Predicting User Intent Through Eye Gaze for Shared Autonomy

Henny Admoni, Siddhartha Srinivasa
The Robotics Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15217
{hadmoni, ss5} @ andrew.cmu.edu

Abstract

Shared autonomy combines user control of a robot with intelligent autonomous robot behavior to help people perform tasks more quickly and with less effort. Current shared autonomy frameworks primarily take direct user input, for example through a joystick, that directly controls the robot’s actions. However, indirect input, such as eye gaze, can be a useful source of information for revealing user intentions and future actions. For example, when people perform manipulation tasks, their gaze centers on the objects of interest before the corresponding movements even begin. This implicit information contained in eye gaze can be used to improve the goal prediction of a shared autonomy system, improving its overall assistive capability. In this paper, we describe how eye gaze behavior can be incorporated into shared autonomy. Building on previous work that represents user goals as latent states in a POMDP, we describe how gaze behavior can be used as observations to update the POMDP’s probability distributions over goal states, solving for the optimal action using hindsight optimization. We detail a pilot implementation that uses a head-mounted eye tracker to collect eye gaze data.

Introduction

Robot teleoperation allows people to operate in inaccessible environments or to extend their physical capabilities beyond their own limitations. However, as robots become more capable of performing real-world tasks, they also become more complex, increasing the degrees of freedom (DOFs) that must be controlled for successful teleoperation. Fully teleoperating a robot can be undesirable, for example, in assistive care, where a user’s physical ability to manipulate an input device is limited (Maheu et al. 2011; Tsui et al. 2008), or in bomb disposal scenarios, where time and user fatigue can lead to catastrophe (Kron et al. 2004). Shared autonomy is one solution to this challenge of teleoperation: by combining human control of a robot with autonomous robot behavior, shared autonomy can help people accomplish tasks more quickly and with less effort.

For example, consider a user with spinal cord injury who has a wheelchair-mounted robotic arm like the Ki-nova JACO (Figure 1). Typically, such a robot is controlled through direct teleoperation using an interface such as a 3 degree of freedom (DOF) joystick (Maheu et al. 2011) or a brain-computer interface (Muelling et al. 2015). However, the JACO has seven DOFs, and the translation from low-dimension input device to high-dimension robot control requires frequent control mode switches and careful physical manipulation. The cognitive and physical loads can be very high.

Shared autonomy addresses this challenge of robot teleoperation by augmenting human control with intelligent robot behavior that seamlessly assists a user in completing their desired action with less control input. However, in shared autonomy systems, robots may not know a priori which goal their human users want to pursue. Therefore, shared autonomy requires a prediction component that continually monitors the user’s control input and outputs a prediction about which goal (or goals) the robot should assist toward.

Previous work in shared autonomy has implemented this as a two step process, first predicting the user’s goal and then assisting when that goal prediction rises above a pre-specified confidence threshold (Fagg et al. 2004; Kofman...
et al. 2005). However, this type of predict-then-act process does not respond well to changing user goals, and the all-or-nothing assistance misses opportunities for assisting toward multiple possible goals simultaneously. Recent work improves on predict-then-act methods by modeling the user’s goal as the latent state of a POMDP, and using hindsight optimization to solve the POMDP in real time (Javdani, Srinivasa, and Bagnell 2015). This system can continually monitor user goals, dynamically adjust its assistance to new predictions, and assist toward multiple goals simultaneously if possible.

Existing shared autonomy systems still rely primarily on user teleoperation input, through joystick or other direct interfaces that send control signals to the robot. Even when the signals are indirect, for example a point-and-click interface that provides high-level action targets (Ciocarlie et al. 2012; Tsui et al. 2008), the user still explicitly communicates their intent through the input device.

In contrast, we investigate how shared autonomy systems can take advantage of the indirect signals people implicitly provide when performing their tasks. In particular, shared autonomy systems could use people’s eye gaze as an additional, implicit signal to predict their goal (Figure 1). Using this indirect eye gaze signal is different than using a direct input signal controlled through an eye gaze interface. Our approach takes advantage of natural gaze behavior rather than repurposing the eyes as an input device.

