Inferring Goals with Gaze during Teleoperated Manipulation

Reuben M. Aronson[1], Nadia Almutlak[2], and Henny Admoni[1]

***Abstract*— Assistive robot manipulators help people with upper motor impairments perform tasks by themselves. However, teleoperating a robot to perform complex tasks is difficult. Shared control algorithms make this easier: these algorithms predict the user's goal, autonomously generate a plan to accomplish the goal, and fuse that plan with the user's input. To accurately predict the user's goal, these algorithms typically use the user's input command (e.g., joystick input) directly. We use another sensing modality: the user's natural eye gaze behavior, which is highly task-relevant and informative early in the task. We develop an algorithm using hidden Markov models to infer goals from natural eye gaze behavior that appears while users are teleoperating a robot. We show that gaze-based predictions outperform goal prediction based on the control input and that our sequence model improves the prediction quality relative to gaze-based aggregate models.**

I. Introduction

Assistive robot manipulators help people with upper motor impairments accomplish activities of daily living [1]. However, manipulator arms are hard to control: users must overcome non-intuitive kinematics and limited input interfaces to achieve the precision required for manipulation [2]. *Shared control* is one approach to overcoming this challenge [3], [4]. These systems work by first predicting what the user is trying to accomplish. Then, the robotic system autonomously constructs a plan to accomplish the goal. Finally, the user's input command is fused with the assistive command. Shared control gives users more control over the system than a fully autonomous approach does while relieving the user of the complexity needed for direct teleoperation [3].

To be effective, shared control systems must infer what the operator is trying to accomplish. Existing approaches infer this goal using the operator's joystick inputs. These systems must handle the confusing commands generated by operators unskilled at controlling the robot. In addition, user input inherently only gives information about how to make incremental progress on the task, rather than informing the system about the user's ultimate goal.

To improve this goal inference, we can use another signal: the user's eye gaze behavior. Psychology research has shown that what people look at in a scene reflects what they are trying to do, particularly when performing a manipulation action by hand [5] and by robot [6]. Eye tracking enables systems to predict people's subgoal during a by-hand manipulation task [7] and their next intentions during food serving [8] and handover [9]. Gaze is especially responsive to new information, since people move their eyes faster than they can either move their hands or control a robot.

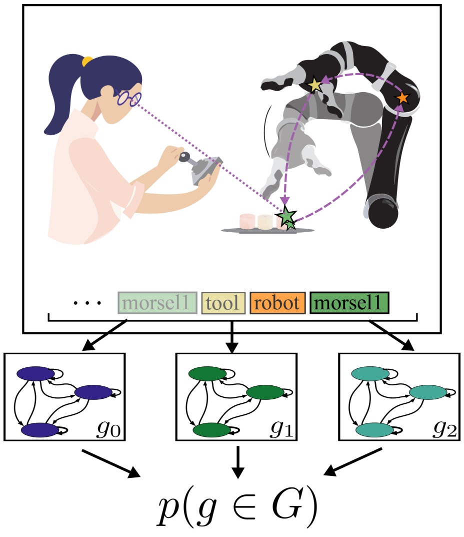

Fig. 1. Overview of the gaze analysis pipeline for goal prediction. While teleoperating a robot, the user's natural gaze towards objects in the scene is collected. This signal is then processed by hidden Markov models trained on each goal candidate g to assign a probability to each goal.

However, the gaze signal itself is complex to use. Approaches that use *intentional* eye gaze (the "gaze as mouse pointer" model) rely on users providing clear, controlled gaze towards specific objects. In *natural* gaze, however, the signal is less clear. While people often look directly at their goals, these goal-directed fixations don't appear reliably. Goal-directed fixations are interspersed among glances at the robot end effector [6], as well as blinks, distractions, and other unexpected behavior. In fact, fixations towards goal objects make up only 16% of all identified fixations in the HARMONIC data set [10], which records the gaze locations of people teleoperating a robot. To enable accurate goal prediction, we must explore more sophisticated models for using gaze. In this work, we combine *fixation labeling* [11], [12] with *scanpath analysis* [13] to derive a new approach for gaze interpretation. Our sequential method outperforms methods that use joystick input and aggregate gaze methods that discard information about the order of fixations.

To validate this model, we present a pipeline for using gaze

*This work was supported by the Paralyzed Veterans of America, the National Science Foundation (IIS-1755823 and IIS-1943072), and the Tang Family Foundation Innovation Fund. Nadia AlMutlak was supported by the King Abdullah University of Science and Technology.

[1] Reuben M. Aronson and Henny Admoni are with the Human and Robot Partners Lab, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, United States of America (email: rmaronson@cmu.edu, henny@cmu.edu).

[2] Nadia Almutlak is with Columbia University, New York, NY, 10027, United States of America (email: na2736@columbia.edu).

for goal prediction during a manipulation task (Fig. 1). We first segment the raw gaze data into individual fixations [14], then assign to each fixation a label indicating the fixated object. We then analyze this sequence using two goal prediction algorithms: an *aggregate* model, which measures how many fixations are directed at each goal, and a *sequential* model, which uses hidden Markov models trained on each goal to use temporal context for detection. We apply the gaze algorithms on the HARMONIC data set [10] and compare the goal predictions from our gaze-based system with existing models that use joystick input. We find that:

- Both natural gaze models predict the user's goals faster and more accurately than the joystick model does, and
- A model using the order in which gaze fixations occur (*scanpath* method) outperforms a model that relies only on counts of object-directed fixations.

