Understanding Teacher Gaze Patterns for Robot Learning

Akanksha Saran
Computer Science Department
University of Texas at Austin
asaran@cs.utexas.edu

Elaine Schaertl Short
Electrical and Computer Engineering
University of Texas at Austin
eshort@utexas.com

Andrea Thomaz
Electrical and Computer Engineering
University of Texas at Austin
athomaz@ece.utexas.edu

Scott Niekum
Computer Science Department
University of Texas at Austin
sniekum@cs.utexas.edu

Abstract: Human gaze is known to be a strong indicator of underlying human intentions and goals during manipulation tasks. This work studies gaze patterns of human teachers demonstrating tasks to robots and proposes ways in which such patterns can be used to enhance robot learning. Using both kinesthetic teaching and video demonstrations, we identify novel intention-revealing gaze behaviors during teaching. These prove to be informative in a variety of problems ranging from reference frame inference to segmentation of multi-step tasks. Based on our findings, we propose two proof-of-concept algorithms which show that gaze data can enhance subtask classification for a multi-step task up to 6% and reward inference and policy learning for a single-step task up to 67%. Our findings provide a foundation for a model of natural human gaze in robot learning from demonstration settings and present open problems for utilizing human gaze to enhance robot learning.

Keywords: Learning from demonstrations, Eye gaze, Kinesthetic Teaching, Learning from observations

1 Introduction

Eye gaze is an important social cue that humans use to convey goals, future actions, and mental load [1, 2] both in verbal and non-verbal settings. 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 [3, 4]. As in human-human interactions, we hypothesize that gaze can play a role in guiding robot learning from humans. To understand what role human gaze plays when humans teach robots, we study eye gaze behaviors in the context of robot learning from demonstrations (LfD) [5], a powerful, natural framework that allows non-experts to communicate rich task knowledge to robots by showing them how to perform a task. We focus on two modalities of LfD for robot manipulation [6]: (1) learning via keyframe-based kinesthetic teaching (KT) in which the joints of a robot are moved by a human teacher through specific points or keyframes while the robot records its joint configurations at these keyframes [7], and (2) learning from observation, specifically video demonstrations, in which a robot can passively observe a human performing the task and learn how the demonstrated actions translate to its own body to achieve the same goal. Video demonstrations are often freely available on the web for many skills needed in offices or households by robots, which makes them a popular choice for robot learning. Learning algorithms for these techniques typically use trajectories of state-action pairs directly or indirectly. In addition to knowledge about actions, information about teacher intent in the form of eye gaze can enhance learning from demonstrations in terms of generalizing to new environments and learning with fewer demonstrations.

To use eye gaze for LfD algorithms, it is necessary to understand gaze behavior during the interaction for a specific demonstration type. The psychological literature has characterized gaze behavior of people performing certain manipulation tasks with their own hands, like moving objects around

3rd Conference on Robot Learning (CoRL 2019), Osaka, Japan.

obstacles [8] or making tea [9]. These studies show that gaze follows the objects involved in the task [10] and that eye gaze precedes hand motion [8, 11]. These insights can be applicable to video demonstrations for a robot. However, our work is the first to study eye gaze for paired kinesthetic teaching (KT) and video demonstrations with an eye toward ways it can be used computationally. Thus in this work, we aim to characterize the gaze behavior of human teachers demonstrating the same task under both teaching paradigms to a robot. We perform a data collection study in which human subjects wear an eye tracker to provide accurate ground truth for gaze fixations. In practice, our findings should be useful even without access to a gaze tracker, through the use of vision based algorithms to predict gaze fixations [12, 13].

In our study, we find that users spend most of their time fixating on objects which are relevant for completing the task. Moreover, human gaze can reveal information about the human's intentions in otherwise ambiguous situations and help predict the reference frame with respect to which certain keyframes or actions are demonstrated. We also show that human gaze is a meaningful feature to distinguish between keyframes which demarcate semantically different actions (step keyframes) versus contiguous keyframes which belong to the same semantic action (non-step keyframes) such as multiple keyframes shaping a pouring motion. These insights open up exciting new avenues for research in learning from human gaze such as automatic segmentation of a task into subtasks, inferring intentions and goals of such subtasks, and efficient robot learning. We observe up to 6% improvement in subtask classification with gaze during a multi-step task for both demonstration types. As another potential application, we show improved goal inference and policy learning for a manipulation task by augmenting Bayesian inverse reinforcement learning with gaze information from the demonstrator. Policy loss improves by 67.4% for video demonstrations compared to 53.75% for KT demonstrations, suggesting that video demonstrations are a richer and more compact source of intention-revealing gaze signals.

