Predicting Intention Through Eye Gaze Patterns

Fatemeh Koochaki
Department of Electrical and Computer Engineering
Rutgers University, Piscataway, NJ 08854
f.koochaki@rutgers.com

Laleh Najafizadeh
Department of Electrical and Computer Engineering
Rutgers University, Piscataway, NJ 08854
laleh.najafizadeh@rutgers.edu

***Abstract*—Eye movement is a valuable (and in several cases, the only remaining) means of communication for impaired people with extremely limited motor or communication capabilities. In this paper, we present a new framework that utilizes eye gaze patterns as input, to predict user's intention for performing daily tasks. The proposed framework consists of two main modules. First, by clustering the eye gaze patterns, the regions of interest (ROIs) on the displayed image are extracted. A deep convolutional neural network is then trained and used to recognize the objects in each ROI. Finally, the intended task is predicted by using support vector machine (SVM) through learning the embedded relationship between recognized objects. The proposed framework is tested using data from 8 subjects, in an experiment considering 4 intended tasks as well as the scenario in which the user does not have a specific intention when looking at the displayed image. Results demonstrate an average accuracy of 95.68% across all tasks, confirming the efficacy of the proposed framework.**

***Index Terms*—Eye Gaze Patterns, Assistive Technologies, Intention Prediction, Convolutional Neural Network, Support Vector Machine.**

I. Introduction

Eyes can be considered as the reflector of the mind. As such, eye movement data has been utilized in various applications. For example, it has been used for determining the user's mental overload when performing sensitive activities [1], in shared autonomy systems to perform cooperative tasks more efficiently [2], and in teleoperation systems [3], [4] or robotics grasping [5], as indirect input to control a robot.

An emerging usage of the eye gaze data is in the development of assistive technologies for paralyzed patients, such as those with amyotrophic lateral sclerosis (ALS), stroke or spinal cord injury (SCI). These groups of patients generally have extremely limited motor or communication capabilities, but still have good cognitive abilities, and preserve awareness and their ability of eye movement. The fact that the eye movement data can be recorded using eye tracker devices in a non-invasive manner is also an added advantage of using this modality in developing assistive technologies.

There exist prior work for inferring the intention through eye gaze patterns. In [6], static features of eye movement such as the number of fixations, duration of fixations, and the last seen object are considered to predict the objects in the displayed image on which the user has focused on. The authors considered a sandwich-making scenario, where a worker has to predict the ingredients the customer desires, based on extracting the static features of the eye gaze. Predicting low-level tasks such as counting, from eye gaze, have also been investigated [7], [8]. In [9], instead of static data, the temporal information of eye movement data is extracted through use of Hidden Markov Model approach, and in [8] by using the temporal information and Fisher Kernel Learning, improved accuracy in predicting tasks has been achieved.

In this paper, we present a new generalizable and scalable framework that uses eye gaze patterns as input, to predict daily tasks intended by the user. Our ultimate goal is to incorporate the proposed framework with assistive devices (such as a robotic arm) to enable patients perform simple activities of their daily life, independently, thereby, significantly, improving their quality of life. An experiment, involving 4 different intended tasks and a scenario where the user has no specific intention when looking at the displayed image, is designed. In this experiment, the users look at the displayed image and express their intended task by looking at the objects in the image that are related to task. The proposed framework receives eye gaze data, and utilizes clustering and convolution neuronal network (CNN) approaches to recognize the objects of interest in the displayed image. Next, by employing support vector machine (SVM) and learning the association of recognized objects, it predicts the intended task.

The rest of the paper is organized as follows. The proposed framework, the experimental paradigm and the data collection procedure are described in Section II. In Section III results are presented and discussed, and finally, the paper is concluded in Section IV.

II. Methods

A. Proposed Framework: An Overview

The proposed framework consists of two main modules: the object detector (OD) module, and the task predictor (TP) module. The OD module receives the raw eye gaze data as the input, identifies clusters of eye gaze patterns on the image, and utilizes CNN to identify the objects in the image based on the output of the clustering algorithm. The TP module receives the identified objects as input, and predicts the intended task. Fig. 1 illustrates the block diagram of the proposed framework.

