Control Input and Natural Gaze for Goal Prediction in Shared Control

Reuben M. Aronson

CMU-RI-TR-22-30

June 9, 2022

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Henny Admoni, *chair*
Artur Dubrawski
Nancy Pollard
Brenna Argall, *Northwestern University*

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Robotics.

Abstract

Teleoperated systems are used widely in deployed robots today, for such tasks as space exploration, disaster recovery, or assisted manipulation. However, teleoperated systems are difficult to control, especially when performing high-dimensional, contact-rich tasks like manipulation. One approach to ease teleoperated manipulation is shared control; this strategy combines the user's direct control input with an autonomous plan to achieve the user's goal, thereby speeding up tasks and reducing user effort. To do so, the system needs a prediction of the user's goal.

One common approach derives this goal prediction from the user's control input itself, as it is already available to the system. Prior work using this prediction source in baseline tasks validates the usefulness of shared control. In this thesis, we prove that the effectiveness of control input for goal prediction is a consequence of how optimal users provide control input. When the user's control input is restricted, however, the assistance may be suboptimal.

To improve on this performance, we turn to another source for goal information: natural gaze. People's natural, unconstrained eye gaze behavior reveals information about their immediate goals and their future tasks. The accuracy and timing of these predictions are different than those provided by the control input pipeline, making it a promising additional source. To effectively use natural gaze for goal predictions and to combine it with control input, we analyze the behavior of each signal and evaluate them in the context of the full assistive system.

In this thesis, we show that control input and eye gaze complement each other for goal prediction during shared control. Control input gives local information about the user's goal, making it particularly effective in simple tasks when people can act optimally but limiting its performance in more complex tasks. On the other hand, eye gaze provides global information about task intentions early, but it does not do so as reliably.

We first formalize evaluation criteria for goal prediction sources and examine how goal prediction using control input, the current state of the art, affects the assistance. We show that the autonomous system does not always need to know the user's specific goal to make progress in the task. One key advantage of control input as a prediction source is that when the user's control is noisily optimal, the parts of the task where the autonomous system requires goal information coincide with those where the user's control input is likely to provide that information. However, when user input is restricted so people cannot act optimally, the user's control input is no longer as informative about the goal; this restriction occurs, for example, when using low-degree of freedom input devices. While the goal information from control input is still reliable when it is available, it may not come early enough in the task, so alternative goal prediction sources may help.

Next, we analyze natural eye gaze as a source of global information to supplement the goal prediction given by control input. We collect a data set of natural gaze during a teleoperated manipulation task and show that while people do look at their goals, more often they look at the robot end-effector, and sometimes they complete tasks without ever looking at the goals. From this analysis, we develop a contextual representation for gaze behavior and use it to predict the user's goal; this signal can give predictions earlier than are available from the control input, but the variability of people's gaze behavior limits the reliability of the signal when used on its own.

Finally, we integrate both signals into a system for online assisted manipulation, and we evaluate the model for the usefulness of each signal in a task that restricts the user's input and requires multidimensional assistance. When using control input for goal prediction, the system reliably provides some assistance, but cannot do so in all dimensions. When we incorporate gaze-based goal prediction, an earlier goal prediction from gaze enables the assistance to act in all dimensions and increases user task performance. However, the assistance using only gaze performs worse than either other condition, so it benefits from the reliability of another goal prediction source like control input.

Developing a model for how different goal prediction sources contribute to assistance quality during shared control enables this assistance strategy to work in more complex situations, such as ones with restricted user input or multipart assistance. The work in this thesis can help ground future explorations of input modalities for goal prediction. With this greater understanding of shared control for effective assistance, this work helps to bring it closer to real-world applications.

Acknowledgments

None of this would have been possible without Prof. Henny Admoni, my thesis advisor, who took in a lost grad student and, through struggles and successes, helped bring all of this to be. Thank you for helping me grow. I hope to do you, and HARP, proud.

I am deeply grateful for the support of my thesis committee, Profs. Brenna Argall, Artur Dubrawski, and Nancy Pollard. Your feedback and conversations from proposal to defense guided me to doing better research and to become a better researcher.

The ideas developed in my and Benjamin Newman's conversations are interwoven throughout this work. Thank you for all of our discussions, and I look forward to continuing them.

The HARP lab has gone from an acronym on a white board to a vibrant, brilliant, and caring research community. I am deeply grateful for the opportunity to be present as we build it, and to everyone who has been a part of it through the years.

I am also indebted to the entire CMU HRI community. Profs. Aaron Steinfeld and Reid Simmons have shaped my thinking through formal and informal conversations. I am especially grateful to Xiang Zhi Tan, who talked me into this HRI thing in the first place, and who founded and ran the HRI Reading Group with me. My thanks also go out to Michal Luria and Samantha Reig, who have continuously challenged me to think beyond my comfortable, technical lens.

I've worked with numerous undergraduates over the years, and I'm grateful for the opportunity to mentor Maggie Collier, Krish Vaswani, Nadia AlMutlak, and Karen Zhang. It has been an honor to participate in your journeys.

Prof. Matt Mason, Robbie Paolini, Ankit Bhatia, Eric Huang, and everyone else in MLab gave me my start in robotics, and while I've changed directions, I'm grateful for the support you gave. I am grateful as well to Rachel Burcin, who gave me supported I didn't know I needed.

I am deeply appreciative of my parents and siblings as I've taking this long journey. You've helped me far more than I've told you.

Research in this dissertation was funded by the Paralyzed Veterans of America, NSF grants IIS-1755823 and IIS-1943072, the Tang Family Foundation Innovation Fund, and an Okawa Foundation Research Grant.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Teleoperated systems are used widely in deployed robots today, for such tasks as space exploration, disaster recovery, or assisted manipulation for users with disabilities [47]. Moreover, even with fully capable autonomous robots, teleoperation remains a valuable control strategy applicable to many scenarios. For example, high-risk activities (such as surgery or space exploration) may be too dangerous to leave to fully autonomous systems, or people may prefer the sense of agency gained by having more control [61].

However, teleoperated systems are difficult to control [99]. Accomplishing manipulation tasks through teleoperation typically requires simultaneously controlling several degrees of freedom while adapting to (possibly dynamic) contact with the environment. Depending on the domain, this problem is further complicated by issues such as controller latency, nonintuitive controller-to-end-effector mappings, and the limitations of visual feedback due to occlusion and the inability to perceive forces. The range of proposed solutions is similarly varied, including novel interface design (such as whole-arm monitoring [95] or haptic joysticks [88]) and controller design (to ensure stability under latency) [47].

One approach that has developed recently to ease teleoperation is *shared control* [28] (Fig. 1.1). Rather than requiring additional equipment, this approach uses planning techniques to ease the complexity of the task and enable users to perform more sophisticated actions with less control input and training. For teleoperated manipulation, shared control combines the user's direct control input with an autonomous command; this process receives a prediction of the user's goal, plans a

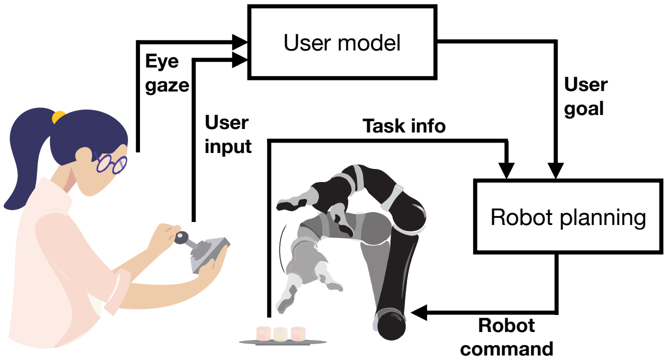


Figure 1.1: Shared control systems use sensing modalities such as the user's control input and natural gaze to infer their goal during a manipulation task. Then, the system can autonomously generate a plan for assisting with achieving the goal.

trajectory for the robot to perform the task, and combines it with the original user command to accomplish the task [28, 72]. This approach is especially promising for assistive applications, since complex controllers are often a barrier for users with disabilities [38, 55].

To predict the user's goal, shared control systems can use the user's control input itself, which is already available to the system [18, 38, 66, 110, 129, 131]. This signal performs well on simple tasks. However, it encounter problems in certain situations, such as when the robot and two goal candidates are collinear [28, 36]. We show that in the context of the full assistive system, using only control input for goal prediction can lead to suboptimal assistance when the user cannot affect the whole system state in a single action.

To improve assistance in these cases, we turn to another signal: natural gaze. When people look around during the task, their behavior can convey their goals or even more sophisticated intentions like their future tasks [6, 37, 98]. Natural gaze is a separate sensing modality, making it a promising candidate to complement the information given by control input. While gaze shows promise, most existing research focuses on by-hand manipulation, leaving gaze behavior during teleoperated robotic manipulation relatively unexplored. Furthermore, it has not been used in shared control specifically, so its suitability and role is unknown. To enable better assistance, we must model the effectiveness of both control input and gaze for goal prediction within the context of the full assistive system.

1.1 Control Input

First, we focus on understanding control input as a source for goal information, as shared control systems regularly use it for that purpose [18, 66, 110, 129, 131]. When the assistance system observes that the user is working towards a particular goal, the system can then autonomously plan an action to achieve that same goal. Control input provides information about the current action that the user wishes to perform. This information is inherently local to the current state of the robot, as the user's control input is interpreted by the robot as an action to apply at that moment.

While control input does not necessarily enable the earliest predictions [11], it works well for assistance, since accurate predictions often arrive exactly when they are needed. To explore this reasoning, consider a task in which a user operates a robot to pick up one of two goal objects. We can divide the state space of the robot based on both the user's most likely action and the optimal robot action at that state. When the user's control input generally differs depending on which object is their goal, the system can infer their goal from this input and can give goal-specific assistance. At states where all goals require the same input, the user's control input likely does not distinguish the goal; however, since the optimal action is the same for all goals, no goal prediction is actually needed. We categorize states by two different features to further explore this logic. Two goals require *different* motion at a state if their optimal *robot* motion depends on the goal; the motion is *identical* otherwise. Similarly, two goals are *distinguishable* at a state if the observed *user* input generally differs based on the goal; they are *indistinguishable* otherwise. Control input provides sufficient information for full assistance as long as indistinguishable states only require identical motion.

Considering the usual assumptions of goal prediction for shared control, we can make a more formal claim: when a user controlling a shared autonomy system [55] provides control input given by $p(u|g) \propto \exp(Q_g(u))$, the expected regret over user actions stays bounded as the cost of taking a suboptimal action increases. In other words, the more important it is for the assistance to take a specific action based on the goal, the more likely the user is to provide the distinguishing input needed to select that action. Therefore, control input is particularly efficient, and sufficient, for simple tasks during which the user can act optimally.

However, users often do not follow this optimal behavior, even accounting for noise. Rather, the scenario itself can prevent the user from acting optimally. Consider a goal that can be split into multiple tasks that the robot could perform in parallel, e.g., moving its end-effector to a desired pose in six independent axes of motion. An optimal user who can control all axes simultaneously will follow the behavior described above and give sufficient information to receive assistance. However, if the user is restricted by the control interface directly, the user cannot necessarily give optimal commands. One common circumstance generating this condition is in joystick-based teleoperation with modal control, in which a low-dimensional joystick can affect only a subset of the directions of end-effector motion at the same time, and the user cycles through modes. The different stages of the task that the user works on in sequence do not all give the same goal information to the robot [39, 40]. A user acting optimally might move first in one mode where two goals share the same optimal action; then, the assistance cannot move the robot in modes where the optimal motion is different, since the user's control input has not yet revealed their goal.

This limitation extends to goal prediction based on immediate user behavior: it cannot work independently from the user. People perform stages of a task in sequence even if a robotic system could perform them in parallel. And when performing each of these stages, the user only reveals information immediately relevant to the current stage through their actions. Even if the system could assist in other tasks at the same time, user actions may not be sufficient to give the information required. For some tasks, this is not a problem. If action on a single task fully determines the user's future goals, optimal assistance can still be provided. However, people will often sequence tasks because of their own preferences, the cost of multitasking, or because the task itself limits the ability of the user to work in parallel. In these cases, their actions may be optimal, but they may not reveal enough information about their subsequent goals to enable full assistance. These situations violate the assumptions that lead to efficient assistance using only control input and benefit from alternate methods for goal prediction.

1.2 Natural Eye Gaze

To enable assistance in these more complex cases, we turn to another sensing paradigm: natural eye gaze. Psychological research in eye gaze behavior indicates that eye gaze is strongly connected to task progression during manipulation [35, 43, 57, 67]. People look at the objects they are reaching for and at the obstacles they are avoiding. People also look at tasks differently depending on their expertise in accomplishing the task [71], their facility with the signal input system [102], their cognitive load [16], etc. These patterns also appear during teleoperated manipulation in specific tasks [6, 12, 37, 98]. By detecting people's eye gaze behavior while they are teleoperating a robotic manipulator, we can build a more complex model of their approach to the task and provide correspondingly complex assistance.

To understand how gaze fits into the overall assistance problem, we first investigate how people's eye gaze behavior during teleoperated manipulation relates to their goals. In two separate studies [12, 84], we collect gaze behavior of participants who teleoperated a robot arm in a food spearing task. From these studies, we find that people reveal their goals by looking directly at them during the task. These goal-directed glances appear at the beginning of the task and more often during translation than rotation. However, the primary use of gaze is to look directly at the robot, so these goal-directed glances are less common. In addition, we find that people look at the locations of failures [8], suggesting that gaze can depend on people's understanding of the broader system state, beyond just their goals.

However, gaze has its own limitations. Unlike in *intentional* gaze applications, in which gaze is an explicit interface, signals during *natural* gaze are much less reliable. While people often look at their goals, they are not guaranteed to do so at any particular time. People can use their peripheral vision or memory to identify the location of their goal without having to look directly at it. People spend most of the time looking at the robot itself. In addition, gaze towards a particular object indicates only that it is relevant but does not identify its role: it could be a goal, an obstacle, or a distraction. While gaze's flexibility and early information makes it a powerful signal, its unreliability makes it behave poorly in isolation.

1.3 Gaze Processing

Next, we use these insights about gaze behavior to develop a pipeline for inferring people's goals online, which we can then use for assistance. Typically, an eye gaze sensor reports the user's gaze as a pixel location on a world camera. We first identify individual fixations within the gaze signal; these periods of looking at the same object reflect the physiological processes of human gaze and allow us to consider multiple samples together. To apply context to the signal, we label each fixation with the object in the scene that is closest to it and explore more sophisticated methods for this semantic gaze labeling process. Finally, we train a hidden Markov model on this timed, labeled sequence and evaluate its goal prediction ability on the HARMONIC data set we collected earlier [84].

With a method for obtaining goal probabilities from the raw eye gaze data, we can perform a quantitative comparison of its predictive ability to that of the control input. We find that the two prediction sources have comparable accuracy but different behavior. While the control input gives a steady increase in accuracy over the task, gaze is bimodal: it cannot be used for goal prediction until the user looks at their goal, at which point it jumps to an accurate, confident prediction. The increase in average accuracy over time occurs as more trials obtain correct predictions. In many trials, though, people do not ever look at their goals [11], leading to no gaze-based prediction at all. We find that gaze often enables earlier goal predictions than the control input does, but it is unreliable.

1.4 Control Input and Gaze Are Complementary

Next, we consider how to use the signals together. Control input provides reliable goal information but is inefficient for more complex tasks. Gaze provides global information that helps in more complex tasks but is unreliable. To explore these contrasts, we develop a task that showcases the relative strength of each signal for goal prediction.

Consider a robot operated through modal control. Aligning the robot to the goal within each mode is an independent task for the user to complete. While the robot can act in all modes simultaneously, the structure of the task restricts the user to only work on one mode at a time. We develop a task in which one mode contains most of the required motion but is independent of the goal, and another requires less motion but does depend on the goal. Users typically start in the first mode; our analysis of control input suggests that this motion will not give enough information for the system to provide assistance in the second. Using gaze for prediction often provides that information, but gaze on its own leads to inconsistent behavior and potentially catastrophic failures.

To evaluate this model, we conduct a user study in which participants teleoperate the robot with assistance that used their gaze, their control input, or both for goal prediction. To account for the COVID pandemic, the study is performed in a hybrid manner: participants receive an experiment kit at home and use it to control the robot in the lab and stream back video of its motion. In the study, we find that while gaze does not reliably give an early goal prediction, when it does so, the overall system performance increases. Specifically, the system can provide assistance to the robot in axes whose motion rely on knowledge of the user's goal earlier than when using control input alone. Using gaze on its own, however, performs worse than either condition, and it even exhibits pathological behavior in response to error. This example showcases how to effectively combine these sources for goal prediction and proposes a framework that extends to more complex tasks.

1.5 Contributions

In this thesis, we show that control input and eye gaze complement each other for goal prediction during shared control. Control input gives *local* information about the user's goal, making it particularly effective in simple tasks when people can act optimally but limiting its performance in more complex tasks. On the other hand, eye gaze provides *global* information about task intentions early, but it does not do so as reliably. To demonstrate this complementarity, we first formalize criteria for goal prediction sources to enable effective assistance, and we prove that control input gives sufficient information in simple tasks, but not for more complex tasks. Next, we analyze natural eye gaze as a source of global information by collecting gaze behavior during a teleoperated manipulation task and showing how it can be used for early goal prediction. Finally, we integrate both signals into a system for online assisted manipulation, and we conduct a user study with a custom-designed task to demonstrate that each signal improves its performance when combined with effective predictions from the other.

We present the following contributions:

- A proof that when using a state-of-the-art shared control assistance algorithm, deriving goal predictions from the control input of noisily rational users results in an overall policy that maintains bounded regret as the cost of taking arbitrary actions increases (Chap. 3; Aronson and Admoni [10])
- Foundational analysis of an existing corpus of natural eye gaze behavior of participants teleoperating a robot arm in a food acquisition task (Chap. 4; Aronson and Admoni [8], Aronson et al. [12])
- Collection of an improved data set of natural eye gaze behavior during a teleoperated manipulation task to obtain higher quality sensing and enable algorithmic processing (Chap. 4; Newman et al. [84])
- Design and evaluation of an algorithm for processing raw gaze data into labeled fixation sequences to incorporate signal context and scene motion (Chap. 5; Aronson and Admoni [9])
- Design and evaluation of a sequence-aware algorithm for predicting user's goals during a manipulation task from the labeled fixation sequences produced by the gaze data (Chap. 6; Aronson et al. [11])
- Validation of the complementarity between control input and gaze behavior for goal inference through a COVID-safe user study in which participants performed a more complex robotic manipulation task with assistance (Chap. 7; Aronson and Admoni [10])

Developing a model for how different goal prediction sources contribute to the overall assistance quality during shared control enables us to extend this assistance strategy to more complex situations, including ones with restricted user input or multipart assistance. This discussion can help ground future questions of whether or not a particular signal modality is worth adding to a system. Now that shared control has shown some promise for effective assistance, this work will help bring it closer to real-world applications.

Chapter 2

Related Work

2.1 Shared Control Paradigms

To ease the problem of robot control, many approaches have been presented to fuse the user's input with an autonomously generated signal. One category of assistance consists of stateless assistance: a robot behavior that can be determined directly from the robot position, environment, and parameters of the task. In this type of assistance, no updating human model is used. Perhaps the most straightforward example of this approach is given by Ramacciotti et al. [97], which proposes an assistive welding system in which the tool frame motion is divided between the user and the robot control. For example, the robot maintains forward motion in x while the user controls the other translational axes y and z. The motion provided by the assistance is determined beforehand by the task parameters and does not change as the task progresses. Similarly, Vu et al. [123] reorients how the user controls map into the rotation axes of the tool frame. This reorientation is a static alteration of the control scheme that does not vary by task circumstance. A more complex type of motion assistance is virtual fixtures [76], which modify the compliance of the robot controller based on its position and the intended direction of motion. For example, motion along the $\pm x$ axis proceeds easily, whereas motion in other directions either moves more slowly (in the open-loop case, where the control gain is reduced) or results in a restoring force back to the desired motion surface (in the closed-loop case).

Another set of stateless control schemes aid the operator in avoiding obstacles.

Crandall and Goodrich [26] presents a system that uses a variant of potential fields to automatically maintain distance from known obstacles, and it shows that adding this automatic obstacle avoidance behavior enables users to focus on another task while controlling a robot. You and Hauser [127] compares several strategies for fusing obstacle avoidance with user input commands. The paper tests three categories of obstacle avoidance: end-effector control with collision rejection, in which commands that lead to detected collision are ignored; potential field control, in which the user command is modeled as an attractive force and environment obstacles as repulsive forces to maintain distance from obstacles; and full motion planning control, in which after receiving a motion command, the robot autonomously plans a collision-free path to the goal position using an RRT and executes that motion. These avoidance techniques all lead to faster task completion with fewer collisions. While these strategies all vary substantially in purpose and complexity, they can all be implemented without using any variable models of human state.

More sophisticated assistance behaviors can be achieved by explicitly modeling otherwise-invisible parts of the operator's internal state. In stateful assistance, systems maintain models of aspects of the user's intentions and update them over time. Aarno and Kragic [1] presents a system for recognizing low-level motions ("gestemes") using layered hidden Markov models (HMMs). If an operator is trying to move the device in a circle, for example, the HMM can recognize this gesture automatically and provide assistance to maintain it. In a more extensible example, Hauser [42] uses a dynamic Bayes net to infer a user's task over a wide variety of task definitions. The assistance system maintains a distribution over likely tasks and enacts different assistance categories based on that task recognition. To assist with bimanual manipulation, Rakita et al. [96] learns a bimanual action vocabulary from motion capture data, designs custom assistance objectives for each action type, and uses an LSTM to recognize the intended action online and activate the appropriate assistance objective.

Often, the internal state required by such assistance is the user's goal, represented as a final robot position. Systems predict the user's own goal among a finite set of choices by autonomously generating plans to achieve each goal and then comparing the user's actual control input with the generated plans. This approach works when the observations used to infer the goal are exactly the user input commands, and it has the advantage that the computed task solution can often be reused in generating an assistive command. This approach has been used to control wheelchairs based on planner results [27] or based on modelling the deviations of the user's actual provided command from a nominal user model [22, 51, 52].

While these models provide motion based on the best-matched user goal according to the user model, more complex uses of the user goal matching are possible. Dragan and Srinivasa [28] proposes thresholding the user goal matching probability and only providing assistance when the system confidence in its model of the user goal exceeds a given value. Trautman [117] broadens the arbitration technique from linear blending into a full probabilistic framework. Building on this approach, shared autonomy [55, 81] plans assistive actions over the user model uncertainty, which allows the algorithm to provide useful assistance even when exact goals are not known. Herlant et al. [45] uses a goal prediction and task information to automatically switch the user's control mode to the optimal one for task progress.

Once an overall shared control system has been selected, the details are still important to refine. Gopinath et al. [38] conducted a study in which users manually controlled the amount of assistance provided by a shared control system. Users generally preferred manual control and slower task completion to more assistance, indicating that users balance their desire for control with using assistance to reduce the task time. This result is echoed in Kim et al. [61], which finds that task metrics do not necessarily correspond to user preferences. Dragan and Srinivasa [28] has success with modulating the amount of assistance based on the algorithm's confidence in its model of the user's goal. Gopinath and Argall [39] changes the starting joystick mode so that the user's initial input is maximally informative about the user's goal, and Gopinath and Argall [40] enables the user to trigger an automatic switch to the most informative mode. Work has also been done to modify the assistance to better match the parts of the task that the user finds more difficult [128].

2.2 Eye Gaze

2.2.1 Eye Gaze During Manipulation

Eye gaze behavior during by-hand manipulation has been a subject of study in psychology for decades. Johansson et al. [57] describes a study in which users were instructed to grasp an object, manipulate it around an obstacle, and place it down elsewhere. While performing that task, people followed consistent eye gaze patterns. They looked at relevant locations before interacting with them: people look at the object until just before they grasp it, look at the obstacle until just before navigating around it, and look at the placement location just before placing the object. In addition, people rarely look at their own hands and their gaze was almost entirely directed towards task-relevant locations. Similar results were found with people performing tasks like making tea [67] and making a sandwich [43]. Perhaps most similar to a teleoperation task is the experiment reported in Sailer et al. [102], in which users are given a non-intuitive mouse controller and instructed to move a cursor around a screen. While learning the controls, users watch the cursor motion on the screen and briefly glance at the goal positions; once they are comfortable, people's eye gaze behavior looks more similar to that in by-hand manipulation.

2.2.2 Eye Gaze for Intent Recognition

Since eye gaze behavior is so closely tied to people's goals, it has been widely studied as a modality for understanding people's mental state in a variety of applications. We restrict this review to analysis techniques more closely related to our manipulation task; see Lukander et al. [74] for a full review. Bader et al. [14] has people perform manipulation actions on a screen simulating a table and uses gaze patterns to predict people's next action (reaching, moving, or releasing) and intended object. Matsuzaka et al. [77] shows that people's gaze predicts their intended grasp object and strategy (one- or two-handed) in a VR manipulation task. In a human-robot interaction study, Huang and Mutlu [48] uses hand-crafted features to predict a person's food order and proactively move a robot manipulator towards their intended food. Duarte et al. [29] shows that people can follow gaze cues when seeing other people perform an object manipulation task, and their understanding persists even when a robot is giving gaze cues.