2 Related Work

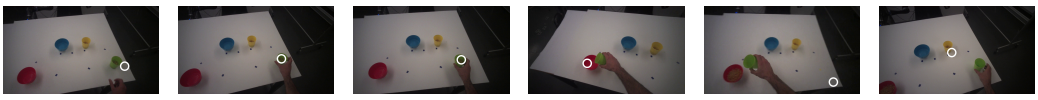
Human gaze and attention are known to be task-dependent and goal-oriented [14]. Flanagan and Johansson [15] demonstrated that adults predict action goals by fixating on the end location of an action before it is reached, both when they execute an action themselves and when they observe someone else executing it. Single fixations have identifiable functions (locating, directing, guiding, and checking) related to the action to be taken. Hayhoe and Ballard [10] show that the point of fixation in a given scenario may not be the most visually salient location, but rather corresponds to a location important for the specifications and spatio-temporal demands of the task. This line of investigation has been used in extended visuo-motor tasks such as driving, walking, sports, and making tea or sandwiches [16, 17]. It has also been found that eye gaze fixations are tightly linked in time to the evolution of the task and very few irrelevant areas are fixated upon [18], implying that the control of fixation comes principally from top-down instructions, not bottom-up salience. Subjects appear to use gaze to select specific information required at a specific point of time in a block manipulation task [8, 19]. These studies suggest that gaze would be helpful in predicting the intention or goal location of human manipulation actions. In our work, we study such human gaze patterns for paired video and KT demonstrations and recover characteristics specific to demonstrations for robots.

There is also a rich body of work on eye gaze for human-robot interaction [20]. Hart et al. [21] use nonverbal cues including gaze to study timing coordination between humans and robots. Gaze information has also been shown to enable the establishment of joint attention between the human and robot partner, the recognition of human behavior and the execution of anticipatory actions [20]. However, these prior works focus on gaze cues generated by the robot and not on gaze cues from humans. More recently, Aronson et al. [22] studied human gaze behavior for shared manipulation, where users controlled a robot arm mounted on a wheelchair via a joystick for assistive tasks of daily living. Novel patterns of gaze behaviors were identified, such as people using visual feedback for aligning the robot arm in a certain orientation and cognitive load being higher for teleoperation versus the shared autonomy condition. However, eye gaze behavior of human teachers has not been studied in the context of robot learning from demonstrations.

Prior research in computer vision has established that task and activity recognition in egocentric videos (similar to video demonstrations in our setup) can benefit from human gaze [23, 24, 25, 26, 27]. While some of these works predict human gaze as an intermediate output of a deep network classifying human activities [24, 26], others either jointly predict gaze and action labels with a

(a) Pouring Task (b) Placement Task (c) Third person view of 2 demonstration modalities

Figure 1: Task completion configurations for: (a) Pouring Task - where pasta from the green cup is poured into the red bowl and from the yellow cup into the blue bowl; (b) Placement Task - where the green ladle is placed either to the right of the yellow bowl or to the left of the red plate (note both instructions refer to the same ambiguous location). (c) A third person view of a KT and a video demonstration provided by the same user.



(a) Reach (b) Grasp (c) Transport (d) Pour (e) Return (f) Release

Figure 2: Semantic keyframes at the start of each pouring subtask with corresponding gaze fixation points (white circle). The reference frames for these subtasks are (i) green cup for reach, grasp, release; (ii) yellow cup for reach, grasp, release; (iii) red bowl for transport and pour; (iv) blue bowl for transport and pour; (v) white table for return.

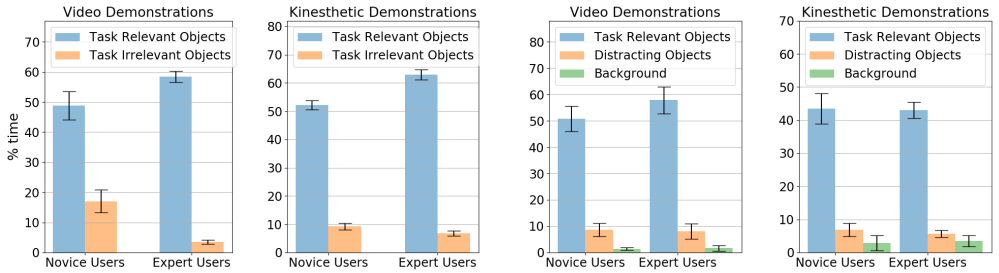
probabilistic generative model explicitly modeling properties of gaze behavior [25] or use human gaze as a weak supervisory signal in a latent SVM learning framework [27]. By contrast, we show a proof of concept for gaze aiding classification of subtasks or actions in egocentric videos of *both* video and KT demonstrations for a multi-step task, which has the potential to further enable task segmentation and policy learning per step.

There has also been some recent work on utilizing human eye gaze for learning algorithms. Penkov et al. [28] used demonstrations from a person wearing an 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. Ravichandar et al. [29] use gaze information as a heuristic to compute a prior distribution of the goal location for reaching motions in a manipulation task. This allows for efficient inference of a multiple-model filtering approach for early intention recognition of reaching actions by pruning model-matching filters that need to be run in parallel. In our work, we show that the use of gaze in conjunction with state-action knowledge can improve reward learning via Bayesian inverse reinforcement learning (BIRL) [30].

3 Data Collection And Analysis

We designed a two-way 2×2 mixed-design human subjects study (user type: novice or expert $\times$ gaze fixation area: task relevant objects or task-irrelevant object/area) for two household tasks relevant to personal robots: pouring and placement. The task layouts and details, which were kept the same across all users, are shown in Fig.1 (a), (b). The order of the tasks and demonstration types were counterbalanced across users. We recruited 20 participants (14 males, 6 females): 10 expert users who had operated or worked with a robot arm, and 10 novice users who had no prior experience operating a robot. Each participant was allowed one practice round for each demonstration type on the task they were assigned to do first. After one round of practicing, participants completed 6 demonstrations (3 KT, 3 video) for the pouring task and 4 demonstrations (2 KT, 2 video) for the placement task. This amounted to a total of $\sim$27 minutes of video demonstration data and $\sim$124 minutes of KT demonstration data.

