Human Gaze Following for Human-Robot Interaction

Akanksha Saran[1], Srinjoy Majumdar[2], Elaine Schaertl Short[2], Andrea Thomaz[2] and Scott Niekum[1]

Abstract— **Gaze provides subtle informative cues to aid fluent interactions among people. Incorporating human gaze predictions can signify how engaged a person is while interacting with a robot and allow the robot to predict a human's intentions or goals. We propose a novel approach to predict human gaze fixations relevant for human-robot interaction tasks—both referential and mutual gaze—in real time on a robot. We use a deep learning approach which tracks a human's gaze from a robot's perspective in real time. The approach builds on prior work which uses a deep network to predict the referential gaze of a person from a single 2D image. Our work uses an interpretable part of the network, a gaze heat map, and incorporates contextual task knowledge such as location of relevant objects, to predict referential gaze. We find that the gaze heat map statistics also capture differences between mutual and referential gaze conditions, which we use to predict whether a person is facing the robot's camera or not. We highlight the challenges of following a person's gaze on a robot in real time and show improved performance for referential gaze and mutual gaze prediction.**

I. INTRODUCTION

Eye gaze is an indicator of engagement, interest and attention when people interact face-to-face with one another [1], [2]. Nonverbal behaviors such as eye gaze, can convey intentions and augment verbal communication [3], [4], [5]. Take the example of people communicating with one another as a team on a collaborative task of assembling furniture together; they may look at parts of the workspace farther away from them to communicate intentions about a nail or hammer they want to be passed by a teammate. Social gaze is also used by humans to regulate turn-taking in two-party [6], [7] and multi-party conversations [8], [9].

In a teacher-learner setup, parents can scaffold a child's learning process by directing their attention using gaze, thereby providing structure to the task [10], [9]. Human gaze fixations can similarly help a robot learner segment a teaching demonstration into steps (e.g. pausing to look at the lid of a jar, then picking it up and screwing it in), and determine the right aspects of the state to pay attention to during the demonstration [11]. For seamless interactions between humans and personal robots of the future, interpreting and reacting to this human social cue will be of significant importance.

We present a real-time approach to predict referential and mutual gaze of a human (gaze directed at an object and gaze directed towards the robot's camera/face respectively) from a robot's perspective embodied in the real world, and show improvement over an existing baseline. Our approach works with a monocular camera attached to a robot without the use of bulky or invasive devices like eye-trackers which can make a user uncomfortable during the interaction. Human gaze has been estimated without the use of eye trackers in images and videos [12], [13], but such approaches are not designed for embodied robots and their human partners interacting with objects in the real world. Prior works for gaze prediction on robots either generate eye gaze vectors when the person faces the camera [14] or use simple head pose estimators [15] which predict a large space as part of the human's attention versus specific objects. These prior works do not address the problem of a robot reliably predicting the object of a human partner's gaze while interacting with real-world objects, but rather focus on either predicting where on a computer screen a person might be focusing or estimating coarsely (large margin of error) a general direction in which the person is facing. By contrast, our approach aims to accurately predict the object of a human partner's attention in the real-world, which can more specifically inform a robot's next action in an interaction scenario. We extend the deep learning pipeline proposed by Recasens et al. [16] to incorporate knowledge about relevant task objects with interpretable parts of the network. We bring together different components such as a state-of-the-art face detector, object detector, and a gaze following algorithm in our system's code, to obtain real-time predictions about which objects a human might be attending to. We also show how this architecture can predict mutual gaze, i.e. when a person is facing the robot versus not, an important aspect of gaze prediction for human-robot interaction.

II. RELATED WORK

In this section we review two areas of prior work–computer vision algorithms for human gaze prediction and use of gaze in human robot interaction tasks.