There are several eye gaze analysis strategies available for performing different kinds of intention recognition. Analysis of the eye gaze dynamics without scene context has been successful at tasks such as identifying whether people are performing free viewing or visual search [20] or which of several different tasks a person is performing [21, 34, 46, 126]. However, this type of context-free analysis performs poorly when trying to recover specific information about *how* a task is performed [15, 114]. For tasks such as predicting gaze behavior during walking [119], driving [58, 112], or combined walking and object manipulation in VR [101], general gaze-based saliency features about the scene were less effective than modeling the dynamics of the actual task.

For predicting specific information about people's behavior during a task, one approach often used is *scanpath analysis* [87]. In this method, the eye gaze signal is treated as timeseries data rather than being reduced to frequency data. Kübler et al. [64] quantizes scanpaths into a small set of regions and used lexical analysis to predict if people will pass a driving test. Kubler et al. [65] goes beyond this approach by dynamically clustering fixations based on SIFT features around the point-of-regard and uses these sequences to determine if people are performing a tea-making task for the first or second time. Chen and Ballard [25] uses a hidden Markov model trained on timeseries gaze and hand position signals to predict which stage of a letter stapling task the participant is executing. [125] uses an LSTM with custom-designed gaze features to detect which task a user is performing, and Fuchs and Belardinelli [37] uses HMMs to identify both the task (pick or place) and the goal (among three objects) from gaze data of users controlling a robot with motion tracking.

2.3 Teleoperated Robots With Eye Gaze

Now that we have described how the eye gaze signal reveals mental state in general, we describe how it has been used for some robotic systems.

2.3.1 Eye Gaze as Direct Input

There has been some research in using eye gaze for direct robot control. The usual strategy presented is to consider the user's eye gaze as a primary input device for an autonomous manipulator system [13, 24, 70, 89, 107, 118, 124]. In these systems, people look at an object they wish to grasp, and the robotic system performs object recognition, maps the gaze to an object in the scene, and autonomously grasps it. Similar systems have been presented for wheelchair navigation [30] or mobile robot navigation [80]. In addition, Tong et al. [116] presents a scheme for using gaze location as a set point for a controller during remote surgery, and McMullen et al. [78] uses gaze as an input method to a screen controller which is paired with a brain-computer interface directing a robot arm. While eye gaze can be used as a direct input as seen here, our approach instead monitors the user's *natural* eye gaze behavior while completing a task.

2.3.2 Natural Eye Gaze for Shared Control

Rather than using eye gaze as direct input, in this thesis we propose using natural eye gaze as an indirect input for shared control. This idea has been discussed by Admoni and Srinivasa [3], which proposes to infer the user's goal based on how close their gaze is to each of the possible goals, and Nikolaidis et al. [86], which proposes a framework for modeling user intention from gaze with naive Bayes updates. Possibly the most similar work is Stolzenwald and Mayol-Cuevas [111], in which people operate a handheld controller to interact with objects on a screen. The user's natural eye gaze behavior is used to predict which object they will interact with next, and they show that assisting towards or against that goal influences the user's task success. We build on these approaches by using more knowledge about the eye gaze signal and the dynamics of the scene to more accurately infer the user's goal. Moreover, we demonstrate additional categories of assistance that rely on additional inference from gaze beyond just looking at the user's goal.

Chapter 3

Control Input

Within shared control systems, the most popular signal for goal prediction is the user's control input itself [18, 66, 110, 129, 131]. This signal effectively predicts user goals in a variety of environments, and the signal is generally already available as part of the teleoperation process. However, the prediction information it provides varies with the state of the robot, the task, and the type of input provided (or available to be provided) by the user [39, 40]. Here, we analyze the signal not just for its predictive properties, but in how its predictions align with the needs of a shared control system. Control input is fundamentally *local* to the current state of the task: the user is giving an immediate command to the system. Even when this information is not predictive, often the assistance itself remains optimal, since the user's control input is tightly coupled to the optimal autonomous command. While this coupling makes control input a particularly efficient goal predictor for simple tasks, it also leads to inefficiency in more complicated tasks in which the user cannot act as freely. These scenarios benefit from alternative methods for goal prediction.

In this chapter, we first give an overview of a popular algorithm for inferring goals from control input, which is our baseline for this work, and the shared autonomy assistance algorithm, which lets us analyze the effectiveness of the signal within the assistance context. Next, we describe the coupling between control input and assistance signal, and we prove that for optimal users, control input continues to enable good assistance even as the cost of suboptimal assistance increases. We then use the task model to understand under what circumstances the assumptions of optimal users are violated so that the control input does not give optimal assistance. We conclude by arguing that this suboptimality extends beyond assisted manipulation: goal inference from observing user actions can show the same inefficiencies in complex, branching tasks. Understanding that control input provides a sufficient signal for simple tasks but can be inefficient in more complex ones shows the importance of investigating complementary sensing modalities for goal prediction.

3.1 Background

3.1.1 Control input for goal prediction

In this section, we summarize the approach for goal prediction and assistance given in Javdani et al. [55]. This method uses the user's control input u to predict their goals, expressed as a probability distribution $p(G)$ over a pre-specified set of goal candidates. To do so, it frames goal inference as an inverse reinforcement learning problem [54, 56, 129, 130] and models the teleoperation problem as a family of Markov decision processes (MDPs) with different, pre-specified cost functions $C_g(x, u)$ for each goal candidate $g \in G$. The system then assumes that the user is noisily optimizing the cost function corresponding to their true goal.

First, this method solves each goal MDP for a goal-specific action value function $Q_g(x, u)$. Then, it assumes that the user's action u at each state x is drawn from a distribution given as

$$p(u|x, g) \propto \exp(Q_g(x, u)). \tag{3.1}$$

Note that this is equivalent to the Boltzmann rational model with $\beta = 1$. Given a sequence of state-action pairs $\xi = (x_0, u_0, \cdots, x_n, u_n)$, the strategy assumes that the user's actions are conditionally independent given their goal. Since ξ is not a trajectory, as the robot acts simultaneously with the user, the method treats only the actions u_i as observations. Using Bayes' rule, it aggregates a goal prediction over time using

$$p(g|u_0, \cdots, u_i) = \frac{p(u_i|g)p(g|u_0, \cdots, u_{i-1})}{\sum_{g'} p(u_i|g')p(g'|u_0, \cdots, u_{i-1})}. \tag{3.2}$$

3.1.2 Shared autonomy

To generate an assistance signal from the goal prediction, this method represents the combined robot-human control problem as a partially observable Markov decision process (POMDP), with the user's goal a hidden parameter. The POMDP augments the system state x with a belief distribution over the user's goal given by $p(g)$ above. The action value function $Q(x, p, a)$ depends on the robot state, next action, and the belief state. Since solving the POMDP is generally computationally prohibitive, it adopts the hindsight optimization assumption, which assumes that the uncertainty expressed by $p(g)$ will resolve in the next step. From here, we can find the optimal assistance policy $\psi(x, p(g))$:

$$\psi(x, p(g)) = \arg\max_{a \in A} \sum_{g} p(g) Q_g(a). \tag{3.3}$$

This assumption replaces the overall value function of the POMDP with the expectation over the goal probabilities of the goal-specific value functions, and it reuses the goal-specific value functions $Q_g(u)$ used in Eqn. 3.1. (We use a here to represent that this action is selected by the robot, as opposed to u which is given by the user.) To compute the overall motion, sum a^* with the user command u directly: $a_{\text{exec}} = a^* + u$.

3.2 Evaluating Prediction Sources

While accuracy and forecast horizon are useful measures to evaluate a prediction of the user's goal, we want to evaluate the assistive system as a whole. Accurate predictions only matter when they improve the quality of the assistance provided. Therefore, rather than measuring the prediction accuracy and timing, we analyze the overall performance of an assistive system using control input for goal prediction. Control input is tightly coupled to the assistance itself:

> *The control input of an optimal user and the autonomous selection of the optimal robot action are parallel functions of the robot state.*

This coupling means that an optimal user's control input leads to particularly efficient assistance, as the control input provides information about the goal precisely when

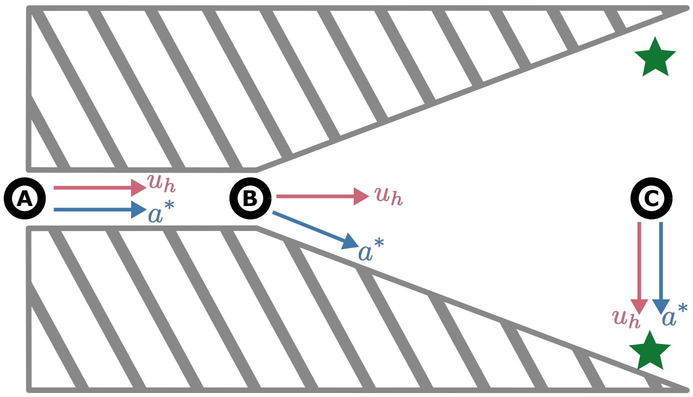


Figure 3.1: Diagram of user input (u) and optimal robot motion (a^*) during an example task. The user moves a point robot to one of the green stars. At **A**, user input and optimal motion are both along $+x$. At **B**, user input is still along $+x$, but the optimal motion is diagonally towards the goal. At **C**, both the user input and the optimal motion point towards the goal. Early prediction improves task performance only at **B**.

the system needs it for its planning.

To explore this coupling between assistance and goal prediction, we start with an example. A planar robot task is shown in Fig. 3.1. The user must move the point robot from **A** to one of the two goals (green stars). At **A**, the only way to make task progress for either goal is to move to the right. The user's expected input is the same for each goal, so it does not yield a goal prediction. However, no prediction is necessary: knowledge of the goal would not change the optimal motion. At **C**, the situation is reversed. The optimal motion is to move either up or down directly towards the user's goal. Here, the system requires a goal prediction to assist. As the user's input depends on the goal, though, the prediction is available.

Location **B** is different. Say the user continues moving in $+x$, which gives no goal information. However, the system can do better. If it knew the goal, it could move diagonally; without goal knowledge, however, it must wait until observing a goal-dependent user input (like at **C**) before it can assist along the y axis. Early, independent goal prediction only improves assistance at points like **B**, where goal information would change the motion but the user input does not provide it.

To formalize this analysis, consider an assistive system with two goal candidates $\{g_1, g_2\}$.

- Two goals require *different* motion at x if their optimal *robot* motion a^* depends on the goal: $a_1^*(x) \neq a_2^*(x)$; the motion is *identical* otherwise.
- Two goals are *distinguishable* at x if the observed *user* input generally differs based on the goal: $u(x|g_1) \neq u(x|g_2)$; they are *indistinguishable* otherwise.

Identical and different motion are properties of the robot's state, whereas distinguishable and indistinguishable goals are determined by the user's input. When the user is acting near-optimally, different motion likely leads to distinguishable goals. Next, we formalize this alignment between optimal users and effective assistance.

3.3 Noisily Optimal Input Bounds Regret

The above analysis suggests that when users give approximately optimal input, the autonomy will likely receive the information needed to provide assistance. If we assume the user follows the model given in Eqn. 3.1, we can evaluate the expected performance of the shared autonomy policy given in Eqn. 3.3. We show that as the importance of taking the optimal robot action (measured by regret) increases, the user's probability of providing a distinguishing input increases faster, such that the overall system has bounded regret.

For simplicity, assume we only have two goal candidates with action value functions Q_1 and Q_2, and assume without loss of generality that the user's goal is g_1. We also assume that the set of actions A is finite and identify actions with the same $Q(a)$. At some state x (which we drop for ease of notation), let Q_1^* be the maximum value of $Q_1(a)$ attained at some action a_1^*. When using Eqn. 3.1 to infer $p(g)$ from control input u as above, the shared autonomy policy $\psi(p(g))$ is in fact a function of u and we write $\psi(u)$. We can then compute the expected regret $R(\psi(u)) = Q_1^* - Q_1(\psi(u))$ of the assistance policy $\psi(u)$ over the user model.

We can measure the importance of taking a_1^* over any other action a' by letting $R_{\min}$ represent the minimum regret over all alternative actions, which is defined for finite A:

$$R_{\min} = \min_{a \neq a_1^*} R(a).$$

We want to understand the behavior of the system as $R_{\min}$ increases. Increasing $R_{\min}$ can be achieved by changing the selected state or the MDP itself. To make this concrete, we consider Q_1 an MDP with reward function $r(x)$. If we scale that reward function, $r'(x) = \lambda r(x), \lambda > 0$, the value function scales similarly, $Q_1'(x, a) = \lambda Q_1(x, a)$. Then, $R'_{\min} = \lambda R_{\min}$, and we can then consider the behavior as λ increases. Similar effects can also occur by changing x or $r(x)$ in other ways that are more complicated to formulate. However the change occurs, increasing values of $R_{\min}$ represent increased importance of taking the optimal action.

We can now determine the expected regret of the assistance policy under a user following Eqn. 3.1.

Proposition.

$$\lim_{\lambda \to \infty} E_u[R(\psi(u))] = 0. \tag{3.4}$$

We sketch a proof in two parts. First, we show that as $\lambda \to \infty$, the assistance action taken when observing the optimal action from the user, $\psi(a_1^*)$, becomes a_1^*:

$$\lim_{\lambda \to \infty} \psi(a_1^*) = a_1^*.$$

By manipulating Eqn. 3.3 and collecting terms in $p(g)$, we find that $\psi(a_1^*) = a_1^*$ is equivalent to, for all $a' \in A$,

$$p(g_1|a_1^*)(Q_1(a_1^*) - Q_1(a')) \geq p(g_2|a_1^*)(Q_2(a') - Q_2(a_1^*)).$$

If $a_1^* \neq a_2^*$, then $p(g_1|a_1^*)$ goes to 1 as the cost of other actions increases, while $p(g_2|a_1^*)$ goes to 0 as $Q_2(a_1^*)$ increases over Q_2^*. Once the importance of taking the optimal action exceeds some threshold, the assistance will take that optimal action whenever it observes it from the user.

The expected regret is given by

$$E_u[R(\psi(u))] = \sum_u R(\psi(u))p(u|g_1).$$

From above, once λ is sufficiently large, $R(\psi(a_1^*)) = R(a_1^*) = 0$. We can therefore break a_1^* out of the sum. If we define $R_{\max} = \max_a R(a)$ analogously, we have

$$\begin{aligned} E_u[R(\psi(u))] &= R(\psi(a_1^*))p(a_1^*|g_1) + \sum_{u \neq a_1^*} R(\psi(u))p(u|g_1) \\ &= \sum_{u \neq a_1^*} R(\psi(u))p(u|g_1) \\ &\leq \lambda R_{\max} p(u \neq a_1^*|g_1). \end{aligned}$$

Finally, we bound the probability of the user giving an action other than the optimal action based on our model of user behavior,

$$\begin{aligned} p(u \neq a_1^*|g_1) &= \frac{\sum_{u \neq a_1^*} \exp \lambda Q_1(u)}{\exp \lambda Q_1^* + \sum_{u \neq a_1^*} \exp \lambda Q_1(u)} \\ &= \frac{\sum_{u \neq a_1^*} \exp(-\lambda R(u))}{1 + \sum_{u \neq a_1^*} \exp(-\lambda R(u))} \\ &\leq \frac{(|A| - 1) \exp(-\lambda R_{\min})}{1 + (|A| - 1) \exp(-\lambda R_{\min})}. \end{aligned}$$

Putting it all together,

$$E_u[R(\psi(u))] \leq \frac{\lambda R_{\max}}{1 + \frac{1}{|A|-1} \exp(\lambda R_{\min})}.$$

Since $R_{\min}$ and $R_{\max}$ are fixed, the result goes to 0 as $\lambda \to \infty$ and the regret is bounded. □

As the importance of taking the optimal action increases, the chance of the user performing that optimal action under the model increases exponentially faster, so the system is more likely to receive the information it needs.

3.4 Control Input Restrictions Require New Prediction Sources

We see from the previous result that noisily-optimal users are particularly easy to assist using input-based goal prediction. If we remove the assumption of optimality —

by assuming, e.g., that the user acts randomly, mistakenly, or adversarially — we no longer have guarantees that the assistance will behave well. However, there is a large class of problems for which the user still acts optimally but the assistance can be arbitrarily ineffective: when the user's action are limited to only a subset of the actions that the autonomous system can take.

> *It is not the user's suboptimality that limits the effectiveness of the system, but the constraints that the system itself puts on the user's behavior.*

One common example of this problem in teleoperation is the use of modal control. In this scheme, the robot can control its end-effector simultaneously in all directions. However, the user has only a 2-D joystick with which to control the robot. They can fully control the robot by cycling through modes, with the joystick controlling x/y, z/yaw, and pitch/roll in turn. If the optimal action does not align with a single control mode, the user cannot perform it. The best the user can do is to provide input in the single most useful mode. And when the robot motion is different but the control input *within the most useful mode* is not distinguishing, assistance does not have enough information to be optimal.

We can return to Fig. 3.1 to explore this limitation further. At **B**, we observe the user giving indistinguishable motion, though the assistance requires different motion per goal. In the noisily rational model, this user action occurs at a lower probability than a distinguishable input. However, if we add the additional restriction that the user can only provide axis-aligned commands, the user's input at **B** is optimal for solving the task themselves (assuming that switching modes incurs a nonzero cost). Even with an optimal user, the assistance does not receive enough information to provide full assistance. In these situations, the system benefits from an alternative, global method for goal prediction that is less reliant on the user's local behavior.

This inefficiency in the face of input limitations extends beyond manipulation to include any task schemes in which the user performs steps sequentially and only some steps provide information about the user's choices. For example, consider a schematic for preparing a hot drink, shown in Fig. 3.2. Assume the task has two variants: the user can prepare tea or coffee. The variants share two subtasks, heating water and taking out a mug, and each has its own variant-specific preparation subtask. While each subtask is contains independent entry points and could theoretically be started

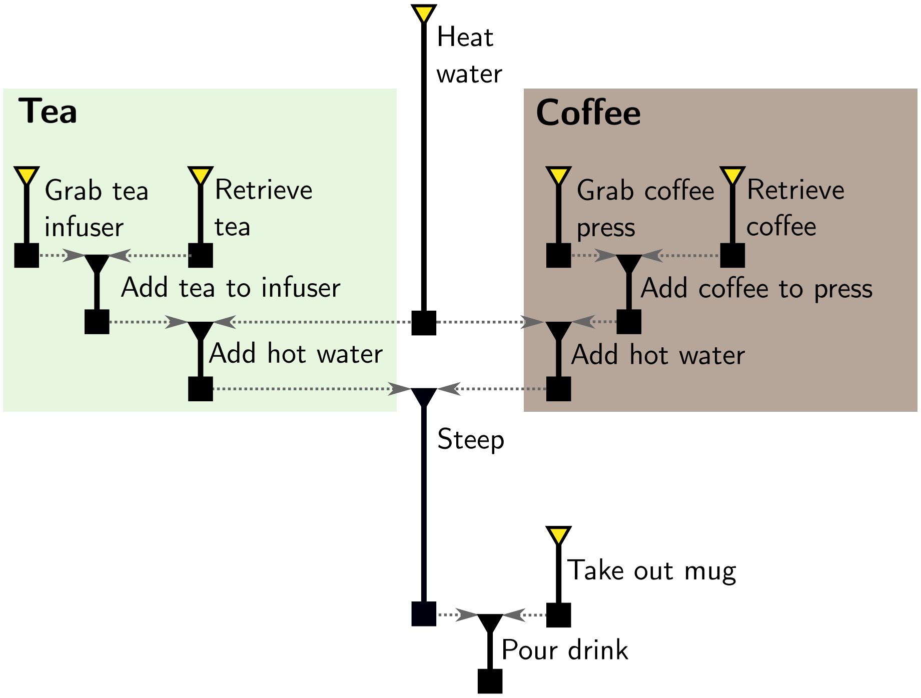


Figure 3.2: Task diagram for preparing a hot drink. A user may be preparing coffee or tea; each of these versions of the task share some steps but differ in others. If the user starts by heating water, the most efficient start, the system can retrieve a mug (necessary for both tasks), but it cannot do more, since the user's actions have not yet revealed whether they are making coffee or tea. The system can only help with task-specific steps after the user has performed an action unique to their task.

simultaneously, based on idle time, the optimal start for the user is to perform an indistinguishable action. A robot assistant that observes the user's actions can always start on the shared subtasks. However, before the robot can provide variant-specific assistance, it must wait to observe the user starting the task, by e.g. retrieving a tea infuser or coffee press. The implicit goal information given by user's actions will eventually give the system enough information to assist, but depending on the user's task sequencing, it might not receive sufficient information early enough to provide maximum assistance.

3.5 Conclusion

For goal inference during assistance, control input provides an easy-to-obtain, effective signal. In simple tasks, in which users can act optimally, this signal is particularly efficient, as it gives local information and is tightly coupled with the state of the task itself. As tasks become more complex and as user actions are more constrained, the signal can become less efficient. Specifically, complex or constrained tasks have states in which the optimal assistance command is different depending on the user's goal, but the user's optimal behavior does not distinguish between them. To achieve better assistance performance in these situations, we look to alternative methods of goal prediction.

This model assumes that the user is completing the task as if they are unassisted and have full control of the task process. However, this is not true in general: as users gain familiarity with the robot's performance, they can act to optimize the system as a whole. In studies of users interacting with this form of assistance, we observe that as they get accustomed to the system, they will often change from executing independently goal-directed actions to informative actions, which are less optimal for the user alone but provide the prediction information needed by the system to enable overall performance increase. The model presented here provides a framework for investigating more sophisticated models of human-robot collaboration.

Chapter 4

Gaze Behavior

To improve the performance of assistance, we look for a prediction source that is less directly tied to the state of the task and can provide *global* information about the user's goal. A particularly appealing signal for that role is the user's natural eye gaze behavior. Gaze reveals people's intentions in a variety of ways in many tasks (see Lukander et al. [74] for a full review), so it is a promising signal for goal inference during assisted manipulation. Unlike intentional eye gaze paradigms, in which users actively use their gaze to provide direct input to a system, natural eye gaze captures the automatic, unconscious eye motion that people perform. This gaze behavior is closely connected to the user's current task in general [114]: in manipulation, people look at important parts of the task (e.g. upcoming goals or obstacles [57]) before moving towards them and ahead to locations that will be important in the future [79]. These patterns suggest that eye gaze can provide the global information about the user's goals to complement the local information provided by the control input.

However, natural gaze during robot teleoperation may not follow the same patterns as people performing tasks by hand. While prior work shows that natural gaze while operating robots gives some information about the tasks being performed [6, 37, 98], translating from by-hand manipulation to teleoperated manipulation changes the problem. To better understand these complications, we characterize gaze behavior during a grasping task in more detail in order to understand both its potential performance for goal prediction and how to use it computationally.

We start by describing two user studies that captured eye gaze during teleoperated manipulation. The first was collected in prior work, which we analyzed; we collected the second using better sensors, an adapted procedure to obtain higher quality data, and a user consent procedure that gave explicit permission to publish the data for other researchers. For each study, 24 participants performed a food acquisition task by teleoperating a robot arm while we recorded their eye gaze patterns. These data sets provided a foundation both for a qualitative understanding of gaze behavior and for training and evaluation of goal inference algorithms.

We find that people look at their goal objects during the task, and they do so more often during particular stages of the task. However, most gaze is directed at the end-effector of the robot, which does not give information about the user's goal. When problems occur during the task, people also tend to look at the locations of the problems; these glances act as additional noise for goal prediction, but may enable more sophisticated modeling of the user's model of the task. Together, these findings suggest that gaze is a powerful signal for predicting people's goals early in the task, but those predictions are not consistently available.

4.1 Data Collection and User Studies

We first describe a data collection study whose results were used in an initial exploration of gaze during manipulation [12, 54]. We next describe a data collection study we designed to improve upon the previous experiment, which includes a more sophisticated eye tracker and additional sensing modalities [84].

4.1.1 Original Study

During the task, participants sit facing a table. In front of them are a robot arm (a Kinova Mico [62]) mounted to the table and a plate holding three morsels of food (we use marshmallows for their ease of spearing). Participants began the study (after informed consent) by receiving an explanation of how to control the robot and then spending 5 minutes practicing with it. Then, for each trial, participants were instructed to select and share which morsel they intended to target. They then used the robot to move the fork held by the robot to a position above their target morsel. They then pressed a button on the joystick to complete the task, at which time the

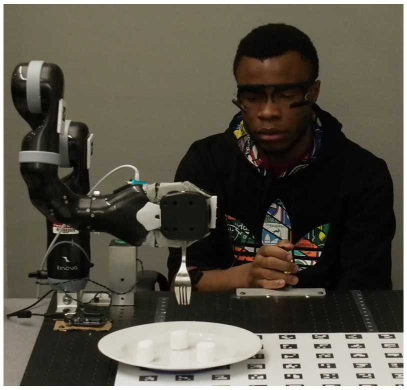
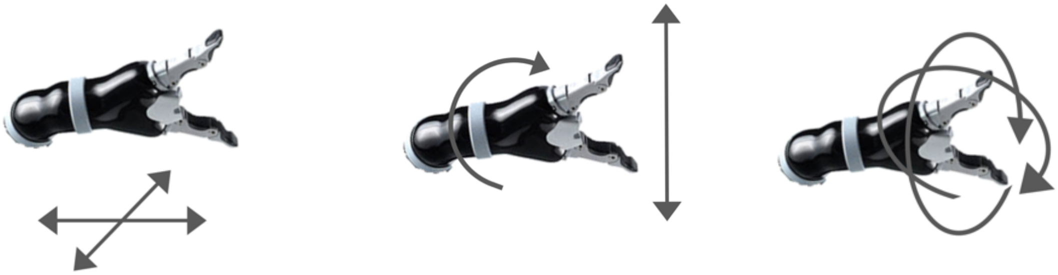


Figure 4.1: The eating task.

robot autonomously moved down to the morsel, then it moved the fork towards the user's mouth to simulate most of an eating motion. The actual spearing was done autonomously so that the minimum robot height could be restricted and we could avoid table collisions.