Users wore the Tobii Pro Glasses 2 eye tracker and provided demonstrations to our robot which has a 7 degree of freedom (DOF) Kinova arm, and a Kinect sensor mounted on its head. KT demonstrations required users to physically move the robot's arm while video demonstrations were given standing in front of the robot, in the robot camera's view (Fig. 1(c)). The eye tracker is equipped with two cameras for each eye to track gaze and one scene camera to record what the user sees. We collected the following data at 50 Hz: (1) raw world camera images in the user's egocentric view,



(a) Pouring Task (b) Placement Task

Figure 3: Avg. % time spent by novice & expert users fixating on task-relevant v/s irrelevant objects.

(2) pixel location of the human's gaze in the egocentric image, (3) gaze time stamps synchronized with keyframe time stamps along the KT demonstration. Raw gaze locations are processed to extract spatio-temporal features of gaze such as fixations and saccades [31, 32]. Visual fixations maintain the focus of gaze on a single location. Fixation duration varies based on the task, but one fixation is typically 100 - 500 ms, although it can be as short as 30 ms [33]. Saccades are rapid, ballistic, voluntary eye movements (usually between 20 - 200 ms) that abruptly change the point of fixation. To filter fixations from raw gaze data, we first detect eye movements with very high speeds (a large distance traversed over a very short period of time is likely a saccade). So we compute object color histograms in a circular area (100-pixel radius) around the 2D eye gaze location obtained from the eye tracker. The object is identified as the focus of attention for that instant if the color of the object is present in a majority of the pixels around the gaze point detected by the eye tracker, since all objects in our tasks are significantly different colors. If gaze remains on one such object for more than 100 ms, we consider it a fixation.

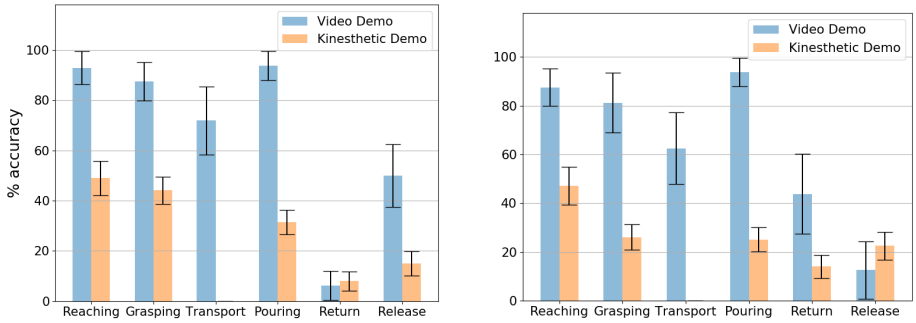
4 Experiments and Results

4.1 Statistical Analysis of Gaze Patterns for LfD

Users rarely fixate on task-irrelevant objects: Consistent with prior work [34] we find that under both tasks and both forms of demonstrations, users fixate more on objects which are relevant to the task. Specifically for the pouring task, the two-way mixed design ANOVA test produces $F(1, 46) = 100.94, p < 0.01$ (video demonstrations) and $F(1, 40) = 762.80, p < 0.01$ (KT demonstrations) showing task relevance of objects of gaze fixation. The main effect is significant for both demonstration types ($p < 0.01$; Fig. 3), i.e. gaze fixations on task-relevant and task-irrelevant objects come from different distributions. In KT demonstrations, there is a significant difference in fixation duration between user type ($F(1, 40) = 20.39, p < 0.01$) and significant interaction effect between task-relevance and user type ($F(1, 40) = 13.62, p < 0.01$). For the placement task as well, significant differences are observed between fixation duration on task relevant and task-irrelevant objects ($p < 0.01$) for each demonstration type. This provides strong evidence that using gaze during demonstrations can help to identify the relative importance of different parts of the workspace.

User fixations can predict the target object of keyframes for video demonstrations: Video demonstrations contain cleaner gaze fixation patterns than KT demonstrations, where object fixations are interspersed with glances at the gripper: the total number of consecutive object-fixation changes across all KT demonstrations for the pouring task are $12\times$ higher than that for video demonstrations. We computed fixation patterns between distinct keyframes for the first trial of video (manually coded) and KT demonstrations (user provided) for each user. In video demonstrations, fixations between keyframes (Fig. 2) signifying the semantic action of reaching, grasping, transport and pouring line up with their target reference frames at least 75% of the time for expert users and at least 70% of the time for novice users (Fig. 4).

Novice robot users attend more to the robot's gripper: For KT demonstrations of the pouring task, we find that novice users on average spend more time fixating on the gripper than expert robot users (Fig. 5(d)), likely because novice users often struggle to manipulate the robot's arm, and thus focus more on moving the gripper. Even though the results are not statistically significant across the entire pouring task ($p = 0.813$) or when observing a single action for the placement task



(a) Expert Users (b) Novice Users

Figure 4: Reference frame detection accuracy for each action of the pouring task.

($p = 0.178$), an overall average difference exists between user types. Both expert and novice users independently still spend more time overall on the task-relevant objects compared to the gripper.