II. Related Work

A. Shared Control

Shared control makes teleoperation easier by controlling a robot using both the user's input signal and a separate autonomous system. Typically, such systems represent the user's goal as a selection from a finite, pre-specified set of possible goals (e.g., pre-grasp positions for relevant objects). The system then calculates a probability distribution over the user's possible goals; approaches include comparing the user's action to the direction towards each goal [4] and fitting the action using a maximum entropy inverse optimal control model [3], [15]. From this goal prediction, the system develops an autonomous motion command through standard manipulation techniques. It then fuses its autonomous command with the user's direct input according to an arbitration method, such as using goal prediction confidence directly [4], [16], via a partially observable Markov decision process [3], or by allowing the user to adjust the level of control [4]. These systems improve task success rates and completion times while requiring less user input [3]. While goal inference has largely used only the user input command, some work explores other modalities such as brain-computer interfaces [17] or fusion of multiple sources [18].

B. Eye Gaze for Goal Inference

In many tasks, people's eye gaze behavior reflects their intentions [19]. We focus here on using natural eye gaze to understand the goal of a user during a manipulation task. Significant psychological evidence (reviewed in [20]) shows that when people are manipulating objects with their hands, they almost exclusively look at objects relevant to the task. This *top-down* model interprets gaze using an information-gathering framework; people direct their gaze to acquire information necessary for the task [21]–[23]. The informativeness of gaze persists during similar activities, such as teleoperating a robot [6], generating robot demonstrations [24], or requesting objects from a robot server [8].

Some systems have been presented to predict users' goals from their gaze during manipulation [25]. Typically, these approaches build from raw gaze data from an off-the-shelf eye tracker (see [26] for a review) to obtain the user's gaze location as a pixel on a scene camera. One approach is to identify the object closest to the gaze location and call that the goal [27]–[32]. This strategy works in simple situations, especially when the user is deliberately using their gaze to indicate an object (rather than natural gaze, which the user produces unconsciously while performing the task). However, it assumes that people look only at their intended goals, which is less true the more complex the task. Other approaches analyze the raw gaze data directly [33]–[38], which can work well in differentiating between different types of tasks, but is relatively poor at identifying the goal of a specific task [22], [23], [39]–[42]. These approaches often use custom aggregate features in the gaze data, but some use sequence-aware models for this identification, in an approach called *scanpath analysis* [7], [13], [43].

To combine these strategies, some approaches process the gaze data to identify individual fixations with objects in the scene, and then process those fixations with other algorithms. [44] achieves 86% object prediction accuracy (among trials with gaze data available) by training a custom support vector machine based on aggregate fixation features such as the fixation duration on each object and whether it was the first object fixated on during a window. [45] uses similar hand-crafted aggregate features to predict actions in a board game. [24] uses the relative amount of time spent fixating on each object to estimate how important each object is in a scene. We build on these results by using the *scanpath analysis* approach to consider the identified fixations in the order in which they occur rather than in aggregate.

III. Problem Statement

A. Goal Inference

Shared control systems rely on accurately predicting people's goals early in the the task. Formally, specify a finite set of possible goals $G = \{g_0, g_1, \cdots, g_n\}$. The *goal inference* problem consists of finding a probability distribution $p(G)$ representing the likelihood that the user is intending to move towards each of the goals.

B. Benchmark: Joystick-based Prediction

We compare our gaze-based approach to *shared autonomy* [15], [17], [46], which uses inverse optimal control to infer the user's goal. This method compares the user's observed command a with the optimal command to achieve each goal. Formally, the user's planning behavior is modeled by a family of Markov decision processes (MDPs), each with a user cost function $C_g(x, a)$ parameterized by the goal $g \in G$ that gives the cost of taking action a in state x Each MDP is solved in advance to obtain a goal-parameterized set of action value functions $Q_g(x, a)$.

We then assume that a user intending goal g provides control inputs a at a state x according to the distribution

$$p(a|x, g) \propto \exp(Q_g(x, a) - Q_g(x, 0)), \tag{1}$$

which computes the reward a user with goal g would gain by producing the observed action in this state compared to providing no action[1]. These probabilities are marginalized over all goal candidates to determine a goal probability $p(g|a,x)$ at each time as

$$p(g|a,x) = \frac{p(a|g,x)}{\sum_{g' \in G} p(a|g',x)},$$

and then accumulated over time using Bayesian combination.

C. Gaze for Goal Inference

In this work, we develop an algorithm for using eye gaze for goal inference. We collect eye gaze data using a head-mounted eye tracker, which records images of each eye and egocentric video. The tracker detects the user's pupils in the eye images, then uses its calibration data to report where in the egocentric video frame the user is looking.

A gaze interpretation pipeline must map this raw stream of gaze locations to a probability distribution over goals. There are several challenges in this process. First, the raw gaze locations are noisy, sensitive to calibration, and include dropouts due to events such as blinks. Second, the raw pixel stream does not include context about the task or environment. To handle these challenges, we use the *semantic gaze labeling* representation, which transforms the raw stream of pixels into labeled fixations.

Another key challenge is that people's gaze behavior does not follow exact patterns. Some gaze-based goal prediction strategies [27], [29]–[32], especially those designed for short interactions or intentional gaze, assume that people look directly at their goal at the beginning of the task. However, in our natural gaze setting, that assumption does not hold. People do tend to look at their goal objects, but they more often look at the robot end-effector and the robot body itself. In the data set used here [10], only 16% of all fixations were directed towards a goal object. We cannot even assume that users *ever* look directly at their goal during the task; in fact, 10% of trials included no identified goal-directed gaze. Users may use their peripheral vision or memory of the scene to localize their goals rather than looking directly at them. Therefore, more sophisticated techniques that can handle inconsistent information are required.

IV. Preprocessing the Gaze Data

Before applying prediction models, we preprocess the raw gaze data to incorporate task context. First, we segment the data into individual fixations. Second, we label each fixation with one of a few pre-specified objects in the scene. This sequence of labeled, timed fixations is then passed into the learning algorithms below.