During each trial, the user controls the robot using a two-axis joystick via modal control, which is typical for these types of arms. During modal control, the two-axis joystick maps to successive pairs of degrees of freedom of the end-effector (x/y, z/yaw, pitch/roll; see Fig. 4.2). Pressing a button on the joystick cycles through the modes. The robotic system can process this joystick input directly into end-effector commands or add an assistance strategy. The five minute practice period helps participants to understand this control strategy, though it remains difficult for many users.



(a) x-y mode

(b) z-yaw mode

(c) pitch-roll mode

Figure 4.2: Modal control for joystick-controlled manipulators. The user controls two axes of motion at a time with a joystick and uses a button to cycle through the three modes.

Participants performed this task five times each for four different assistance conditions, fully counterbalanced over 24 participants. In this study, the assistance conditions consisted of full teleoperation, in which the users had complete control; shared autonomy, a policy-based assistance blending strategy (elaborated on in Sec. 3.1.2); policy blending, an alternate, more conservative assistance strategy; and full autonomy, in which the robot selected a morsel and planned a trajectory itself while ignoring all user input. While eye gaze was collected for all conditions, we focused on the teleoperation and shared autonomy assistance conditions primarily, as they represent the conditions most related to our intended use. Eye gaze data was collected over a total of 120 trials per condition. Eye gaze data was collected using a Pupil Labs Pupil monocular eye tracker [93].

4.1.2 HARMONIC Study

After publishing the results of the previous study, we decided to repeat the study in order to amass a higher quality data set that would be appropriate for quantitative analysis. To do so, we repeated the study described above, but with a few enhancements. First, we upgraded the eye gaze sensor to a binocular sensor, which gives much higher quality sensing for three-dimensional gaze. We also recorded more of the internal gaze sensor data for later postprocessing. Second, we altered the assistance algorithm slightly, by replacing the *shared autonomy* and *blend* conditions with two different levels of shared autonomy, and replacing the *autonomous* condition with a mode in which goal intention was derived from the user input but actual control direction was supplied entirely by the autonomous system. Finally, we added an electromyography sensor on the user's wrist. This data set has been made publicly available at `harp.ri.cmu.edu/harmonic` and is published as Newman et al. [84].

4.2 Eye Gaze Behavior

To understand the broad patterns of eye gaze behavior during teleoperation, we examine the participant's eye gaze behavior during one teleoperated trial (Fig. 4.3). In this run, the participant begins by performing a *planning glance*: the user looked at the end-effector of the robot, then their target object, then back to the end-effector,

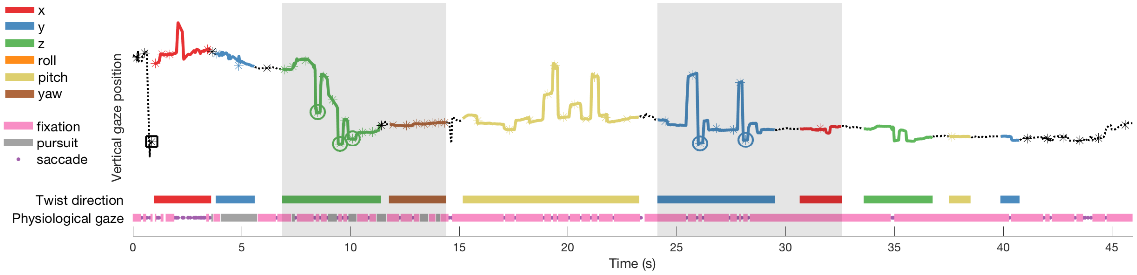


Figure 4.3: Vertical position of gaze points in the world image over time from a representative trial. Twist direction colors indicate which DOF is being controlled by the participant through the joystick; physiological gaze colors and dots indicate detected fixations, smooth pursuits, and saccades (see Sec. 5.2). Plate glances are outlined with either a black square (planning glance) or colored circle (monitoring glance). Shaded sections highlight two examples of repeated monitoring glances.

all without moving the robot. Next, the participant moves the robot in the x/y mode to the approximate location of the end-effector above the morsel, and they watch the end-effector through the entire process. The participant then toggles to the z/yaw mode and lowers the robot in z, while performing a *monitoring glance*: alternating focus between the end-effector and the target while the robot is moving. Then, the participant aligns the fork vertically above the morsel while moving in yaw, pitch, and roll; they look at different places on the end-effector but do not glance at their target. Finally, the participant performs fine alignment in x, y, and z, performing *monitoring glances* throughout.

From this example, we derive some generalizations about eye gaze during teleoperated manipulation:

People spend a lot of time looking at the end-effector of the robot. Unlike in by-hand manipulation, people look at the end-effector of the robot throughout the trial. Specifically, 68.1 ± 2.1% of the fixations during each trial were at the end-effector or tool. Presumably, this gaze difference is due to people needing visual feedback to determine the location of the robot end-effector, whereas during by-hand manipulation, people can use their own proprioception to determine their hand position.

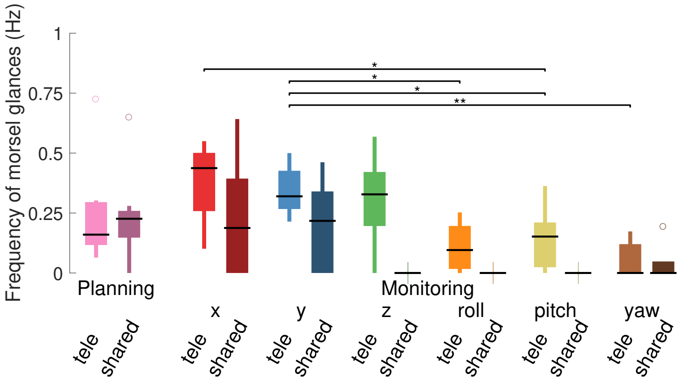


Figure 4.4: Mean frequency of planning and monitoring glances to the plate during each robot assistance mode. Monitoring glances are subdivided by joystick control direction. * indicates significance at the $\alpha = 0.05$ level; ** at $\alpha = 0.01$.

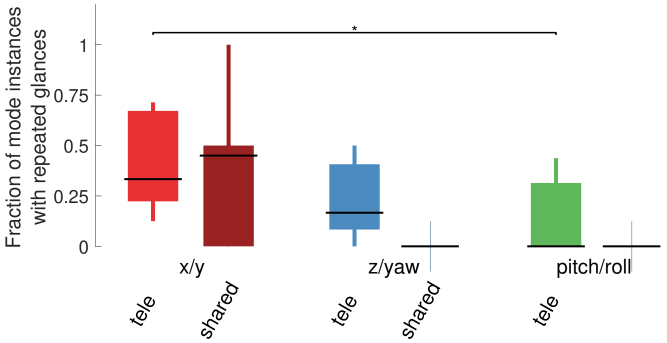


Figure 4.5: Proportion of joystick control sequences of the same mode that contained multiple (≥ 2) monitoring glances, subdivided by their control mode. * indicates significance at the $\alpha = 0.05$ level.

People look at their goals based on the status of the task. As indicated above, two eye gaze patterns recurred: *planning glances*, in which people held the robot stationary and alternated their focus between the end-effector of the robot and their goal object, and *monitoring glances*, in which people moved the robot while looking back and forth between it and their goal position. These patterns were frequent, with planning glances appearing in 76% of trials, and have also been observed in teleoperation tasks using motion control [37]. In addition, morsel monitoring glances were significantly more frequent during translation than during rotation (Fig. 4.4). Repeated morsel monitoring glances, in which participants checked the morsel position more than once while they watched the robot end-effector, also occurred more often when in the x/y translation mode than in the pitch/roll pure rotation mode (Fig. 4.5). This distinction between rotation and translation may be because participants find rotation harder [128] or because an external reference for the target is less necessary during rotation. In both cases, we find that the timing of meaningful plate glances is highly related to the dynamics of the task.

Another key insight from this study is how people behave when they encounter problems in the task [8]. While these incidents occur infrequently in the dataset, we can nevertheless examine some case studies to understand how people's eye gaze changes when something goes wrong. To illustrate this phenomenon, we describe two case studies that appeared in our HARMONIC data set (see Sec. 4.1.2).

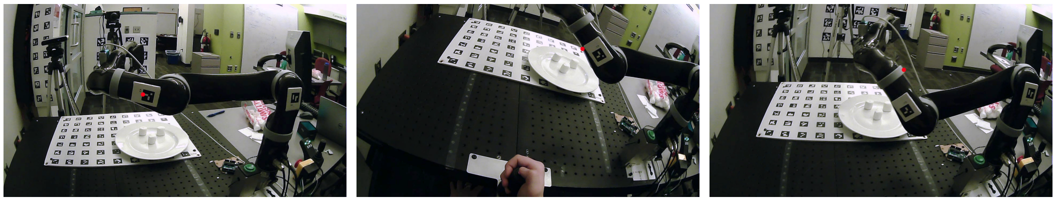


Figure 4.6: When the robot occludes the goal morsels, people move their heads for a better view.

People move their heads to compensate for robot occlusion (Fig. 4.6). In several cases, people moved the robot into a configuration where the robot itself occluded their view of the target morsel. In these situations, people often moved their heads significantly more than usual in order to get a better view. While this pattern is not revealed through semantic gaze analysis (see Sec. 5.3), it can be obtained from the raw gaze signal and the head motion. Knowledge that the operator is struggling to see a particular object can be used to trigger contextual assistive actions.

People look at robot joints during kinematic failure (Fig. 4.7). As noted above, people often look at the end-effector of the robot. People rarely look anywhere else on the robot, except for one notable case: when the robot goes into a problematic kinematic configuration, people will look at the joint that is causing them an issue. For example, the robot has two general configurations while being teleoperated downward, which we can call *elbow-down* and *elbow-up* (see Fig. 4.7). In the *elbow-down* configuration, the robot cannot be moved towards the table, as the robot joint

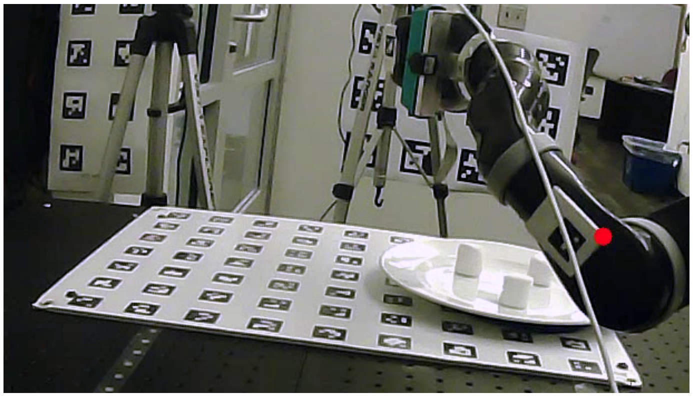


Figure 4.7: When the robot moves into a problematic joint configuration, people look at the joint that is causing problems. The red dot represents the participant's gaze location.

collides with the table before the target position is reached. It is difficult to fix this problem with only end-effector control, as the motion required is by definition within the nullspace of the joint Jacobian. When this problem occurs, people often look at the location of the joint that is causing them a problem. Again, with contextual information, this eye gaze pattern can indicate to an assistive system that corrective behavior would be especially useful.

4.3 Conclusion

These results build a theoretical foundation for using eye gaze as a signal for goal inference during assistive manipulation. We show that people indeed look at their goal objects while teleoperating a robot, so this signal can be used to predict the operator's goal. However, we find that this signal is not consistent, though it follows some clear patterns. First, gaze is mostly directed at the end-effector of the robot, meaning that goal-directed glances are relatively rare. They do occur more often during translation than in rotation, which suggests the usefulness of context-aware algorithms for interpretation. On the other hand, people can perform the task without ever glancing at their goals, especially during repeated tasks in the same environment; this limitation restricts the performance of algorithms that require goal-directed gaze to distinguish the user's goals and suggests that gaze is inherently unreliable for goal prediction. This finding motivates both the development of sequence-aware algorithms for gaze interpretation and our focus on using gaze alongside more reliable information sources like the user's control input. Finally, the data set collected enables the training and evaluation of different gaze analysis methods on actual participant data.

Looking beyond goal inference, moreover, we find that natural gaze can reveal more information about the user during the task. For example, kinematic failures lead to people looking at the location of the failure, and gaze behavior changes as the robot occludes the goals. This signal may also capture similar situations, such as detecting user distraction when they look away from the task or sensing errors when users look at locations of objects that the system failed to detect. Showing that eye gaze can be used for broader failure detection and user behavior inference suggests that developing a richer understanding of gaze during complex interactions is an important area of future work.

Chapter 5

Gaze Processing

One of the challenges of extracting information from eye gaze behavior is incorporating the scene context into the signal. While statistical analysis of the signal has been effective for such problems as activity recognition or expertise measurement [14, 20, 21, 23, 34, 46, 71, 126], the information we are concerned with is intrinsically tied with the specific objects of the task. Therefore, we first process the data to incorporate semantic information about what objects people are fixating.

One strategy for incorporating context, which inspires our approach, is to label the user's gaze according to the object in the environment nearest their gaze. This labeling process has been explored for both 2-D [17, 75] and 3-D [41, 70, 90–92, 108] gaze, though these approaches generally focus more on the labeling process itself than on applying it to real-world systems. Many projects that use gaze for robotic systems [13, 69, 70, 107, 118, 124] do not describe the gaze processing in detail and (implicitly or explicitly) rely on on the user's *intentional* gaze to correct the system towards a better outcome. We most closely follow the approach given in Huang et al. [50], which splits the gaze into individual fixations, labels them with pre-specified keypoints, and uses these label sequences to predict user goals.

However, this labeling process is more complex than it seems. Gaze data is inherently noisy, and becomes more so when trying to use more cost-effective sensors in complex, 3-D, moving environments (Sec. 5.1). For goal prediction, incorrect labels can cause complex feedback loops with system behavior, and using natural gaze means that the system cannot rely on the user changing their input behavior to account for these errors. To enhance performance in these situations, we formalize the *semantic gaze labeling* process and present algorithms for improving labeling accuracy even without clean underlying data.

In this chapter, we describe the pipeline for collecting raw gaze data, segmenting it into individual fixations, and labeling these fixations with corresponding objects in the scene (Fig. 5.1). We begin with an overview of the raw gaze signal and sensing mechanisms. Next, we describe the physiological properties of gaze and describe an algorithm for fixation detection.

We then introduce our formalism for semantic gaze labeling (Sec. 5.3) and present three features to use for the labeling algorithm: *position features*, a baseline that measures the distance between the fixation and the label (Sec. 5.4); *velocity features*, which compare the motion between fixations to the motion between keypoints (Sec. 5.5); and *pursuit features*, which compare the motion within fixations to the motion within keypoints (Sec. 5.6). Accurately processing the raw gaze signal into labeled sequences of fixations, using the full pipeline described in Fig. 5.1, enables goal prediction to use scene context, which underlies our gaze prediction algorithms described in the next chapter.

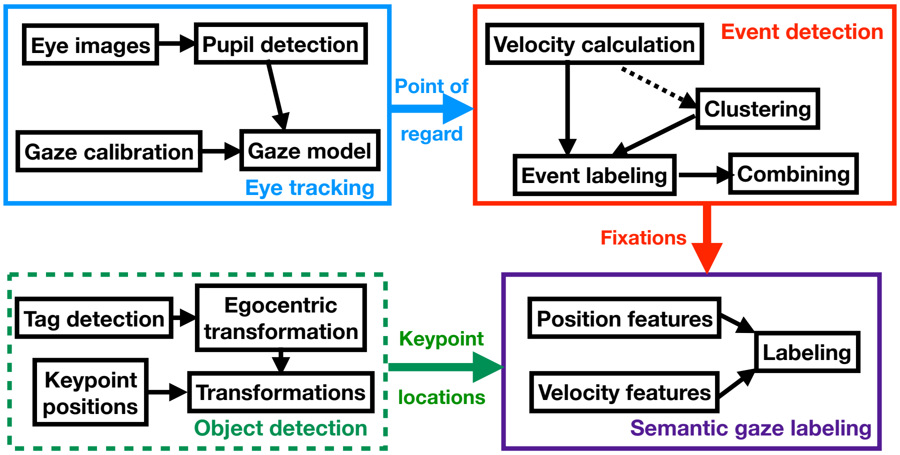
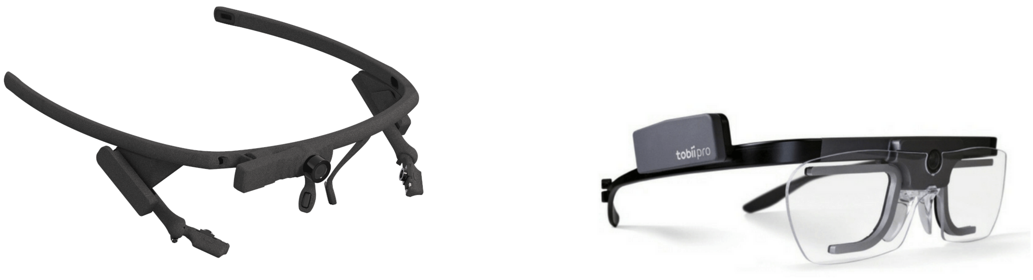


Figure 5.1: Flow chart of the gaze analysis pipeline. The eye tracking system provides the user's point of regard at 30-200 Hz (depending on the sensor; Sec. 5.1). This signal is then passed into the event detection framework (Sec. 5.2), which uses a pre-trained clustering model (dashed arrow) to output a sequence of fixations at ∼1-2 Hz. Finally, the semantic gaze labeling process labels these fixations with keypoints supplied by an external tracking system (dashed rectangle; Sec. 5.3).

5.1 Eye Gaze Sensors

To understand the gaze processing pipeline, we begin with an overview of the gaze sensors themselves. Developments in eye tracking technology over the last several years mean that tracking gaze direction in a real-world environment at high sample rates (90-200 Hz or higher) is now possible. In this work, we use both 2D screen-based gaze trackers and 3D mobile eye trackers. Screen-based trackers such as the Tobii 4C [115] typically sit at the base of a monitor and emit gaze location of a user as a pixel on the screen; the relative consistency of the user's head location makes the system fairly reliable, though it cannot be used for 3D settings. For these settings, mobile eye trackers, such as the Pupil Core [93] or the Tobii Pro Glasses 2 [115] (Fig. 5.2), are appropriate. These sensors typically consist of a glasses-like frame worn by the user, on which a number of cameras are mounted. One or two IR cameras are mounted above the users' eyes (corresponding to a monocular or binocular setup) and record high-frequency video of the eyes themselves. In addition, a forward-mounted ("egocentric" or "world") camera captures the scene from the point of view of the user.



(a) Pupil Core

(b) Tobii Pro Glasses 2

Figure 5.2: Mobile eye trackers.

For screen-based eye trackers, proprietary algorithms use the sensor's cameras along with a 5-point calibration process to produce the user's gaze location as a 2D pixel location on a screen. Mobile eye trackers are similar, but calibration must take into account depth of field [32]. For mobile tracker work, we use a 27-point calibration procedure: the user looks at a calibration target, which is moved to each point in a 3-by-3-by-3 grid encompassing the workspace of interest.

The output of the Pupil Core mobile eye tracker includes two 120Hz streams of pupil pixel location in the eye cameras and one 30Hz stream of gaze target pixel location in the world camera. Similarly, the Tobii 4C screen tracker produces a pixel location at 90 Hz. This raw data is processed in the rest of our pipeline.

5.2 Event Detection

To simplify the semantic gaze labeling process, we can take advantage of the physiological characteristics of human gaze. Rather than having free control over eye gaze direction, people follow consistent eye gaze patterns, which consist primarily of *fixations* ($\sim$ 500-2000 ms stationary periods) separated by *saccades* (rapid $\sim$ 100-500 ms ballistic trajectories between fixation locations). Controlled eye gaze motion typically only occurs in two situations: *smooth pursuits*, when someone looks at a moving object and follows it with their eyes, and *vestibulo-ocular reflex* (VOR), when people move their heads and their eye gaze moves to compensate while focusing on the same object. Therefore, when determining what object people are looking at, we need not perform the classification problem at the sample rate of the eye gaze sensor. Instead, we can perform saccade and fixation segmentation (known in the literature as *event detection*) and assume that during a fixation, smooth pursuit, or vestibulo-ocular reflex, the entire period is spent focusing on the same object.

There are two traditional algorithms for event detection: dispersion thresholding (I-DT) and velocity thresholding (I-VT) [103]. Both depend on noting that the point-to-point velocity during saccades are generally much larger than during fixations (or VOR, or pursuits). In I-DT, a measure of dispersion (e.g., variance) is calculated over windows of the eye gaze signal. Windows less than a manually-chosen value are determined to be fixations, while windows above the value are labeled saccades. In I-VT, the point-to-point velocity of the signal (the numerical derivative) is computed, and each point is labeled a fixation or saccade if it is below or above a custom threshold. Then, successive fixation labels that exceed a minimum fixation time are fused together and determined to be a single fixation.

For this work, we use a variant of I-VT, known as I-BMM [60, 104]. This method similarly calculates the velocity of the eye gaze signal (by angle), but learns a dynamic threshold by fitting a 2-component Gaussian mixture model to a sample of eye gaze data. Then, adjacent fixation labels are fused, and fusions that exceed a specified minimum time are labeled as fixations. A Python implementation that works both offline and online was written and made open-source[1].

Our event detection algorithm varies slightly from the standard eye tracking approach due to the dynamics of our task. First, we do not distinguish between fixations, pursuits, and vestibulo-ocular reflexes. In the eye gaze signal, fixations appear as periods of zero velocity, whereas pursuits and VORs appear as periods of small but nonzero motion. However, since all three categories involve the user focusing on a single object, for labeling objects of interest the distinction is not important. Therefore, in our fixations, there may be some internal motion, which may make distinguishing between saccades and fixations more difficult. However, saccade motions are still significantly significantly faster than pursuit or VOR motions. When motion is fast enough that pursuits are split into several consecutive fixations, the labeling procedure (below) labels each with the same object, so a corrective procedure can recombine adjacent identically-labeled fixations into a single fixation.

5.3 Semantic Gaze Labeling

Now that we have obtained a sequence of separate fixations, we proceed to the labeling step. Building on previous work, especially Huang et al. [50], we can describe this labeling process in more detail. To do so, we start by explicitly laying out two assumptions:

Users focus on one object at a time. In particular, people look directly at objects of interest. This assumption generally holds during by-hand manipulation [43, 57, 68]. While peripheral vision does provide some assistance for by-hand manipulation [109], the central assumption that people direct their gaze directly at specific, task-relevant objects is strongly supported in the psychology literature [15].

Users look at task-relevant objects. Manipulation tasks tend to involve interacting with a limited number of objects that stay unchanged during the task. Since glanced objects are task-relevant [43, 67], we assume that the the majority of informative glances are to the set of objects relevant to the manipulation task. Any

[1] http://github.com/HARPLab/ibmmpy

off-object glances are assumed to be noise or a signal that (for example) the user is not paying attention to the task, so they can safely be combined into a single category. Note, however, that this assumption does not hold as strongly during less structured tasks [114].

These two assumptions enable us to incorporate context into the gaze signal by labeling gaze data with the object that the user is most likely looking at during that time. Then, rather than raw gaze data, we use sequences of object labels as input to intent inference systems; we assume these object locations can be obtained from a separate sensing process. This processing pipeline converts the raw gaze data to a format where context is already included, which eases the development of intent inference systems.

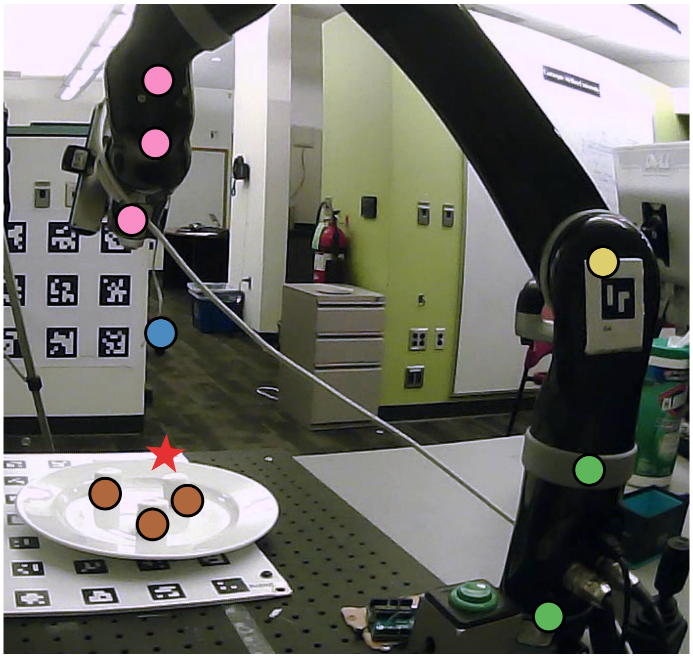


Figure 5.3: Fixations were manually assigned to one of ten scene keypoints, which were the three goal morsels and each robot joint. For bulk comparison, scene keypoints were also grouped as shown in the colors of the scene.

5.3.1 Formal definition

Formally, we describe the problem as such:

Given: a sequence of eye gaze locations segmented into fixations $I_t = (\tau_0^t, \cdots, \tau_{m_t}^t)$, where I_t is an index set and all such sets partition the original gaze sample sequence g_τ. These fixation subsequences are derived using an event detection algorithm, as described in Sec. 5.2. Here, we represent gaze locations g_τ as unit vectors pointing from the sensor frame into the world.