Natural eye gaze is highly predictive of object goals in manipulation tasks, because gaze during task completion is rarely directed outside of the objects required for the task (Hayhoe and Ballard 2005), and it is closely linked to hand motion (Pelz, Hayhoe, and Loeber 2001). Previous work has identified intentions and actions based on gaze behavior (Huang et al. 2015; Yi and Ballard 2009) and used these predictions to provide robotic assistance in a human-robot collaboration (Huang and Mutlu 2016).

In this paper, we extend a previous shared autonomy framework (Javdani, Srinivasa, and Bagnell 2015) with indirect inputs in the form of eye gaze. Our main contribution is this formalization of gaze-based goal prediction which can be integrated into the shared autonomy model. We describe this formalization and outline an initial data collection study to learn how eye gaze predicts goals in a shared autonomy interaction.

**Related Work**

This paper builds on existing shared autonomy models by describing how eye gaze input can help augment user goal predictions. Therefore, this work is relevant to both the shared autonomy domain, which tends to includes machine learning and robotics, as well as the eye gaze domain, which includes cognitive psychology, computer vision, and human-robot interaction.

**Shared Autonomy**

Shared autonomy (alternatively: shared control, assistive teleoperation) combines human teleoperation of a robot with intelligent robotic assistance. In assistive care, shared autonomy systems have been developed for wheelchair-mounted robot arms that use a variety of interfaces, such as an “eye-mouse” on a computer screen (Bien et al. 2004), natural language commands (Volosyak, Ivlev, and Graser 2005), and laser pointers (Lim, Lee, and Kwon 2003). These interfaces usually provide explicit action commands directly to the robot (such as opening the gripper). In contrast, the current system uses the implicit signal from natural eye gaze behavior to predict what a user wants the robot to do.

The first step of shared autonomy is to predict the human’s goal, often through machine learning methods that are trained from human demonstrations. Goal prediction has been achieved using object affordances with conditional random fields to model the spatio-temporal structure of tasks (Koppula and Saxena 2016); through a probabilistic model of observed human movements that represents intention as a latent variable (Wang et al. 2013); and using Gaussian Mixture Autoregression, which uses statistical features of 2D control inputs (mouse movements) to estimate task types and parameters (Hauser 2013).

In prior work that we build from in this paper (Javdani, Srinivasa, and Bagnell 2015), goal prediction is achieved via maximum entropy inverse optimal control (MaxEnt IOC), which derives a user’s utility function from demonstrations using the principle of maximum entropy (Ziebart et al. 2008). MaxEnt IOC has been used to infer the trajectories of pedestrians (Ziebart et al. 2009) and to predict pointing trajectories in user interfaces (Ziebart, Dey, and Bagnell 2012), as well as for shared autonomy in a manipulation task (Javdani, Srinivasa, and Bagnell 2015).

The second step of shared autonomy is to provide assistance. This often involves blending the user’s input with the robot assistance calculated to achieve the predicted goal. Blending methods range from allowing the robot to take full control once a threshold prediction confidence is reached (Fagg et al. 2004; Kofman et al. 2005) to combining human and robot trajectories using a motion planner (Dragan and Srinivasa 2013). Assistance can also provide task-dependent guidance (Aarno, Ekvall, and Kragic 2005) or mode switches (Herlant, Holladay, and Srinivasa 2016) that adjust based on the detected goal.

In the prior work that forms the basis for this paper (Javdani, Srinivasa, and Bagnell 2015), assistance is provided as a blending between user input and robot control, based on the confidence of the robot’s goal prediction. The notable advancement of this prior work is that assistance is automatically provided to all goals simultaneously if possible, removing the need to separate goal prediction and assistance phases. In this paper, we build off this contribution by including eye gaze as an input to the goal prediction.

