Prediction of Visual Search Target

Ramvinojen Narayana Perumal
University of Stuttgart
Stuttgart, Germany
st159851@stud.uni-stuttgart.de

Shyamnarayan Sankarasubramanian
University of Stuttgart
Stuttgart, Germany
st159284@stud.uni-stuttgart.de

Vijeth Kumar
University of Stuttgart
Stuttgart, Germany
st158777@stud.uni-stuttgart.de

ABSTRACT

Prior works on the prediction of visual search targets focused only on the stationary camera settings where the users' movement, blurring and the viewing angle of the fixation were not taken into account. In this work we go beyond our predecessors by working in an open-world setting using a mobile eye tracker. We present a dataset containing gaze data of seventeen participants searching for a target from eight categories in realworld-like setting. A simple and effective method to pre-process the output of the mobile eye tracker is presented. An intuitive feature extraction technique is used to feed the gaze and scene data into the developed models for the prediction of search targets. In this paper we propose novel search target prediction algorithms and also a comparison study to quantify the effectiveness of the developed models. The results of our experiments show that it is possible to train a machine learning model to predict search targets with high categorical accuracy.

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction (HCI).*

KEYWORDS

Mobile eye tracker, Open World, Regression, Convolutional Neural Network, Dense Neural Network, Linear Discriminant Analysis, Feature Extraction, Feature Reduction, Fixation, Gaze Information, Classifier, Pupil Labs Eye Tracker

Description	Number
Participant	17
Total DataSet	286
Training Dataset	190
Validation Dataset	48
Testing Dataset	48

Table 1: Dataset details

Super Class	Class
Car	Sedan
	Hatchback
	Pickup
Truck	Fire Truck
	Cement Truck
	Road Roller
Bus	Public Bus
	School Bus
Distractors	Crane
	Cement Trailer
	Tractor

Table 2: Class details

INTRODUCTION & RELATED WORK

There has been huge progress in techniques for the discovery/prediction of a person's thoughts, or goals. These researches mainly used the neural activity of the person to infer a person's thoughts. Our project focuses on behavioral decoding in the context of a visual search task, here the category of a person's search target is the decoding goal. This technique of decoding uses behavioral measures to infer a person's thoughts. Yarbus [6] in his work showed that the visual behavior when looking at a visual scene is closely linked to the search task. Thus, in order to make predictions about user behavior we analyze the visual behavior metrics. In this project we choose fixation points and fixation duration as behavior metric. The fixations on non-target objects or distractors made during the search are not random. The more similar these objects are to the target category, the more likely they are to be fixated first or fixated longer compared to less target-similar objects, i.e task is influenced by fixation patterns.

Many works have been conducted to recreate Yarbus' [6] findings and to extend them to predicting the observers' tasks. A series of researches followed to recreate and also incrementally improve the prediction methods. These works clearly proved that the observers' tasks could be successfully predicted with just the gaze information. Few other works including Bulling [1] investigated the recognition of everyday(office) activities such as reading, taking hand-written notes, or browsing the web based on long-term eye movement recordings. Further they were also able to infer high-level contextual cues, such as social interactions or mental activity from visual behavior.

Only a few of these previous works focused on visual search and the prediction of search targets from gaze. Zelinsky et al. [5] were able to predict gaze patterns during categorical search tasks. A series of experiments were conducted, in which participants had to search and find two categorical search targets among few distractors which were very much similar to the search targets. They predicted the number of fixations made, prior to search judgements. Borji et al. [6] focused on predicting search targets from fixations. Additionally, their research showed that as the complexity of the search target increases, the participants were clearly guided more by finer sub-patterns rather than the whole pattern.

The works of Bulling et al. [2] and Sattar et al. [1] are most related to ours. However, both works only considered simplified visual stimuli or synthesised natural scenes in a closed-world setting and worked with stationary mobile tracker. In that setting, all potential search targets were part of the training set and fixations for all of these targets were observed. In contrast, in our work we address both the open-world setting and use mobile eye tracker which very much increases the generalization of the algorithm for a real life application.