Given: A set of timeseries keypoint locations k_τ^i, determined from an object detection algorithm. Each keypoint k^i is a semantically relevant object in the workspace (determined manually).

Goal: Assign to each fixation t a label $\ell_t \in (1, \cdots, n)$ representing which keypoint the user is likely to be looking at for that particular fixation. Then, the gaze can be represented as a sequence $(f_1, \cdots, f_n)$ where $f_t = (\ell_t, d_t)$ represents both the fixation's label and its duration $(d_t = \tau_{m_t}^t - \tau_0^t)$.

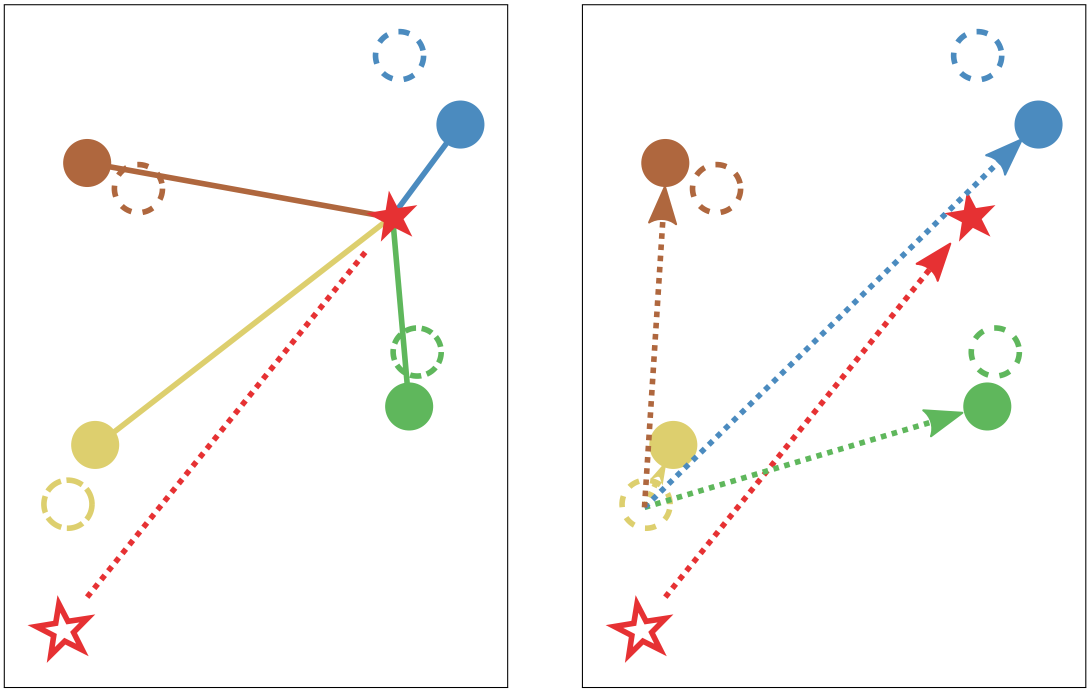
To perform the labeling process, we present three features to use. Position features, which calculate the distance between the fixation center and the object's location, represent the standard approach. We introduce velocity features, which compare the motion *between* fixations with the motion of each keypoint, and show that they improve accuracy over position features alone [9]. Finally, we propose pursuit features, which compare the motion of the gaze *within* each fixation to the motion of each keypoint.

To assign a label to each fixation, we use a simple feature weighting procedure:

$$\ell_t = \arg\min_i \sum_j w_j q_j(g_t, k_t^i) \tag{5.1}$$

in which the label ℓ_t is assigned based on a weighted linear combination of each feature function q_j with weights w_j. While more sophisticated classification algorithms may improve the data, the natural meaningfulness of the features chosen combined with the desire to evaluate the features as directly as possible motivates the simple method. In addition, if we wish to extend the labeling procedure to produce a distribution over keypoints rather than a keypoint itself, we can instead use the softmax

$$p(\ell_t = i) \propto \exp(-\sum_j w_j q_j(g_t, k_t^i)).$$



(a) Position features.

(b) Velocity features.

Figure 5.4: A schematic representation of the calculated features. Colored circles represent keypoints. Filled circles represent keypoint positions at the current time. Outlined circles represent keypoint positions at the previous time. The filled red star represents the average fixation location at the current time, and the outlined star the previous fixation location. Figure 5.4a represents the position features at the current time; the closest keypoint to the fixation is the blue one, but the distance is similar to the green distance due to a constant offset. Figure 5.4b represents the velocity features; the relative motion that the fixation would have taken between the previous time and the current time is represented by a dashed arrow for each keypoint and the observed relative motion by the dashed red arrow. The high similarity between the blue arrow and the red arrow leads to a small velocity feature for the blue keypoint independent of the constant offset.

5.4 Position features

One straightforward way to compare the fixation subsequence to each keypoint to determine how well they align is to use the distance between them averaged over the entire fixation (see Fig. 5.4a). In particular, let

$$c_t = \mathrm{mean}_{\tau \in I_t}\, g_\tau$$

represent the average gaze point during the fixation, and

$$k_t^i = \mathrm{mean}_{\tau \in I_t}\, k_\tau^i$$

represent the average keypoint location for each keypoint i during each fixation t.

In the case of an actual fixation, the gaze is roughly stationary and this sequence should have small variance. During pursuit or vestibulo-ocular reflex, the point will move during the sequence so the average is a poor measure, but by assumption the corresponding keypoint moves similarly, so any error induced by taking the mean will be matched by the error in the keypoint.

Once the means are calculated, we can determine a position feature by computing the distance between the fixation mean and the position mean,

$$p_t^i = d_p(c_t, k_t^i), \tag{5.2}$$

where d_p is a distance function over gaze points g_τ (here the cosine distance between the two 3-D vectors).

5.5 Velocity features

One issue with the position features described above is that they tend to be highly susceptible to static error. Many of the errors in the gaze signal, such as calibration error, appear as large, slow-changing position errors. To counteract the effect of constant errors, we can draw inspiration from signal processing and take the derivative of the comparison (see Fig. 5.4b). In particular, if we assume that

$$c_t = k_t^{\ell_t} + \epsilon_t,$$

where ϵ is a roughly constant error term ($\frac{\partial \epsilon}{\partial t}$ is small), then if we subtract the same equation for the previous fixation we get

$$c_t - c_{t-1} = k_t^{\ell_t} - k_{t-1}^{\ell_{t-1}} + (\epsilon_t - \epsilon_{t-1}).$$

By assumption, $\epsilon_t - \epsilon_{t-1} \approx \frac{\partial \epsilon}{\partial t}\Delta t$ is small. Therefore, if ℓ_t is the correct label,

$$(c_t - c_{t-1}) - (k_t^{\ell_t} - k_{t-1}^{\ell_{t-1}}) \approx 0,$$

so this feature value should be smaller for correct values of ℓ_t. Thus, we want to compare how the change of gaze target *between fixations* compares to the change of keypoint locations between fixations.

Formally, define

$$\delta f_t = \overrightarrow{d_p}(f_{t-1}, f_t)$$

to represent the vector change between f_{t-1} and f_t. Then, we can determine the vector change between keypoints i and j during fixations $t-1$ and t respectively by computing

$$\delta k_t^{ij} = \overrightarrow{d_p}(k_{t-1}^i, k_t^j).$$

Finally, we compute the velocity feature v_t^i that measures how well fixation t matches keypoint i as

$$v_t^i = d_v(\delta p_t, \delta k_t^{\ell_{t-1} i}),$$

where ℓ_{t-1} represents the label (keypoint index) assigned to the previous fixation and d_v is a distance function over velocities. Here, we define $\overrightarrow{d_p}$ to be the quaternion representing rotation between two 3-D unit vectors. Then d_v represents geodesic distance between quaternions. Note that this velocity feature term relies on the label assigned to the previous fixation, ℓ_{t-1}. This dependency makes this feature vulnerable to stacking error: if the previous label ℓ_{t-1} is incorrect, the value of this feature is meaningless. Therefore, velocity features work best when paired with other sources of information.

5.5.1 Evaluation

To evaluate this algorithm, we use the simple feature weighting in Eqn. 5.1, using only to the position and velocity weights (as q_1 and q_2 respectively). For the first algorithm to compare, we use the position features only and discard the velocity features; that is, we set $w_1 = 1, w_2 = 0$. Next, since the feature v_t^i depends on the value of ℓ_{t-1}, we derive this previous value in two different ways. In one variant, the *true position-velocity* variant, we use the true value of ℓ_{t-1} when finding the value of ℓ_t. This variant is not representative of how this algorithm would be used in practice, as it requires access to the true label. However, it isolates the classification of each data point so that an incorrect classification does not cause later problems, and thus it more directly represents the ideal power of these features. We also evaluate the *sequential position-velocity* variant, in which we use the labeled value of ℓ_{t-1}. This method represents how this algorithm would be used in practice, but it is more susceptible to error stackup. These three approaches are compared in the following sections.

5.5.2 Validation on Synthetic Data

To determine the usefulness of the velocity features described above, we perform two evaluations. For the first, we generate synthetic eye gaze data so that the gaze error can be more properly controlled. For the second, we hand-labeled the fixations for the HARMONIC dataset (see Sec. 4.1.2) and evaluated the semantic gaze labeling procedure on that data. Throughout, we use $w_1 = 0.8$ for the position features and $w_2 = 0.2$ for the velocity features (determined through cross-validation). In both cases, we find that using velocity features as part of the classification strategy increases the robustness of the classification accuracy to larger differences between gaze and keypoint positions.

We generated a synthetic dataset to ensure that the velocity features were useful in an idealized case. To build this dataset, we first generate trajectories for four keypoints in an image frame by having them follow random Gaussian walks starting from uniformly random starting positions. Then, we generate a randomized gaze signal by first generating a random segmentation of the time period following typical eye gaze dynamics to generate fixations and then assigning each of these fixation to one of the keypoints at random. Finally, the gaze signal during each fixation is determined by adding random noise around the corresponding keypoint position. This procedure is repeated to generate 200 keypoint and gaze trajectory sets of length 33.3 seconds each (1000 samples). To simulate the effect of a constant offset, an additional error of fixed magnitude and direction randomized per trajectory is added to the gaze signal. This simulated gaze signal is then processed according to the semantic gaze labeling procedure described above.

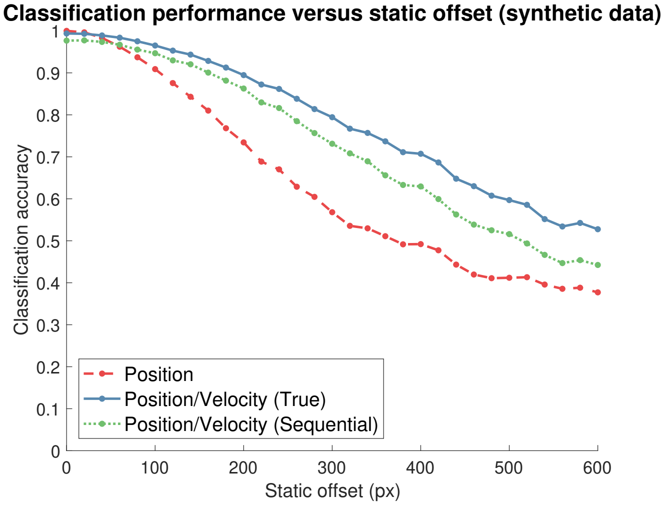


Figure 5.5: Classification accuracy on synthetic dataset.

The results of the algorithm on this synthetic dataset is shown in Fig. 5.5. This figure plots the overall classification accuracy on the synthetic dataset as a function of the magnitude of the offset added to the data. As expected, when the offset is very small, the position and velocity features perform similarly. However, as the magnitude of the offset increases, the velocity-based classification strategies stay more accurate, whereas the position-only strategy decreases in accuracy. Thus, the velocity features succeed at being more robust towards constant offsets. In addition, the true position-velocity strategy outperforms the sequential true position-velocity strategy, but even the sequential strategy gives benefits over the position-only strategy.

5.5.3 Validation on HARMONIC Data Set

Next, we evaluate these algorithms on the HARMONIC dataset (see Sec. 4.1.2). For keypoint locations, we used a tag grid present in the egocentric video frame to compute the egocentric camera extrinsics, which we then smoothed using a Kalman filter. Then, object positions were projected into the egocentric camera using these extrinsics and a prior tag grid location calibration step.

To obtain ground-truth labels, four coders examined each fixation that occurred in the 120 teleoperation-only trials and assigned it a label corresponding to each robot joint and morsel target, or -1 if the fixation was determined to be noise. In addition, all coders coded the same randomly-selected 10% of the trials, and the average pairwise Cohen's kappa (inter-rater reliability) score was 0.645, indicating acceptable agreement. Determining dependence on the error magnitude is more difficult than in synthetic data, as the offset is not controllable. To measure this dependency, we first calculated the angle distance between each fixation and the mean position of its true label, then we binned fixations based on this value with a width of 0.6° and discarded bins with fewer than 20 members.

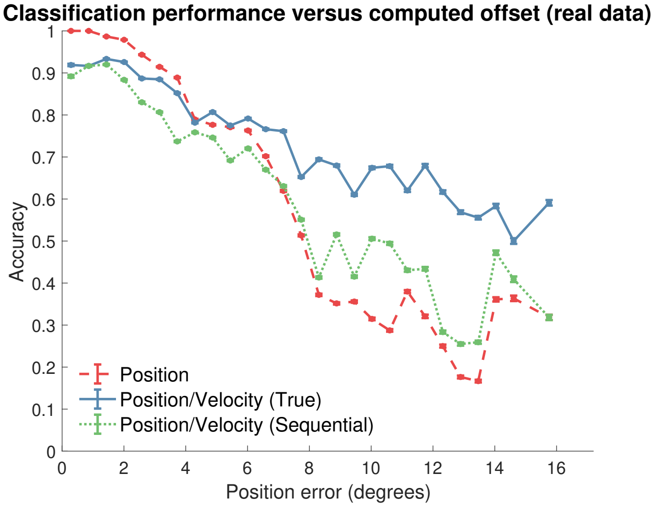


Figure 5.6: Classification accuracy on the HARMONIC dataset.

Fig. 5.6 shows the accuracy of our classification strategies applied on the HARMONIC dataset as a function of offset bin. As in the synthetic data, all methods have good accuracy for small errors, though the position-only method slightly outperforms

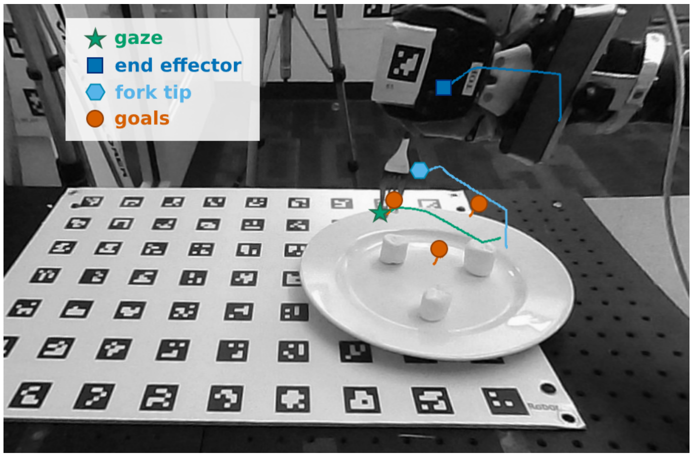


Figure 5.7: Still of egocentric video during the food grasping task. While the gaze (green star) is closest to a goal keypoint (orange circle), the motion of the gaze during the fixation, given by the paths from each symbol, more closely matches the motion of the fork tip.

the others. As the offset increases, the position-only and sequential methods drop off, whereas the true velocity method maintains its performance. Thus, velocity features are indeed useful for improving the accuracy of semantic labeling.

5.6 Pursuit Features

In addition to the position and velocity features presented here, we also hypothesize that another feature, the *simultaneous motion* of the gaze and the objects during the fixation, will help in semantic gaze labeling. While people are fixating on a particular object, their eyes move to track its motion during smooth pursuit. Though this feature cannot distinguish between stationary objects, it can distinguish gaze among objects that are moving in different ways, even if they are positioned close together. This approach has been successfully used to develop calibration-free screen-based gaze target selection [120–122] with motion control.

A particularly tempting use case in our scenario is to use pursuit features to distinguish between gaze at the end-effector of the robot from gaze at a goal, particularly when the robot is close to the goal. Fig. 5.7 gives an example of a fixation for which this ability would be useful. While the user's gaze is nominally closest to a goal object, its motion much more clearly matches the motion of the end-effector. This feature can remove spurious glances towards incorrect goals and may enhance the goal predictive quality of gaze. We can further validate the usefulness of pursuit features by examining the pairwise correlation between each pair of keypoints averaged over all fixations, shown in Fig. 5.8. This plot shows the theoretical usefulness of pursuit features. While it cannot distinguish between goals on its own, or between goals and stationary parts of the robot, it should separate moving robot joints from those stationary objects and from each other.

However, the data captured here is difficult to use with pursuit features due to noise in both the 3D gaze signal and the localization of the tracker relative to the scene. In addition, the lack of ground-truth labels for gaze confounds an evaluation, as manual coders used these features for labeling. Further work includes capturing a more controlled gaze data set to evaluate all of these features together.

5.7 Conclusion

The pipeline given in this chapter describes how to take the raw gaze signal as obtained by a gaze sensor and process it into a timed sequence of foveated scene objects drawn from a prespecified set so that it incorporates scene context. This form enables us to take advantage of the patterns in gaze during teleoperation discussed previously and develop algorithms for inferring people's goals in a teleoperated manipulation task from their gaze behavior.

While this work relies on some strong assumptions about the simplicity of the scene, we anticipate that this approach will be applicable in additional situations. Our data confirms that people do generally look at pre-defined, task-relevant objects: about 63% of the fixations in our data set could be labeled as specific objects in the scene, and many of the others were errors in sensing. In addition, many manipulation setups, at least in the lab, will already have access to information about the manipulation environment. This assumption does, however, limit the usefulness of this method in fast, rapidly-changing environments like driving, during which people rely on peripheral vision, and less task-driven viewing tasks like free-viewing or visual search [114]. Future work can focus on extending this object-driven contextual model with integrated sensing or statistical attention models for more general use.

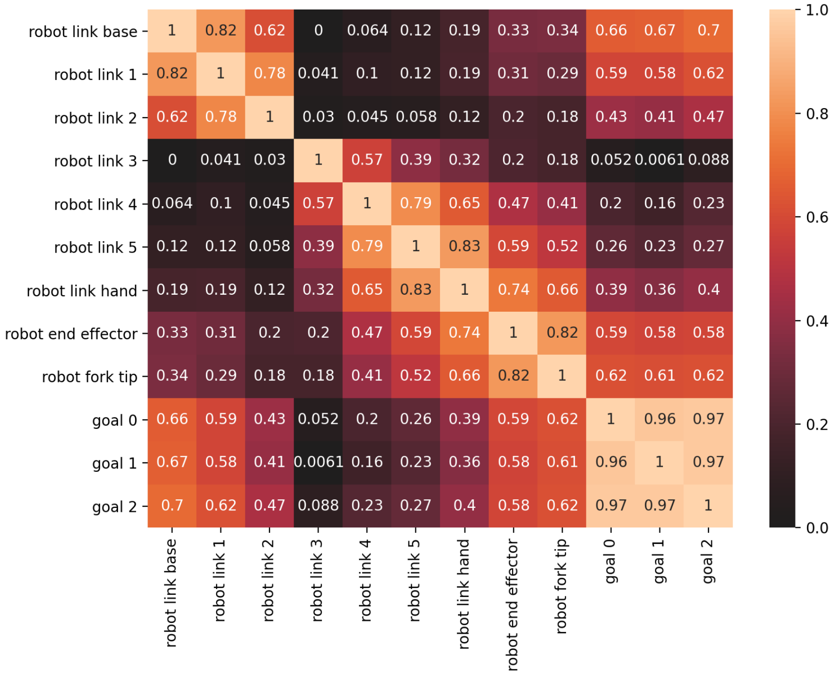


Figure 5.8: Pairwise correlation of the motion of keypoints during each fixation, averaged over all fixations. Lighter blocks, such as the three goal points, are nearly perfectly correlated and thus are difficult to distinguish with pursuit features. In contrast, joints of the robot generally move more during the task, so they are more likely to be distinguishable from each other and from the stationary objects.

Chapter 6

Gaze for Goal Prediction

Now that we have developed a representation of gaze that includes contextual information, we can use that signal to predict the user's goal. Our earlier work on natural gaze behavior during manipulation, described in Chap. 4, shows that goal-directed glances do occur with consistent patterns, but they occur relatively rarely, and gaze is mostly directed at the robot's end-effector. In the HARMONIC data set [84], 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, we explore more sophisticated algorithms for goal prediction that can better interpret the signal.

In this section, we develop two algorithms for goal prediction from gaze and validate them on the HARMONIC data set [84]. First, we present an *aggregate* gaze model which uses only the counts of fixations labeled as goals and discards sequence information (Sec. 6.1). This aggregate model represents standard 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 (Sec. 6.2). This method generates earlier and more confident predictions than the aggregate method, but it performs worse when no distinguishing gaze is available. Using this method, gaze can provide early, global goal prediction independent of the state of the task, but the lack of a consistent signal makes it unreliable as a sole source of information.

6.1 Aggregate Gaze Method

For gaze-based goal prediction, we take as input the sequences of labeled fixations derived in Chap. 5 above. The gaze data in each trial i consists of a sequence of fixations $f^i = (f^i_0, \cdots, f^i_T)$, each of which contains a start time, duration, and label $f = (s, d, \ell)$, as input and a reported goal g_i as output. Existing work [13, 70, 107, 118, 124] 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

$$p^{\text{agg}}(g_k|(f_0, \cdots, f_T)) \propto \exp \sum_{t=0}^{T} \delta(\ell_t = \ell_{g_k}), \tag{6.1}$$

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.

6.2 Sequential Gaze Prediction 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.

6.2.1 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 $f = (s_t, d_t, \ell_t)$, generate the new sequence f' as

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

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. (See Fig. 5.3 for label identification.)

6.2.2 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 keypoint 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

$$s_k(\ell_0, \cdots, \ell_T) = \log p(\ell_0, \cdots, \ell_T; \mathrm{HMM}_k). \tag{6.2}$$

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

$$p(g_k | \ell_0, \cdots, \ell_T) = \frac{\exp s_k(\ell_0, \cdots, \ell_T)}{\sum_{k'} \exp s_{k'}(\ell_0, \cdots, \ell_T)}. \tag{6.3}$$

All HMM operations were performed using the `hmmlearn` package[1]. We set the number of hidden states $n = 3$, quantization parameter $\Delta t = 250\,\mathrm{ms}$, and cutoff value $N_{\max} = 3$ 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-stage tasks. Moreover, it does not require that the goal objects themselves be identified among the labels in advance.

6.3 Comparing Sequential and Aggregate Prediction

6.3.1 Data for Evaluation

To evaluate the goal prediction algorithms, we used the HARMONIC data set [84] we previously collected (see Sec. 4.1.2), which contains eye gaze and joystick input from participants teleoperating a robot arm using a joystick to spear one of three marshmallows on a plate. For this work, we include only trials in the direct teleoperation condition in which the user completed the task successfully. This filtering left 64 trials, with an average of 60 fixations per trial.

Predictions by the joystick method (described in Sec. 3.1.1) were provided in the data set. Predictions from the aggregate model was computed directly. For the

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

	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

Table 6.1: Algorithm accuracy metrics.

sequential predictions, the trials were divided into five sets with goal frequencies balanced, and the goal prediction for each trial was computed using a model trained on the four other sets that did not include the trial.

6.3.2 Metrics

We compare the algorithms on several metrics.

Overall accuracy 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 6.1.

Confidence We compute the set of final probabilities assigned to the *correct* goal at the end of each trial i, i.e., $\{\forall i : p(g_i)|f^i)\}$. We report the *mean probability*, the mean of this set, as a measure of the confidence of that prediction. Since these probabilities are highly non-Gaussian (see Sec. 6.3.3), we also report the *median probability.* These results also appear in Table 6.1.

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 6.1.

Prediction over time We consider the sequence of prediction confidence over the course of each trial. Given a subset of the data $(0, \cdots, t \leq T)$, we compute the probability of the correct goal derived from that subset $(p(g_i|(f_0^i, \cdots, f_t^i))$. This probability is a function of time T, which is normalized to the length of the trial.

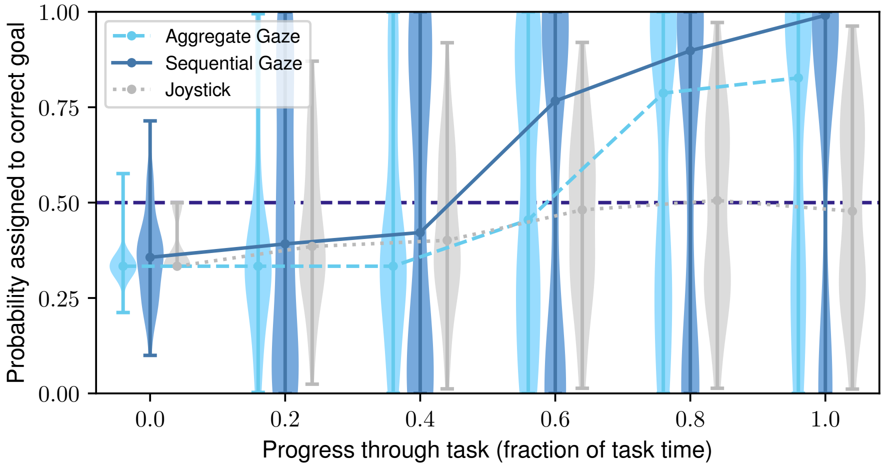


Figure 6.1: Distributions of probability assigned to the correct goal by each algorithm during the evolution of each trial. The first data point (at $t = 0$) uses the first fixation, so the initial probability is not uniform. Lines connect probability medians. When the probability assigned to the correct goal is above 0.5 (denoted by the horizontal dashed line), the classifier is guaranteed to be correct at that time.