A. Fixation Segmentation

While matching gaze to objects in the scene can be done frame by frame, the physiology of gaze suggests an easier approach. Gaze follows specific patterns: people look at stationary objects (during *fixations*) or follow moving objects (during *smooth pursuits*) [14]. These single-focus spans are separated by rapid motions (*saccades*). By first segmenting the gaze data into individual fixations[2], we group together consecutive gaze points and label them jointly. This step ensures that assigned labels are inherently smoothed and provides additional data for labeling.

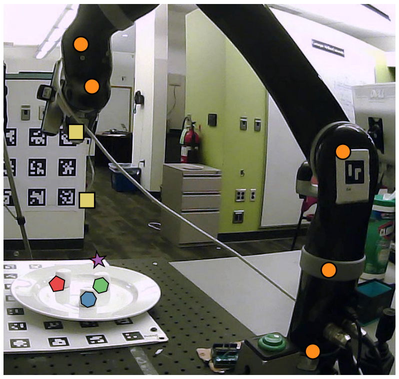

Fig. 2. Keypoint labels assigned during semantic gaze labeling. First, each fixation is mapped to one of the keypoints, which consist of each robot joint and each possible goal morsel. Next, the labels are grouped, with keypoints in the same groups indicated by matching markers. Fixations that could not be labeled accurately were given a group `none`. The purple star denotes the user's current gaze location, which here is labeled `morsel 1`.

Fixation segmentation was performed using I-BMM [47]; software is available at `https://github.com/HARPLab/ibmmpy`. I-BMM calculates the gaze velocity for each frame, then fits these velocities to a two-component Gaussian mixture model. Velocity samples that match the larger component are labeled *saccades*, and those in the smaller component, *fixations*. Consecutive *fixation* samples are then merged into fixations.

B. Semantic Gaze Labeling

After individual fixations have been identified, they are then matched with objects in the scene. This process assumes that gaze during manipulation tasks is directed at relevant objects [48]. Furthermore, it requires that scene-relevant objects are tracked, which our manipulation architecture provides. We first identify a set of *keypoints* in the scene as possible gaze targets. For this task, we chose each robot joint and each goal morsel (Fig. 2). We manually labeled each fixation with its most relevant keypoint [49]. For each trial, we then have a sequence of tuples (s_t, d_t, ℓ_t) corresponding to its start time, duration, and label among a selection of pre-selected possibilities. This sequence of labeled fixations is used as input to our learning algorithms.

V. Algorithms for Goal Prediction

In this section, we discuss algorithms used for gaze-based goal prediction. First, we present an aggregate gaze model which uses only the counts of fixations labeled as goals and discards sequence information. This aggregate model

[1] This equation is given in [15] as $p(a|x,g) \propto \exp(Q_g(x,a) - V_g(x))$, where $V_g(x)$ is the state value function. The data set [10] alters the formulation to remove any bias for user inaction: if the user does nothing, the probability does not update, wherever the robot is located.

[2] For this analysis, whether or not the object is moving when fixated is not relevant, so we use *fixation* to include both traditional fixations and smooth pursuits.

represents methods that only consider goal-directed fixations and treat other data points as noise. Second, we present a novel sequential gaze model, which learns hidden Markov models from the sequences of labeled fixations.

A. Baseline: Aggregate Gaze Method

For gaze-based goal prediction, we take as input the sequences of labeled fixations derived above. The gaze data in each trial i consists of a sequence of start times, durations, and labels (s_t^i, d_t^i, ℓ_t^i) of length n_i as input and a reported goal g_i as output. Existing work [27], [29]–[32] uses this signal by only considering goal-based glances. Typically, the entire gaze is assumed to be directed only at the goal, and other gaze information is discarded as noise. Then, they predict the goal as the one closest to the user's gaze location.

For this baseline, we use an aggregate probability function that counts the number of fixations directed towards each goal. Specifically, we set

$$p^{\text{agg}}(g_k | (s_t, d_t, \ell_t)) \propto \exp \sum_{t=0}^{T} \delta(\ell_t = \ell_{g_k}), \tag{2}$$

with normalization performed over the three possible goals g_k. Here, $\delta(a = b)$ evaluates to 1 if the arguments are equal and 0 otherwise. This method requires the specification of ℓ_{g_k}, the label corresponding to each goal g_k. However, it requires no training.

B. Sequential Method via Hidden Markov Models

We now present a sequential method which, unlike the aggregate method, takes into account both the *order* in which fixations appear as well as fixations directed towards *non-goal objects*. Thus, this algorithm is able to improve on recognition speed and confidence. Building this model consists of two steps: sequence processing and model learning.

1) Sequence processing: We use a hidden Markov model, which operates on untimed sequences of categorical observations. Therefore, the first step is to transform our timed sequence into an equivalent untimed sequence. A simple way to do so would be to just drop the timing elements entirely. However, that method removes information conveyed by the fixation durations. Instead, we repeat each label a number of times based on its duration. This untimed sequence is suitable for use in a Markov model, but its expansion retains a representation of the fixation durations using repetition counts.

Specifically, given a sequence $q = (s_t, d_t, \ell_t)$, generate the new sequence q' as

$$q' = (\underbrace{\ell_0, \cdots, \ell_0}_{N(d_0)}, \cdots, \underbrace{\ell_i, \cdots, \ell_i}_{N(d_i)}, \cdots, \ell_n),$$

where each individual label ℓ_i is repeated based on a multiplicity function $N(d_i)$. We set

$$N(d) = \text{clamp}(\left\lfloor \frac{d}{\Delta t} \right\rfloor; 1, N_{\text{max}}),$$

where Δt is a fixed time quantization parameter, N_{max} is the maximum number of repeats of a single fixation, and `clamp` forces the result within the range specified. Smaller values of Δt mean that fixation durations are more faithfully represented but that the observed sequences are longer. N_{max} enforces a cutoff value for long fixations so they do not overwhelm the data.