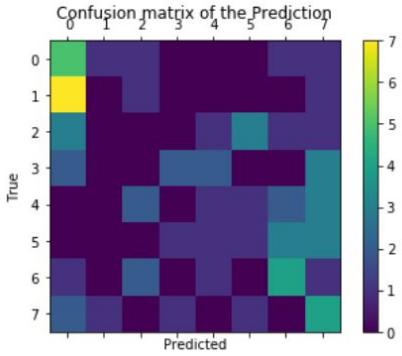

Figure 1: Confusion matrix for Model 1 with LDA Classifier

Class	Precision	Recall	Support
0	0.25	0.56	9
1	0.00	0.00	9
2	0.00	0.00	9
3	0.50	0.22	9
4	0.17	0.11	9
5	0.17	0.11	9
6	0.36	0.44	9
7	0.24	0.44	9
Avg/Total	**0.21**	**0.24**	**72**

Table 3: Results of Model 1 with LDA Classifier

DATA COLLECTION STUDY

Dataset available from previous works for visual target prediction are collected using a stationary camera setting. Hence, we conducted our own data collection study. Eight classes were selected as show in the Table 2. Also, to explore the impact of the similarity in appearance between target and search-set, as many distractors as possible were used. We decided to use virtual objects instead of real objects. The virtual objects were color printouts of the eight class images. Our goal is to collect gaze data and scene information correspoonding to various search tasks for all the participants.

Calibration

The mobile eye tracker needs to be calibrated for each participant. There are various calibration methods. Initially, we used screen marker calibration technique to calibrate the tracker. In screen marker calibration the participant would be asked to look at the moving marker in computer screen. This technique is quite effective for stationary setting but was not useful for mobile setting due to the limited vision range in the actual search task. Secondly, we tried using Manual marker calibration technique for improving the vision range. In Manual marker calibration, the marker was moved manually on the wall. The participant was asked to look at the marker at different positions on the wall. The vision range improved when compared with the previous calibration method but the required precision was not achieved for the actual task. Finally, we tried with natural scene calibration technique. In this method, custom printed tiny markers around 30 to 50 were pasted on wall to cover the entire vision range of the participant. Then the participants were asked to look at the markers sequentially and the fixation points were sampled on the Pupil capture tool. The obtained vison range and the precision drastically increased compared to previous calibration techniques.

Apparatus

To record the search task, a mobile eye tracker from Pupil Labs was used. It provides gaze data at a sampling frequency of 200Hz and a scene data at a sampling frequency of 30Hz for HD resolution, 1920x1080 Pixels. About 120 printouts for eight classes of virtual objects with three distractor classes were used. For each search task, close to sixty virtual objects were posted on the walls of a room.

Procedure

Participants were asked to search for a target belonging to one of the eight stimulus types from a set of around sixty displayed images which also includes distractors. The images were amply spaced out on the wall to differenciate between the fixations. As soon as the participant found the target image, the video recording was stopped. This procedure is repeated two to three times for eight different targets resulting in a total of 18 search tasks. For each search task, images were shuffled so that

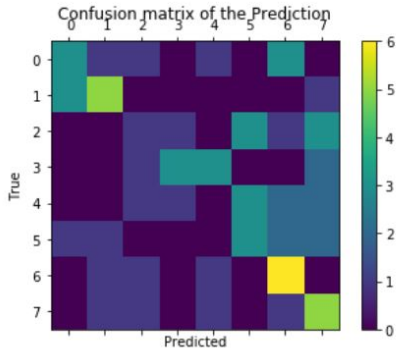

Figure 2: Confusion matrix for Model 1 with Neural Network Classifier

Class	Precision	Recall	Support
0	0.43	0.33	9
1	0.56	0.56	9
2	0.17	0.11	9
3	0.60	0.33	9
4	0.00	0.00	9
5	0.33	0.33	9
6	0.40	0.67	9
7	0.33	0.56	9
Avg/Total	**0.35**	**0.36**	**72**

Table 4: Results of model 1 with Neural Network Classifier

the participant doesn't memorize the images. The participants were given freedom to start from the position of their own interest for every search task. The dataset consists of recorded fixation data for 18 participants with different nationalities and aged between 22 and 35 years. Details are shown in Table 1.

PREPROCESSING

The raw search task video and the gaze information obtained from the mobile eye tracker has to be preprocessed before being fed into any model. The gaze data consists of the gaze co-ordinate and timestamp, this information is used to detect fixations. In our experiment we empirically arrived at the value 100ms as our minimum fixation duration for participants. Thereby, implying that every fixation will have more than 3 frames. Now, the best frame for every fixation is extracted from the gaze information by comparing the Laplacian of the Gaussian values of various image frames. The Fixation image patches are obtained from the selected frame by cropping a fixed region around the normalized gaze co-ordinate for that particular time stamp. A fixed-region-crop is more than sufficient because we are more interested in extracting the features around the gaze point rather than the entire object itself. Also, if necessary, the cropping region can also be made dynamic by using suitable object recognizing algorithm. The image patches are still prone to noise and slight motion blur. In order to improve the quality of the feature extraction step, these patches are subjected to few enhancing techniques including contrast adjustments, brightness adjustments, under sampling.