Fig. 6.1 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 goal at that time. Mass at larger y values indicates more confident correct predictions, and data at smaller x values represent prediction confidence earlier in the trial.

6.3.3 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. 6.1 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. 6.1 shows that both gaze models demonstrate strong bimodal behavior. This finding evokes the result found in Huang et al. [50], 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, the system detected 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 (i.e., 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 6.1 better represent each algorithm's quality than the means do.

6.3.4 Understanding the Sequential Model

To understand the benefits of the sequential model, we examine a single trial in detail. Fig. 6.2 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. 6.3 shows how this model evaluates a single trial. Eye gaze is mostly directed toward the tool, particularly at the start of the trial (Fig. 6.3, middle), and the HMM largely stays in `S1` in response (Fig. 6.3, 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. 6.3, 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.

6.4 Characteristics of Gaze-based Goal Inference

With a goal prediction pipeline established, we now turn to understanding how this prediction behaves in comparison to the control input-based prediction discussed

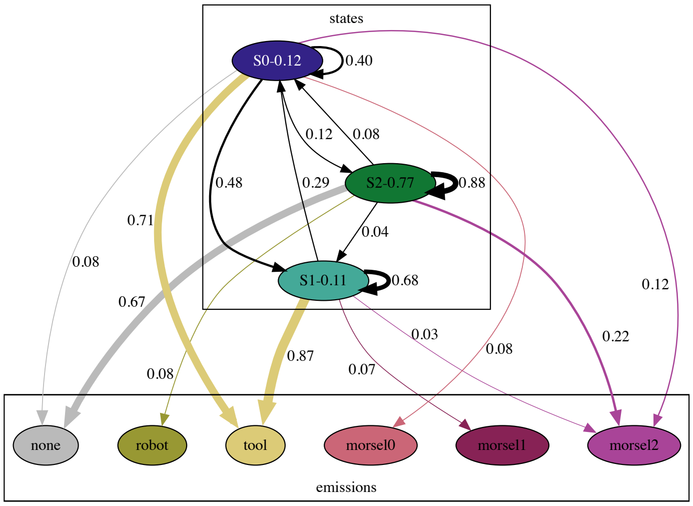


Figure 6.2: Graph representation of a learned hidden Markov model for morsel 2. State labels include their prior probabilities, edges between states represent transition probabilities, and edges to emissions represent emission probabilities. Edges with $p < 0.02$ were omitted for clarity.

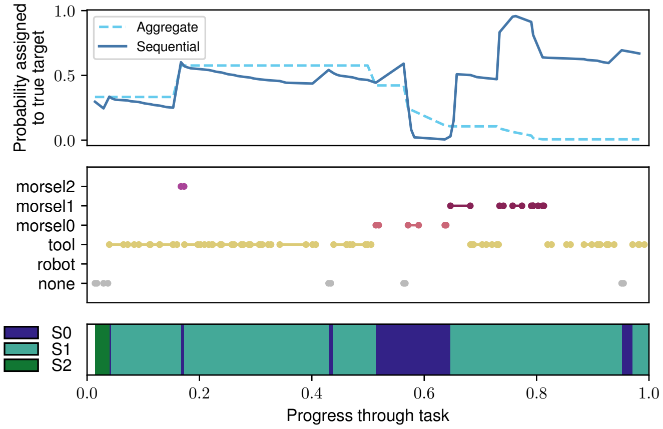


Figure 6.3: Sample trial comparing aggregate and sequential prediction performance. The top plot shows the output probability assigned to morsel 2 (the true target) during the trial. The second plot shows the gaze labels supplied to each algorithm. The bottom bar shows the hidden state predicted by the 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.

in Chap. 3. We find that since the gaze is highly bimodal, it is capable of giving accurate predictions earlier than the control input does. However, it is significantly less reliable: it does not reliably give this earlier prediction, and it can often fail to provide any goal information at all. This behavior suggests it is best used as a signal of opportunity and combined with a more reliable prediction source like control input rather than being used on its own.

6.4.1 Gaze vs. Joystick: Forecasting Horizon

First, we 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 locate them [12], whereas joystick input is similar when all goals are in the same direction from the robot's current pose. Particularly for the object spearing task, the robot trajectory is similar for the first half of the trial (as participants reorient the robot 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 the start of the trial. 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. While restrictive, these exclusion criteria remove the effect of the bias of the data before receiving any information. Results appear in Fig. 6.4.

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

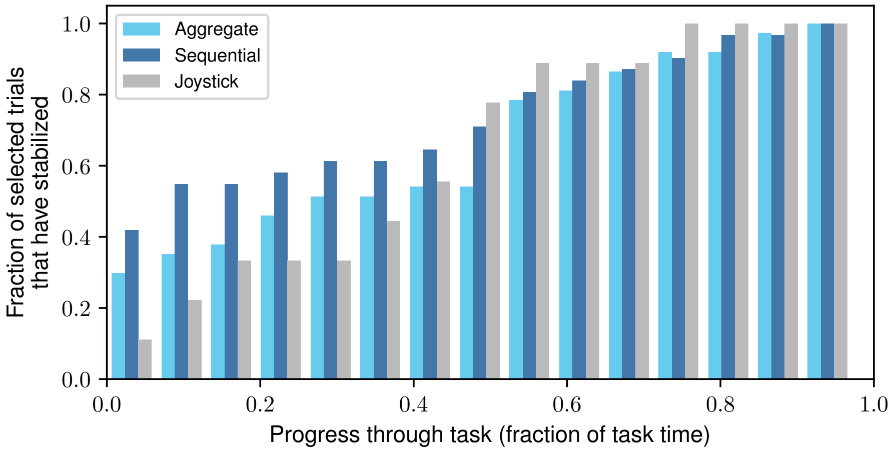


Figure 6.4: Fraction of selected trials that have stabilized on the correct prediction by progress through the task. Selected trials include those where the algorithm is incorrect in its first guess without any information, but it is correct at the end. The exclusions resulted in $n_{\text{aggregate}} = 37, n_{\text{sequential}} = 31, n_{\text{joystick}} = 9$. 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.

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. 6.4 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.

6.4.2 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 makes this combination especially appealing. The signals can combine their predictions together using Bayesian combination [53], 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 [57]. 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.

6.5 Conclusion

In this chapter, we present a novel algorithm for processing the sequences of labeled fixations developed earlier into predictions of the user's goal in the teleoperated manipulation setting. We find that gaze generally gives a relatively early goal prediction, especially when compared with methods using control input. In addition, using a sequential method to process the gaze encodes the signal context and improves performance over baseline aggregate methods. However, gaze-based goal prediction is unreliable, and this unreliability is rooted in the signal itself: people can go whole trials without performing any distinguishing gaze behavior. Therefore, we predict that the signal is best used in combination with another goal prediction source like the user's control input. We develop and evaluate the complementarity of these signals in the next chapter.

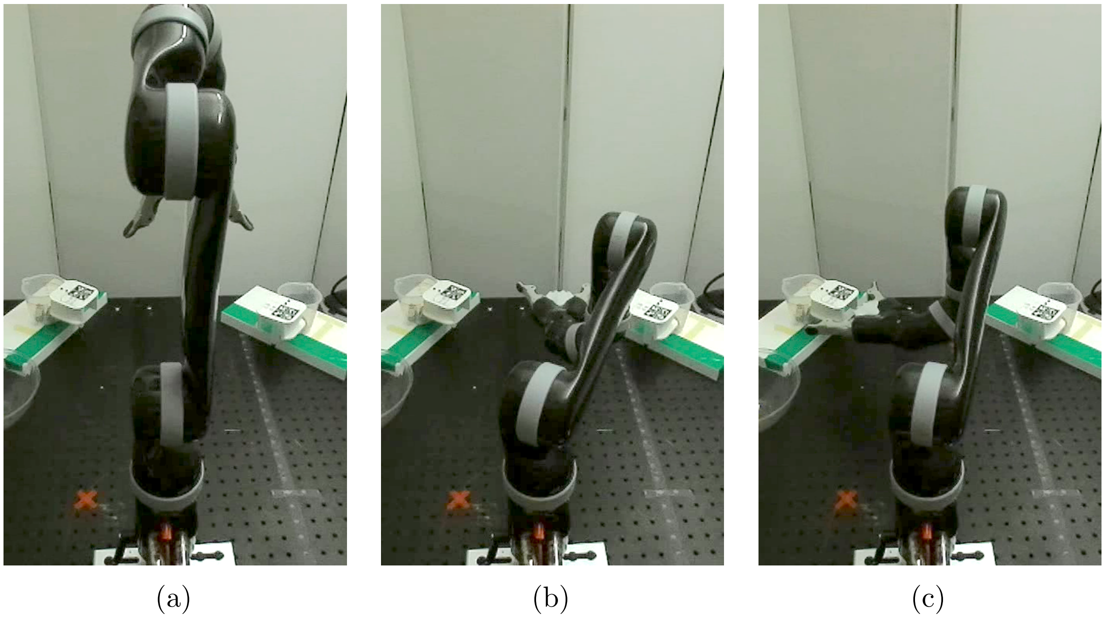
Chapter 7

Control Input and Gaze Are Complementary

The previous discussion of goal inference using either control input or gaze enables us to consider the role of each signal in the overall assistance process. In Chap. 3, we see that control input is sufficient for simple tasks, but it may not give sufficient information for optimal assistance when the user's control is constrained. In contrast, natural gaze can give goal information to enable assistance at any point during the task without the input restriction; however, it cannot do so reliably. We see then that these signals complement each other: control input provides a reliable fallback, while gaze enhances performance when it is available.

To demonstrate this complementarity, we implement an assistive teleoperation system that incorporates goal prediction using both the user's control input and their eye gaze behavior. We use this system to evaluate each prediction source in a real-world, COVID-safe user study. In the study, participants teleoperated a robot manipulator using modal control to pick up one of two cups while our system provided assistance. The scenario was designed so the user could not act optimally, so their control input was unlikely to yield optimal assistance. During each trial, the assistance relied on goal prediction based on their joystick input, their gaze behavior, or both.

We find that for this experimental scenario, assistance based on joystick input alone is delayed relative to using both joystick and gaze, but only when the gaze



(a) (b) (c)

Figure 7.1: Evolution of cup grasping task. First (a), users generally reorient the robot so that the gripper is coplanar with the grasp points of the cups (b). Next, users translate and rotate the robot to align with their specific goal (c). If the robot knows the user's goal in stage (a), it can provide goal-specific motion in x and roll.

prediction arrives sufficiently early. In the cases with early gaze predictions, trials finished more quickly and users supplied less control input. Specifically, early gaze leads to earlier assistance exactly on the axes for which the goal positions differ, and the assistance is the same otherwise, matching our theoretical analysis. However, we confirm that gaze-based predictions are inherently less reliable, as many trials never gave sufficient information for accurate goal prediction, and feedback loops led to arbitrarily poor performance in some cases. This work explores a fundamental limitation of input-based goal prediction for assistance and shows that eye gaze provides the global information required for systems to provide as much assistance as possible.

7.1 Task Development

As discussed in Sec. 3.4, we expect that only some tasks benefit from early prediction. We design a task such that at some state typically reached, the assistance required is *different* but the user's command is *indistinguishable*. The task is a 6-dimensional, 3-mode analogue for the example in Fig. 3.1, in which the user can control only one axis at a time.

We start from an object spearing task used in our prior assisted manipulation work [12, 54, 84], but modify it into a cup grasping task. The robot starts at a neutral position, and the user must teleoperate it with modal control to grasp one of two cups. From prior work with this task, we observe that users generally start by moving the robot forward ($+y$) to close the distance to all goals and reorienting the end-effector to face forward (pitch) before performing goal-specific motion. Therefore, we change the initial robot position to start midway between the goals in the x axis, so initial left-right motion is different based on the goal. We add an additional, goal-specific constraint along the roll axis by orienting the cup handles differently; to grasp a cup, the user must rotate the end-effector to align with its handle, another motion that depends on the goal. The stages of the new task appear in Fig. 7.1. While the user is moving the robot in y and pitch, the system does not get any information about their goal from their control input; early, gaze-based goal prediction enables assistance in x and roll before the user begins providing goal-specific input.

7.2 Goal Prediction

7.2.1 Control Input

To provide real-time goal prediction from the user's control input, we follow Sec. 3.1.1 in assuming an observation model

$$p(u|x, g) \propto \exp Q_g(x, u),$$

with u representing the user's joystick command signal, and solving for

$$\begin{aligned} p(g|\text{joystick}) &= p(g|x, u) \\ &= \frac{\exp Q_g(x, u)}{\sum_{g' \in G} \exp Q_{g'}(x, u)} \end{aligned} \tag{7.1}$$

using Bayes' rule.

7.2.2 Sequential Goal Prediction

We start by processing raw eye gaze behavior into a sequence of semantic labels, described in Chap. 5. The sequence is next passed into a hidden Markov model for processing, as described in Sec. 6.2. To apply the method for a new setting with two goals instead of three, we modify the sequence processing algorithm and retrain the HMM on the HARMONIC data set [84] using the modified algorithm.

To make the algorithm agnostic to the number of goals, we change the sequence labels so that each sequence is formatted for a one-vs-rest classifier, and we use a single HMM to match all goal candidates by instead varying the sequence representation depending on the goal. First, the labels are remapped to higher-level groups: $\{\texttt{none}, \texttt{robot}, \texttt{tool}\} \cup G$, where G represents the set of all possible goals. We then remap the label ℓ again for each goal candidate $g \in G$ according to a function f_g:

$$f_g(\ell) = \begin{cases} \texttt{my\_goal}, & \ell = g \\ \texttt{other\_goal}, & \ell \in G \setminus \{g\} \\ \ell, & \text{else.} \end{cases}$$

Then, the sequences encode the goal choice directly. We can now use a single HMM, which is trained on sequences transformed by $f_{g_{\text{true}}}$, as a one-vs-rest classifier. We obtain an observation probability of each sequence given a goal g by applying f_g to the sequence and computing the HMM probability, so

$$p(\ell_0, \cdots, \ell_n | g) = p_{\text{HMM}}(f_g(\ell_0), \cdots, f_g(\ell_n)).$$

To obtain goal probabilities, we marginalize over all goal candidates assuming a uniform prior:

$$\begin{aligned} p(g|\text{gaze}) &= p(g|\ell_0, \cdots, \ell_n) \\ &= \frac{p_{\text{HMM}}(f_g(\ell_0), \cdots, f_g(\ell_n))}{\sum_{g' \in G} p_{\text{HMM}}(f_{g'}(\ell_0), \cdots, f_{g'}(\ell_n))}. \end{aligned} \tag{7.2}$$

We train the model on the semantic gaze label sequences from all trials by first determining the user's intended goal in that trial g_{true}, then by remapping the sequence to $(f_{g_{\text{true}}}(\ell_0), \cdots, f_{g_{\text{true}}}(\ell_n))$ according to that goal. When applying this new method in the same analysis as in Sec. 6.3.3, we obtain comparable results to the original sequential modeling approach: 57.8% accuracy (vs. 33% chance), 63.2% mean probability assigned to the correct goal at the end of the trial, and 92.0% median probability assigned to the correct goal at the end of the trial. With this alteration, we can apply the same HMM trained on the previous data set with three goals to this new task with only two.

7.2.3 Combined Prediction

To combine the joystick and gaze predictions, we follow Jain and Argall [53] and assume that the prediction based on each source is independent conditioned on the goal. Assuming a uniform prior, we compute

$$p(g|\text{joystick}, \text{gaze}) = \frac{p(g|\text{joystick})p(g|\text{gaze})}{\sum_{g' \in G} p(g'|\text{joystick})p(g'|\text{gaze})}.$$

$p(g|\text{joystick})$ is given in Eqn. 7.1, and $p(g|\text{gaze})$ is given in Eqn. 7.2. Combining the probabilities ensures that the assistance command is always providing the maximum effort based on the system knowledge, so conflicting information between the signals leads to full movement to a neutral position.

7.3 User Study

We conducted a user study in which participants performed this cup grasping task. The study was performed within subjects and fully counterbalanced, with three conditions {*joystick*, *gaze*, *merged*} corresponding to which prediction strategy was used for the assistance.

Because of the COVID-19 pandemic, the user study was performed in a hybrid remote-local fashion. The robot and a stationary camera were set up in the lab. Each participant received a laptop, eye gaze sensor (Tobii Eye Tracker 4C, a screen-based tracker), joystick, and computer paraphernalia at their home. Participants assembled the equipment with remote experimenter supervision. They then connected the laptop to the lab via OpenVPN. Using ROS and a custom interface, the laptop displayed a

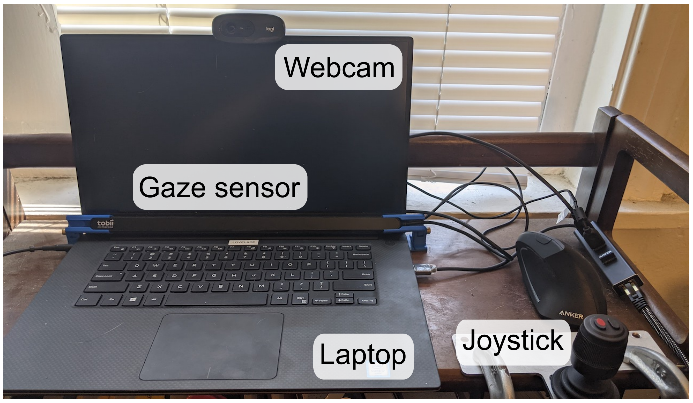


Figure 7.2: Study setup that participants prepared at home.

live video feed of the robot and transmitted the user's joystick command and gaze data. In this way, participants controlled the robot without indoor contact.

7.3.1 Procedure

After filling out a consent form and reporting demographic information, participants received an explanation of the task while observing an autonomous grasp by the robot. Next, participants were instructed on how to control the robot and practiced for approximately five minutes. During this time, camera parameters were adjusted to compensate for latency; the resulting delay was typically $50 - 70$ ms. In addition, the fixation segmentation algorithm [113] was trained on their eye gaze data. Next, the participant performed four trials with no assistance. Finally, the participant performed four trials each of the three conditions (*joystick*, *gaze*, and *merged*), fully counterbalanced. To accustom participants to the assistance, they performed an additional trial in their first assisted condition which was omitted from analysis. Participants filled out a questionnaire after each condition and another questionnaire at the end.

7.3.2 Participants

The study was conducted with 12 participants (6 male, 6 female, 0 other). Ages of participants were 6 aged 18-24, 4 aged 25-30, and 2 aged 30-40. For familiarity with operating robots, 2 reported lots of familiarity, 6 reported some familiarity, and 4 reported no familiarity. Participants received $20 compensation for their participation, which took approximately 1.5-2 hours including setup and teardown. The study was approved by the university IRB office. Since the study required lending materials to participants, recruitment was limited to university posting and word of mouth.

7.3.3 Evaluation Metrics

Algorithmic metrics Within each trial, we compute the *prediction strength*, which is the probability assigned to the correct goal during the course of the trial.

Trial metrics For each trial, we compute the *trial duration* and the *active fraction*. *Trial duration* refers to how long it took the user to complete the task, and *active fraction* refers to what fraction of the trial the joystick command was non-zero; i.e., the user was explicitly providing input. Shorter trials and trials with less joystick input were considered better.

Subjective metrics After completing trials for each condition, participants answered questions on a seven-point Likert scale, following Javdani et al. [55] (emphasis added; emphasized words act as references for reporting results):

- I felt in **control** while using this system.
- I was able to accomplish the tasks **quickly** while using this system.
- The robot did what I **wanted** while using this system.
- My goals were perceived **accurately** by this system.
- If I were going to teleoperate a robotic arm, I would **like** to use this system.

Participants also answered two open-response questions:

- Did you use any particular strategies while operating the robot?
- What are your comments about this system?

7.3.4 Hypotheses

H1 *Eye gaze can predict the user's goal earlier than joystick input can.* This hypothesis follows the observation in Chap. 6 that gaze often gives an earlier prediction horizon, which underlies our model for task improvement. We do not require (or

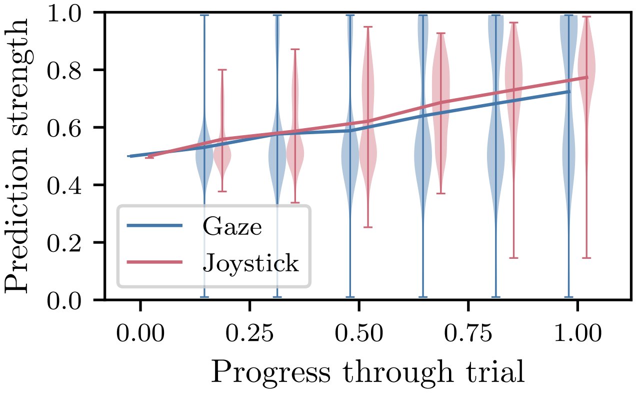


Figure 7.3: Distributions of prediction strength given by gaze and joystick methods over all trials, normalized by trial duration. While the median prediction strength over time is similar between the two, the distributions are different. The joystick prediction for each trial smoothly increases over time. The gaze prediction, however, is bimodal, and the median gaze prediction strength increases as more trials transition from $p \approx 0.5$ to $p \approx 1$ at different times.

expect) the gaze prediction to *always* precede joystick prediction; rather, we need it to do so sufficiently often to evaluate how the assistance changes.

H2 *When the assistance system receives a prediction from gaze before a distinguishable state, trial metrics will improve and goal-specific assistance will appear earlier.* By considering only trials in which gaze yielded a prediction and analyzing when the prediction was received, we evaluate the model of when joystick-based assistance is improved.

7.4 Results

7.4.1 Gaze Gives Early Predictions

Our model for gaze improving assistance requires that it gives earlier predictions than the joystick input does. Figure 7.3 shows the prediction strength of gaze and joystick over the course of each trial. While gaze and joystick prediction medians behave similarly, they follow different distributions. Gaze-based prediction is bimodal, which

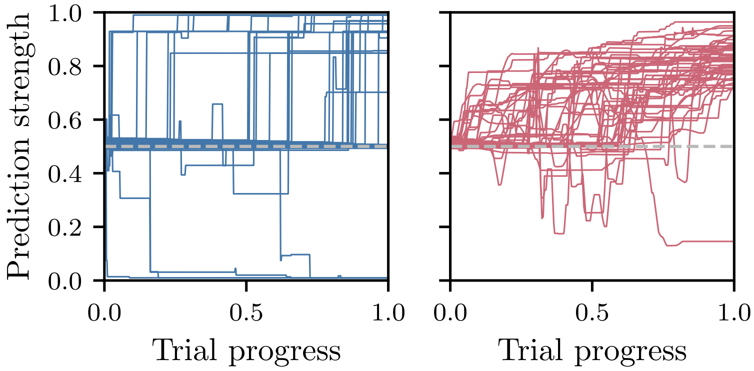


Figure 7.4: Prediction strength for each condition over all trials, normalized by trial duration. The gaze predictions (left) generally transition sharply between $p \approx 0.5$ (no prediction) and $p \approx 1$ (confident, accurate prediction). The joystick predictions (right) smoothly increase over time.

agrees with results from Chap. 6. While the joystick prediction strength steadily increases throughout each trial, the gaze prediction strength increases by shifting probability mass from $p = 0.5$ to $p \approx 1$. Fig. 7.4 shows traces of all runs in the gaze and joystick conditions. The gaze prediction generally starts at 0.5 and jumps to $p \approx 1$ at some point. This jump occurs at the first identified fixation on one of the goals. While the effect is not consistent, we do find that gaze is capable of providing earlier predictions than the joystick can, so *H1* is supported.

7.4.2 Early Gaze Improves Trial Performance

Next, we assess how early goal prediction from gaze affects trial performance. First, we consider only trials in which the gaze gave a prediction at all. We divide this set into those that gave an *early* prediction and those that gave a *late* prediction. Early trials gave a stable, consistent prediction from a threshold time T_c to the end of the task. Specifically, we require:

$$\forall t, t \geq T_c : |p(g|\text{data}_0, \cdots, \text{data}_t) - 0.5| \geq 0.1.$$

Since there are only two goals, either goal can be used for this calculation. These criteria mirror the ones given in Sec. 3.2: the gaze must give a prediction when the

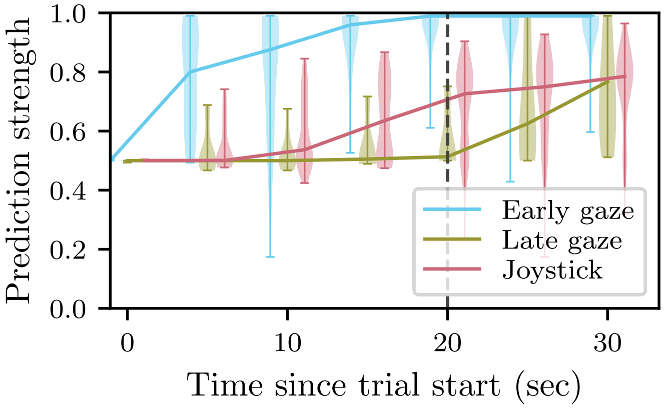


Figure 7.5: Distributions of prediction strength over all trials for early gaze, late gaze, and joystick. The x-axis here is not normalized by trial time. The dashed line at 20s indicates the cutoff time T_c for early gaze prediction.

optimal motion is different for each goal, but the user's command is still indistinguishable. To choose this threshold, we observe that the goal-independent assistance generally finishes about $T_c = 20$ seconds into the task. The remaining trials that gave a prediction were labeled late. Of the 47 trials in the merged condition, 21 (45%) were early and 9 (19%) were late. (The remaining 17 (37%) did not give a prediction.) We compare the early and late gaze prediction strength with the joystick prediction strength in Fig. 7.5 to confirm that this threshold generally aligns with when the joystick gives a goal prediction.