To handle labels with low prevalence in the data, we mapped the labels into larger categories. In particular, fixations towards to either the end-effector or the tool were relabeled as `tool` fixations, and fixations to elsewhere on the robot were relabeled as `robot` fixations (Fig. 2).

2) Goal prediction: For sequence modeling, we use a hidden Markov model (HMM), a powerful technique for representing sequence structures. We apply these HMMs to the processed sequences. Let the set of emissions be the set of possible labels K. For each goal possibility, we select all sequences corresponding to trials with that goal. We then train a hidden Markov model from this subset of the sequences. This process yields one HMM for each goal possibility g_k.

To perform goal inference on a data sequence, we compute the score s_k of the observed sequence $(\ell_0, \cdots, \ell_T)$ given by each pre-trained HMM as

$$s_k(\ell_0, \cdots, \ell_T) = \log p(\ell_0, \cdots, \ell_T; \text{HMM}_k). \tag{3}$$

Then, a goal probability is found by marginalizing over all the known goals and assuming a uniform prior,

$$p(g_k | \ell_0, \cdots, \ell_T) = \frac{\exp s_k(\ell_0, \cdots, \ell_T)}{\sum_{k'} \exp s_{k'}(\ell_0, \cdots, \ell_T)}. \tag{4}$$

All HMM operations were performed using the `hmmlearn` package[3]. We set the number of hidden states $n = 3$, quantization parameter $\Delta t = 250\,\text{ms}$, and cutoff value $N_{\text{max}} = 3$ through cross-validation.

While this method requires specifying the number of goals in advance, it can be extended to different numbers of goals with appropriate training data. It can also be expanded to identify intermediate goals for multi-staged tasks. Moreover, it does not require that the goal objects themselves be identified among the labels in advance.

VI. Results

A. Data for Evaluation

To evaluate the goal prediction algorithms, we used the HARMONIC data set [10], which contains eye gaze and joystick input from participants performing a robot teleoperation task. Participants teleoperated a Kinova Mico robot arm using a joystick to spear one of three marshmallows on a plate. Participants first reported which morsel they intended to grasp, then operated the robot using modal control: the two axes of the joystick mapped to robot x/y, z/yaw, and roll/pitch in turn, and users cycled through the modes by pressing a button on the joystick. The study consisted of 24 participants, each of whom performed five trials in four different assistance conditions. For this work, we include only trials without assistance that succeeded in the task. This filtering left 64 trials, with an average of 60 fixations per trial.

[3] https://hmmlearn.readthedocs.io/en/latest/

TABLE I

Algorithm Accuracy Metrics

	Accuracy	Mean probability	Median probability
Aggregate gaze	0.578	0.637	0.827
Sequential gaze	**0.671**	**0.643**	**0.991**
Joystick	0.531	0.486	0.478
Sequential gaze (by participant)	0.594	0.591	0.986

Predictions by the joystick method (Sec. III-B) were provided in the data set. The aggregate model was computed directly. The sequence model was evaluated using five-fold cross-validation with actual goals balanced.

B. Comparisons

We compare the algorithms on several metrics. First, we measure the overall *accuracy* of each algorithm given all of the trial data. An algorithm is marked correct on a trial if the probability assigned to the correct goal given all of the data is strictly larger than the probability assigned to each other possible goal. If the algorithm assigns the maximum probability to more than one goal (e.g. the aggregate method with no goal glances), its prediction is marked incorrect. Accuracy for each algorithm appears in Table I.

Second, we measure how *confident* each algorithm is in its correct predictions. We compute the set of final probabilities assigned to the *correct* goal at the end of each trial i, i.e., $\{\forall i : p(g^i_{\text{correct}}|\text{data}^i)\}$. We report the *mean probability*, the mean of this set. Since these probabilities are highly non-Gaussian (see Sec. VII-A), we also report the *median probability*. These results also appear in Table I.

To validate that this result extends to new participants, we compute these evaluations for the sequential gaze method using a different test/train split such that each participant's data appears in only one fold. These results, which are comparable to the results that measure across participants, appear at the end of Table I.

Finally, we determine how each algorithm's *accuracy* evolves during the trial. We consider the sequence of *partial probabilities* of the correct goal. Given a subset of the data $(0, \cdots, T \leq n_i)$, we compute the probability of the correct goal derived from that subset $(p(g^i_{\text{correct}}|\text{data}^i_{0,\cdots,T})$. This partial probability is a function of time T, which is normalized to the length of the trial. Fig. 3 shows how each algorithm's partial probability evolves during the course of the trial. For each time bin, the width of the bar represents the proportion of partial probabilities of the correct trial goal at that particular time. Mass at larger y values indicates more confident correct predictions, and lines at smaller x values represent prediction confidence earlier in the trial.

VII. Discussion

In this section, we examine the usefulness of gaze, particularly sequential gaze. First, we discuss the strong bimodality that appears in the probabilities assigned to the correct goals when using gaze. We then show that the sequential method allows for earlier goal predictions than the other methods do. We analyze the sequential method evaluated on a single trial. Finally, we discuss in what cases the gaze-based methods fail and how to use them effectively.

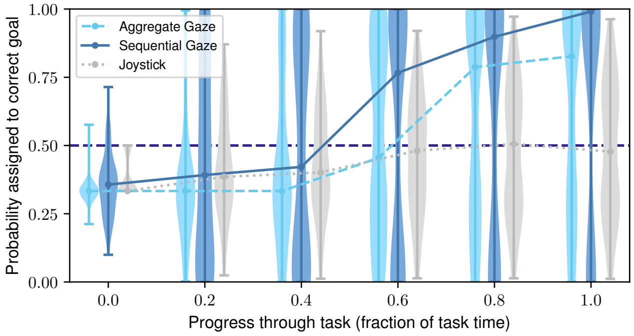

Fig. 3. Distributions of probability assigned to the correct goal by each algorithm during the evolution of each trial. The first data point (at $t = 0$) uses the first fixation, so the initial probability is not uniform. Lines connect probability medians. When the probability assigned to the correct goal is above 0.5 (denoted by the horizontal dashed line), the classifier is guaranteed to be correct at that time.