A. Gaze Prediction with Computer Vision

Recasens et al. [16] developed a deep network architecture for monocular images to predict human gaze fixations. They use manually-labeled data to learn where the person is focusing their attention, given that both the person's head and the object of attention are visible in the image. It is meant to work primarily for a single 2D image, and does not use specific contextual task information such as knowledge

[1] Akanksha Saran and Scott Niekum are with the Department of Computer Science, University of Texas at Austin, Austin, TX 78712, USA. {asaran,sniekum}@cs.utexas.edu

[2] Srinjoy Majumdar, Elaine Short and Andrea Thomaz are with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712, USA. {srinjoy.majumdar,elaine.short}@utexas.edu, athomaz@ece.utexas.edu

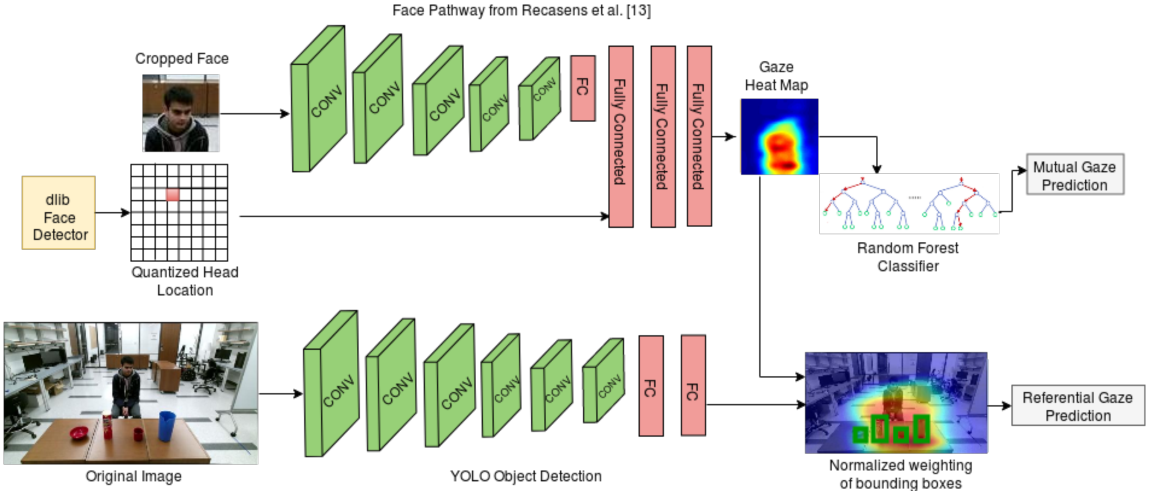


Fig. 1. The pipeline for the algorithm presented in this paper.

about objects relevant to the task, which might be available apriori in a human-robot interaction scenario. It also does not resolve prediction of mutual gaze (the person looking towards the camera). We extend this work to address these shortcomings, create a real-time gaze tracking system and evaluate it on a robot designed for interacting with humans.

Baltrusaitis et al. [14] developed an open-source toolkit for facial behavior analysis including gaze vector prediction, head pose prediction, facial landmark detection and facial action unit recognition. They show state-of-the-art results in all these domains. However, their gaze prediction algorithm does not give referential and mutual gaze estimations. Rather, they train their models on facial data where subjects look at different parts of a computer screen. Prior work has used head pose as a simple and coarse indicator of gaze [15], but we also incorporate task relevant information into our predictions.

Other prior work has modeled gaze prediction for egocentric videos by leveraging the camera wearer's head motion and hand location from the video and combining them to estimate where the eyes look [17]. This requires the user to wear a camera rather than interact with a robot that has a camera. Gaze following has been shown to work in videos where a person in one frame is looking at objects in a different frame [12]. Estimates of saliency, gaze pose, and geometric relationships between views are computed with a deep network using gaze as supervision. This requires the video to be post-processed (access to frames before and after the current frame) and hence is not suitable for real-time evaluation. Krafka et al. [18] developed an eye tracking software for commodity hardware such as mobile phones and tablets, without the need for additional sensors or devices using a convolutional neural network in real time (10-15 fps). However, this can only predict what part of the mobile screen the person is looking at and does not follow the gaze of the person while attending to other objects in the real world. Vasudevan et al. [13] incorporate the approach of Krafka et al. [18] in their work for estimating human gaze along with using appearance and motion cues for localizing objects referred to in natural language. This work is also restricted to using gaze estimates only when people stare at a computer screen (while watching a video to describe an object), and not when humans interact with objects in the real world.

B. Use of Gaze for Human-Robot Interaction

There is a rich body of work on eye gaze for human-robot interaction. The survey paper by Admoni et al. [5] outlines previous work which used gaze for human-robot interaction. Gaze information enables the establishment of joint attention between the human and robot partner, the recognition of human behavior [19] and the execution of anticipatory actions [20]. These prior works focus on gaze cues generated by the robot, however, in our work, we intend to interpret 'human' gaze cues during an interaction with the robot. This can enhance future applications with a robot responding in accordance with the human partner's intentions [21], [22].