FEATURE EXTRACTION

The enhanced image patches need to be transformed to a set of meaningful features to be used in the prediction model. One method to obtain features from the image patches is by using image gradient techniques such as SIFT and KAZE. The use of pre-trained neural network classifier such as Resnet, VGG models is prominent in this field of research. Thus, we choose Resnet50 pre-trained with ImageNet Database. Resnet50 is a multi-layered Neural network, the last layer of the Resnet50 model is a simple classification layer. When the last layer is removed, the output from the previous layers gives us a 2048 values long feature vector which can be used for classification. The obtained 2048 values long feature vector is used in the further steps.

Every search task consists of a set of fixation images and one target image. These image patches are fed into the Resnet50 and the corresponding feature vectors are obtained. Each fixation feature vector is weighted with normalized fixation duration and combined to obtain a single combined feature vector. Now, every search task consists of combined feature vector and a target feature vector of feature length 2048, this makes up the processed dataset. These combined feature vectors are now used in conventional machine learning and deep learning approaches.

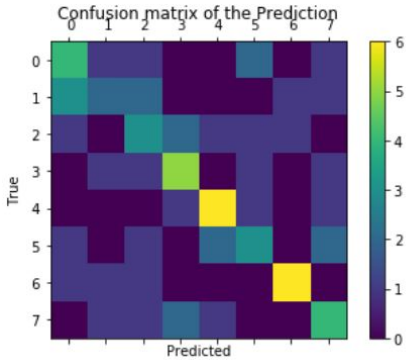

Figure 3: Confusion matrix for Model 2 prediction

Class	Precision	Recall	Support
0	0.40	0.44	9
1	0.33	0.22	9
2	0.30	0.33	9
3	0.50	0.56	9
4	0.60	0.67	9
5	0.38	0.33	9
6	0.75	0.67	9
7	0.40	0.44	9
Avg/Total	**0.46**	**0.46**	**72**

Table 5: Results of Model 2

MODEL 1: CONVENTIONAL MACHINE LEARNING MODEL

The target images are fed into ResNet50 to obtain target feature vector of size 2048. The combined fixation feature vector and the target feature vector are given to linear regression model with least squares as cost function. The output of the regression model is given to Linear Discriminant Analysis and Neural Network classifier as seen in the Figure 4.

The Neural Network classifier is trained with augmented image patches. The augmentation techniques used are Image tilting and Gaussian noise addition. Evaluation study was performed by comparing the results obtained from the different classifiers. Precision and Recall were used as the metrics for comparision study.

Figure 1 shows the confusion matrix of the Linear Discriminant Analysis prediction. It can be observed that , there are huge number of misclassifications. Table 3 shows the precision and recall score of the 8 major classes. Classes 1,3 and 6 were predicted better when compared to the remaining classes.

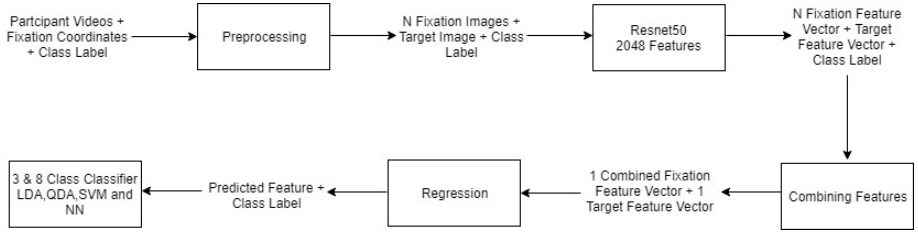

Figure 4: Block diagram for Model 1

Figure 2 shows the confusion matrix of the Neural Network prediction. It can be observed that, the number of misclassifications are reduced when compared to the Linear Discriminant Analysis classifier. The precision and recall score of majority of the classes has shown significant improvement as seen in Table 4.

The link for the code is given below:
/nramvinojen/Programs/Workbench/24Jan2019/MainCode/1_Resnet2048_Regression_Classifier

MODEL 2: DEEP LEARNING MODEL

The deep learning model consisting of dense layers is trained with datasets(combined fixation feature vector and class labels) . Adam optimizer was used with relu activations in hidden layers and softmax for the output layer .The combined fixation feature vector is given to the deep neural network and prediction is done for eight classes as depicted in Figure 5.