A. Sequential Gaze vs. Aggregate Gaze

First, we compare our novel sequential method with the aggregate baseline. The sequential model has slightly higher accuracy than the baseline. In addition, Fig. 3 shows that the sequential model has higher *confidence*: when it is correct, its reported correct probability is nearer to one, and when incorrect, that probability is nearer to zero. In contrast, the aggregate model is more indecisive, with more trials ending in equal probabilities assigned to all goals.

Additionally, Fig. 3 shows that both gaze models demonstrate strong bimodal behavior. This finding evokes the result found in [44], in which the gaze-based algorithm did not predict any intention in approximately 30% of cases and performed well otherwise. Examining the data suggests that the two modes may be related to the availability of goal-directed gaze data. In 10% of trials, users made no goal-directed fixations at all, making classification based on gaze difficult. In addition, this observation explains the discrepancy in worst-case performance between the aggregate and sequential algorithms. When none of the user's fixations are directed towards goals, the aggregate model makes no prediction (emits uniform probability $p = 0.33$ over all goals), while the sequential algorithm generates a prediction anyway and performs poorly ($p \approx 0$ for the correct goal). Thus, the median probability results in Table I better represent each algorithm's quality than the means do.

B. Gaze vs. Joystick: Forecasting Horizon

We next explore whether gaze provides information faster than the joystick does. Intuitively, we would expect this result: people look at their targets early in the trial to localize them [6], whereas joystick input is similar when all goals are in the same direction from the robot's current pose. Particularly for this task, the robot trajectory is similar for the first half of the trial (as participants reorient the robot

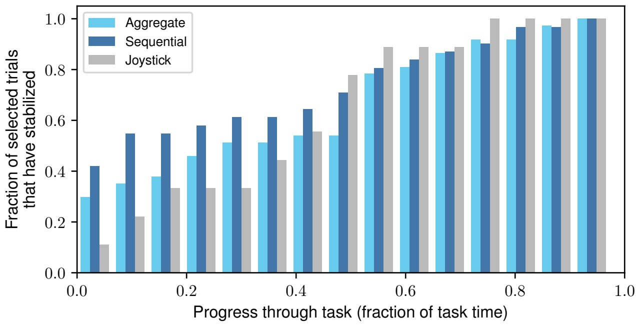

Fig. 4. Fraction of selected trials that have stabilized on the correct prediction by progress through the task. Selected trials include those where the algorithm is incorrect in its first guess without any information, but is correct at the end. The exclusions resulted in $n_{\text{aggregate}} = 37, n_{\text{sequential}} = 31, n_{\text{joystick}} = 9$. Trials have *stabilized* when their (correct) prediction stays the same from that point in the trial through the end. While the small n for the joystick method precludes strong conclusions, this plot suggests that distinguishing evidence occurs later in the joystick method than for the aggregate method, and it comes fastest for the sequential method.

into a spearing position), and the joystick information only diverges in the second half of the trial.

To measure this prediction horizon, we compute how long it takes for a correct prediction to *stabilize*: if a trial is ultimately correct, what is the earliest time such that the (correct) prediction persists through the end of the trial? This measure shows when each algorithm has obtained enough information to make its final prediction. If more of the trials have stabilized earlier, we conclude that that algorithm gets sufficient information early to make a decision. However, some of the algorithms have inherent priors from the structure of the data. For example, the sequential method, given no information, arbitrarily predicts goal 2. If the true value is equal to this prior, the stabilization time measure usually shows that the prediction is correct from before the trial starts. Therefore, we omit trials that predicted the correct result before receiving any data. *Stabilization time* measures how much trial time it takes for the algorithm to have enough confidence to switch from its initial prediction to the correct goal. Results appear in Fig. 4.

We find that the sequential method outperformed the other methods on stabilization time. Median stabilization time (as fraction of trial time; lower is better) is 32% for the aggregate method, 8.6% for the sequential method, and 45% for the joystick method. Unfortunately, our exclusion criteria left relatively few trials for the joystick case ($n = 9$), so it is difficult to draw clear conclusions. However, Fig. 4 suggests that in general, the joystick method does not begin to stabilize until about halfway through the trial. The gaze methods, and especially the sequential gaze, can get to the correct conclusion much faster. This evidence reinforces the idea that gaze can detect goals earlier in the trial, but more investigation is required.

C. Understanding the Sequential Model

To understand the benefits of the sequential model, we examine a single trial in detail. Fig. 5 shows the trained HMM for recognizing goal 2, represented by the gaze label `morsel 2`. States `S0` and `S1` activate on fixations directed towards labels `tool` and `morsel 2`. In addition, these states activate slightly when they see `morsel 0` or `morsel 1` respectively. This HMM has some possibility of producing fixations towards the other goal options, so it can incorporate them in its prediction.