Prior work incorporating human gaze cues for Human-Robot Interaction have used specialized hardware like eye-trackers [21], [23], [24]. Even though use of such hardware provides robust estimates of human gaze, they deviate from a natural interaction between humans and robots in the real world, where such hardware might not be available or make the user uncomfortable during the interaction. Recently, Penkov et al. [24] used demonstrations from a person wearing specialized eye tracking hardware along with an egocentric camera to simultaneously ground symbols to their instances in the environment and learn the appearance of such object instances.

Most closely related to our work is that of Lemaignan et al. [15], which uses head orientation and the visual focus of attention of humans to estimate gaze in real-time. Human gaze estimation is used to evaluate the human's focus of attention with the concept of "with-meness“ (to what extent the user is "with" the robot). Similar to our work, it makes use of contextual task information, such as attentional targets that are expected by the robot *a priori*. However, they use a very wide field of visual attention around a vector for head pose direction (a cone spanning 40 degrees). Everything that lies in that span of attention and on the planar task surface is used as the outcome of the referential gaze predictor. Our approach is comparatively fine-grained in the sense that it selects a single object as the focus of the human partner's attention at a given point of time.

III. Approach

Our approach uses a deep neural network to follow the gaze of a human from a 2D image, provided both the person and object of attention are visible in the image. In our experiments, we use a Kinect at the location of the robot's eyes/face, as both the person and objects are in the field of view from that mounting point. However, the camera could be placed anywhere on the robot as long as both the person's face and the objects of attention are visible in the image frame. We build on the work by Recasens et al. [16], who propose a deep network with two pathways: one for estimating the head direction and another for salient objects in the image. The face pathway for head direction takes as input the location of the head and the image of the person's face coming from any face detection algorithm to estimate which part of the image the person might be looking at. The saliency pathway takes as input the entire image to detect salient objects visible in the image, irrespective of the person's location. Combining these two pathways, they predict the most likely point in the image where the person might be fixating.

At the end of these pathways, fully connected layers output a 169-dimensional feature which can be visualized as a 13×13 heatmap overlaid with the image as shown by Recasens et al. [16]. Visualizing heatmaps in both these pathways for a video, we found that the saliency pathway had insignificant variance and the distribution of the heatmap values do not change over time when the person fixates at different objects, if the camera position is fixed. Most of the changes happen in the face pathway as the person moves around in a video. To improve gaze predictions, we utilize the face pathway's heatmap from this network and combine it with the knowledge of objects available in the workspace that a human is likely to be fixating at (Section III-B). Also, Recasens et al. [16] do not deal with cases when a person looks straight at the camera/robot (mutual gaze), but works only with referential gaze. We again use the face pathway heatmap to estimate whether a person is facing the robot or not (Section III-C). The entire pipeline for the proposed approach is shown in Fig. 1.

A. System Integration

To build a real-time system for gaze following, we incorporated several different components. The face pathway in the deep network requires the face location as input. We make use of a *dlib* based face detector [25]. We found this deep learning based face detector to be the most robust in terms of variation with head pose, compared to other open source face detection implementations. We smooth over the face location in the real-time video using a 1D Gaussian filter on the past three image frames. We also use Yolo [26], a state-of-the-art deep learning approach for object detection to incorporate contextual task knowledge into our method. We fine-tune this detector for 4 objects from the YCB dataset [27] under different configurations, lighting conditions, and environments. At test time, it outputs bounding box coordinates for each of the detected objects.

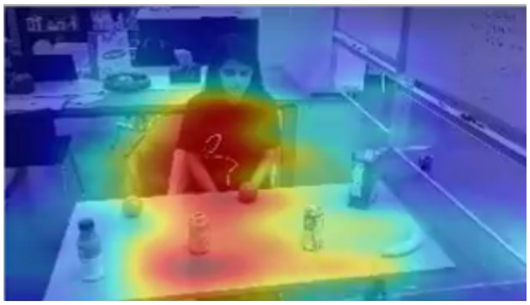


Fig. 2. Gaze heat map from the face pathway of the gaze following deep network by Recasens et al. [16].