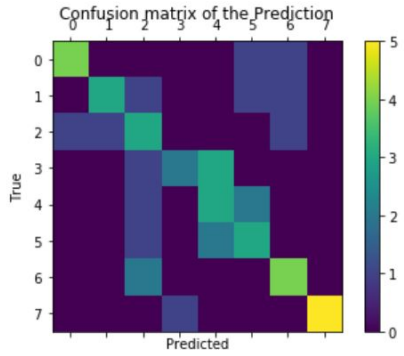

Figure 6: Confusion matrix for Model 3 prediction

Class	Precision	Recall	Support
0	0.80	0.67	6
1	0.75	0.50	6
2	0.33	0.50	6
3	0.67	0.33	6
4	0.38	0.50	6
5	0.43	0.50	6
6	0.57	0.67	6
7	1.00	0.83	6
Avg/Total	**0.62**	**0.56**	**48**

Table 6: Results of Model 3

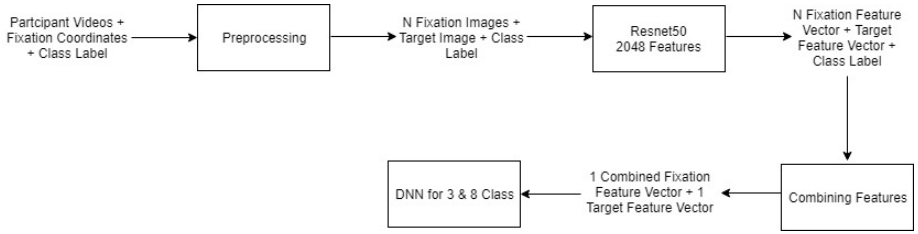

Figure 5: Block diagram for Model 2

Figure 3 shows the confusion matrix of the Dense Neural Network prediction. It can be observed that the number of misclassifications is very less when compared to conventional models. Table 5 shows the precision and recall score of the eight classes. It can be inferred that the overall accuracy and individual class accuracy is much better when compared to conventional machine learning framework.

Comparing Model 1 and Model 2 shows that, classes 1 and 2 were not at all predicted in Model 1 whereas the precision scores were 33% and 30% respectively in Model 2.

The link for the code is given below:

/nramvinojen/Programs/Workbench/24Jan2019/MainCode/2_Resnet2048FV_DNN_withVal

MODEL 3: FEATURE REDUCTION

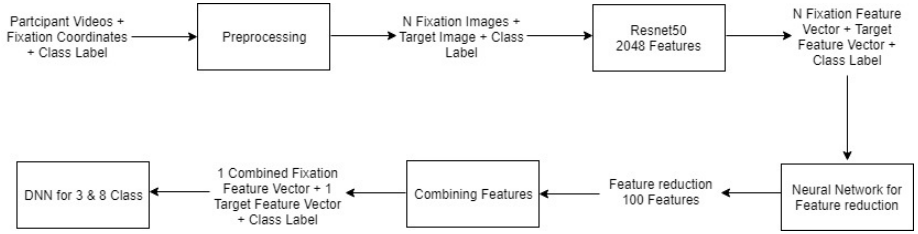

Figure 7: Block diagram for Model 3

In the earlier models, the dimension of feature vector for fixations and target is 2048, which is too high. Therefore, neural networks trained with augmented fixation data was used for reducing the feature size from 2048 to 100. The reduced feature vector is then combined to get the combined fixation feature vector which is fed into the deep neural network trained datasets(combined fixation feature vector and class labels). Adadelta optimizer was used with relu activations in hidden layers and softmax for the output layer .The combined fixation feature vector is given to the deep neural network and prediction is done for eight classes as show in the Figure 7.

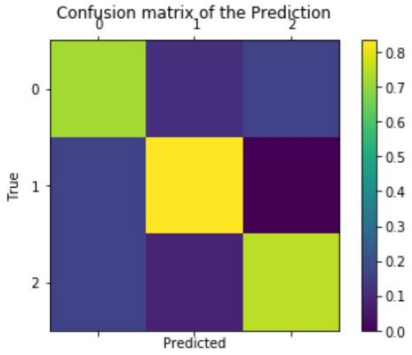

Figure 8: Confusion matrix for 3 class prediction

Class	Precision	Recall	Support
0	0.72	0.72	18
1	0.83	0.83	18
2	0.75	0.75	12
Average/Total	**0.77**	**0.77**	**48**

Figure 9: Results of 3 class prediction

Figure 6 shows the confusion matrix of Model 4 prediction. It can be observed that the number of misclassifications is significantly less when compared with Model 2. The individual class accuracy and the overall prediction accuracy is much better when compared to Model 2. In Model 2 precision score of class 7 was 40% whereas it is 100% in Model 3 as seen in the Table 6.