**Eye Gaze**

People’s eye gaze reveals their intentions and future actions by indicating their direction of attention, particularly in manipulation tasks. In a natural pick-and-place task, hand and eye movements are strongly correlated (Pelz, Hayhoe, and Loeber 2001). When completing such tasks, gaze is focused on objects related to the task at hand, and gaze typically shifts to an object about 600ms before that object is grasped (Hayhoe and Ballard 2005; Land and Hayhoe 2001).
Prior work has looked at predicting where people’s attention will be directed. Some work takes a neurobiological approach, modeling the underlying neuronal processes to identify salient regions of a visual scene (Borji and Itti 2013; Itti and Koch 2001; Vig, Dorr, and Cox 2014). Other work uses features of body motion—for example that there is a center bias in egocentric gaze (when the camera is mounted on a viewer’s head), or that gaze is predictably distributed around the hands during manipulation—to predict gaze patterns in video (Li, Fauth, and Rehg 2013).

Eye gaze has also been used to automatically recognize user intent and actions. For example, an HMM that was trained on eye-head-hand coordination, which included gaze tracking, can recognize actions on objects, like “unscrewing a jar” (Yu and Ballard 2002). A dynamic Bayes network trained with head, hand, and eye gaze has been used to predict which steps a person is engaged in during a complex (sandwich-making) task (Yi and Ballard 2009). In another sandwich-making task, an SVM-based classifier predicted the intended target of a speaker’s request for ingredients using just eye gaze data with 76% accuracy (Huang et al. 2015). This classifier was then used in a system that provided anticipatory robotic assistance in real time from gaze-based intent predictions (Huang and Mutlu 2016).

In contrast with this previous work in human-robot collaboration, the system presented in this paper uses the goal prediction to provide assistance via shared control rather than using the goal prediction for autonomous robot behavior. Additionally, the current framework is robust to scenario changes as it does not need to be trained on object appearances or locations. Instead, it uses features of the scene (namely object distances from gaze fixations) to predict the user’s goal.

**Approach**

In this section, we summarize the formalization for goal prediction and assistance from previous work (Javdani, Srinivasa, and Bagnell 2015) and define the addition of eye gaze as an input. In summary, there exist a set of discrete goals (e.g., grasp poses for a set of objects on a table). A human user supplies both direct and indirect inputs to the robot (e.g., joystick control and eye gaze, respectively) to achieve one of these goals (e.g., grasping the target object).

The shared autonomy system begins with no knowledge of the user’s particular goal, but predicts that goal through the user’s inputs. Simultaneously, it selects robot actions that, given the goal prediction, minimize the expected cost of completing that goal.

**Predicting Human Goals**

The robot has continuous states \( x \in X \) (e.g., position, velocity) and continuous actions \( a \in A \) (e.g., velocity, torque). The robot is modeled as a deterministic dynamical system with transition function \( T : X \times A \rightarrow X \).

A user can supply continuous direct inputs \( j \in J \) through a control interface (e.g., joystick commands), which map to robot actions through a known deterministic function \( D : J \rightarrow A \). The function \( D \) corresponds to a user fully teleoperating the robot.

In this paper, we define a second set of (possibly discrete) indirect user inputs \( e \in E \) produced through behavior like eye gaze. This indirect input is not used to drive the robot, so there is no analogous transition function to \( D \). To simplify the goal prediction, we assume \( j \) and \( e \) are independent. The joint pair of user actions is denoted as \((j, e) = u \in U\).

There exist a known, discrete set of goals \( g \in G \), one of which is the user’s target. To predict the user’s next action given this target, the system has a stochastic user policy for each goal \( \pi_u^g(x) = p(u|x, g) \), which produces a distribution over actions for each robot state \( x \) and goal \( g \). This user policy is typically learned from demonstration. Following previous work (Javdani, Srinivasa, and Bagnell 2015), we model this user policy using the MaxEnt IOC framework (Ziebart et al. 2008), which assumes the user is optimizing a cost function for their target goal \( g \), \( C_g : X \times U \rightarrow \mathcal{R} \).

Because the user’s goals are known to themselves, the user model corresponds to a Markov Decision Process (MDP) for the specified goal, defined by tuple \((X, U, T, C_g^u)\).