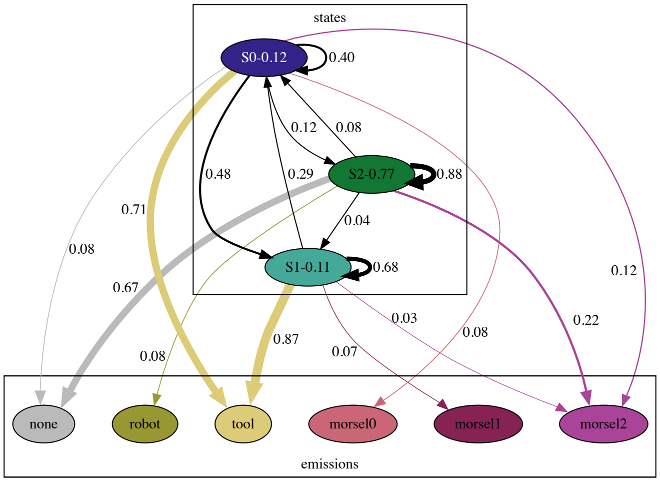

Fig. 5. Graph representation of a learned hidden Markov model for morsel 2. State labels include their prior probabilities, edges between states represent transition probabilities, and edges to emissions represent emission probabilities. Edges with $p < 0.02$ were omitted for clarity.

Fig. 6 shows how this model evaluates a single trial. Eye gaze is mostly directed toward the tool, particularly at the start of the trial (Fig. 6, middle), and the HMM largely stays in `S1` in response (Fig. 6, bottom). When it encounters fixations labeled with non-goal `morsel 0` at about 50% of the way through the trial, the model transitions to `S0` and incorporates those fixations smoothly. With additional fixations labeled `morsel 1` at 65% through the trial, the model re-enters `S1` and correctly predicts the goal for the remainder of the trial (Fig. 6, top). In contrast, the aggregate method is unable to handle these glances towards incorrect goal candidates, so it fails to recognize morsel 2 as the goal.

D. Limitations of Gaze-based Prediction

While gaze is a powerful signal for goal recognition, there are two key complications for using it in practice. First, clear gaze information is not always available. In our data set, six trials (10% of the data) included no goal-directed glances at all. People may use other strategies for identifying their goals, such as their peripheral vision or their memory of the object location from a previous task. Therefore, gaze may be better used as a *signal of opportunity*. In its absence, we must fall back to an alternate method, such as the joystick-based model. To explore this possibility, we measure whether the gaze and joystick algorithms are correlated in their accuracy. Trials where the gaze gives a correct prediction and those where the joystick gives the correct prediction show no strong correlation ($\chi^2(1) = 0.0336, p = 0.854$). The results comparing the joystick and aggregate methods are similar ($\chi^2(1) = 0.00628, p = 0.937$). Thus, the complementarity of these methods make this combination especially appealing.

The signals can combine their predictions together using Bayesian combination [18], or we can use alternative methods that are more sensitive to the data.

Second, gaze information depends on task context. While the semantic labeling procedure encodes scene information, it does not clarify *why* people are looking at a particular object. People look both at objects they intend to interact with and objects they are trying to avoid [50]. Therefore, gaze *alone* cannot distinguish the role of an object in a task. Gaze shows that an object is relevant, but other systems are needed to interpret that relevance. While our approach avoids this issue, more complex tasks will require handling it.

E. Extensions to More Complex Tasks

While the gaze analysis method presented here succeeds on this task, we observe that the gaze signal itself may require more sophisticated analysis to extend to more complex tasks. If the possible goals are known in advance, the method here can be extended to an arbitrary number of goals with only a single HMM by using label remapping in a one-vs-rest framework. Specifically, train a single HMM on all labeled runs with the label corresponding to the ground-truth goal mapped to g_{target} and all other goal labels mapped to g_{other}. Given a new sequence, inference is performed by performing the transformation once for each possible goal and evaluating the HMM's score on each new sequence. While this method loses some accuracy by aggregating all goals in a single model and requires advance specification of the goal candidates, it is adaptable to any number of goals without retraining. If the goals are not known in advance, however, an entirely new model must be trained.

For more complex tasks, such as those involving scene objects that play different roles like obstacles or partial tasks, the method here will struggle to provide additional information. This method identifies a particular object that the user considers important in the task. However, it is unable to determine *why* that object is important. For these tasks, this method can be part of a broader algorithm that also incorporates information about task context.

VIII. Conclusions and Future Work

In this paper, we present a pipeline for using people's eye gaze data during teleoperated manipulation to predict their goals. We show that gaze data benefits from analysis using sequential methods, which use the sequential nature of gaze and the non-goal fixations as additional context. We also find that gaze-based prediction outperforms joystick-based prediction on average. While the overall gaze accuracy is high, the signal quality is bimodal, with excellent predictive quality when goal-directed fixations are available and poor prediction quality otherwise. Fortunately, the trials where the gaze performs poorly are not correlated with those where the joystick performs poorly, so the two prediction methods complement each other.

Future work from this project can focus on combining the gaze and joystick signals presented here, along with other signals. Building a full system that can anticipate when each signal will perform well may lead to even better goal prediction. In addition, this sequential approach can be applied to other human signals, like body pose, when using them for intention prediction.

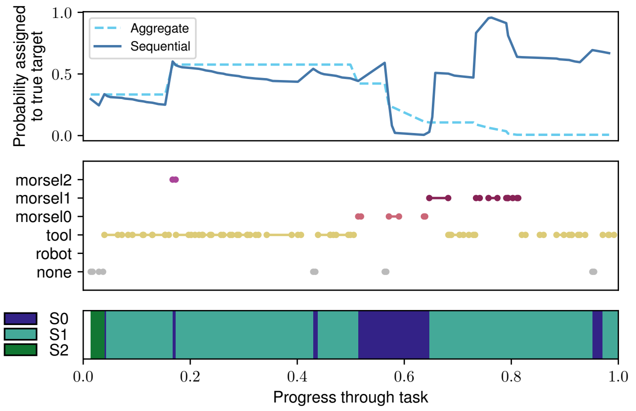

Fig. 6. Sample trial comparing aggregate and sequential prediction performance. The top plot shows the output probability assigned to morsel 2 (the true target) during the trial. The second plot shows the gaze labels supplied to each algorithm. The bottom bar shows the hidden state predicted by the sequential model as calculated from the entire data run. While the HMM mostly maintains a single state, the presence of fixations towards morsel 1 (near $0.5 < t < 0.65$) triggers a different hidden state. This flexibility enables the HMM to incorporate the misleading gaze information.