Gaze can identify intent for ambiguous actions: In the placement task, the ladle is placed at roughly the same location on the table for 2 instructions (left of the red plate or right of the yellow bowl) given to the user (Fig. 1(b)). A different spatial reference frame for placing the ladle should change the user's internal objective function and we expect this to be reflected in the amount of time spent fixating on the object representing the reference frame. For the instruction relative to the red plate, we find that all users on average fixate more on the plate in comparison to the bowl in both video and KT demonstrations (Fig. 5). They similarly fixate relatively more on the yellow bowl for the instruction relative to the bowl. Our results for video demonstrations and for novice users of KT demonstrations are statistically significant ($p < 0.01$). This finding aligns with past research on understanding gaze for natural manipulation behavior, showing strong promise to be utilized as an additional signal for inferring reward functions from demonstrations.

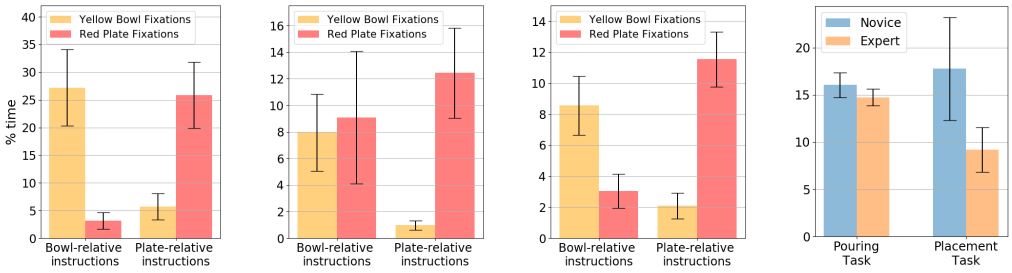
Gaze patterns differ between step and non-step keyframes: We refer to keyframes of KT demonstrations which mark the boundaries of semantically different actions (such as Fig. 2) as step keyframes. We hypothesize gaze fixations before and after such keyframes will more likely constitute different objects of attention compared to non-step keyframes. The object of attention on which a user spends the maximum time fixating 3 seconds before and 3 seconds after every keyframe is computed. For novice users, the target object of attention is different before and after 19.44% of non-step keyframes, and 24.51% of step keyframes. For expert users, the target object of attention is different before and after 15.79% of non-step keyframes, and 27.85% of step keyframes. This implies an average of 6% and 12% more of step keyframes constitute a change in the object of attention compared to non-step keyframes for novice and expert users respectively. Even though gaze alone might not be sufficient in distinguishing between step and non-step keyframes, gaze can be a useful feature in addition to other features for this classification task. We propose the use of gaze for keyframe classification as an open problem.

4.2 Utilizing Human Gaze for Learning

4.2.1 Subtask Prediction

Many LfD methods focus on the case in which the robot learns a monolithic policy from a demonstration of a simple task with a well-defined beginning and end [5]. However, this approach often fails for complex tasks that are difficult to model with a single policy. Several household tasks require multiple steps comprising of different actions involving different goals, objects and features. It is, therefore, important to segment a complex task into simpler subtasks, and then learn subsequent policies for each subtask. It has been shown that learning a separate policy for each step of a task can lead to improved generalization [35, 36].

Gaze fixation patterns accumulated and analyzed over subtasks reveal that gaze can predict their target reference frames well, especially for video demonstrations (Section 4.1). Motivated by this finding, we show in a proof-of-concept experiment that gaze can improve automatic subtask classification for multi-step demonstrations, as an intermediate step to multi-step policy learning. We use



(a) All Users (VD) (b) Expert Users (KD) (c) Novice Users (KD) (d) Gaze on Robot Gripper

Figure 5: % of time spent fixating on the red plate and yellow bowl during placement demonstrations for (a) all users during video demonstrations (VD), (b) expert users and (c) novice users during KT demonstrations (KD). (d) Proportion of time by user expertise spent fixating on the gripper relative to task objects in the pouring and placement tasks during KT demonstrations.

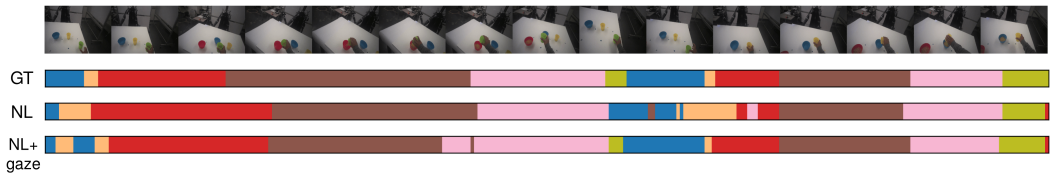


Figure 6: Visualization of sub-task prediction for a video demo of the pouring task. Each color represents a different subtask label. The rows below the demonstration images show ground truth labels (GT), labels from a non-local neural network not using gaze (NL), labels from a non-local neural network using gaze (NL + gaze).

two different model architectures for subtask classification: (i) Non-Local (NL) neural network [37] and (ii) Compact Generalized Non-Local (CGNL) neural network [38], which have been shown to work well for activity recognition. We use the ResNet-50 architecture with one NL or one CGNL block, weights initialized from a pre-trained ImageNet model. To incorporate gaze information, we use the normalized gaze coordinates as input before the last fully connected layer for both these networks. Egocentric videos from the eye tracker are sub-sampled to generate $\sim$3K images for video demonstrations and $\sim$12K images for KT demonstrations. In a 10-fold cross validation experiment, action labels are predicted per frame by each model. Incorporating gaze improves accuracy of subtask prediction with both models for both demonstration types (Table 1). An example of the action labels predicted is shown in Fig. 6. Even though action classification at the snippet-level does not necessarily yield clean, contiguous activity segment labels, Goo et al. [39] showed that even with moderate action label noise, reward inference and subsequent policy learning can improve versus policy learning on entire videos without any segmentation.