Though the user knows their own goal, the system does not know the user’s goal. Therefore, we model the system’s knowledge as a Partially Observable Markov Decision Process (POMDP), where the user’s goal is the unobservable state. A POMDP is similar to an MDP in that actions taken in one state lead to a new state as defined by the transition function, but it differs from an MDP in that we cannot fully observe the current state. Instead, the POMDP maintains a distribution over states, known as the belief \( b \), and it maps beliefs (rather than states) to actions.

The system state \( s \in S \) is now defined by both a robot state and a user goal, \( s = (x, g) \), and \( S = X \times G \). We assume the robot state is known, and uncertainty is entirely over the user’s goal. The POMDP transition function \( T^s : S \times A \rightarrow S \) corresponds to transitioning between robot states but maintaining the same user goal.

Observations in our POMDP correspond to user inputs \( u \). Our POMDP has an observation model \( \Omega \) that infers a probability distribution over states \( s \) given a sequence of observations, which are time-linked state-input pairs \((s_t, u_t)\). Since uncertainty is only over user goals, inferring the probability distribution over states is equivalent to inferring the probability over user goals. Let \( \xi^{0\rightarrow t} = \{s_0, u_0, \ldots , s_t, u_t\} \) be a sequence of observations from time 0 to time \( t \). To compute the probability of a goal given the observed sequence up to time \( t \), we apply Bayes’ rule:

\[
p(g|\xi^{0\rightarrow t}) = \frac{p(\xi^{0\rightarrow t}|g)p(g)}{\sum_g p(\xi^{0\rightarrow t}|g)p(g)}
\]

To select an action, the system uses a cost function \( C^g : S \times A \times U \rightarrow \mathcal{R} \), which defines the cost of taking robot action \( a \) from state \( s \) when the observed user input is \( u \). The POMDP is thus defined by tuple \((S, A, T, C^g, U, \Omega)\).

**Modeling Eye Gaze**

To integrate gaze into this shared autonomy model, we extend the observation model from our POMDP to include gaze input. Our challenge is to formalize how eye gaze predicts goals. Given some sequence of gaze inputs over time,
behavior. Formally, the likelihood of an object being focused on an object related to the task, we devise a simple model for goal prediction from gaze is non-trivial. Unlike direct joystick input, which (in an ideal case) continually moves the robot end effector closer to the target, eye gaze behaviors more stochastically. People look at their goals, but they also use gaze to monitor obstacles, confirm their body position, and engage socially.

Despite these complexities, eye gaze does reveal goals during task-based action. People’s gaze is temporally linked to the task at hand, and task-related objects are fixated with the eyes even before the hand starts moving toward them (Land and Hayhoe 2001). When operating on a sequence of objects, people shift their gaze toward the next object in the sequence before they finish using the previous one, revealing future action. For example, gaze will shift away from an object that is about to be grasped just as the hand closes around it (Johansson et al. 2001). Objects that are not involved in the task are rarely gazed at (Hayhoe and Ballard 2005). People are skilled at using gaze to predict future action: when a person looks at an object as they begin a verbal reference, viewers respond quickly to identify that object; when gaze is occluded, that response time increases significantly (Boucher et al. 2012).

Building on this idea that task-based gaze is primarily focused on an object related to the task, we devise a simple model for how eye gaze predicts goals. More complex models could be learned, which is a planned extension of this work. In our simple model, the likelihood of an object being the goal is proportional to its distance from the center of a gaze fixation. Formally,

$p(g_t | I_t) \propto \exp(-\text{dist}(o_t, z))$

where $o_t$ is a particular object, $p(g_t)$ is the probability of that object being the goal, and $\text{dist}(o_t, z)$ is the Euclidean distance in image coordinates from the center of the object $o_t$ in the image to the gaze point $z$. In other words, the likelihood that an object is the goal falls off exponentially with its distance from the gaze fixation. Figure 2 visualizes this probability distribution as a heat map averaged over several nearby fixations over several seconds.

Because eye gaze does not solely fixate on the target object, each individual image frame $I_t$ may have widely varying probability distributions. We can use a smoothing function like a Kalman filter to add the probability distribution from each new image to the overall probability distribution $p(g | I)$ to account for noise in the sensors and in the gaze behavior.