Finally, the full gaze-based shared control system should be validated in a user study. A full user study requires an online version of the proposed algorithm and bringing in study participants to perform the task. While this evaluation is out of scope of this paper, it is essential for understanding the effects of gaze-based prediction on robot assistance.

References

[1] C. S. Chung, H. Wang, and R. A. Cooper, "Functional assessment and performance evaluation for assistive robotic manipulators: Literature review," *Journal of Spinal Cord Medicine*, vol. 36, no. 4, pp. 273–289, 7 2013.

[2] L. V. Herlant, R. M. Holladay, and S. S. Srinivasa, "Assistive teleoperation of robot arms via automatic time-optimal mode switching," in *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2016-April. IEEE, 3 2016, pp. 35–42.

[3] S. Javdani, H. Admoni, S. Pellegrinelli, S. S. Srinivasa, and J. A. Bagnell, "Shared autonomy via hindsight optimization for teleoperation and teaming," *The International Journal of Robotics Research*, vol. 37, no. 7, pp. 717–742, 6 2018.

[4] D. Gopinath, S. Jain, and B. D. Argall, "Human-in-the-loop optimization of shared autonomy in assistive robotics," *IEEE Robotics and Automation Letters*, vol. 2, no. 1, pp. 247–254, 1 2017.

[5] M. Land, N. Mennie, and J. Rusted, "The Roles of Vision and Eye Movements in the Control of Activities of Daily Living," *Perception*, vol. 28, no. 11, pp. 1311–1328, 11 1999.

[6] R. Aronson, T. Santini, T. Kübler, E. Kasneci, S. Srinivasa, and H. Admoni, "Eye-Hand Behavior in Human-Robot Shared Manipulation," in *ACM/IEEE International Conference on Human-Robot Interaction*, vol. Part F1350, 2018.

[7] Y. Chen and D. H. Ballard, "Learning to recognize human action sequences," in *Proceedings - 2nd International Conference on Development and Learning, ICDL 2002*, 2002.

[8] 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)*. IEEE, 3 2016, pp. 83–90.

[9] E. C. Grigore, K. Eder, A. G. Pipe, C. Melhuish, and U. Leonards, "Joint action understanding improves robot-to-human object handover," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 11 2013, pp. 4622–4629.

[10] B. A. Newman, R. M. Aronson, S. S. S. Srinivasa, K. Kitani, and H. Admoni, "HARMONIC: A Multimodal Dataset of Assistive Human-Robot Collaboration," *ArXiv e-prints*, 7 2018.

[11] R. Mantiuk, B. Bazyluk, and R. K. Mantiuk, "Gaze-driven Object Tracking for Real Time Rendering," vol. 32, no. 2, 2013.

[12] M. Bernhard, E. Stavrakis, M. Hecher, and M. Wimmer, "Gaze-to-Object Mapping during Visual Search in 3D Virtual Environments," *ACM Transactions on Applied Perception*, vol. 11, no. 3, pp. 1–17, 8 2014.

[13] D. Noton and L. Stark, "Eye Movements and Visual Perception," *Scientific American*, vol. 224, no. 6, pp. 34–43, 1971.

[14] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the symposium on Eye tracking research & applications - ETRA '00*, 2000, pp. 71–78.

[15] S. Javdani, S. S. Srinivasa, and J. A. Bagnell, "Shared autonomy via hindsight optimization," in *Robotics: Science and Systems*, vol. 11. MIT Press Journals, 2015.

[16] A. D. Dragan and S. S. Srinivasa, "A policy-blending formalism for shared control," in *International Journal of Robotics Research*, vol. 32, no. 7. SAGE PublicationsSage UK: London, England, 6 2013, pp. 790–805.

[17] K. Muelling, A. Venkatraman, J. S. Valois, J. E. Downey, J. Weiss, S. Javdani, M. Hebert, A. B. Schwartz, J. L. Collinger, and J. A. Bagnell, "Autonomy infused teleoperation with application to brain computer interface controlled manipulation," *Autonomous Robots*, vol. 41, no. 6, pp. 1401–1422, 2017.

[18] S. Jain and B. Argall, "Probabilistic Human Intent Recognition for Shared Autonomy in Assistive Robotics," *ACM Transactions on Human-Robot Interaction*, vol. 9, no. 1, pp. 1–23, 12 2019.

[19] J. F. Boisvert and N. D. Bruce, "Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features," *Neurocomputing*, vol. 207, pp. 653–668, 9 2016.

[20] J. R. Flanagan, M. C. Bowman, and R. S. Johansson, "Control strategies in object manipulation tasks," *Current Opinion in Neurobiology*, vol. 16, no. 6, pp. 650–659, 12 2006.

[21] N. Sprague and D. Ballard, "Eye movements for reward maximization," in *Advances in Neural Information Processing Systems*, 2004.

[22] C. A. Rothkopf, D. H. Ballard, and M. M. Hayhoe, "Task and context determine where you look," *Journal of Vision*, vol. 7, no. 14, 12 2007.

[23] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard, "Eye guidance in natural vision: reinterpreting salience," *Journal of Vision*, vol. 11, no. 5, p. 5, 5 2011.

[24] A. Saran, E. S. Short, A. Thomaz, and S. Niekum, "Understanding Teacher Gaze Patterns for Robot Learning," in *3rd Conference on Robot Learning (CoRL 2019)*, Osaka, Japan, 2019.