We now consider how the timing of the prediction affects trial metrics. Fig. 7.6 show task metrics for early and late trials compared to trials in the joystick condition. A one-way ANOVA evaluated on the log of the data shows significance for both trial duration ($F(2,76) = 6.78, p < 0.002$) and active fraction ($F(2,76) = 4.32, p < 0.013$). Post-hoc analysis with the Tukey HSD test shows that early gaze has shorter trials than both late gaze ($p < 0.006, 95\%$ CI $= [0.14, 0.93]$) and joystick alone ($p < 0.008, 95\%$ CI $= [0.077, 0.60]$). In addition, early gaze takes less joystick effort than does joystick alone ($p < 0.02, 95\%$ CI $= [0.077, 0.93]$). The benefit of early gaze specifically relative to both late gaze and joystick show that *H2* is supported.

We also consider the magnitude of the assistance over time, shown in Fig. 7.7. As described in Sec. 7.1, the task is designed such that the optimal motion is different depending on the user's goal along the x axis throughout the task, but it is identical

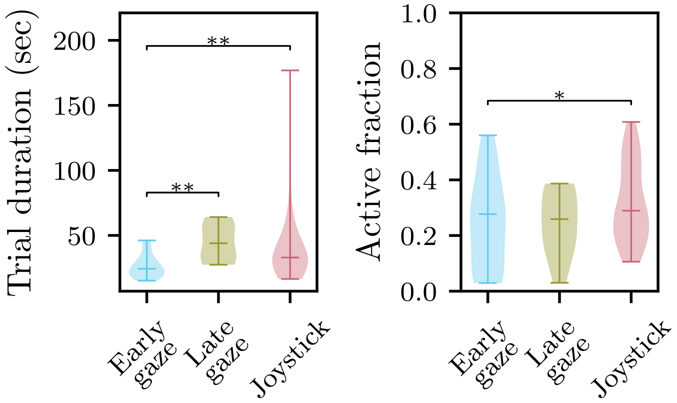


Figure 7.6: Trial metrics for early gaze, late gaze, and joystick. * indicates significance at $p < 0.05$ and ** at $p < 0.01$. Early gaze trials are shorter than both other conditions and require less input than the joystick.

along the y axis. We see that the early gaze allows earlier assistance in x than late gaze or joystick do, since the latter conditions can only assist once the user input becomes distinguishing. In contrast, the assistance along the y axis is similar for all cases; receiving a goal prediction does not change the autonomous action. This observation aligns with the reasoning given in Sec. 3.2.

7.4.3 Gaze Alone Performs Poorly

To further explore the usefulness of natural gaze for goal prediction, we measure how effective the gaze signal is for assistance on its own. We report overall trial metrics in Fig. 7.8 for each condition. A one-way ANOVA evaluated on the log of the data shows significance only for trial duration ($F(2, 142) = 12.7, p < 10^{-5}$). Post-hoc analysis using the Tukey HSD test on the log shows that the gaze condition alone takes longer than both the merged condition ($p = 0.001, 95\%$ CI $= [-0.71, -0.25]$) and the joystick condition ($p < 0.002, 95\%$ CI $= [-0.59, -0.12]$). In addition, in subjective responses (visualized in Fig. 7.9 and analysis reported in Tab. 7.1), people generally rated the gaze-alone condition worse than either of the others.

Gaze suffers because goal-directed gaze does not occur in every trial. Familiarity with the scene from previous trials, adjusting goal-independent factors such as robot rotation, and peripheral vision all contribute to the unreliability of distinguishing gaze

Question	$\chi^2(2)$	p	Conditions	U	Corrected p
Control	6.3	0.042*	gaze-merged	45.5	n.s.
			gaze-joystick	29	0.013*
			merged-joystick	65.5	n.s.
Quickly	5.9	n.s. (0.054)			
Wanted	8.5	0.014*	gaze-merged	35.5	n.s. (0.050)
			gaze-joystick	25.5	0.0094**
			merged-joystick	65.5	n.s.
Accurately	8.0	0.019*	gaze-merged	28.5	0.014*
			gaze-joystick	35.5	0.046*
			merged-joystick	67.0	n.s.
Like	7.1	0.029*	gaze-merged	31.5	0.026*
			gaze-joystick	37	n.s. (0.058)
			merged-joystick	63.5	n.s.

Table 7.1: Statistical analysis of participant answers to questions. Significance testing was performed first with a Kruskal-Wallis test for overall significance, and post-hoc analysis was done using the Mann-Whitney U test ($n_1 = n_2 = 12$) with Bonferroni correction for multiple comparisons. * indicates significance at $p < 0.05$, ** at $p < 0.01$. Marginally significant values ($p < 0.1$) are shown in parentheses. n.s. means “not significant” at $\alpha = 0.05$.

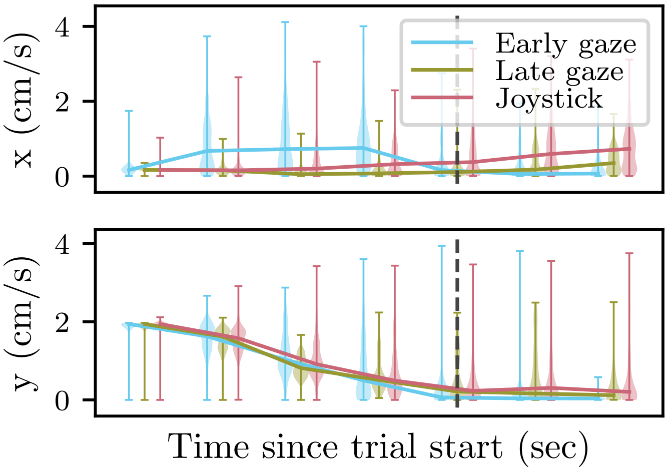


Figure 7.7: Robot assistance over time in x (top) and y (bottom). Early gaze enables assistance in x before the $T_c = 20$ sec. cutoff, while late gaze and joystick do not assist in this axis until after T_c. In y, the assistance is the same for all conditions, as no goal information is required.

behavior [6, 12]. In fact, 33/95 (35%) of trials exhibited no goal-directed fixations at all. In these cases, assistance was provided for the first part of the trial (when it is identical for each goal), but subsequent motion is unassisted.

Incorrect predictions were even worse than no predictions at all. If the gaze prediction selects the incorrect goal early in the trial, it was nearly impossible for users to correct it. For example, if the user glances at one goal while trying to navigate to the other (due to, e.g., wandering attention or an error in gaze detection), the gaze-based assistance moves the robot directly to that goal. When the user attempts to maneuver the robot arm away from that goal, they look at the robot end-effector and at the incorrect goal to avoid collision, reinforcing the incorrect prediction. This self-reinforcing behavior was nearly impossible for participants to correct. Participants described this condition as “adversarial” and “like trying to hold onto a slimy eel while it attempts to wriggle away,” and even changed their goals to “accept its whimsy ways.” This behavior is analogous to the adversarial conditions in Stolzenwald and Mayol-Cuevas [111] and Newman et al. [82]. While this issue can arise when a system using control input approaches collinear goals [36], when gaze is the only prediction source, even maximum input to the other goal does not fix the

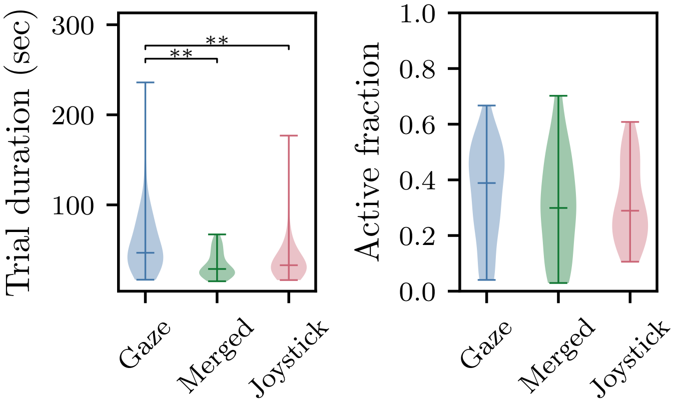


Figure 7.8: Trial metrics per condition. * indicate significance at $p < 0.05$, and ** at $p < 0.01$. Gaze takes significantly longer than either condition, and there is no distinction within active fraction.

problem. The simplicity of the gaze model, and the focus on object identification without an understanding of object role, illustrates the fragility of this method for goal prediction in even a simple task.

7.4.4 Adding Gaze Does Not Provide Overall Improvement

While adding gaze improves on tasks metrics when the gaze provides an early prediction, we consider the overall impact of adding gaze. The merged condition, which uses both gaze and joystick predictions, does not show improvement over using joystick alone in trial metrics (Fig. 7.8) or subjective metrics (see supplementary material). While 45% of merged trials contained early gaze and thus better performance, the effect may not have been sufficiently large or occur frequently enough to make an overall difference. In addition, the downsides of poor gaze may have led to frustrating behavior that counteracted the benefit gained from early gaze.

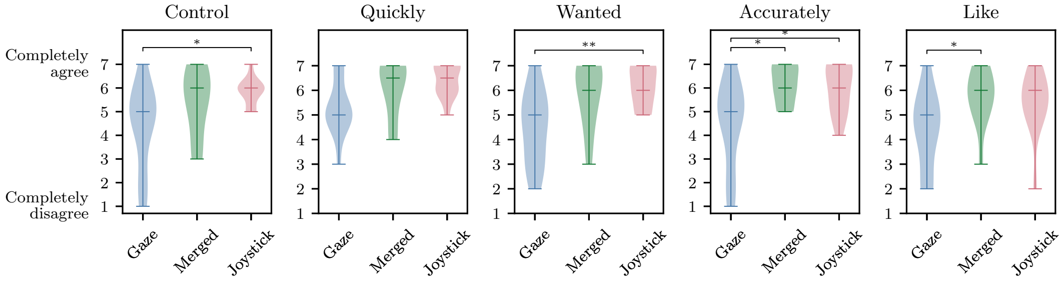


Figure 7.9: Participant answers to the post-condition Likert questions. Significance testing per question was performed with a Kruskal-Wallis test with $\alpha = 0.05$, and when significance was achieved, a Mann-Whitney U test was used for post-hoc evaluation. Conditions annotated with * indicate significance at $p < 0.05$, and ** at $p < 0.01$. Overall, participants disliked the gaze condition, while showing no clear preference between the other two.

7.5 Discussion

7.5.1 Natural vs. Intentional Gaze

This study proposed to evaluate *natural* gaze for goal prediction. Unlike during passive data collection, the system responded actively to participants' gaze behavior. Therefore, participants may have noticed that the system responded to their gaze and chosen to use their gaze as an explicit input. To determine if the gaze was indeed natural, participants were asked after each condition if they used any particular strategies to control the robot. In addition, in the final questionnaire, they were asked to select trials in which the robot was responsive to their gaze. Of the 12 participants, 8 reported that they did not notice gaze responsiveness in any system, 2 incorrectly labeled the joystick condition as gaze-responsive, 1 identified the merged condition but not the gaze condition, and 1 labeled the conditions correctly. Several participants expressed surprise at the question and during the subsequent debrief, saying they had forgotten about the gaze sensing entirely or assumed that it was only for passive collection. Therefore, much of the gaze captured seems to be natural rather than intentional.

7.5.2 Remote Robot Control

As described in Sec. 7.3 above, the study was performed in a hybrid manner, in which a participant at their home controlled a robot in the lab, which led to some challenges. The primary challenge mentioned by participants was using a single, stationary camera to judge the robot's position. Participants often reported struggling with depth perception, particularly when aligning the robot gripper with the goal handle. Early practice during the unassisted trials did help, though, and the assistance further contributed to success. Few participants reported latency problems; when they did, modifying the video streaming resolution mitigated the problem. In addition, using a stationary viewpoint made the gaze detection problem significantly easier, as it eliminated head motion, 3D gaze detection, and parallax. Pre-trial practice and counterbalancing likely removed many of the condition-agnostic effects, and while the reliability of the gaze tracking system may have been affected, the results should at least bound the usefulness of gaze. Ultimately, the remote study worked well enough to validate our system on a physical robot using eye tracking in the loop despite the restrictions imposed by the COVID-19 pandemic.

7.5.3 Extension to More Complex Tasks

The gaze-based method can be extended to include additional goals, with the caveat that gaze discrimination becomes noisier as the goals get closer together. For more complex tasks, however, gaze prediction will require more sophisticated analysis. In particular, it is difficult using gaze itself to determine the *role* that any particular object has in a task: users can look at one object since it is a goal, and another since it is an obstacle. More detailed analysis such as stronger task models [25] or analysis of gaze locations within an object [6, 57] may help for more general tasks.

In addition, this work assumes that a grasp is the only possible interaction with a goal. However, both control input [59] and natural gaze [125] can be used to infer information about the intended task of the user. We believe that task inference may follow similar patterns as goal inference, with task-specific control input revealing only the immediate task and gaze providing earlier task information when available. Extending this work to more varied tasks is an important aim of future work.

Finally, this work assumes that the user's goal is one of a pre-specified set of objects already known to the assistance system. While this assumption is standard [55], it represents a significant gap between the experimental conditions and a full, deployed system. We look forward to expanding the goal inference process to more general settings.

7.6 Conclusion

From these results, we validate our model of control input for goal prediction in an overall assistance system, as presented in Chap. 3. While control input reliably provides goal predictions for assistance, the modal constraints on user input during the task mean that assistance using only control input is not optimal, and the task benefits from an alternate goal prediction source.

We also find that natural gaze can provide global goal prediction information, and when it does so, the assistance is more effective. This result confirms that even though control input alone is sufficient for simpler tasks, adding an independent, global prediction source like gaze is worth exploring in more complex tasks.

In contrast, gaze performs poorly when used alone. The gaze pipeline used here, and the gaze signal itself, does not provide goal information consistently. Only 21/47 (45%) of trials using gaze alongside the joystick gave accurate predictions sufficiently early to outperform trials with only joystick-based assistance. Though better gaze analysis can improve performance, the lack of any goal-directed fixations during some trials fundamentally limits its predictive ability. Therefore, combining gaze with a more reliable signal, like control input, stabilizes the goal prediction and leads to better assistance. The separate role of each signal showcases the complementarity of different source of information for goal prediction in an assistive setting.

To further improve on the combined system, future work can explore alternate strategies for merging the two prediction methods. Since we find that gaze only helps when it is available before the joystick prediction, we can use gaze for an initial prediction, but switch to the joystick method and entirely omit gaze once distinguishing input becomes available. In addition, other tasks that are more sensitive to early prediction may show greater improvement using gaze. By analyzing the specific role and of each prediction source, we can combine multiple signals in a more nuanced way and achieve better overall performance.

Chapter 8

Discussion

In this work, we analyze the behavior and utility of the user's control input and natural gaze for goal prediction within an assistive task. We show that control input provides a particularly effective source of goal information when the user can act optimally. Natural gaze can enhance the system when users cannot be optimal, but its unreliability makes it perform poorly on its own. This work acts as a foundation for assessing new signals that can be used for goal prediction and for expanding shared control to more complex tasks. Here, we discuss several of the assumptions and limitations underlying the current work and propose future research directions.

8.1 Limitations

Both the goal prediction process and the full assistance system require a fair amount of information about the task. The goal prediction method requires knowledge of all objects that the user might want to interact with and what types of interactions they would perform. The assistance further requires that each possible interaction can be represented as a prespecified policy. Similarly, the semantic gaze labeling process outlined in Chap. 5 requires information about the location of each task-relevant object in the scene. While these assumptions make the problem tractable, they limit its applicability in fully deployed systems. These challenges are similar to those faced by fully autonomous manipulation in unstructured environments, so we expect that adopting research on sensing, task representations, and general manipulation will enable assisted manipulation to expand similarly. Shared control for manipulation has been extended to more generalizable approaches to task models such as deep reinforcement learning [100, 106], latent representations [59, 73] or expandable custom representations [94], as well as identifying when the user is performing a task that is not modeled [132].

The quality of the gaze detection and analysis process represents another challenge. As discussed in Chap. 5, collecting high-quality gaze data in uncontrolled 3D environments is difficult and noisy. Gaze behavior may also depend on the particular situation of the task or the capabilities of the user themselves; for example, users with disabilities who use assistive devices may have different gaze patterns than the participants in our experiments. Similarly, the gaze processing and analysis strategies used here can be improved upon in future work, which may lead to better gaze performance. Despite the challenges in capturing, processing, and understanding people's gaze behavior, our results show that robots can monitor gaze to determine people's goals. More sophisticated analysis techniques will only improve the effectiveness of this signal.

However, gaze is itself a highly complex signal that people use for many different reasons, meaning that guaranteeing the availability of a gaze-based goal prediction is impossible even with perfect sensing. As the task and environment become more complex, understanding people's gaze behavior requires consideration of more possible explanations. As seen in the gaze-only condition in Chap. 7, a simplistic gaze model that cannot reason about the possible roles of a fixated object leads to pathological behavior when presented with richer gaze data. A particularly challenging addition is to incorporate gaze in social contexts. Gaze serves an essential role socially for both general communication and task-oriented collaboration [2]. People instinctively use their gaze while speaking to moderate conversational flow and intimacy [7]. In collaborative tasks, people perform task-directed gaze for their individual tasks, social gaze for conversations, and specific gaze behaviors like joint attention for collaborative tasks [4, 49]. In collaborative environments, people may interleave all of these gaze behaviors, making gaze analysis especially difficult. While the context here enables an investigation of simpler gaze behavior, more general tasks, especially in social contexts, must be able to reason over many possible interpretations. The multiple uses of gaze show that the inconsistent timing of goal-informative gaze behavior is a fundamental property of the signal itself rather than an artifact of simplistic processing. We discuss additional strategies for using gaze and other signals in Sec. 8.2.1.

When discussing how asymmetries between the user's and robot's tasks limit the effectiveness of assistance using control input, we only evaluate the limitations of 2D modal control for 6D end-effector motion. We posit that the discussion extends to more general input modalities. For teleoperation, restricted and non-intuitive interfaces are common [99], due to both the correspondence problem between human and robot kinematics [95] and to the physical limitations of the user [38, 61]. This limitation is not restricted to teleoperation settings. If the Markov decision processes modeling the human and robot behavior are different, as described in Chap. 3, one can likely generate a task for which the assistance is suboptimal even for an optimal user. For example, we can generate a task tree for making a hot beverage, shown in Fig. 3.2. If the human and robot are independently performing actions to move down the tree, the robot assistant cannot predict the appropriate branch selection to perform early enough to ensure effective assistance because the optimal initial task, heating water, does not identify which beverage they are preparing. While we only evaluate this limitation for modal control in a teleoperation task, the argument generalizes to any goal prediction using action observations. See Sec. 8.2.2 for more discussion of how to navigate this problem.

In addition, the assistance makes a strong assumption about the user behavior: that behavior does not change in the presence of assistance. Understanding the complex interplay between users and robotic assistance is an important area of research in itself [83]. The assumption of a stationary user simplifies the problem, as models of the interaction dynamics between the user and a robot quickly explode in complexity. In reality, however, people are indeed reacting to the assistance, and hints of their reactions arise during studies and followup responses. People appear to adapt their behavior to improve the overall system performance at the cost of their individual efficiency. Specifically, people focus on providing distinguishing input early in our manipulation task, even if it is not yet necessary or would even be suboptimal if performed without the assistance. This change in input strategy becomes sharper as users gain experience with the assistance system. While counterbalancing likely prevented this effect from altering results, it also may have led to better overall assistive performance than would be expected under the model assumptions. While some work discusses these interaction dynamics explicitly [82, 85], the user behavior observed here suggests that it is an important phenomenon to investigate and design for. In Sec. 8.2.3, we reflect on general behavior of assistive robots and discuss future projects to better understand this dynamic.

8.2 Future Work

8.2.1 Nonverbal Signals for Intention Inference

Predicting users' immediate goals is only a fraction of the information that gaze reveals about an operator. People's gaze patterns can reveal aspects of themselves like their expertise on a task [31] and their cognitive load [12, 16, 19], and broader information about the task like what type of manipulation activity they are trying to perform [33, 125], what object they will interact with next [79], or when something unexpected has occurred [8]. In addition, if we step back from the contextual representation used in this work, gaze gives global information about a direction of interest, which may be useful for broader analysis. Our work posits that gaze acts as an unreliable but early signal for global information about the task, and we speculate that this description remains useful when incorporating these additional types of information.

The wide array of information available from gaze can be used for many variations of assistance. Knowledge of the user's level of expertise or cognitive load may be useful in aligning the degree of assistance provided with the preferences of the user [38]. Information about what task the user wants to perform or what action they will do next enables a rich variety of assistive actions: rather than only providing assistance in positioning the end-effector at specified pre-grasp positions, the assistance framework can be merged with broader structures of task and motion planning to provide more complex behavior [125]. Gaze towards unexpected locations can also indicate problems with the task. In Sec. 4.2, we see that people look at internal joints of the robot only when they are in a problematic configuration, and people adjust their gaze and viewing angle to compensate for occlusion by the robot. An assistance system can detect these gaze anomalies and automatically perform a relevant action such as stopping the robot, reconfiguring the robot joints, or modifying the level of assistance provided. Finally, the directional information provided by gaze, along with the assumption that people look at relevant parts of the scene, can be used to improve learning from demonstration by highlighting the relevant parts of a task [105]. This use of gaze for object detection rather than selection may also lessen the requirement that relevant objects are known beforehand, thereby expanding the usefulness of our work. The work presented here lays a foundation for future projects that incorporate gaze into assistance in more creative and effective ways.

Looking beyond gaze, our work provides a framework for considering additional signals for goal prediction within a shared control system and for more sophisticated, role-aware signal fusion techniques. Many aspects of people's behavior, from body pose to gestures to speech, can provide goal and task information, and our analysis demonstrates how these signals could be analyzed. For example, industrial manufacturing uses audio and vibration signals to detect anomalies in the process [44], and future work can use this framework to understand the role of a similar monitoring signal for assisted teleoperation. We can also build on source-agnostic signal fusion approaches [53] to incorporate information about the signal's role and behavior. For example, our work suggests that gaze-based prediction is only useful before the control input prediction becomes available, so it may be better to ignore the gaze source after that point. Looking further, this work suggests the importance of future research into evaluating the dynamics of the goal prediction itself and how it affects the performance of the overall assistance system.

8.2.2 Implicit Communication for Assistance

Work on implicit communication during collaborative tasks suggests another way to bypass the limits of action-based intention prediction described in Chap. 3: repair [5]. This concept in conversation analysis describes how people work together effectively despite the possibility of miscommunication. Rather than explicitly stating all relevant information, people assume that common ground is maintained and only give clarifications when they observe evidence that there is a mismatch between the parties' understandings. Knepper et al. [63] models this behavior by considering the probability of an observed action by the partner given the current model of their mental state. Actions out of distribution indicate a high likelihood of mismatch, and new knowledge can be added implicitly by generating hypotheses and selecting those that cause the probability of the observed action to increase. With this strategy, people act based on their own understanding and rely on their partner to communicate if a mismatch is detected. Collaborators whose mental states lead to the same actions do not need to communicate at all. This concept suggests an alternate strategy in the shared control setting. Rather than moving to a neutral position with no goal information, the system could select a goal at random. If the selection is accurate, the user never needs to provide control input at all, and if the selection is incorrect, even low-information responses like a single control action, a confused glance, or an utterance can indicate that the system is incorrect and transition it to the correct goal. This behavior parallels the concept of branch prediction in computer architecture, in which microprocessors speed up code execution by guessing the outcome of a conditional and processing the result in advance. When the system predicts correctly, execution is much more efficient, and when the prediction is incorrect, it only costs slightly more time than if the system had waited for the answer before starting. This collaboration strategy may improve the efficiency of shared control systems.

8.2.3 The Nature of Robot Assistance

Another promising direction for future work concerns broadening our understanding of when and how robot assistance is useful, both by designing new paradigms and by better understanding the tradeoffs and structure inherent to assistive robotics. In work outside this dissertation, we develop a definition of robotic assistance, in which the robot is autonomous in action but subordinate to the target of the assistance in goal [83]. We also outline three key design axes for assistive robots. First, assistive robots need to plan out how to consider the people participating in the interaction: some people are targets of assistance whose goals are prioritized, while some are interactants involved with the robot but not targets, and the robot must determine its goals based on those roles and adjudicate conflicts. Second, since the overall task is performed collaboratively with a user, the same task can be assisted in multiple domains, from assisting with gathering information or planning, to physically affecting the user's body, to performing tasks independently. Finally, robots can vary their level of initiative in their autonomy: proactive robots perform tasks without explicit cues, which minimizes user effort but can lead to problematic robot behavior, and reactive robots are more directly controlled, which makes them less likely to act unpredictably at the cost of requiring more user input. This framing helps to develop a cross-domain understanding of robotic assistance.