Once we have established the probability distribution over goals based on eye gaze, we combine this with the probability distribution over goals based on joystick input to attain the joint probability distribution based on inputs, $p(g | \xi_0^{\rightarrow t}, \eta_0^{\rightarrow t})$. We currently use a simple product,

$p(g | \xi_0^{\rightarrow t}, \eta_0^{\rightarrow t}) = p(g | \xi_0^{\rightarrow t}) p(g | \eta_0^{\rightarrow t})$

though more sophisticated methods with weights could be learned from demonstration. This new joint probability is then used to solve for the optimal action of the POMDP using hindsight optimization.

Selecting Assistive Actions via Hindsight Optimization

Once we’ve established the distribution of goals, solving for the optimal action is performed as in previous work (Javadi, Srinivasa, and Bagnell 2015), but using the joint probability distribution. The optimal solution for the POMDP is a robot action that minimizes the expected accumulated cost $\Sigma \{ s_t, C^t(s_t, a_t, u_t) \}$. Solving this POMDP is computationally intractable, however, so we use the QMDP approximation (Littman, Cassandra, and Kaelbling 1995), also known as hindsight optimization (Chong, Givan, and Chang 2000; Yoon et al. 2008), to select actions.

Intuitively, the process assumes that full observability will be attained at the next timestep, and estimates the cost-to-go of the belief based on this assumption. This allows for a computationally efficient system even in continuous state and action spaces.

The cost-to-go is defined by the action-value function $Q(b, a, u)$, which returns the cost of taking action $a$ when in belief state $b$ with user input $u$, and acting optimally from that point forward. The QMDP approximation (Littman, Cassandra, and Kaelbling 1995) is

$Q(b, a, u) = \sum_{g} b(g) Q_g(x, a, u)$

where $Q_g(x, a, u)$ represents the estimated cost-to-go if the robot in state $x$ with user input $u$ took action $a$. Because uncertainty is only over the user goals $g \in G$,

$b(s) = b(g) = p(g | \xi_0^{\rightarrow t}, \eta_0^{\rightarrow t})$
We approximate $Q_g$ by assuming that user input will cease in the next timestep, which allows us to use the cost function $C^*(s, a, u) = C^*(s, a, 0)$. We can analytically compute the value of this cost function.

**Implementation and Future Work**

Egocentric gaze data is collected via a head-mounted eye tracker (Figure 1). We use the Pupil Labs Pupil headset (Kassner, Patera, and Bulling 2014), an open-source eye tracker that is worn like eyeglasses. The outward-facing world camera provides a “user’s eye view” of the scene, while the inward-facing eye camera records the pupil position and uses it to identify the gaze location on the world camera image (Figure 3).

To identify object and eye gaze locations, objects in the scene and the table are tagged with visual fiducial markers. Object positions and orientations can thus be calculated in 3D relative to the table. Additionally, the world camera extrinsics can be identified automatically, and correspondingly, the user’s head is localized in the same 3D frame as the objects. Therefore, though our goal detection gaze algorithm functions on 2D images, robot assistance can be provided in 3D space once the goal object is identified.

This work is currently in the data collection phase. We plan to collect examples of users completing everyday manipulation tasks with their own hands (such as unscrewing a jar, dialing a phone, or pouring a glass of water from a pitcher), which will provide gaze trajectories that can inform a more sophisticated model for goal prediction via gaze. Since the gaze-based goal prediction is implemented, we can evaluate the system by comparing user performance on robotic manipulation tasks in three conditions: full teleoperation, which acts as a baseline; shared autonomy without gaze prediction, as in prior work; and shared autonomy that integrates gaze into its goal prediction.

**Acknowledgments**

Thank you to Laura Herlant and EJ Eppinger for their help collecting and visualizing gaze data. This work has been partially funded by DARPA SIMPLEX (ARO contract #67904LSDRP), NSF CPS (#1544797), NSF NRI (#1227495), and ONR (#N00014-16-1-2084).

**References**