[25] T. Pfeiffer, P. Renner, and N. Pfeiffer-Leßmann, "EyeSee3D 2.0," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications - ETRA '16*. New York, New York, USA: ACM Press, 2016, pp. 189–196.

[26] M. Cognolato, M. Atzori, and H. Müller, "Head-mounted eye gaze tracking devices: An overview of modern devices and recent advances," *Journal of Rehabilitation and Assistive Technologies Engineering*, vol. 5, p. 205566831877399, 1 2018.

[27] L. Shi, C. Copot, and S. Vanlanduit, "Application of Visual Servoing and Eye Tracking Glass in Human Robot Interaction: A case study," in *2019 23rd International Conference on System Theory, Control and Computing (ICSTCC)*. IEEE, 10 2019, pp. 515–520.

[28] R. Siegfried, B. Aminian, and J. M. Odobez, "ManiGaze: A dataset for evaluating remote gaze estimator in object manipulation situations," in *Eye Tracking Research and Applications Symposium (ETRA)*, 2020.

[29] R. Atienza and A. Zelinsky, "Intuitive human-robot interaction through active 3D gaze tracking," *Springer Tracts in Advanced Robotics*, vol. 15, pp. 172–181, 2005.

[30] K. M. Tsui, A. Behal, D. Kontak, and H. A. Yanco, "I want that: Human-in-the-loop control of a wheelchair-mounted robotic arm," *Applied Bionics and Biomechanics*, vol. 8, no. 1, pp. 127–147, 2011.

[31] S. Li, X. Zhang, and J. D. Webb, "3-D-Gaze-Based Robotic Grasping Through Mimicking Human Visuomotor Function for People with Motion Impairments," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 12, pp. 2824–2835, 12 2017.

[32] M. Y. Wang, A. A. Kogkas, A. Darzi, and G. P. Mylonas, "Free-View, 3D Gaze-Guided, Assistive Robotic System for Activities of Daily Living," in *IEEE International Conference on Intelligent Robots and Systems*. Institute of Electrical and Electronics Engineers Inc., 12 2018, pp. 2355–2361.

[33] D. Novak, X. Omlin, R. Leins-Hess, and R. Riener, "Predicting targets of human reaching motions using different sensing technologies," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 9, pp. 2645–2654, 2013.

[34] A. Borji and L. Itti, "Defending yarbus: Eye movements reveal observers' task," *Journal of Vision*, vol. 14, no. 3, 2014.

[35] P. Hevesi, J. A. Ward, O. Amiraslanov, G. Pirkl, and P. Lukowicz, "Analysis of the Usefulness of Mobile Eyetracker for the Recognition of Physical Activities," in *UBICOMM 2017: The Eleventh International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, 2017.

[36] W. Yi and D. Ballard, "Recognizing behavior in hand-eye coordination patterns," *International Journal of Humanoid Robotics*, vol. 6, no. 3, pp. 337–359, 2009.

[37] A. Fathi, Y. Li, and J. M. Rehg, "Learning to Recognize Daily Actions Using Gaze," in *Proceedings of the 12th European conference on Computer Vision - Volume Part I*. Springer-Verlag, 2012, pp. 314–327.

[38] A. Bulling, J. A. Wardz, H. Gellersenz, and G. Tröstery, "Eye movement analysis for activity recognition," in *UbiComp*, 2009, pp. 41–50.

[39] D. H. Ballard and M. M. Hayhoe, "Modelling the role of task in the control of gaze," *Visual Cognition*, vol. 17, no. 6-7, pp. 1185–1204, 8 2009.

[40] K. A. Turano, D. R. Geruschat, and F. H. Baker, "Oculomotor strategies for the direction of gaze tested with a real-world activity," *Vision Research*, vol. 43, no. 3, pp. 333–346, 2 2003.

[41] B. T. Sullivan, L. Johnson, C. A. Rothkopf, D. Ballard, and M. Hayhoe, "The role of uncertainty and reward on eye movements in a virtual driving task," *Journal of Vision*, vol. 12, no. 13, 2012.

[42] L. Johnson, B. Sullivan, M. Hayhoe, and D. Ballard, "Predicting human visuomotor behaviour in a driving task," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1636, 2 2014.

[43] T. C. Kübler, E. Kasneci, and W. Rosenstiel, "SubsMatch: Scanpath similarity in dynamic scenes based on subsequence frequencies," in *Eye Tracking Research and Applications Symposium (ETRA)*. Association for Computing Machinery, 2014, pp. 319–322.

[44] 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, 7 2015.

[45] R. Singh, T. Miller, J. Newn, E. Velloso, F. Vetere, and L. Sonenberg, "Combining gaze and AI planning for online human intention recognition," *Artificial Intelligence*, vol. 284, p. 103275, 7 2020.

[46] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey, "Maximum Entropy Inverse Reinforcement Learning," in *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, ser. AAAI'08. AAAI Press, 2008, p. 1433–1438.

[47] E. Tafaj, G. Kasneci, W. Rosenstiel, and M. Bogdan, "Bayesian online clustering of eye movement data," *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*, p. 285, 2012.

[48] M. M. Hayhoe, A. Shrivastava, R. Mruczek, and J. B. Pelz, "Visual memory and motor planning in a natural task," *Journal of Vision*, vol. 3, no. 1, p. 6, 2 2003.

[49] R. M. Aronson and H. Admoni, "Semantic gaze labeling for human-robot shared manipulation," in *Eye Tracking Research and Applications Symposium (ETRA)*. Association for Computing Machinery, 2019.

[50] R. S. Johansson, G. Westling, A. Bäckström, and J. R. Flanagan, "Eye–Hand Coordination in Object Manipulation," *Journal of Neuroscience*, vol. 21, no. 17, pp. 6917–6932, 9 2001.