	Video Demos		KT Demos	
	NL	CGNL	NL	CGNL
Without Gaze	86.95	82.42	61.95	63.00
With Gaze	**87.63**	**88.88**	**62.71**	**64.11**

Table 1: 10-fold cross validation accuracy of subtask prediction.

4.2.2 Reward Learning

We hypothesize that differences in the amount of time spent looking at an object of interest can arise from the intent or internal reward of the demonstrator. The role of internal reward in guiding eye and body movements has been observed in neurophysiological studies [10]. Specifically, neural recordings have shown that vast areas of the brain's gaze computation system exhibit sensitivity to reward signals [10]. To investigate this hypothesis, we examine the possible role of gaze in an inverse reinforcement learning (IRL) setting. IRL offers an intuitive means to specify robot goals by providing demonstrations from which the robot can recover the reward function to optimize. One method for IRL is Bayesian inverse reinforcement learning (BIRL) [30], which models the

posterior distribution, $P(R|D) \propto P(D|R)P(R)$, over reward functions R, given demonstrations D. BIRL estimates the probability of state-action pairs of a demonstration set, given a reward, to infer this distribution. It assumes the demonstrator follows a softmax policy, resulting in the following likelihood function:

$$P(D|R) = \prod_{(s,a)\in D} \frac{e^{cQ_R^*(s,a)}}{\sum_{b\in A} e^{cQ_R^*(s,b)}} \tag{1}$$

where c is a parameter representing the degree of confidence we have in the demonstrator's ability to choose the optimal actions [30], and Q_R^* denotes the Q-function of the optimal policy under reward R. Markov Chain Monte Carlo (MCMC) sampling is used to obtain samples from the posterior, from which an estimate of the maximum a posteriori (MAP) reward function R_{MAP} or the mean reward function $\overline{R}$ can be extracted.

Additional information, such as the gaze of the demonstrator, is typically ignored in IRL algorithms. We recover the reward function for different instructions of the placement task (placement with respect to red plate or yellow bowl) using gaze with BIRL. By incorporating gaze information G as a prior into this framework, we formulate the posterior as follows:

$$P(R|D,G) \propto P(D|R)P(R|G) \tag{2}$$

where we model $P(R|G) = -\sum_{i,j} I_{ij} \frac{f_i}{f_j}$ and I_{ij} is an indicator function which is 1 when $w_i < w_j$ and $f_i > f_j$. f_i is the time spent fixating at object i and w_i is the sum of the 5 RBF kernel weights: 4 RBFs are placed around (top-right, top-left, bottom-right, bottom-left) and 1 is placed at the center of the object i. The RBFs are used to capture spatial information relative to objects [40], and the ratio of fixation times captures relative attention given to objects. The indicator function is 1 if the ranking of RBF weight magnitudes for a pair of objects does not match the ranking of the magnitude of fixation times on the respective objects. Given k items of interest on the table, we assume the reward for placement location x is given by:

$$R(x) = \sum_{i=1,j=1}^{i=k,j=5} w_{ij} \cdot rbf(x, c_{ij}, \sigma_i^2) \tag{3}$$

with

$$rbf(x, c, \sigma^2) = exp(-||x - c||^2/\sigma^2). \tag{4}$$

This formulation downweights reward functions in which the relative time spent fixating near two objects does not match the relative weights assigned to their RBF kernels (i.e. we expect features to have larger magnitude weights and influence the reward function more when they are defined relative to objects that were looked at more frequently).

We hypothesize that incorporating gaze in BIRL will help to identify the important object-relations for the task, thereby imposing preferences over reward functions that might otherwise appear equally good when looking at demonstrations without gaze information. We find an improvement in policy learning (Table 2, 3) after performing reinforcement learning on the inferred reward function. Gaze fixation times on the yellow bowl and red plate across 5 ambiguous video and KT demonstrations are used to determine how well incorporating the fixation time performs relative to ignoring the gaze information (standard BIRL algorithm). Given the instruction for placement, we formulate a ground truth reward in which the ladle should be placed as instructed (e.g.: for the instruction to place the ladle on the right of the bowl, weights of RBFs on the top-right and bottom-right of the bowl are set to 0.5 each and all remaining RBF weights are set to 0). With the demonstrated placement location and gaze fixation time, the underlying reward function is recovered. The placement policy is computed by picking the best position via gradient ascent with random restarts. Generalization is measured under 100 different configurations of the bowl and plate in simulation.