Three class prediction was performed on the reduced feature vectors which gave a precision score of 77%. In the earlier models, the prediction accuracy of class 3, 4 and 5 were very less as they were relatively overlapping classes. Hence, when all the overlapping classes amongst the eight classes were combined 2, a three class classification could be performed which gave a precision score of 83 %. Figure 8 show the confusion matrix of the 3 class prediction model. Table 9 shows the individual class and overall precision and recall score.

The link for the code is given below:

/Programs/Workbench/24Jan2019/MainCode/3_Resnet2048FV_Reduce_100FV_DNN_withVal

OTHER APPROACHES

Instead of feeding the combined fixation feature vector of size 2048, reduced feature vector size of 48 is fed into the already existing regression model. The results of this model were below par when compared to Models 1,2 and 3. The poor results might be due to the loss of important feature information during feature reduction.

In the next approach, a custom build convolutional neural network is used for feature extraction from the image patches. This model didn't yield better results, this can be attributed to the limited training set for the CNN model.

In another approach, the feature vector corresponding to the longest fixating image patch and the combined feature vector obtained from the remaining image patches are concatenated to form a feature vector of length 200. This new feature vector is now fed to Models 1, 2 and 3. This method did not improve the prediction accuracy. Infact the doubling of the feature length only decreased the precision of prediction for few classes.

The link for the code is given below:

/nramvinojen/Programs/Workbench/24Jan2019/MainCode/5_Resnet_Reduce48FV_Regression_DNN

APPLICATIONS

This work applies to a wide spectrum of applications in the field of human machine interaction. The ability of the system to understand and predict the users targets, goals and state of mind creates the possibility to assist and enhance the user experience even without any explicit interaction with the machine. Few use cases outlining the capabilites of this approach are,

Figure 10: Virtual Search Assistant in Store

Figure 11: Virtual Search Assistant for Drivers

1) Virtual Search Assistant in Store : The above idea can be extended to a smart stores where a customer neither requires shopkeeper for assistance nor does he need to refer to a floor plan. The involuntary visual cues during a search task can be captured to provide automatic virtual assistance.

2) Virtual Search Assistant for Drivers : The visual cues of the drivers could be used to provide navigation assistance. For example,when a tourist enters a new city, free parking facillity, key buildings suggestions can be provided by seamlessly connecting with the smart parking grid of the city.

INFERENCE AND CONCLUSION

In this work, we proposed the method to predict the category of visual search targets from gaze data. To this end, we proposed a novel method to combine the normalized fixation feature vectors obtained from a pre-trained neural network that facilitates the visual eye task using a mobile eye tracker. We believe that the proposed idea can be extended to any real world problems which involves search tasks. We have presented a sound conventional machine learning and deep learning framework that takes set of videos and gaze information during a search task as input and provides the prediction results of the participant. In addition to this, we also propose an intutive idea for feature reduction, which can be used to improve prediction accuracy.

ACKNOWLEDGEMENT

The research was supported by Institute for Human Computer Interaction and Computer Vision. We also thank Andreas Bulling, Sven Mayer, Huy Viet Le, for their valuable feedback and guidance throughout the course of the project.

REFERENCES

[1] H. Sattar, S. Müller, M. Fritz, and A. Bulling. 2015. Prediction of search targets from fixations in open-world settings. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 981–990. https://doi.org/10.1109/CVPR.2015.7298700

[2] Julian Steil, Michael Xuelin Huang, and Andreas Bulling. 2018. Fixation Detection for Head-mounted Eye Tracking Based on Visual Similarity of Gaze Targets. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA '18)*. ACM, New York, NY, USA, Article 23, 9 pages. https://doi.org/10.1145/3204493.3204538

[3] Ali Borji, Andreas Lennartz, and Marc Pomplun. 2015. What Do Eyes Reveal About the Mind? *Neurocomput.* 149, PB, 788–799. https://doi.org/10.1016/j.neucom.2014.07.055

[4] H. Sattar, A. Bulling, and M. Fritz. 2017. Predicting the Category and Attributes of Visual Search Targets Using Deep Gaze Pooling. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2740–2748. https://doi.org/10.1109/ICCVW.2017.322

[5] Gregory J Zelinsky, Yifan Peng, and Dimitris Samaras. 2013. Eye can read your mind: Decoding gaze fixations to reveal categorical search targets. *Journal of vision* 13. https://doi.org/10.1167/13.14.10

[6] Ali Borji and Laurent Itti. 2014. Defending Yarbus: Eye movements reveal observers' task. *Journal of vision* 14. https://doi.org/10.1167/14.3.29