This deeper understanding of assistance encourages a focus on the relationship between the user and the assistive robot. As we have seen here, users are not passive recipients of assistance. Rather, people will learn how the system works and control it to achieve their goals, whatever the system's model of the user may be. This mutual adaptation between a robot and a human in a sustained interaction remains an essential topic for future work. Even in this simple task and with only limited interaction episodes, people clearly adapt to the robot's behavior, and they did so in individual ways and to varying degrees. While in this case, this adaptation improved system performance, considering only performance metrics obscures key features of the system. For example, this change of strategy could have increased cognitive load, changed the user's choice of goal based on aspects of the system's performance, or caused users to overfit to a particular task. More problematically, adaptive robot systems may simultaneously learn from the observed user behavior, leading to unpredictable dynamics. Understanding the dynamics of mutual adaptation and designing effective assistance systems sensitive to them is essential.

This user adaption also calls into question the implicitness of the interaction. A key motivation of shared control is that it works seamlessly with a user who is completing the task independently. However, this is not how people use the system. Rather, with increased familiarity, people use their control inputs to explicitly communicate their goals by moving the robot in early, goal-specific directions. From the user's perspective, then, the system consists of a goal selection interface with an awkward input modality, and its advantage over traditional explicit interfaces like touchscreens or speech is less clear. One possible perspective, then, is to view shared control as a system for simultaneous direct input and high-level explicit control: users can both select a goal and adjust the robot's location with the same interface. Gaze input can be viewed similarly: while participants did not generally use intentional gaze to control the robot, the system enabled that behavior if they chose. This explicit communication mitigates the limitation on action-based goal prediction described in Chap. 3, since that system assumed that the user was optimal in their individual actions. Rather, users seem to be optimizing for the overall system performance, and their actions may not make sense without the context of the full system.

In this interface perspective, we deemphasize the contrast between implicit and explicit input and instead view the entire assistance process as a black box, just as the user does. Rather than framing this effect as a limitation, we should see it as an opportunity. People will often use assistive devices for years, and thus they become experts at how to use the devices to accomplish their goals, whatever the design of the system. Therefore, future work can frame the shared control problem as an interface design problem and prioritize systems with high skill ceilings that people can learn to use effectively. Our work on the usefulness of different sources of information for goal prediction during assistance only begins to explore the richness of shared control as an interface, a collaboration, and a relationship between a person and a robot.

Chapter 9

Conclusion

In this thesis, we show that control input and eye gaze complement each other for goal prediction during shared control. Control input gives *local* information about the user's goal, making it particularly effective in simple tasks when people can act optimally but limiting its performance in more complex tasks. On the other hand, eye gaze provides *global* information about task intentions early, but it does not do so as reliably. To demonstrate this complementarity, we first formalize criteria for goal prediction sources to enable effective assistance, and we prove that control input gives sufficient information when provided by noisily optimal users in simpler tasks, but it is not as successful in more complex tasks when the user cannot fully control the system. Next, we analyze natural eye gaze as a source of global information by collecting gaze behavior during a teleoperated manipulation task and showing how it can be used for early goal prediction. Finally, we integrate both signals into a system for online assisted manipulation, and we conduct a user study with a custom-designed task to demonstrate that each signal improves its performance when combined with effective predictions from the other.

Developing a model for how these goal prediction sources contribute to the overall assistance quality during shared control enables us to reason theoretically about when to include these signals and how to use them effectively. This thesis can help ground future questions of whether or not a prediction source, such as gaze or control input, is worth adding to a system. Now that shared control has shown some promise for effective assistance, this work will help bring it closer to real-world applications.

Bibliography

[1] Daniel Aarno and Danica Kragic. Motion intention recognition in robot assisted applications. *Robotics and Autonomous Systems*, 56(8):692–705, 8 2008. ISSN 09218890. doi: 10.1016/j.robot.2007.11.005. URL http://www.sciencedirect.com/science/article/pii/S0921889007001704. Cited on page 10.

[2] Henny Admoni and Brian Scassellati. Social Eye Gaze in Human-Robot Interaction: A Review. *Journal of Human-Robot Interaction*, 6(1), 2017. doi: 10.5898/jhri.6.1.admoni. Cited on page 82.

[3] Henny Admoni and Siddhartha Srinivasa. Predicting user intent through eye gaze for shared autonomy. In *AAAI Fall Symposium - Technical Report*, volume FS-16-01 -, pages 298–303, 2016. ISBN 9781577357759. URL https://www.aaai.org/ocs/index.php/FSS/FSS16/paper/viewPaper/14137. Cited on page 14.

[4] Henny Admoni, Anca Dragan, Siddhartha S. Srinivasa, and Brian Scassellati. Deliberate delays during robot-to-human handovers improve compliance with gaze communication. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 49–56, New York, NY, USA, 3 2014. IEEE Computer Society. ISBN 9781450326582. doi: 10.1145/2559636.2559682. URL https://dl.acm.org/doi/10.1145/2559636.2559682. Cited on page 82.

[5] Saul Albert and J. P. de Ruiter. Repair: The Interface Between Interaction and Cognition. *Topics in Cognitive Science*, 10(2):279–313, 4 2018. ISSN 17568757. doi: 10.1111/tops.12339. URL https://onlinelibrary.wiley.com/doi/10.1111/tops.12339. Cited on page 85.

[6] Sai Krishna Allani, Brendan John, Javier Ruiz, Saurabh Dixit, Jackson Carter, Cindy Grimm, and Ravi Balasubramanian. Evaluating human gaze patterns during grasping tasks: Robot versus human hand. In *Proceedings of the ACM Symposium on Applied Perception, SAP 2016*, pages 45–52, 2016. ISBN 9781450343831. doi: 10.1145/2931002.2931007. URL http://dx.doi.org/10.1145/2931002.2931007. Cited on pages 2, 5, 25, 75, and 78.

[7] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. Conversational gaze aversion for humanlike robots. In *ACM/IEEE International Conference*

on Human-Robot Interaction, pages 25–32, New York, NY, USA, 3 2014. IEEE Computer Society. ISBN 9781450326582. doi: 10.1145/2559636.2559666. URL https://dl.acm.org/doi/10.1145/2559636.2559666. Cited on page 82.

[8] Reuben M Aronson and Henny Admoni. Gaze for Error Detection During Human-Robot Shared Manipulation. In *Fundamentals of Joint Action workshop, Robotics: Science and Systems*, 2018. Cited on pages 5, 8, 31, and 84.

[9] Reuben M. Aronson and Henny Admoni. Semantic gaze labeling for human-robot shared manipulation. In *Eye Tracking Research and Applications Symposium (ETRA)*. Association for Computing Machinery, 6 2019. ISBN 9781450367097. doi: 10.1145/3314111.3319840. Cited on pages 8 and 41.

[10] Reuben M. Aronson and Henny Admoni. Gaze Complements Control Input for Goal Prediction During Assisted Teleoperation. In *Robotics: Science and Systems*, 2022. Cited on page 8.

[11] Reuben M Aronson, Nadia Almutlak, and Henny Admoni. Inferring Goals with Gaze during Teleoperated Manipulation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021. Cited on pages 3, 6, and 8.

[12] R.M. Aronson, T. Santini, T.C. 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*, volume Part F1350, 2018. ISBN 9781450349536. doi: 10.1145/3171221.3171287. Cited on pages 5, 8, 26, 59, 65, 75, and 84.

[13] Rowel Atienza and Alexander Zelinsky. Intuitive human-robot interaction through active 3D gaze tracking. *Springer Tracts in Advanced Robotics*, 15: 172–181, 2005. ISSN 16107438. doi: 10.1007/11008941{_}19. URL http://link.springer.com/10.1007/11008941_19. Cited on pages 13, 35, and 52.

[14] Thomas Bader, Matthias Vogelgesang, and Edmund Klaus. Multimodal integration of natural gaze behavior for intention recognition during object manipulation. In *ICMI-MLMI'09 - Proceedings of the International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interfaces*, pages 199–206, 2009. ISBN 9781605587721. doi: 10.1145/1647314.1647350. Cited on pages 12 and 35.

[15] Dana H. Ballard and Mary M. Hayhoe. Modelling the role of task in the control of gaze. *Visual Cognition*, 17(6-7):1185–1204, 8 2009. ISSN 13506285. doi: 10.1080/13506280902978477. URL http://www.tandfonline.com/doi/abs/10.1080/13506280902978477. Cited on pages 13 and 39.

[16] Jackson Beatty. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 91(2):276–292, 3 1982.

ISSN 00332909. doi: 10.1037/0033-2909.91.2.276. Cited on pages 5 and 84.

[17] Matthias Bernhard, Efstathios Stavrakis, Michael Hecher, and Michael Wimmer. Gaze-to-Object Mapping during Visual Search in 3D Virtual Environments. *ACM Transactions on Applied Perception*, 11(3):1–17, 8 2014. ISSN 15443558. doi: 10.1145/2644812. URL http://doi.acm.org/10.1145/2644812. Cited on page 35.

[18] Andreea Bobu, Dexter R.R. Scobee, Jaime F. Fisac, S. Shankar Sastry, and Anca D. Dragan. LESS is more: Rethinking probabilistic models of human behavior. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 429–437, New York, NY, USA, 3 2020. IEEE Computer Society. ISBN 9781450367462. doi: 10.1145/3319502.3374811. URL https://dl.acm.org/doi/10.1145/3319502.3374811. Cited on pages 2, 3, and 15.

[19] Joseph Bolarinwa, Iveta Eimontaite, Tom Mitchell, Sanja Dogramadzi, and Praminda Caleb-Solly. Assessing the Role of Gaze Tracking in Optimizing Humans-In-The-Loop Telerobotic Operation Using Multimodal Feedback. *Frontiers in Robotics and AI*, 8:265, 10 2021. ISSN 2296-9144. doi: 10.3389/frobt.2021.578596. URL https://www.frontiersin.org/articles/10.3389/frobt.2021.578596/full. Cited on page 84.

[20] Ali Borji and Laurent Itti. Defending yarbus: Eye movements reveal observers' task. *Journal of Vision*, 14(3), 2014. ISSN 15347362. doi: 10.1167/14.3.29. Cited on pages 12 and 35.

[21] Andreas Bulling, Jamie A. Wardz, Hans Gellersenz, and Gerhard Tröstery. Eye movement analysis for activity recognition. In *UbiComp*, pages 41–50, 2009. ISBN 9781605584317. doi: 10.1145/1620545.1620552. Cited on pages 13 and 35.

[22] Tom Carlson and Yiannis Demiris. Collaborative control in human wheelchair interaction reduces the need for dexterity in precise manoeuvres. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, number March, pages 59–66, 2008. Cited on page 11.

[23] Monica S. Castelhano, Michael L. Mack, and John M. Henderson. Viewing task influences eye movement control during active scene perception. *Journal of Vision*, 9(3), 3 2009. ISSN 15347362. doi: 10.1167/9.3.6. Cited on page 35.

[24] Sean Chen, Jensen Gao, Siddharth Reddy, Glen Berseth, Anca D Dragan, and Sergey Levine. ASHA: Assistive Teleoperation via Human-in-the-Loop Reinforcement Learning. In *IEEE International Conference on Robotics and Automation*, 2022. URL http://arxiv.org/abs/2202.02465. Cited on page 13.

[25] Yu Chen and D. H. Ballard. Learning to recognize human action sequences. In *Proceedings - 2nd International Conference on Development and Learning, ICDL 2002*, 2002. ISBN 0769514596. doi: 10.1109/DEVLRN.2002.1011726.

Cited on pages 13 and 78.

[26] Jacob W. Crandall and Michael A. Goodrich. Characterizing efficiency of human robot interaction: A case study of shared-control teleoperation. In *IEEE International Conference on Intelligent Robots and Systems*, volume 2, pages 1290–1295, 2002. doi: 10.1109/irds.2002.1043932. Cited on page 10.

[27] Eric Demeester, Alexander Hüntemann, Dirk Vanhooydonck, Gerolf Vanacker, Alexandra Degeest, Hendrik Van Brussel, and Marnix Nuttin. Bayesian estimation of wheelchair driver intents: Modeling intents as geometric paths tracked by the driver. In *IEEE International Conference on Intelligent Robots and Systems*, pages 5775–5780, 2006. ISBN 142440259X. doi: 10.1109/IROS.2006.282386. Cited on page 11.

[28] Anca D Dragan and Siddhartha S Srinivasa. A policy-blending formalism for shared control. In *International Journal of Robotics Research*, volume 32, pages 790–805. SAGE PublicationsSage UK: London, England, 6 2013. doi: 10.1177/0278364913490324. URL http://journals.sagepub.com/doi/10.1177/0278364913490324. Cited on pages 1, 2, and 11.

[29] Nuno Ferreira Duarte, Mirko Rakovic, Jovica Tasevski, Moreno Ignazio Coco, Aude Billard, and Jose Santos-Victor. Action Anticipation: Reading the Intentions of Humans and Robots. *IEEE Robotics and Automation Letters*, 3(4): 4132–4139, 2018. ISSN 23773766. doi: 10.1109/LRA.2018.2861569. URL https://ieeexplore.ieee.org/abstract/document/8423498. Cited on page 12.

[30] Mohamad A. Eid, Nikolas Giakoumidis, and Abdulmotaleb El Saddik. A Novel Eye-Gaze-Controlled Wheelchair System for Navigating Unknown Environments: Case Study with a Person with ALS. *IEEE Access*, 4:558–573, 2016. ISSN 21693536. doi: 10.1109/ACCESS.2016.2520093. URL http://ieeexplore.ieee.org/document/7394111/. Cited on page 13.

[31] Shahram Eivazi, Roman Bednarik, Markku Tukiainen, Mikael Von Und Zu Fraunberg, Ville Leinonen, and Juha E. Jääskeläinen. Gaze behaviour of expert and novice microneurosurgeons differs during observations of tumor removal recordings. In *Eye Tracking Research and Applications Symposium (ETRA)*, pages 377–380, New York, New York, USA, 2012. ACM Press. ISBN 9781450312257. doi: 10.1145/2168556.2168641. URL http://dl.acm.org/citation.cfm?doid=2168556.2168641. Cited on page 84.

[32] Carlos Elmadjian, Pushkar Shukla, Antonio Diaz Tula, and Carlos H. Morimoto. 3D gaze estimation in the scene volume with a head-mounted eye tracker. In *Proceedings - COGAIN 2018: Communication by Gaze Interaction*, pages 1–9, New York, New York, USA, 2018. ACM Press. ISBN 9781450357906. doi: 10.1145/3206343.3206351. URL http://dl.acm.org/citation.cfm?doid=3206343.3206351. Cited on page 37.

[33] Alireza Haji Fathaliyan, Xiaoyu Wang, and Veronica J. Santos. Exploiting three-dimensional gaze tracking for action recognition during bimanual manipulation to enhance human-robot collaboration. *Frontiers Robotics AI*, 5(APR):25, 4 2018. ISSN 22969144. doi: 10.3389/frobt.2018.00025. URL http://journal.frontiersin.org/article/10.3389/frobt.2018.00025/full. Cited on page 84.

[34] Alireza Fathi, Yin Li, and James M. Rehg. Learning to recognize daily actions using gaze. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7572 LNCS, pages 314–327. Springer-Verlag, 2012. ISBN 978-3-642-33717-8. doi: 10.1007/978-3-642-33718-5{_}23. URL http://link.springer.com/10.1007/978-3-642-33718-5_23. Cited on pages 13 and 35.

[35] J Randall Flanagan, Miles C Bowman, and Roland S Johansson. Control strategies in object manipulation tasks. *Current Opinion in Neurobiology*, 16(6):650–659, 12 2006. ISSN 09594388. doi: 10.1016/j.conb.2006.10.005. URL http://www.sciencedirect.com/science/article/pii/S0959438806001450. Cited on page 5.

[36] Matthew Fontaine and Stefanos Nikolaidis. A Quality Diversity Approach to Automatically Generating Human-Robot Interaction Scenarios in Shared Autonomy. In *Robotics: Science and Systems*, 2021. doi: 10.15607/rss.2021.xvii.036. URL https://roboticsconference.org/program/papers/036/. Cited on pages 2 and 75.

[37] Stefan Fuchs and Anna Belardinelli. Gaze-Based Intention Estimation for Shared Autonomy in Pick-and-Place Tasks. *Frontiers in Neurorobotics*, 15 (647930), 4 2021. ISSN 16625218. doi: 10.3389/fnbot.2021.647930. URL https://dx.doi.org/10.3389/fnbot.2021.647930. Cited on pages 2, 5, 13, 25, and 30.

[38] Deepak Gopinath, Siddarth Jain, and Brenna D. Argall. Human-in-the-loop optimization of shared autonomy in assistive robotics. *IEEE Robotics and Automation Letters*, 2(1):247–254, 1 2017. ISSN 23773766. doi: 10.1109/LRA.2016.2593928. URL http://ieeexplore.ieee.org/document/7518989/. Cited on pages 2, 11, 83, and 84.

[39] Deepak E Gopinath and Brenna D Argall. Mode switch assistance to maximize human intent disambiguation. In *Robotics: Science and Systems*, volume 13, 2017. ISBN 9780992374730. doi: 10.15607/rss.2017.xiii.046. URL http://www.roboticsproceedings.org/rss13/p46.pdf. Cited on pages 4, 11, and 15.

[40] Deepak E. Gopinath and Brenna D. Argall. Active Intent Disambiguation for Shared Control Robots. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(6):1497–1506, 2020. ISSN 15580210.

doi: 10.1109/TNSRE.2020.2987878. URL https://ieeexplore.ieee.org/document/9066939/. Cited on pages 4, 11, and 15.

[41] Kakeru Hagihara, Keiichiro Taniguchi, Irshad Abibouraguimane, Yuta Itoh, Keita Higuchi, Jiu Otsuka, Maki Sugimoto, and Yoichi Sato. Object-wise 3d gaze mapping in physicalworkspace. In *ACM International Conference Proceeding Series*. ACM, 2018. ISBN 9781450354158. doi: 10.1145/3174910.3174921. URL https://doi.org/10.1145/3174910.3174921. Cited on page 35.

[42] Kris Hauser. Recognition, prediction, and planning for assisted teleoperation of freeform tasks. In *Autonomous Robots*, volume 35, pages 241–254. Springer US, 11 2013. doi: 10.1007/s10514-013-9350-3. URL http://link.springer.com/10.1007/s10514-013-9350-3. Cited on page 10.

[43] Mary M. Hayhoe, Anurag Shrivastava, Ryan Mruczek, and Jeff B. Pelz. Visual memory and motor planning in a natural task. *Journal of Vision*, 3(1):49–63, 2 2003. ISSN 15347362. doi: 10.1167/3.1.6. URL http://www.ncbi.nlm.nih.gov/pubmed/12678625. Cited on pages 5, 12, and 39.

[44] Patricia Henriquez, Jesus B. Alonso, Miguel A. Ferrer, and Carlos M. Travieso. Review of automatic fault diagnosis systems using audio and vibration signals, 2014. ISSN 21682232. Cited on page 85.

[45] Laura V. Herlant, Rachel M. Holladay, and Siddhartha S. Srinivasa. Assistive teleoperation of robot arms via automatic time-optimal mode switching. In *ACM/IEEE International Conference on Human-Robot Interaction*, volume 2016-April, pages 35–42. IEEE, 3 2016. ISBN 9781467383707. doi: 10.1109/HRI.2016.7451731. URL http://ieeexplore.ieee.org/document/7451731/. Cited on page 11.

[46] Peter Hevesi, Jamie A Ward, Orkhan Amiraslanov, Gerald Pirkl, and Paul 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. doi: ISBN9781612085982. URL https://discovery.ucl.ac.uk/id/eprint/10039438/. Cited on pages 13 and 35.

[47] Peter F. Hokayem and Mark W. Spong. Bilateral teleoperation: An historical survey. *Automatica*, 42(12):2035–2057, 12 2006. doi: 10.1016/J.AUTOMATICA.2006.06.027. URL http://www.sciencedirect.com/science/article/pii/S0005109806002871. Cited on page 1.

[48] Chien-Ming Huang and Bilge Mutlu. Anticipatory robot control for efficient human-robot collaboration. In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 83–90. IEEE, 3 2016. ISBN 978-1-4673-8370-7. doi: 10.1109/HRI.2016.7451737. URL http://ieeexplore.ieee.org/

document/7451737/. Cited on page 12.

[49] Chien-Ming Huang and Andrea L Thomaz. Joint Attention in Human-Robot Interaction. In *2010 AAAI Fall Symposium Series*, 2010. URL https://www.aaai.org/ocs/index.php/FSS/FSS10/paper/viewPaper/2173. Cited on page 82.

[50] Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology*, 6: 1049, 7 2015. ISSN 1664-1078. doi: 10.3389/fpsyg.2015.01049. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01049/full. Cited on pages 35, 39, and 56.

[51] Alexander Hüntemann, Eric Demeester, Gerolf Vanacker, Dirk Vanhooydonck, Johan Philips, Hendrik Van Brussel, and Marnix Nuttin. Bayesian plan recognition and shared control under uncertainty: Assisting wheelchair drivers by tracking fine motion paths. In *IEEE International Conference on Intelligent Robots and Systems*, pages 3360–3366, 2007. ISBN 1424409128. doi: 10.1109/IROS.2007.4399524. Cited on page 11.

[52] Alexander Hüntemann, Eric Demeester, Marnix Nuttin, and Hendrik Van Brussel. Online user modeling with Gaussian Processes for Bayesian plan recognition during power-wheelchair steering. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, pages 285–292, 2008. ISBN 9781424420582. doi: 10.1109/IROS.2008.4651040. Cited on page 11.

[53] Siddarth Jain and Brenna Argall. Probabilistic Human Intent Recognition for Shared Autonomy in Assistive Robotics. *ACM Transactions on Human-Robot Interaction*, 9(1):1–23, 12 2019. ISSN 2573-9522. doi: 10.1145/3359614. URL http://dl.acm.org/citation.cfm?doid=3375676.3359614. Cited on pages 61, 67, and 85.

[54] Shervin Javdani, Siddhartha S. Srinivasa, and J. Andrew Bagnell. Shared autonomy via hindsight optimization. In *Robotics: Science and Systems*, volume 11. MIT Press Journals, 2015. ISBN 9780992374716. doi: 10.15607/RSS.2015.XI.032. Cited on pages 16, 26, and 65.

[55] Shervin Javdani, Henny Admoni, Stefania Pellegrinelli, Siddhartha S. Srinivasa, and J. Andrew Bagnell. Shared autonomy via hindsight optimization for teleoperation and teaming. *The International Journal of Robotics Research*, 37 (7):717–742, 6 2018. ISSN 0278-3649. doi: 10.1177/0278364918776060. URL https://doi.org/10.1177/0278364918776060. Cited on pages 2, 3, 11, 16, 69, and 79.

[56] Hong Jun Jeon, Smitha Milli, and Anca Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. In *Advances in Neural*

Information Processing Systems, volume 2020-Decem, 2020. Cited on page 16.

[57] Roland S Johansson, Göran Westling, Anders Bäckström, and J Randall Flanagan. Eye–Hand Coordination in Object Manipulation. *Journal of Neuroscience*, 21(17):6917–6932, 9 2001. ISSN 1529-2401. URL http://www.jneurosci.org/content/jneuro/21/17/6917.full.pdf. Cited on pages 5, 11, 25, 39, 61, and 78.

[58] Leif Johnson, Brian Sullivan, Mary Hayhoe, and Dana Ballard. Predicting human visuomotor behaviour in a driving task. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1636), 2 2014. ISSN 09628436. doi: 10.1098/rstb.2013.0044. Cited on page 13.

[59] Hong Jun Jeon, Dylan Losey, and Dorsa Sadigh. Shared Autonomy with Learned Latent Actions. In *Robotics: Science and Systems*. Robotics: Science and Systems Foundation, 6 2020. doi: 10.15607/rss.2020.xvi.011. URL https://doi.org/10.15607/rss.2020.xvi.011. Cited on pages 78 and 82.

[60] Enkelejda Kasneci, Gjergji Kasneci, Thomas C. Kübler, and Wolfgang Rosenstiel. The applicability of probabilistic methods to the online recognition of fixations and saccades in dynamic scenes. In *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '14*, pages 323–326, New York, New York, USA, 2014. ACM Press. ISBN 9781450327510. doi: 10.1145/2578153.2578213. URL http://dl.acm.org/citation.cfm?doid=2578153.2578213. Cited on page 38.

[61] Dae Jin Kim, Rebekah Hazlett-Knudsen, Heather Culver-Godfrey, Greta Rucks, Tara Cunningham, David Portée, John Bricout, Zhao Wang, and Aman Behal. How Autonomy Impacts Performance and Satisfaction: Results from a Study with Spinal Cord Injured Subjects Using an Assistive Robot. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, 42(1):2–14, 2012. ISSN 15582426. doi: 10.1109/TSMCA.2011.2159589. Cited on pages 1, 11, and 83.

[62] Kinova Robotics, Inc. Robot arms, 2020. URL http://www.kinovarobotics.com/assistive-robotics/products/robot-arms/. Cited on page 26.