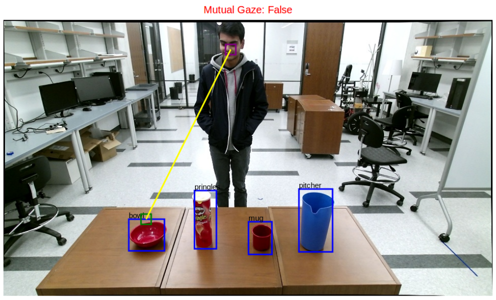


Fig. 3. Gaze prediction in terms of object of a person's attention, working in conjunction with object detection and face detection.

The gaze predictor, face detector and Yolo object detector all run in parallel, and we use the last update from each component on each frame (Fig. 3). We developed a web-based GUI for real-time visualization of the output of the system, and to annotate ground truth during data collection. Using a GTX 1070 NVIDIA GPU, we obtain an image processing rate between 7–15 FPS for the overall system. Such a rate might not detect very fast eye movements or saccades, but is sufficient to track eye fixations. All components of this system rely on the use of a GPU. Apart from gaze prediction with convolutional neural networks, state of the art face detectors and object detectors also typically use deep learning and hence require GPUs to run at real-time. For low-compute platforms, options such as cloud-based GPU computation, a modular GPU component externally used with a robot or compressed networks meant to reduce computation demands can be used as alternatives.

B. Referential Gaze

We use the intermediate output of the face pathway from [16] in our approach (Fig. 2), i.e. the output of the 4th fully connected layer, a 13x13 map (p_{ij}). This gaze map when overlayed on the image, helps visualize what parts of the image the network is giving more weight to for the gaze following prediction. We observe this map updates significantly as the person looks in different directions and at different objects. Recasens et al. [16] combine it with the saliency map from the saliency pathway with element-wise product followed by fully connected layers to get coordinates of the gaze fixation point in the image. However, when we visualized the saliency maps for different users, we found that it did not show significant variation as the person moved their head/gaze in the same scene. We expected the hot spots in the saliency heat map to focus on different objects as the person looked in different directions, but the hot spots were found to be concentrated on specific regions in the scene throughout. We even replaced the saliency map with a custom map at the end of the saliency pathway, placing more weight on the objects detected by yolo compared to the rest of the image, and found it did not improve gaze following accuracy compared to the default pipeline. This is plausible because the network internally learns weights from the training data to focus on specific parts of the image, which the authors visualized and called saliency, but it may not be suitable to interpret that exactly as task relevant saliency in the scene. Instead, we process the gaze map and obtain a likelihood score for each of the objects being the target of the person's gaze. We compute the sum of the upsampled gaze map's values inside the object bounding boxes, normalized by the area of the bounding boxes. This gives a score for each bounding box proportional to how likely it is to be the object of fixation. The object o_k with the highest score is chosen to be the object of human's attention (obj):

$$obj = \arg\max_k \frac{\sum_{(i,j)\in o_k} p_{ij}}{area_{o_k}}. \tag{1}$$

To compare against baseline methods, we snap the prediction of the default network by Recasens et al. [16] to an object which is the closest to the predicted gaze coordinates. Although the baseline gives a coordinate on the image frame, it rarely falls inside the bounding box. Hence, we make their approach snap to a bounding box for which any one of the four corners are closest to the predicted coordinates. This provides a simple, fast, way of calculating this snapping, but other methods could be used in this framework. Both the baseline version and the proposed approach then predict the outcome of referential gaze as one of the pre-determined objects known ahead of time. We also compare against another baseline, Open Face [14], which predicts 3D eye gaze vectors from the camera image. It provides 3D vectors for both the eyes. We take their average to get a single gaze direction vector which we project to the image plane. We then compute this projected vector's normal distance to the Yolo [26] bounding boxes. The bounding box of the object with the shortest distance is chosen as the object of attention.

C. Mutual Gaze

We process the face pathway's output to determine whether the person is looking at the robot using a random forest classifier. We found that the gaze heat map is concentrated with most of its weight in one region of the image in cases where the user is fixating on an object (referential gaze) and the heat map is more evenly distributed throughout the image when the user is facing the camera (mutual gaze) as shown in Fig. 4. The random forest is trained to distinguish between these two patterns with data collected in our lab. Since we only use intermediate gaze heat maps from the network as input to the random forest, we hypothesize that it will generalize well to other users that are not part of the training data. We use a random forest classifier as opposed to a deep network which requires large amounts of data to train well, because of the limited amount of training data collected with users.