To evaluate our experiment, we use two metrics: the policy loss and placement loss. The policy loss of executing a policy π under the reward R is given by the Expected Value Difference:

$$EVD(\pi, R) = V_R^{\pi^*} - V_R^{\pi} \tag{5}$$

$\pi = \pi_{MAP}$ is used as the robot's best guess of the optimal policy under the demonstrator's reward, where π_{MAP} is the optimal policy corresponding to R_{MAP}, the maximum a posteriori reward given the demonstrations so far. π^* is the optimal policy corresponding to the ground truth reward. The placement loss is computed using the difference between the ground truth placement location and the placement location estimated by π_{MAP}. We find that both policy loss and placement loss are lower when gaze is incorporated into the learning framework. Even with a single ambiguous demonstration, gaze improves performance. Since video demonstrations contain richer gaze signals (Sec. 4.1), there is an overall greater improvement in both metrics when incorporating gaze from video demonstrations. We envision that the use of gaze information in other learning algorithms would also result in better generalization performance, which we pose as an open problem for future work.

	5 KT Demos		5 Video Demos		1 KT Demo		1 Video Demo	
Instruction relative to	Bowl	Plate	Bowl	Plate	Bowl	Plate	Bowl	Plate
Without Gaze	0.619	0.081	0.678	0.036	0.073	0.666	0.486	0.184
With Gaze	**0.329**	**0.032**	**0.046**	**0.021**	**0.043**	**0.225**	**0.098**	**0.120**
Avg improvement w/ Gaze	53.7%		**67.4%**		53.6%		**57.3%**	

Table 2: Average policy loss w/ and w/o gaze information in BIRL for the placement task.

	5 KT Demos		5 Video Demos		1 KT Demo		1 Video Demo	
Instruction relative to	Bowl	Plate	Bowl	Plate	Bowl	Plate	Bowl	Plate
Without Gaze	0.494	0.102	0.536	0.064	0.087	0.492	0.383	0.160
With Gaze	**0.291**	**0.063**	**0.068**	**0.045**	**0.066**	**0.191**	**0.102**	**0.122**
Avg improvement w/ Gaze	39.7%		**58.5%**		42.7%		**48.6%**	

Table 3: Average placement loss w/ and w/o gaze information in BIRL for the placement task.

5 Conclusion

In this work, we showed that human gaze behavior during teaching is informative in a variety of ways. We find that gaze behaviors exhibited during video demonstrations and KT demonstrations are similar in that users mostly fixate on objects being manipulated or objects with respect to which manipulation occurs. These demonstration modalities differ in terms of gaze fixations of video demonstrations lining up more accurately with target objects for a semantic action, and being more informative to improve reward inference with Bayesian IRL for a simple placement task. Consistent with previous findings is the notion that gaze of a user reflects their internal reward function. Particularly during ambiguous demonstrations, when it is unclear from a single demonstration what feature in the workspace the user is trying to optimize, gaze can reveal intentions which are not directly observable from the action alone.

We also discover eye gaze patterns specific to demonstrations for robots. Specifically, human gaze fixations during demonstrations differ for step and non-step keyframe segments; and between users with different robot-expertise. Also, gaze fixations can identify target objects of subtasks part of a multi-step video demonstration. Motivated by this finding, we show utilizing gaze leads to an improvement in subtask classification from egocentric videos of both demonstration types. Most importantly, our results show an existence proof on the informativeness of gaze data and related open problems for the research community. We envision that gaze information will increasingly be used to improve applications including automatic task segmentation, policy learning, video and kinesthetic demonstration alignment, keyframe classification, and reference frame detection.

Acknowledgments

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 the NSF (IIS-1724157, IIS-1638107, IIS-1749204). SIM research is supported in part by NSF (IIS 1724157, IIS 1638107) and ONR (N000141612835, N000141612785). We thank Dr. Garrett Warnell (Army Research Laboratory and University of Texas at Austin) for access to the Tobii Pro Glasses 2 eye tracker and Ruohan Gao (University of Texas at Austin) for advice about it's use.

References

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

[2] M. Argyle. Non-verbal communication in human social interaction. 1972.

[3] 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 *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction*, pages 201–208, 2008.

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

[5] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

[6] O. Kroemer, S. Niekum, and G. Konidaris. A Review of Robot Learning for Manipulation: Challenges, Representations, and Algorithms. *arXiv preprint arXiv:1907.03146*, 2019.

[7] B. Akgun, M. Cakmak, J. W. Yoo, and A. L. Thomaz. Trajectories and keyframes for kinesthetic teaching: A human-robot interaction perspective. In *Proceedings of the seventh annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 391–398, 2012.

[8] R. S. Johansson, G. Westling, A. Bäckström, and J. R. Flanagan. Eye–hand coordination in object manipulation. *Journal of Neuroscience*, 21(17):6917–6932, 2001.

[9] M. F. Land, N. Mennie, and J. Rusted. Eye movements and the roles of vision in activities of daily living: making a cup of tea. *Perception*, 28(4):1311–1328, 1999.

[10] M. Hayhoe and D. Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188–194, 2005.

[11] M. F. Land and M. Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision research*, 41(25-26):3559–3565, 2001.

[12] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*, pages 199–207, 2015.

[13] E. Chong, N. Ruiz, Y. Wang, Y. Zhang, A. Rozga, and J. M. Rehg. Connecting Gaze, Scene, and Attention: Generalized Attention Estimation via Joint Modeling of Gaze and Scene Saliency. *arXiv preprint arXiv:1807.10437*, 2018.