B. Experimental Design

To collect data for the training of the TP module and evaluating the performance of the proposed framework an experiment is designed. Several possible scenarios can be

978-1-5386-3603-9/18/$31.00 ©2018 IEEE

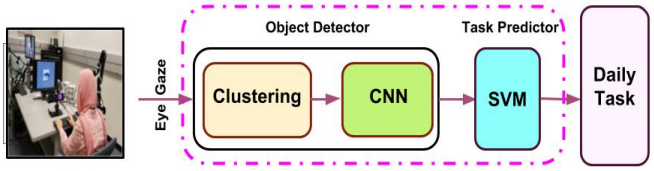


Fig. 1: Block diagram of the proposed framework.

considered as daily tasks. In this work, the kitchen environment is chosen, where a variety of tasks can be defined. We considered four tasks: preparing coffee, making smoothie, cooking, and washing. Completing each of these tasks involves the use of a couple of main and optional objects that are present in the kitchen environment. For example, preparing coffee involves using the coffee maker, coffee and cup as main objects and milk as an optional object. Note that there could be commonality in objects used in different tasks. For example, milk can be used for making smoothie as well as for making coffee. Additionally, we considered the possible scenario in which the user has no specific intention when looking at the image, and just explores the environment. Therefore, a total of five tasks (four intentional and one unintentional) with 12 objects are considered. Table I summarizes the tasks and their main and optional objects.

C. Data Acquisition

Eight volunteers (mean age of 26 ± 5) participated in the study. Written informed consents approved by the Rutgers IRB were obtained prior to experiments. Subjects sat in a quiet room, in front of a monitor. Eye gaze patterns were recorded using the EyeLink 1000 plus [10]. At the beginning of each experiment calibration was performed. Two of the participants wore eyeglasses and the calibration was done while they were wearing their glasses.

The experiment consists of 45 trials. In each trial, an image, selected randomly from a pool of 8 images of different kitchen environments, is displayed to the participant. The main and optional objects are positioned randomly in different images.

During the experiment, subjects sat in front of a monitor. For the 40 first trials, subjects are notified about the type of the task prior to each trial, and are instructed that once the image is displayed, they look at the main and optional objects involved in the mentioned task, based on their preference. For the last 5 trials, subjects selected the type of the task, themselves. At the end of each of these trials, subjects were asked to verbally mention the type of their intented task. The experiment was repeated two times for each subject, with five minutes of rest time in between. Calibration was performed for each subject, and before each experiment.

III. Results

A. Object Detector

As mentioned before, the OD module receives the raw eye gaze data as its input, and returns a 12 (total number of main and optional objects) dimensional presence probability vector as its output. Each element in the vector represents the presence probability for each object listed in Table I. The OD module consists of two submodules. Details for each submodules are described below.

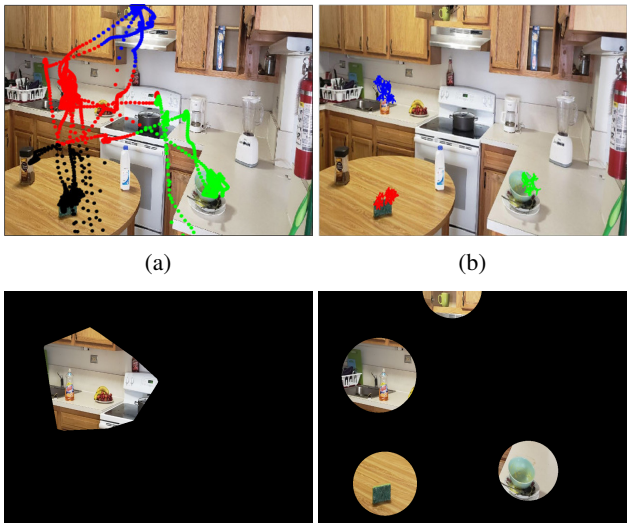
TABLE I: Description of the tasks and the list of main and optional objects involved in each task.

Task	Main Objects	Optional Objects
Making Coffee	Coffee Maker, Coffee, Cup	Milk
Making Smoothie	Blender, Fruits, Milk	Cup
Cooking	Pot, Spaghetti, Ketchup	Dish
Washing	Sponge, Liquid Soap	Cup, Dish, Pot
Unintentional	None	None

1) Clustering: In each trial, subjects look at the objects relevant to the intended task, and the eye tracker records their eye gaze movements. By overlaying the scanpath recorded from the eye tracker with the displayed image, specific areas in the image with more eye gaze points can be identified (see Fig 2). These areas are considered as regions of interest (ROIs). The aim of the clustering submodule is to identify and extract these ROIs from displayed images.

Fig. 2: An example of eye gaze patterns overlayed with the displayed image.

There are various unsupervised clustering approaches that can be utilized to identify the ROIs. Here, we considered using k-means [11] and DBSCAN [12] as clustering methods. k-means is a partition-based clustering approach, while DBSCAN is a density-based clustering method. An example of clustering results using both approaches is shown in Fig. 3. Given that the number of objects (hence clusters) can vary from task to task (and hence the true value for k is not known in advance in k-means), and that k-means attempts to cluster all the available gaze points into k clusters without recognizing the outliers, (and hence leading to clusters containing more than one object (see Figs. 3(a) and 3(d))), we settled for the DBSCAN approach. The DBSCAN clustering approach automatically finds the number of clusters based on the density of the eye gaze points, and can distinguish the outliers which are the points in low density regions (see Fig. 3(b)). After clustering is performed using DBSCAN, in order to extract the ROIs from the image, a convex hull [13] around each cluster is found, and a circle is inscribed around the convex hull (see Fig. 3(d)).