We also compare our method against the Open Face eye gaze vectors [14]. Open Face provides a 3D vector with respect to each eye of the user. We compute the angle between each eye gaze vector and the eye to camera vector. We average the angles for the left and right eye and use a threshold on it to estimate whether the person is looking towards the robot's camera or away from it (towards an object).

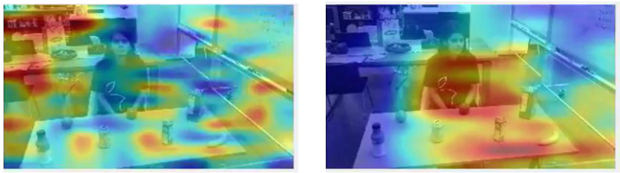

(a) Mutual Gaze
(b) Referential Gaze

Fig. 4. Heat map from Recasens et al. [16] for (a) mutual gaze condition being true (b) mutual gaze condition being false. The hotspots in the heatmap are more spread out when the user is looking at the robot compared to the case when fixating at an object.

IV. Evaluation

To evaluate our proposed referential and mutual gaze algorithms, we capture data of users gazing at different objects in front of a robot. We collected data from 10 participants: 5 males, 5 females. The task set up includes a robot with a Kinect mounted on its head, across the table from a human subject (Fig. 5). Each subject is given instructions to stare at specific objects or at the robot for a fixed duration of 5 seconds as prompted by an observer on the side (out of the Kinect's field of view). The order of object placement and order of staring at different things is randomized for each subject. Ground truth on where the subject is instructed to look is annotated as a bounding box on a graphical user interface, by the observer on the side. The images are streamed for annotation via ethernet from the robot. We use a set of 4 objects (Fig. 5) from the YCB dataset [27] placed roughly 15 inches apart from one another.

Evaluation for referential gaze computation is done under 5 different conditions (as shown in Fig. 6): (1) the user is sitting across the table and objects are in a straight line, (2) the user is standing across the table and objects are placed

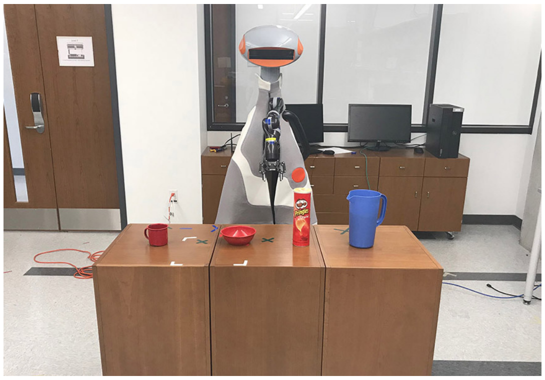
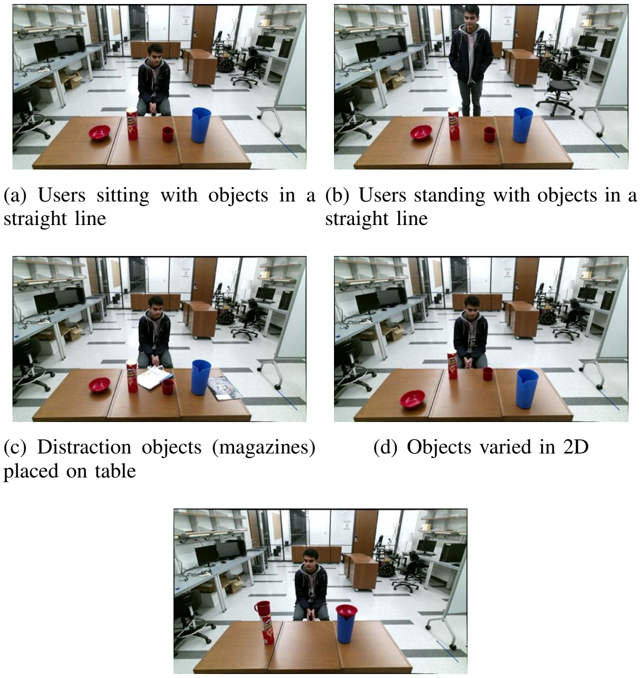


Fig. 5. Experimental setup with 4 YCB objects placed on a table between a person and a robot.

in a straight line, (3) the user is sitting across the table and a couple of distraction objects (magazines) are placed on the table to add clutter, (4) the user is sitting and the objects are scattered on the table and not placed in a straight line, (5) the user is sitting and the objects are stacked on top of one another (object stacking). Mutual gaze computation is done under two conditions where the user changes configuration, i.e. standing versus sitting. These varied conditions were chosen to test the limits of the system under a range of user and object configurations that are plausible while interacting with a robot. The subjects were not asked to stand or sit at a specific distance from the robot, rather whatever the users felt comfortable with. Also, the objects were roughly placed at uniform distances based on the conditions, and not at exactly the same marked locations.



