for user 1.

The euclidean distances outputted from the Siamese CNN model indicate the similarities between the search targets and each of the book covers with fixation points. Figure 6 demonstrates the values of the euclidean distances from one collage corresponding to search target 4. If the user found the search target, the last euclidean distance is 0 since there is no difference between the target and the point. Because the SVM would quickly learn that this is the search target instead of learning attributes from other fixation points, we discard the last fixation point. So, in figure 6 there were 25 book covers with fixation points and the first 24 are illustrated. Although there is no pattern between the euclidean distances throughout each of the fixation points, we learned that the second to last fixation point from the collages had the lowest values. This indicates that the second to last book cover contains the most similar features to the search target. Then we calculated the average euclidean distances per search target for each of the six participants. Figure 7 could indicate that Amazon collages were the most difficult to find the particular search target; however a euclidean distance is not necessarily the best comparison factor. The mugshots collage would show low euclidean distances because the images are the most similar to each other, which could make it harder. Finally, we calculated the average number of fixations per search target for each of the three types of collages for user 1 and similar graphs are generated for each of the six participants. Figure 8 illustrates that search target 5 seemed the most difficult to search for all three collages, given that the number of fixa-

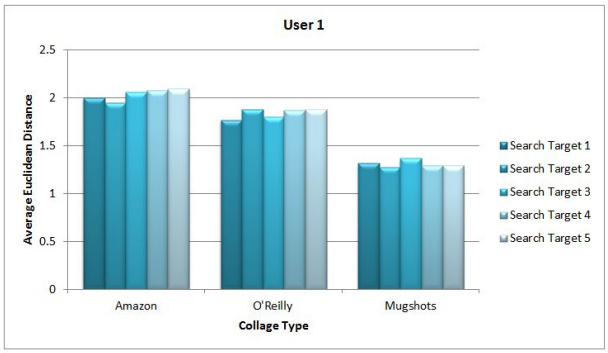


Figure 7. The average euclidean distance per search target for each of the three collage types for user 1.

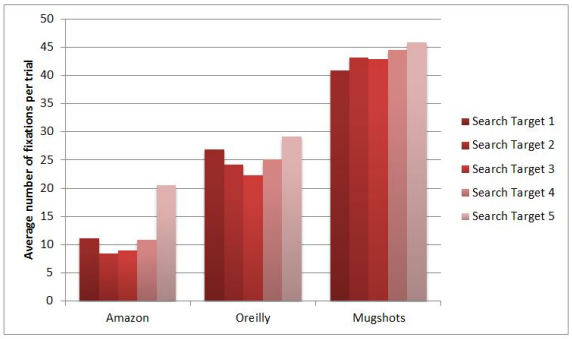


Figure 8. The average number of book covers with fixations on them per search target per collage type for user 1.

tions were the highest. This figure also demonstrates that the mugshots were the most difficult by the same reasoning.

VI. Discussion

This project has set up the baseline for the support vector machine. In combination with the annotations about gaze information from the Max Planck Institute we have been able to classify the collages. The Amazon book covers provide the most distractors (i.e. longest lingering time) but also two significant features (structure and color) for the users to rely on, Mugshots proved most difficult given the highly similar images within the collages. It took the six users longer to find the search target - given the large number of fixations and lowest lingering time. In the future, we will calculate a classification accuracy to demonstrate if the collage and associated search target do output the lowest euclidean distance. Using this accuracy we can train the SVM to predict in closed-world settings first and then expand to open-world. Further, we will add weights to the fixation points based on the gaze duration. If a participant fixated on a particular book cover for a longer period of time, this indicates that that book cover has more significant features. Last, we will expand the data set to include a series of collages that has a distinct structure within a grey-scaled image. In this case, participants will have to rely only on the shape of the target to find it within the collage.

VII. Acknowledgements

We would like to thank Dr. Mubarak Shah, Dr. Niels Lobo for their guidance and support through this project. This work is funded by the National Science Foundation (NSF) Research Experience for Undergraduates (REU) summer program at the University of Central Florida. Thank you to Dr. Hosna Sattar from Max Planck Institute for Informatics, Saarbrucken, Germany for her valuable insight with direction of the project and the guidance using the dataset.

References

Alexander, R. G., & Zelinsky, G. J. (2011, jul). Visual similarity effects in categorical search. *Journal of Vision*, *11*(8), 9–9. Retrieved from http://dx.doi.org/10.1167/11.8.9 doi: 10.1167/11.8.9

Boisvert, J. F., & Bruce, N. D. (2016, may). Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features. *Neurocomputing*. Retrieved from http://dx.doi.org/10.1016/j.neucom.2016.05.047 doi: 10.1016/j.neucom.2016.05.047

Glickstein, M. (1988, sep). The discovery of the visual cortex. *Sci Am*, *259*(3), 118–127. Retrieved from http://dx.doi.org/10.1038/scientificamerican0988-118 doi: 10.1038/scientificamerican0988-118

Haji-Abolhassani, A., & Clark, J. J. (2014, oct). An inverse yarbus process: Predicting observers' task from eye movement patterns. *Vision Research*, *103*, 127–142. Retrieved from http://dx.doi.org/10.1016/j.visres.2014.08.014 doi: 10.1016/j.visres.2014.08.014

Sattar, H., Muller, S., Fritz, M., & Bulling, A. (2015, jun). Prediction of search targets from fixations in open-world settings. In *2015 IEEE conference on computer vision and pattern recognition (CVPR)*. Institute of Electrical & Electronics Engineers (IEEE). Retrieved from http://dx.doi.org/10.1109/cvpr.2015.7298700 doi: 10.1109/cvpr.2015.7298700

Zagoruyko, S., & Komodakis, N. (2015, jun). Learning to compare image patches via convolutional neural networks. In *2015 IEEE conference on computer vision and pattern recognition (CVPR)*. Institute of Electrical & Electronics Engineers (IEEE). Retrieved from http://dx.doi.org/10.1109/cvpr.2015.7299064 doi: 10.1109/cvpr.2015.7299064

Zelinsky, G., Zhang, W., & Samaras, D. (2010, mar). Eye can read your mind: Decoding eye movements to reveal the targets of categorical search tasks. *Journal of Vision*, *8*(6), 380–380. Retrieved from http://dx.doi.org/10.1167/8.6.380 doi: 10.1167/8.6.380