(a) (b) (c) (d)

Fig. 3: An example of results from the clustering module using k-means and DBSCAN: (a) Clustering with k-means (k=4); (b) Clustering with DBSCAN; (c) One extracted partition with k-means; (d) Extracted segments with DBSCAN.

2) Object Recognition by Deep Learning: Once the ROIs are extracted, the object in each region must be recognized. Note that due to possible calibration error and also individual variability in visual behavior, a complete image of the objects in extracted ROIs may not be visible. It is possible that a given ROI contains a partial view of an object, or overlapped objects.

In the traditional object detection approaches hand-crafted features like SIFT [14] are extracted, and these features are used for training a classifier. Since in these approaches features are not trained, the extracted features play an important role in setting the detection accuracy. In the case of partially observable or overlapped objects, these features can not represent the object perfectly. Consequently, the classifier fails to recognize the object correctly.

Since the goal of this work is to have a generalizable, scalable and robust object detector which can precisely recognize a wide variety of objects, a deep convolutional neural networks (CNN)-based approach [15] appears to be a good choice for the implementation of this submodule.

There exist several CNN-based architectures, such as Alexnet [16], Resnet [17], and others [18], for object classification. In this work, we used Resnet in our framework for object detection. Resnet allows to go deep instead of increasing the width of the network, addressing challenges such as slow training and reducing the vanishing gradient decent [17].

Given the small number of available images in our experiment, for training and validation of the model, images from imagenet [19] were used. 80% of the images were used for training and 20% for validation, and the trained model achieved 99% accuracy for the validation dataset.

For the testing phase, we used data from our own experiment. For all subjects, the extracted ROIs obtained from the clustering submodule were labeled, and used as input to the trained model. Since the ROIs are extracted from real experiment, they contain partially visible or overlapped objects. For our defined tasks, there is a total of 12 objects. The accuracy results for each object is shown in Fig. 4, demonstrating achieving larger than 91% for majority of the objects (10 out of 12).

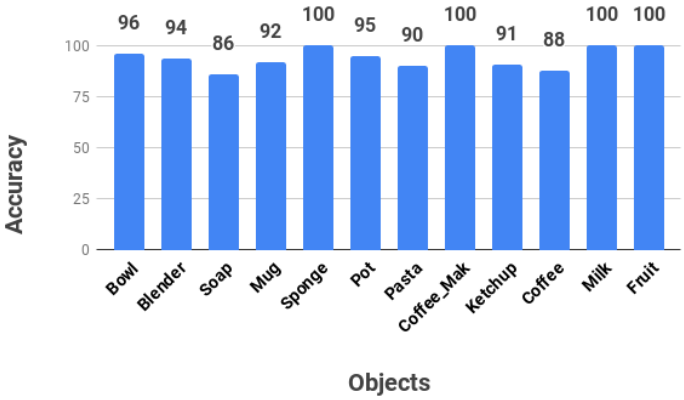


Fig. 4: Accuracy results for each object, from the outcome of the Object detector module.

B. Task Predictor

As discussed in Section II, each task involves two or three main objects and one or two optional objects. For this module, a support vector machine (SVM) [20] is utilized to learn the embedded relationship between objects, based on their presence probability (as identified from the OD submodule), and to predict the intended task.

The data for training the SVM is from our experiment. The challenge here is the existence of variability across individuals and experimental conditions. The eye gaze patterns are variable across individuals, due to error in camera calibration, variability in personal characteristics and visual behavior [21]. As an example, Fig. 5 shows results from the clustering submodule for three subjects who looked at same image, and intended the washing task. As can be seen, because of different eye gaze patterns among individuals, the clustering results cover different parts of the image across subjects. If subjects select one unrelated object, this variability can affect the final accuracy result from SVM.

The SVM model was trained and tested using the data from all subjects. Data for the first 40 trials of each subject was used for training, and data from the remaining trials was used for testing the model. Table II summarizes the results for accuracy, sensitivity and specificity, for each task. It can be seen that an average accuracy of 95.68% with a standard deviation of 3.43 is achieved across all tasks.