(a) Users sitting with objects in a straight line

(b) Users standing with objects in a straight line

(c) Distraction objects (magazines) placed on table

(d) Objects varied in 2D

(e) Objects varied in 3D

Fig. 6. Different experimental conditions under which gaze following is evaluated.

A. Referential Gaze

To evaluate referential gaze predictions, we compare our proposed method against modified versions of the Open Face eye gaze vectors [14] and the approach of Recasens et al. [16] on all conditions as described in Section III-B. These methods are modified to snap to a single object of interest to compare against our work. By themselves, these baselines do not directly aim to snap at a relevant object, but rather provide a general gaze direction or unconstrained gaze fixation coordinates.

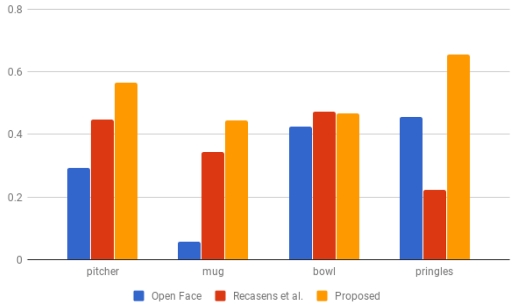


Fig. 7. Average F1 score (across 10 users) of referential gaze computation for 4 YCB objects placed in a straight line when users are sitting across a table from the robot.

For the first condition where the users are sitting across a table from the robot when objects are placed in a straight line, the average F1 scores of detecting each object of attention are shown in Fig. 7 (higher numbers are better). We find that across all objects, we get an overall improvement of 20% in the average F1 score, and an improvement in each object category as well. Particularly, our approach improves performance on the ‘pringles’ and ‘pitcher’ objects, which are the two largest objects among the set. This highlights the strength of our approach which constrains the system to only focus on objects of relevance. We also find an overall improvement of about 16% (from 37.7% to 53.7%) in the average F1 score of the 4 horizontal locations (regardless of the object placed there) for this condition.

For all the five experimental conditions, the average F1 scores of object prediction across all users and all objects are shown in Table I (higher numbers are better). We find that for four of the five conditions, our approach outperforms the baseline methods. When the users are standing and objects are placed in a straight line, the ‘mug’ and ‘pringles’ objects get the maximum boost in performance with our approach. When distraction objects like magazines are placed in between the objects in a straight line, the ‘mug’, ‘pitcher’ and ‘pringles’ gain in prediction accuracy. For the case when objects are scattered on the table, the baseline from Recasens et al. [16] does slightly better, particularly for the ‘pitcher’ and ‘bowl’ whereas we perform better for the ‘pringles’. For when objects are stacked on top of each other, we find that only one of the objects in the stack is predicted more commonly, which highlights a specific area of improvement for any gaze prediction algorithm. In general, we find that our algorithm fails more often for smaller objects and favors the 'pringles' and 'pitcher' more which cover a larger area in the heat map. It is important to note that there is a possibility of error in object detection to get propagated to our approach and the Open Face baseline. We found that the Open Face implementation [28] failed to detect one of our user's face reliably for a couple of conditions. This user was discarded from the evaluation for Open Face.

TABLE I

AVERAGE F1 SCORES FOR REFERENTIAL GAZE PREDICTION ACROSS ALL USERS AND ALL OBJECTS UNDER DIFFERENT CONDITIONS

Condition	Open Face [14]	Recasens et al. [16]	Ours
Users Sitting	0.301	0.372	**0.533**
Users Standing	0.266	0.369	**0.474**
Distraction Objects	0.298	0.382	**0.530**
Objects Scattered	0.311	**0.605**	0.577
Objects Stacked	0.356	0.302	**0.345**

The average distance error in prediction, in terms of the number of hops of misclassified object from the ground truth is shown in Table II (lower numbers are better). If the misclassified object is adjacent to the ground truth object, the hop would be 1. The number of hops are averaged over all image frames and all users. Our proposed approach outperforms the baselines in three of the five conditions. However, there is scope for improvement in performance even with the proposed approach. The two conditions where we do worse are when the user is standing, and when the objects are scattered. When the user is standing, potentially a wider field of view can be covered by the hot spots in the heatmap. This can lead to more variation in the error of prediction. When objects are scattered, hot spots in the heat map are not able to capture the variation in distance of the object from the person, causing our approach to provide larger distance error in misclassification. This demonstrates the challenges of following human gaze accurately without specialized eye tracking hardware and we hope this encourages other researchers to work towards this problem.