[14] A. L. Yarbus. Eye movements during fixation on stationary objects. In *Eye movements and vision*, pages 103–127. Springer, 1967.

[15] J. R. Flanagan and R. S. Johansson. Action plans used in action observation. *Nature*, 424 (6950):769, 2003.

[16] M. Hayhoe. Vision using routines: A functional account of vision. *Visual Cognition*, 7(1-3): 43–64, 2000.

[17] M. M. Hayhoe, A. Shrivastava, R. Mruczek, and J. B. Pelz. Visual memory and motor planning in a natural task. *Journal of vision*, 3(1):6–6, 2003.

[18] M. F. Land. Vision, eye movements, and natural behavior. *Visual neuroscience*, 26(1):51–62, 2009.

[19] D. H. Ballard, M. M. Hayhoe, and J. B. Pelz. Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, 7(1):66–80, 1995.

[20] H. Admoni and B. Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017.

[21] J. W. Hart, B. Gleeson, M. Pan, A. Moon, K. MacLean, and E. Croft. Gesture, gaze, touch, and hesitation: Timing cues for collaborative work. In *HRI Workshop on Timing in Human-Robot Interaction, Bielefeld, Germany*, page 21, 2014.

[22] R. M. Aronson, T. Santini, T. C. Kübler, E. Kasneci, S. Srinivasa, and H. Admoni. Eye-hand behavior in human-robot shared manipulation. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 4–13. ACM, 2018.

[23] H. R. Tavakoli, E. Rahtu, J. Kannala, and A. Borji. Digging Deeper Into Egocentric Gaze Prediction. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 273–282. IEEE, 2019.

[24] M. Lu, Z.-N. Li, Y. Wang, and G. Pan. Deep Attention Network for Egocentric Action Recognition. *IEEE Transactions on Image Processing*, 2019.

[25] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, pages 314–327. Springer, 2012.

[26] Y. Huang, M. Cai, Z. Li, and Y. Sato. Mutual Context Network for Jointly Estimating Egocentric Gaze and Actions. *arXiv preprint arXiv:1901.01874*, 2019.

[27] N. Shapovalova, M. Raptis, L. Sigal, and G. Mori. Action is in the eye of the beholder: Eye-gaze driven model for spatio-temporal action localization. In *Advances in Neural Information Processing Systems*, pages 2409–2417, 2013.

[28] 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.

[29] H. C. Ravichandar, A. Kumar, and A. Dani. Gaze and motion information fusion for human intention inference. *Int. J. of Intelligent Robotics and Applications*, 2(2):136–148, 2018.

[30] D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. *Urbana*, 51(61801): 1–4, 2007.

[31] E. Kasneci, T. Kübler, K. Broelemann, and G. Kasneci. Aggregating physiological and eye tracking signals to predict perception in the absence of ground truth. *Computers in Human Behavior*, 68:450–455, 2017.

[32] M. Nyström and K. Holmqvist. An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior research methods*, 42(1):188–204, 2010.

[33] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford, 2011.

[34] N. Mennie, M. Hayhoe, and B. Sullivan. Look-ahead fixations: anticipatory eye movements in natural tasks. *Experimental Brain Research*, 179(3):427–442, 2007.

[35] S. Niekum, S. Osentoski, G. Konidaris, and A. G. Barto. Learning and generalization of complex tasks from unstructured demonstrations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5239–5246, 2012.

[36] G. Konidaris, S. Kuindersma, R. Grupen, and A. Barto. Robot learning from demonstration by constructing skill trees. *The International Journal of Robotics Research*, 31(3):360–375, 2012.

[37] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[38] K. Yue, M. Sun, Y. Yuan, F. Zhou, E. Ding, and F. Xu. Compact Generalized Non-local Network. *Advances in Neural Information Processing Systems*, 2018.

[39] W. Goo and S. Niekum. One-Shot Learning of Multi-Step Tasks from Observation via Activity Localization in Auxiliary Video. *IEEE Int. Conf. on Robotics and Automation (ICRA), 2019*.

[40] D. S. Brown, Y. Cui, and S. Niekum. Risk-aware active inverse reinforcement learning. *Conference on Robot Learning (CoRL)*, 2018.

[41] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *The Symposium on Eye tracking research & applications*, 2000.

A Procedure for User Study

We recruited 20 participants (14 males, 6 females) within our university – 10 expert users who had operated or worked with a robot arm, and 10 novice users who had no prior experience operating a robot. The details of the two goal-directed tasks used in our study are as follows:
Pouring task: Two cups (green and yellow) filled with pasta and two empty bowls (blue and red) are placed at pre-specified locations on a table in front of the robot. Users pour pasta from the green cup to the red bowl, followed by pouring from the yellow cup to the blue bowl. We ask users to provide 3 demonstrations of each type for this task, where demonstration type is counter-balanced across users.
Placement task: Four objects are placed on the table in pre-determined locations – a purple cup, a yellow bowl filled with yellow pasta, a red plate and an orange cup. A large green ladle with a light blue foam support on the edge of its handle (to ease grasping by the robot gripper), is held in the hand of the demonstrator for a video demonstration or gripped by the robot for the kinesthetic demonstration. The user is given an instruction to place the spoon on a relative location on the table with respect to other objects. Each user is given two instructions (order of instructions is counter-balanced across users) for each demonstration type – (1) place the green ladle to the left side of the red plate, and (2) place the green ladle to the right side of the yellow bowl. The red plate and the yellow bowl are adjacent to one another on the table with a gap in between them to place the spoon. If the ladle is placed between these two objects on the table, it can be ambiguous to determine which instruction was followed by the user. There were 4 demonstrations provided by each user in this task (2 instructions x 2 demonstration types).