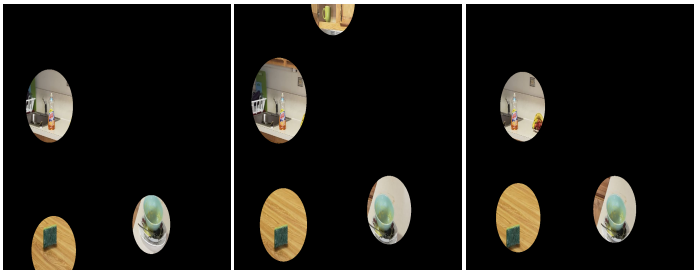


Fig. 5: DBSCAN results on eye gaze points for three subjects when they do the same task using the same kitchen image display.

TABLE II: Results for accuracy, specificity and sensitivity for the proposed framework, for each task.

Metric	Unintentional	Coffee	Smoothie	Cooking	Washing
Accuracy	91.1	97.3	97.7	94.6	97.7
Specificity	92.8	99.4	98.9	97.2	98.3
Sensitivity	84.4	89.5	93	84.7	95.4

IV. Conclusion

In this paper, we presented a generalized and scalable framework for predicting the user's intention through eye gaze patterns. The proposed framework consists of two main modules. First, an automatic density-based clustering approach is used to cluster the patterns of the eye gaze, and extracts the corresponding ROIs in the image. Next, utilizing ResNet, a deep convolutional neural network, which is generalizable, scalable and robust, the objects in ROIs are classified. Using CNN for object classification, resulted in high accuracy in object detection, even with partially observed or overlapped views of the objects. In the final step, the intended task is predicted using a SVM classifier. Experimental results demonstrated achieving an average accuracy of 95.68% across all considered tasks, including the scenario where the subject does not have a specific intention. While in this work we considered the kitchen environment, the proposed framework is easily generalizable to other environments, if the CNN model is trained using proper images. Future work involves taking into account the order of the selection of seen objects when predicting tasks.

Acknowledgment

This work was supported in part by DARPA under grant 68799LDSRP.

References

[1] S. N. McClung and Z. Kang, "Characterization of visual scanning patterns in air traffic control," *Computational intelligence and neuroscience*, 2016.

[2] H. Admoni and S. Srinivasa, "Predicting user intent through eye gaze for shared autonomy," in *Proceedings of the AAAI Fall Symposium Series: Shared Autonomy in Research and Practice (AAAI Fall Symposium)*, 2016, pp. 298–303.

[3] J. D. Webb, S. Li, and X. Zhang, "Using visuomotor tendencies to increase control performance in teleoperation," in *American Control Conference (ACC), 2016*, 2016, pp. 7110–7116.

[4] C.-M. Huang and B. Mutlu, "Anticipatory robot control for efficient human-robot collaboration," in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, 2016, pp. 83–90.

[5] V. Azizi, A. Kimmel, K. Bekris, and M. Kapadia, "Geometric reachability analysis for grasp planning in cluttered scenes for varying end-effectors," in *Automation Science and Engineering (CASE), 2017 13th IEEE Conference on*, 2017, pp. 764–769.

[6] C.-M. Huang, S. Andrist, A. Sauppé, and B. Mutlu, "Using gaze patterns to predict task intent in collaboration," *Frontiers in psychology*, vol. 6, p. 1049, 2015.

[7] M. R. Greene, T. Liu, and J. M. Wolfe, "Reconsidering yarbus: A failure to predict observers task from eye movement patterns," *Vision research*, vol. 62, pp. 1–8, 2012.

[8] C. Kanan, N. A. Ray, D. N. Bseiso, J. H. Hsiao, and G. W. Cottrell, "Predicting an observer's task using multi-fixation pattern analysis," in *Proceedings of the symposium on eye tracking research and applications*, 2014, pp. 287–290.

[9] A. H. Abolhassani and J. J. Clark, "Visual task inference using hidden markov models." in *IJCAI*, 2011, pp. 1678–1683.

[10] "https://www.sr-research.com/products/eyelink-1000-plus/."

[11] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[12] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[13] E. Yazdi, V. Azizi, and A. Nourollah, "Finding the convex hull of a simple polygon." in *CSC*, 2010, pp. 159–163.

[14] D. G. Lowe, "Object recognition from local scale-invariant features," in *The proceedings of the 7th IEEE international conference on Computer vision,*, vol. 2, 1999, pp. 1150–1157.

[15] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, M. Hasan, B. C. Van Esesn, A. A. S. Awwal, and V. K. Asari, "The history began from alexnet: A comprehensive survey on deep learning approaches," *arXiv preprint arXiv:1803.01164*, 2018.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[20] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[21] A. Coutrot, J. H. Hsiao, and A. B. Chan, "Scanpath modeling and classification with hidden markov models," *Behavior research methods*, vol. 50, no. 1, pp. 362–379, 2018.