B. Mutual Gaze

We evaluate our mutual gaze random forest with 5-fold cross validation over 10 users (i.e. use data from 8 users to train the random forest and test on the remaining 2 for each fold). Our random forest has 20 trees each with a maximum depth of 10. The classifier is evaluated under the two conditions where the user changes their configuration (standing versus sitting). Comparisons against the Open Face eye gaze vectors [14] are shown in Table III (higher numbers are better). We find that our approach obtains an F1 score of 72.3% over the two classes (mutual versus non-mutual) for the sitting user condition and 74.2% for the standing user condition. There is an improvement of more than 20% over the baseline for both conditions. The high variance (up to 105.88 degrees across both conditions and labels) of the angles between the eye gaze vector and the head to camera vector generated by Open Face , lower it's overall F1 scores. This shows that our approach is able to generalize well to new users because the gaze heat maps from the deep network are used to train the classifier instead of the original image. We expect a user-specific classifier would further improve performance.

TABLE II

AVERAGE LOCATION ERROR IN REFERENTIAL GAZE PREDICTIONS ACROSS ALL USERS AND ALL OBJECTS UNDER DIFFERENT EXPERIMENTAL CONDITIONS

Condition	Open Face [14]	Recasens et al. [16]	Ours
Users Sitting	0.778	0.656	**0.430**
Users Standing	0.726	**0.465**	0.579
Distraction Objects	0.598	0.666	**0.432**
Objects Scattered	0.710	**0.313**	0.459
Objects Stacked	0.635	0.687	**0.587**

TABLE III

AVERAGE F1 SCORES FOR MUTUAL GAZE PREDICTION ACROSS ALL USERS UNDER DIFFERENT EXPERIMENTAL CONDITIONS

Condition	Open Face [14]	Ours
Users Sitting	0.504	**0.723**
Users Standing	0.534	**0.742**

V. CONCLUSION

In this work, we present an integrated system for the challenging problem of a robot following a human partner's gaze, i.e. object of their attention in real-time. We focus on the idea of predicting human gaze without the use of any specialized eye trackers, which either make the interaction uncomfortable or cannot estimate the gaze when the user is not fixating on a fixed resolution screen. Our approach assumes that both the person's face and the object of attention are visible in the image frame. We evaluated our system with 4 objects on a table, and the person sitting or standing at an approximate fixed distance from the objects, along the orthogonal axis of a static camera on a robot. We propose a method that builds on top of a state-of-the-art gaze prediction algorithm to predict both referential and mutual gaze. We show improved performance for fixating on an object over baseline algorithms. It is important to note that while the baseline algorithms provide a smooth trajectory as the person moves their head, they perform poorly when snapping to the closest object. Our work specifically looks at the case where we care about the specific object being fixated upon, instead of a general direction or coordinate for gaze fixation on the image frame. We find that objects bigger in size gain prediction performance with our approach. This warrants the need for algorithms which focus on smaller objects which might be of importance in a human-robot collaboration task. Our work highlights the challenges of capturing gaze outputs accurately in real time.

We automated the evaluation process by recording video data and annotating the objects on a user interface. Manually annotating would be more accurate but also more time consuming. The unconstrained environment of the experiment, with just a general instruction for participants to stare at objects, can lead to errors during the evaluation including inconsistency in following the experimenter's instruction, errors in face detection and object detection, and slowing of the system due to data being written to disk. We work under the constraints of these possible errors, which potentially provide a lower performance during evaluation.

Further evaluations of this integrated system on a moving robot platform would reveal its robustness to disturbance in the image and change of viewpoints, which makes the gaze-following problem more complex. Furthermore, predicting gaze when users are not fixating on an object but rather quickly glancing at objects during a task, will be more challenging. Future work also entails extending this system to multiple people for teams of humans or robots to work together. Our work highlights the challenge of predicting gaze without specialized hardware for natural interactions, as there is scope for improvement in prediction performance and particularly when objects are stacked on top of each other. In all our conditions, none of the objects were occluded, which could add further challenges to following human gaze. This work is a first step towards integrating embodied task knowledge with gaze tracking to enable robots to engage in more seamless collaborations with humans by understanding their gaze behavior.