The eye tracker was individually calibrated once at the beginning of the study for each user, and subsequently calibrated in-between demonstrations if there was data loss over the network or the user wanted to take a break. Each participant was allowed one practice round for each demonstration type on the task they were assigned to do first. After one round of practicing, participants completed 6 demonstrations (3 KT, 3 video) for the pouring task and 4 demonstrations (2 KT, 2 video) for the placement task. All trials of each task and demonstration type pair were completed sequentially, and the order of these pairs was fully counterbalanced across the participant pool. A trial of a video demonstration lasted between 1.5 seconds to about 25 seconds, and a KT demonstration lasted between 2 minutes to 7 minutes.

For two users the tracker did not calibrate well after multiple trials, hence only a part of the study could be conducted with them. Due to noisy observations or loss of data transmission, we eliminated data from such users, which left us with 8 expert and 8 novice users for the pouring task and 9 novice and 7 expert users for the placement task. This amounted to a total of $\sim$27 minutes and $\sim$124 minutes of video and KT demonstration data, respectively.

The tracker is equipped with two cameras for each eye to track the gaze and one scene camera to record what the user sees. The tracker technology ensures automatic compensation for slipping and makes it possible to track eye gaze reliably in dynamic environments. It has a simple calibration process in which users stare for a few seconds at the center of a calibration card with concentric circles, provided by the manufacturer. The first-person view and corresponding eye tracking data from the glasses (Fig. 2) are recorded at 50 Hz. The data is saved onto a pocket-sized recording unit that allows the participant to move around unrestricted.

For KT demonstrations, keyframes are provided by the users themselves. However, they do not attach a semantic meaning to them. A single experimenter logs keyframes along with a semantic label for the action in that keyframe as per their judgment. The semantic labels in the order they are used during the pouring task are: (1) start, (2) reach, (3) grasp, (4) transport, (5) pour, (6) return, (7) release, (8) reach, (9) grasp, (10) transport, (11) pour, (12) return, (13) release, (14) end (Fig. 2). These keyframes are then synchronized with the gaze data time stamps to recover at what point in the first person video the keyframes lie. Users provide different levels of granularity in their keyframe segmentations. For example, some users break down the pouring action into multiple keyframes in which the gripper is being rotated at different angles until all the pasta falls out, and some only rotate the wrist joint once and mark the end of the pouring action as a keyframe. Video demonstrations are relatively much shorter in duration, as the user is able to complete the task within seconds. Therefore, we manually annotate videos with a fixed number of semantically meaningful keyframes for the pouring task. Examples of keyframes for the pouring task for a video demonstration type are

shown in Fig. 2. The placement task is relatively simple, as it only requires logging a single action of placing the ladle on the table.

B Gaze Fixation Filtering

Eye gaze movements can be characterized as: (a) Fixations, (b) Saccades, (c) Smooth pursuits, and (d) Vestibulo-ocular movements. Visual fixations maintain the focus of gaze on a single location. Fixation duration varies based on the task, but one fixation is typically 100 - 500 ms, although it can be as short as 30 ms [33]. Saccades are rapid, ballistic voluntary eye movements (usually between 20 - 200 ms) that abruptly change the point of fixation. Smooth pursuit movements are slower tracking movements of the eyes that keep a moving stimulus on the fovea. Such movements are voluntary in that the observer can choose to track a moving stimulus, but only highly trained people can make smooth pursuit movements without a target to follow. Smooth pursuit movements are minimally present in our trials and are preserved after filtering for saccades. Vestibulo-ocular movements stabilize the eyes relative to the external world to compensate for head movements. These reflex responses prevent visual images from slipping on the surface of the retina as head position changes. The accelerometer and gyroscope sensors of the eye tracker glasses differentiate between head and eye movements which eliminates the impact of head movements on eye tracking data.

Salvucci et al. [41] proposed a novel taxonomy of fixation identification algorithms and evaluated existing algorithms in the context of this taxonomy. They identify two characteristics—spatial and temporal— to classify different algorithms for fixation identification. For spatial characteristics, three criteria distinguish primary types of algorithms: velocity-based, dispersion-based, and area-based. For temporal characteristics, they include two criteria: whether the algorithm uses duration information, and whether the algorithm is locally adaptive. The use of duration information is guided by the fact that fixations are rarely less than 100 ms and often in the range of 100-500 ms. In our work, we use velocity-based and area-based criteria under spatial characteristics and duration based criteria under temporal characteristics to filter out fixations from saccades. We first filter out eye movements with very high speeds (a large distance traversed over a very short period of time is likely a saccade). Then we compute object color histograms in a circular area (100-pixel radius) around the 2D eye gaze location obtained from the eye tracker. The object is identified as the focus of attention for that instant if the color of the object is present in a majority of the pixels around the gaze point detected by the eye tracker. If the eye gazes at one such object for more than 100 ms, we declare it a fixation.