[63] Ross A. Knepper, Christoforos I. Mavrogiannis, Julia Proft, and Claire Liang. Implicit Communication in a Joint Action. In *ACM/IEEE International Conference on Human-Robot Interaction*, volume Part F127194, pages 283–292, New York, NY, USA, 3 2017. IEEE Computer Society. ISBN 9781450343367. doi: 10.1145/2909824.3020226. URL https://dl.acm.org/doi/10.1145/2909824.3020226. Cited on page 85.

[64] Thomas C. Kübler, Enkelejda Kasneci, and Wolfgang Rosenstiel. SubsMatch: Scanpath similarity in dynamic scenes based on subsequence frequencies. In

Eye Tracking Research and Applications Symposium (ETRA), pages 319–322. Association for Computing Machinery, 2014. ISBN 9781450327510. doi: 10.1145/2578153.2578206. URL https://dl.acm.org/doi/10.1145/2578153.2578206. Cited on page 13.

[65] Thomas C. Kubler, Dennis R. Bukenberger, Judith Ungewiss, Alexandra Worner, Colleen Rothe, Ulrich Schiefer, Wolfgang Rosenstiel, and Enkelejda Kasneci. Towards automated comparison of eye-tracking recordings in dynamic scenes. In *EUVIP 2014 - 5th European Workshop on Visual Information Processing*. Institute of Electrical and Electronics Engineers Inc., 1 2015. ISBN 9781479945726. doi: 10.1109/EUVIP.2014.7018371. Cited on page 13.

[66] Minae Kwon, Erdem Biyik, Aditi Talati, Karan Bhasin, Dylan P Losey, and Dorsa Sadigh. When humans aren't optimal: Robots that collaborate with risk-aware humans. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 43–52, 2020. ISBN 9781450367462. doi: 10.1145/3319502.3374832. URL https://doi.org/10.1145/3319502.3374832. Cited on pages 2, 3, and 15.

[67] Michael Land, Neil Mennie, and Jennifer Rusted. The Roles of Vision and Eye Movements in the Control of Activities of Daily Living. *Perception*, 28(11): 1311–1328, 11 1999. doi: 10.1068/p2935. URL http://journals.sagepub.com/doi/10.1068/p2935. Cited on pages 5, 12, and 39.

[68] Michael F. Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, 41(25):3559–3565, 2001. ISSN 00426989. doi: 10.1016/S0042-6989(01)00102-X. Cited on page 39.

[69] Yann Seing Law-Kam Cio, Maxime Raison, Cedric Leblond Menard, and Sofiane Achiche. Proof of Concept of an Assistive Robotic Arm Control Using Artificial Stereovision and Eye-Tracking. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(12):2344–2352, 12 2019. ISSN 15580210. doi: 10.1109/TNSRE.2019.2950619. URL https://pubmed.ncbi.nlm.nih.gov/31675337/. Cited on page 35.

[70] Songpo Li, Xiaoli Zhang, and Jeremy D. Webb. 3-D-Gaze-Based Robotic Grasping Through Mimicking Human Visuomotor Function for People with Motion Impairments. *IEEE Transactions on Biomedical Engineering*, 64(12): 2824–2835, 12 2017. ISSN 15582531. doi: 10.1109/TBME.2017.2677902. URL https://ieeexplore.ieee.org/document/7870669/. Cited on pages 13, 35, and 52.

[71] Yan Liu, Pei Yun Hsueh, Jennifer Lai, Mirweis Sangin, Marc Antoine Nüssli, and Pierre Dillenbourg. Who is the expert? Analyzing gaze data to predict expertise level in collaborative applications. In *Proceedings - 2009 IEEE International Conference on Multimedia and Expo, ICME 2009*, pages 898–901, 2009. ISBN

9781424442911. doi: 10.1109/ICME.2009.5202640. Cited on pages 5 and 35.

[72] Dylan P. Losey, Craig G. McDonald, Edoardo Battaglia, and Marcia K. O'Malley. A review of intent detection, arbitration, and communication aspects of shared control for physical human–robot interaction, 1 2018. ISSN 00036900. URL http://asmedigitalcollection.asme.org/appliedmechanicsreviews/article-pdf/70/1/010804/5964415/amr_070_01_010804.pdf. Cited on page 2.

[73] Dylan P. Losey, Krishnan Srinivasan, Ajay Mandlekar, Animesh Garg, and Dorsa Sadigh. Controlling Assistive Robots with Learned Latent Actions. In *Proceedings - IEEE International Conference on Robotics and Automation*, pages 378–384. Institute of Electrical and Electronics Engineers Inc., 5 2020. ISBN 9781728173955. doi: 10.1109/ICRA40945.2020.9197197. Cited on page 82.

[74] Kristian Lukander, Miika Toivanen, and Kai Puolamäki. Inferring intent and action from gaze in naturalistic behavior: A review, 10 2017. ISSN 19423918. Cited on pages 12 and 25.

[75] R. Mantiuk, B. Bazyluk, and R. K. Mantiuk. Gaze-driven object tracking for real time rendering. *Computer Graphics Forum*, 32(2):163–173, 5 2013. ISSN 14678659. doi: 10.1111/cgf.12036. URL http://doi.wiley.com/10.1111/cgf.12036. Cited on page 35.

[76] Panadda Marayong, Ming Li, Allison M. Okamura, and Gregory D. Hager. Spatial motion constraints: Theory and demonstrations for robot guidance using virtual fixtures. In *Proceedings - IEEE International Conference on Robotics and Automation*, volume 2, pages 1954–1959, 2003. doi: 10.1109/robot.2003.1241880. Cited on page 9.

[77] Ayaka Matsuzaka, Liu Yang, Chuangyu Guo, Takuya Shirato, and Akio Namiki. Assistance for Master-Slave System for Objects of Various Shapes by Eye Gaze Tracking and Motion Prediction. In *2018 IEEE International Conference on Robotics and Biomimetics, ROBIO 2018*, pages 1953–1958. IEEE, 12 2018. ISBN 9781728103761. doi: 10.1109/ROBIO.2018.8664898. URL https://ieeexplore.ieee.org/document/8664898/. Cited on page 12.

[78] David P. McMullen, Guy Hotson, Kapil D. Katyal, Brock A. Wester, Matthew S. Fifer, Timothy G. McGee, Andrew Harris, Matthew S. Johannes, R. Jacob Vogelstein, Alan D. Ravitz, William S. Anderson, Nitish V. Thakor, and Nathan E. Crone. Demonstration of a semi-autonomous hybrid brain-machine interface using human intracranial EEG, eye tracking, and computer vision to control a robotic upper limb prosthetic. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(4):784–796, 2014. ISSN 15344320. doi: 10.1109/TNSRE.2013.2294685. Cited on page 14.

[79] Neil Mennie, Mary Hayhoe, and Brian Sullivan. Look-ahead fixations: Anticipatory eye movements in natural tasks. *Experimental Brain Research*, 179 (3):427–442, 5 2007. ISSN 00144819. doi: 10.1007/s00221-006-0804-0. URL http://link.springer.com/10.1007/s00221-006-0804-0. Cited on pages 25 and 84.

[80] Ikuhisa Mitsugami, Norimichi Ukita, and Masatsugu Kidode. Robot navigation by eye pointing. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3711 LNCS, pages 256–267, 2005. ISBN 3540290346. doi: 10.1007/11558651{_}26. Cited on page 14.

[81] Katharina Muelling, Arun Venkatraman, Jean-Sebastien Sebastien Valois, John E Downey, Jeffrey Weiss, Shervin Javdani, Martial Hebert, Andrew B Schwartz, Jennifer L Collinger, and J. Andrew Bagnell. Autonomy infused teleoperation with application to brain computer interface controlled manipulation. *Autonomous Robots*, 41(6):1401–1422, 8 2017. ISSN 15737527. doi: 10.1007/s10514-017-9622-4. URL https://doi.org/10.1007/s10514-017-9622-4. Cited on page 11.

[82] Benjamin A. Newman, Abhijat Biswas, Sarthak Ahuja, Siddharth Girdhar, Kris K. Kitani, and Henny Admoni. Examining the Effects of Anticipatory Robot Assistance on Human Decision Making. In *International Conference on Social Robotics (ICSR)*, volume 12483 LNAI, pages 590–603. Springer, Cham., 11 2020. ISBN 9783030620554. doi: 10.1007/978-3-030-62056-1{_}49. URL https://doi.org/10.1007/978-3-030-62056-1_49. Cited on pages 75 and 84.

[83] Benjamin A. Newman, Reuben M. Aronson, Kris Kitani, and Henny Admoni. Helping People Through Space and Time: Assistance as a Perspective on Human-Robot Interaction. *Frontiers in Robotics and AI*, 8:410, 1 2022. ISSN 22969144. doi: 10.3389/frobt.2021.720319. Cited on pages 83 and 86.

[84] Benjamin A. Newman, Reuben M. Aronson, Siddartha S. Srinivasa, Kris Kitani, and Henny Admoni. HARMONIC: A Multimodal Dataset of Assistive Human-Robot Collaboration. *The International Journal of Robotics Research*, 41(1): 3–11, 7 2022. ISSN 17413176. doi: 10.1177/02783649211050677. URL http://journals.sagepub.com/doi/10.1177/02783649211050677. Cited on pages 5, 6, 8, 26, 28, 51, 54, 65, and 66.

[85] Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *International Journal of Robotics Research*, 36(5-7):618–634, 6 2017. ISSN 17413176. doi: 10.1177/0278364917690593. URL https://journals.sagepub.com/doi/full/10.1177/0278364917690593. Cited on page 84.

[86] Stefanos Nikolaidis, Enkelejda Kasneci, and Siddhartha Srinivasa. Leveraging Eye Tracking and Physiological Signals for Fluent Human Robot Collaboration. *CARIS workshop at IROS*, 2017. Cited on page 14.

[87] David Noton and Lawrence Stark. Eye Movements and Visual Perception. *Scientific American*, 224(6):34–43, 1971. ISSN 00368733, 19467087. URL http://www.jstor.org/stable/24922750. Cited on page 13.

[88] A. M. Okamura. Methods for haptic feedback in teleoperated robot-assisted surgery. *Industrial Robot*, 31(6):499–508, 2004. ISSN 0143991X. doi: 10.1108/01439910410566362. URL https://pubmed.ncbi.nlm.nih.gov/16429611/. Cited on page 1.

[89] Pavel Orlov, Ali Shafti, Chaiyawan Auepanwiriyakul, Noyan Songur, and A. Aldo Faisal. A Gaze-contingent Intention Decoding Engine for human augmentation. In *Eye Tracking Research and Applications Symposium (ETRA)*, pages 1–3, New York, New York, USA, 2018. ACM Press. ISBN 9781450357067. doi: 10.1145/3204493.3208350. URL http://dl.acm.org/citation.cfm?doid=3204493.3208350. Cited on page 13.

[90] Lucas Paletta, Katrin Santner, Gerald Fritz, Heinz Mayer, and Johann Schrammel. 3D Attention: Measurement of Visual Saliency Using Eye Tracking Glasses. In *Conference on Human Factors in Computing Systems - Proceedings*, volume 2013-April, pages 199–204, New York, New York, USA, 2013. ACM Press. ISBN 9781450318990. doi: 10.1145/2468356.2468393. URL http://dl.acm.org/citation.cfm?doid=2468356.2468393. Cited on page 35.

[91] Thies Pfeiffer and Patrick Renner. EyeSee3D: a low-cost approach for analyzing mobile 3D eye tracking data using computer vision and augmented reality technology. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '14, pages 369–376, New York, New York, USA, 2014. ACM Press. ISBN 9781450327510. doi: 10.1145/2578153.2628814. URL http://doi.acm.org/10.1145/2578153.2628814. Cited on page 35.

[92] Thies Pfeiffer, Patrick Renner, and Nadine Pfeiffer-Leßmann. EyeSee3D 2.0: Model-based Real-time Analysis of Mobile Eye-tracking in Static and Dynamic Three-dimensional Scenes. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, ETRA '16, pages 189–196, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4125-7. doi: 10.1145/2857491.2857532. URL http://doi.acm.org/10.1145/2857491.2857532. Cited on page 35.

[93] Pupil Labs, Inc. Pupil Labs - Pupil, 2017. URL https://pupil-labs.com/pupil/. Cited on pages 28 and 37.

[94] Gabriel Quere, Annette Hagengruber, Maged Iskandar, Samuel Bustamante, Daniel Leidner, Freek Stulp, and Jorn Vogel. Shared Control Templates for Assistive Robotics. In *Proceedings - IEEE International Conference on Robotics and Automation*, pages 1956–1962, 2020. ISBN 9781728173955. doi: 10.1109/ICRA40945.2020.9197041. Cited on page 82.

[95] D Rakita, B Mutlu, and M Gleicher. A Motion Retargeting Method for Effective Mimicry-Based Teleoperation of Robot Arms. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pages 361–370, 2017. Cited on pages 1 and 83.

[96] Daniel Rakita, Bilge Mutlu, Michael Gleicher, and Laura M. Hiatt. Shared control–based bimanual robot manipulation. *Science Robotics*, 4(30), 5 2019. ISSN 24709476. doi: 10.1126/scirobotics.aaw0955. URL https://www.science.org/doi/abs/10.1126/scirobotics.aaw0955. Cited on page 10.

[97] Marco Ramacciotti, Mario Milazzo, Fabio Leoni, Stefano Roccella, and Cesare Stefanini. A novel shared control algorithm for industrial robots. *International Journal of Advanced Robotic Systems*, 13(6):1–10, 12 2016. ISSN 17298814. doi: 10.1177/1729881416682701. URL http://journals.sagepub.com/doi/10.1177/1729881416682701. Cited on page 9.

[98] Yosef Razin and Karen Feigh. Learning to predict intent from gaze during robotic hand-eye coordination. In *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, volume 31, pages 4596–4602, 2 2017. Cited on pages 2, 5, and 25.

[99] Daniel J Rea and Stela H Seo. Still Not Solved: A Call for Renewed Focus on User-Centered Teleoperation Interfaces. *Frontiers in Robotics and AI*, 9, 2022. doi: 10.3389/frobt.2022.704225. Cited on pages 1 and 83.

[100] Siddharth Reddy, Anca D Dragan, and Sergey Levine. Shared Autonomy via Deep Reinforcement Learning. In *Robotics: Science and Systems*, 2018. ISBN 9780992374747. doi: 10.15607/rss.2018.xiv.005. URL http://www.roboticsproceedings.org/rss14/p05.pdf. Cited on page 82.

[101] Constantin A. Rothkopf, Dana H. Ballard, and Mary M. Hayhoe. Task and context determine where you look. *Journal of Vision*, 7(14), 12 2007. ISSN 15347362. doi: 10.1167/7.14.16. Cited on page 13.

[102] Uta Sailer, J. Randall Flanagan, and Roland S. Johansson. Eye–Hand Coordination during Learning of a Novel Visuomotor Task. *Journal of Neuroscience*, 25(39), 2005. URL http://www.jneurosci.org/content/25/39/8833.long. Cited on pages 5 and 12.

[103] Dario D Salvucci and Joseph H Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the symposium on Eye tracking research & applications - ETRA '00*, pages 71–78, 2000. ISBN

1581132808. doi: 10.1145/355017.355028. URL https://dl.acm.org/doi/10.1145/355017.355028. Cited on page 38.

[104] Thiago Santini, Wolfgang Fuhl, Thomas Kübler, Enkelejda Kasneci, Thomas Kubler, and Enkelejda Kasneci. Bayesian identification of fixations, saccades, and smooth pursuits. In *Eye Tracking Research and Applications Symposium (ETRA)*, volume 14, pages 163–170. Association for Computing Machinery, 3 2016. ISBN 9781450341257. doi: 10.1145/2857491.2857512. URL https://dl.acm.org/doi/10.1145/2857491.2857512. Cited on page 38.

[105] Akanksha Saran, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. Understanding Teacher Gaze Patterns for Robot Learning. In *3rd Conference on Robot Learning (CoRL 2019)*, Osaka, Japan, 2019. URL http://arxiv.org/abs/1907.07202. Cited on page 85.

[106] Charles Schaff and Matthew Walter. Residual Policy Learning for Shared Autonomy. In *Robotics: Science and Systems*, 2020. doi: 10.15607/rss.2020.xvi.072. Cited on page 82.

[107] Lei Shi, Cosmin Copot, and Steve 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)*, pages 515–520. IEEE, 10 2019. ISBN 978-1-7281-0699-1. doi: 10.1109/ICSTCC.2019.8886064. URL https://ieeexplore.ieee.org/document/8886064/. Cited on pages 13, 35, and 52.

[108] Ronal Singh, Tim Miller, Joshua Newn, Eduardo Velloso, Frank Vetere, and Liz Sonenberg. Combining gaze and AI planning for online human intention recognition. *Artificial Intelligence*, 284:103275, 7 2020. ISSN 00043702. doi: 10.1016/j.artint.2020.103275. Cited on page 35.

[109] Barbara Sivak and Christine L. MacKenzie. Integration of visual information and motor output in reaching and grasping: The contributions of peripheral and central vision. *Neuropsychologia*, 1990. ISSN 00283932. doi: 10.1016/0028-3932(90)90143-C. Cited on page 39.

[110] Arjun Sripathy, Andreea Bobu, Daniel S. Brown, and Anca D. Dragan. Dynamically Switching Human Prediction Models for Efficient Planning. In *Proceedings - IEEE International Conference on Robotics and Automation*, volume 2021-May, pages 3495–3501. Institute of Electrical and Electronics Engineers (IEEE), 10 2021. ISBN 9781728190778. doi: 10.1109/ICRA48506.2021.9561430. Cited on pages 2, 3, and 15.

[111] Janis Stolzenwald and Walterio W Mayol-Cuevas. Rebellion and Obedience: The Effects of Intention Prediction in Cooperative Handheld Robots. In *IEEE International Conference on Intelligent Robots and Systems*, volume

abs/1903.0, pages 3012–3019, 2019. ISBN 9781728140049. doi: 10.1109/IROS40897.2019.8967927. URL https://ieeexplore.ieee.org/abstract/document/8967927. Cited on pages 14 and 75.

[112] Brian T. Sullivan, Leif Johnson, Constantin A. Rothkopf, Dana Ballard, and Mary Hayhoe. The role of uncertainty and reward on eye movements in a virtual driving task. *Journal of Vision*, 12(13), 2012. ISSN 15347362. doi: 10.1167/12.13.19. Cited on page 13.

[113] Enkelejda Tafaj, Gjergji Kasneci, Wolfgang Rosenstiel, and Martin Bogdan. Bayesian online clustering of eye movement data. *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*, page 285, 2012. doi: 10.1145/2168556.2168617. URL http://dl.acm.org/citation.cfm?id=2168556.2168617. Cited on page 68.

[114] Benjamin W. Tatler, Mary M. Hayhoe, Michael F. Land, and Dana H. Ballard. Eye guidance in natural vision: reinterpreting salience. *Journal of Vision*, 11(5):5, 5 2011. ISSN 15347362. doi: 10.1167/11.5.5. URL http://jov.arvojournals.org/Article.aspx?doi=10.1167/11.5.5. Cited on pages 13, 25, 40, and 49.

[115] Tobii. Improving your research with eye tracking since 2001 - Tobii Pro. URL https://www.tobiipro.com/. Cited on page 37.

[116] Irene Tong, Omid Mohareri, Samuel Tatasurya, Craig Hennessey, and Septimiu Salcudean. A retrofit eye gaze tracker for the da Vinci and its integration in task execution using the da Vinci Research Kit. In *IEEE International Conference on Intelligent Robots and Systems*, volume 2015-Decem, pages 2043–2050. IEEE, 9 2015. ISBN 9781479999941. doi: 10.1109/IROS.2015.7353648. URL http://ieeexplore.ieee.org/document/7353648/. Cited on page 14.

[117] Pete Trautman. Assistive Planning in Complex, Dynamic Environments: A Probabilistic Approach. In *Proceedings - 2015 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015*, pages 3072–3078. Institute of Electrical and Electronics Engineers Inc., 1 2016. ISBN 9781479986965. doi: 10.1109/SMC.2015.534. Cited on page 11.

[118] Katherine M. Tsui, Aman Behal, David Kontak, and Holly A. Yanco. I want that: Human-in-the-loop control of a wheelchair-mounted robotic arm. *Applied Bionics and Biomechanics*, 8(1):127–147, 2011. ISSN 17542103. doi: 10.3233/ABB-2011-0004. Cited on pages 13, 35, and 52.

[119] Kathleen A Turano, Duane R Geruschat, and Frank H Baker. Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research*, 43(3):333–346, 2 2003. ISSN 00426989. doi: 10.1016/S0042-6989(02)00498-4. URL http://www.ncbi.nlm.nih.gov/pubmed/12535991. Cited on

page 13.

[120] Eduardo Velloso, Marcus Carter, Joshua Newn, Augusto Esteves, Christopher Clarke, and Hans Gellersen. Motion Correlation. *ACM Transactions on Computer-Human Interaction*, 24(3):1–35, 4 2017. doi: 10.1145/3064937. URL http://dl.acm.org/citation.cfm?doid=3086563.3064937. Cited on page 48.

[121] Eduardo Velloso, Flavio Luiz Coutinho, Andrew Kurauchi, and Carlos H Morimoto. Circular orbits detection for gaze interaction using 2D correlation and profile matching algorithms. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications - ETRA '18*, pages 1–9, New York, New York, USA, 2018. ACM Press. ISBN 9781450357067. doi: 10.1145/3204493.3204524. URL http://dl.acm.org/citation.cfm?doid=3204493.3204524. Cited on page 48.

[122] Mélodie Vidal, Andreas Bulling, and Hans Gellersen. Pursuits. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing - UbiComp '13*, page 439, New York, New York, USA, 2013. ACM Press. ISBN 9781450317702. doi: 10.1145/2493432.2493477. URL http://dl.acm.org/citation.cfm?doid=2493432.2493477. Cited on page 48.

[123] Dinh-Son Son Vu, Ulysse Côté Allard, Clément Gosselin, François Routhier, Benoit Gosselin, and Alexandre Campeau-Lecours. Intuitive adaptive orientation control of assistive robots for people living with upper limb disabilities. In *IEEE International Conference on Rehabilitation Robotics*, pages 795–800. IEEE Computer Society, 8 2017. ISBN 9781538622964. doi: 10.1109/ICORR.2017.8009345. URL http://graal.ift.ulaval.ca/ulysse/articles/2017ICORR.pdf. Cited on page 9.

[124] Ming-Yao Yao Wang, Alexandros A. Kogkas, Ara Darzi, and George P. Mylonas. Free-View, 3D Gaze-Guided, Assistive Robotic System for Activities of Daily Living. In *IEEE International Conference on Intelligent Robots and Systems*, pages 2355–2361. Institute of Electrical and Electronics Engineers Inc., 12 2018. ISBN 9781538680940. doi: 10.1109/IROS.2018.8594045. Cited on pages 13, 35, and 52.

[125] Xiaoyu Wang, Alireza Haji Fathaliyan, and Veronica J. Santos. Toward Shared Autonomy Control Schemes for Human-Robot Systems: Action Primitive Recognition Using Eye Gaze Features. *Frontiers in Neurorobotics*, 14:66, 10 2020. ISSN 16625218. doi: 10.3389/fnbot.2020.567571. URL https://www.frontiersin.org/article/10.3389/fnbot.2020.567571/full. Cited on pages 13, 78, and 84.

[126] Weilie Yi and Dana Ballard. Recognizing behavior in hand-eye coordination patterns. *International Journal of Humanoid Robotics*, 6(3):337–359, 2009. doi:

10.1142/S0219843609001863. URL http://www.ncbi.nlm.nih.gov/pubmed/20862267. Cited on pages 13 and 35.

[127] Erkang You and Kris Hauser. Assisted teleoperation strategies for aggressively controlling a robot arm with 2D input. In *Robotics: Science and Systems*, volume 7, pages 354–361, 2012. ISBN 9780262517799. doi: 10.7551/mitpress/9481.003.0050. Cited on page 10.

[128] Michael Young, Christopher Miller, Youyi Bi, Wei Chen, and Brenna D Argall. Formalized Task Characterization for Human-Robot Autonomy Allocation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6044–6050, 2019. doi: 10.1109/icra.2019.8793475. URL https://ieeexplore.ieee.org/document/8793475. Cited on pages 11 and 31.

[129] Brian D Ziebart, Andrew Maas, J Andrew Bagnell, and Anind K Dey. Maximum Entropy Inverse Reinforcement Learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, page 1433–1438. AAAI Press, 2008. ISBN 9781577353683. Cited on pages 2, 3, 15, and 16.

[130] Brian D Ziebart, Andrew Maas, J Andrew Bagnell, and Anind K Dey. Human behavior modeling with maximum entropy inverse optimal control. In *AAAI Spring Symposium - Technical Report*, volume SS-09-04, pages 92–97, 2009. ISBN 9781577354116. URL http://www.cs.cmu.edu/~bziebart/publications/human-behavior-bziebart.pdf. Cited on page 16.

[131] Brian D. Ziebart, Nathan Ratliff, Garratt Gallagher, Christoph Mertz, Kevin Peterson, J. Andrew Bagnell, Martial Hebert, Anind K. Dey, and Siddhartha Srinivasa. Planning-based prediction for pedestrians. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3931–3936. IEEE, 10 2009. ISBN 9781424438044. doi: 10.1109/IROS.2009.5354147. URL http://ieeexplore.ieee.org/document/5354147/. Cited on pages 2, 3, and 15.

[132] Matthew Zurek, Andreea Bobu, Daniel S Brown, and Anca D Dragan. Situational Confidence Assistance for Lifelong Shared Autonomy. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2783–2789, 2021. doi: 10.1109/icra48506.2021.9561839. Cited on page 82.