ACKNOWLEDGMENT

This work has taken place in the Personal Autonomous Robotics Lab (PeARL) and the Socially Intelligent Machines (SIM) Lab at The University of Texas at Austin. PeARL research is supported in part by NSF (IIS-1638107, IIS-1724157). SIM research is supported in part by ONR (N000141612835, N000141612785) and NSF (1564080, 1724157).

REFERENCES

[1] M. Argyle and M. Cook, "Gaze and mutual gaze." 1976.

[2] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," in *5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010.

[3] M. Argyle, "Non-verbal communication in human social interaction." 1972.

[4] R. J. Jacob and K. S. Karn, "Eye tracking in human-computer interaction and usability research: Ready to deliver the promises," in *The Mind's Eye*. Elsevier, 2003, pp. 573–605.

[5] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: A review," *Journal of Human-Robot Interaction*, vol. 6, no. 1, pp. 25–63, 2017.

[6] Y. Kuno, K. Sadazuka, M. Kawashima, K. Yamazaki, A. Yamazaki, and H. Kuzuoka, "Museum guide robot based on sociological interaction analysis," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2007.

[7] A. Yamazaki, K. Yamazaki, Y. Kuno, M. Burdelski, M. Kawashima, and H. Kuzuoka, "Precision timing in human-robot interaction: coordination of head movement and utterance," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2008.

[8] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, "Footing in human-robot conversations: how robots might shape participant roles using gaze cues," in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction (HRI)*, 2009.

[9] J. G. Trafton, M. D. Bugajska, B. R. Fransen, and R. M. Ratwani, "Integrating vision and audition within a cognitive architecture to track conversations," in *3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2008.

[10] J. V. Wertsch, N. Minick, and F. J. Arns, "The creation of context in joint problem-solving." 1984.

[11] Y. Nagai, C. Muhl, and K. J. Rohlfing, "Toward designing a robot that learns actions from parental demonstrations," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2008.

[12] A. Recasens, C. Vondrick, A. Khosla, and A. Torralba, "Following gaze in video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] A. B. Vasudevan, D. Dai, and L. Van Gool, "Object referring in videos with language and human gaze," *arXiv preprint arXiv:1801.01582*, 2018.

[14] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[15] S. Lemaignan, F. Garcia, A. Jacq, and P. Dillenbourg, "From real-time attention assessment to with-me-ness in human-robot interaction," in *The 11th ACM/IEEE International Conference on Human Robot Interaction (HRI)*, 2016.

[16] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba, "Where are they looking?" in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[17] Y. Li, A. Fathi, and J. M. Rehg, "Learning to predict gaze in egocentric video," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[18] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," *arXiv preprint arXiv:1606.05814*, 2016.

[19] W. Yi and D. Ballard, "Recognizing behavior in hand-eye coordination patterns," *International Journal of Humanoid Robotics*, vol. 6, no. 03, pp. 337–359, 2009.

[20] C.-M. Huang and B. Mutlu, "Anticipatory robot control for efficient human-robot collaboration," in *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016.

[21] L. K. Pinpin, R. S. Johansson, C. Laschi, and P. Dario, "Gaze interface: Utilizing human predictive gaze movements for controlling a hbs," in *2nd IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics*, 2008.

[22] A. S. Clair, R. Mead, M. J. Matarić, *et al.*, "Monitoring and guiding user attention and intention in human-robot interaction," in *ICRA-ICAIR Workshop, Anchorage, AK, USA*, 2010.

[23] 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.

[24] S. Penkov, A. Bordallo, and S. Ramamoorthy, "Physical symbol grounding and instance learning through demonstration and eye tracking," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[25] O. M. Parkhi, A. Vedaldi, A. Zisserman, *et al.*, "Deep face recognition." in *British Machine Vision Conference (BMVC)*, 2015.

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[27] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The YCB object and model set: Towards common benchmarks for manipulation research," in *International Conference on Advanced Robotics (ICAR)*, 2015.

[28] "openface_ros," https://github.com/interaction-lab/openface_ros.