UNIVERSITY OF CALIFORNIA

Los Angeles

Data-driven Approaches to Enhance the Human Experience in Human-Robot Systems: Leveraging Eye Gaze for Intention Recognition and Trust Calibration

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Mechanical Engineering

by

Alireza Haji Fathaliyan

2022

ABSTRACT OF THE DISSERTATION

Data-driven Approaches to Enhance the Human Experience in Human-Robot Systems: Leveraging Eye Gaze for Intention Recognition and Trust Calibration

by

Alireza Haji Fathaliyan
Doctor of Philosophy in Mechanical Engineering
University of California, Los Angeles, 2022
Professor Veronica J. Santos, Chair

With the fast-paced proliferation of robots in our daily lives, it is especially important that the human experience is prioritized for human-robot interactions and automated systems. We investigated how human eye movements can be leveraged in real-time by robotic systems to enhance the experience of humans that interact with and/or observe semi-autonomous robots. Specifically, we studied spatiotemporal relationships between human eye gaze, human intent, and human-automation trust for activities of daily living.

In Study #1, we identified features from 3D gaze behavior for use by machine learning classifiers for the purpose of action recognition. We investigated gaze behavior and gaze-object interactions as participants performed a bimanual activity of preparing a powdered drink. We generated 3D gaze saliency maps and used characteristic gaze object sequences to demonstrate an action recognition algorithm.

In Study #2, we introduced a classifier for recognizing action primitives, which we defined as triplets having a verb, “target object,” and “hand object.” Using novel 3D gaze-related features, a recurrent neural network was trained to recognize a verb and target object. The gaze object angle and its rate of change enabled accurate recognition and a reduction in the observational latency of the classifier. Using a non-specific approach for indexing objects, we demonstrated potential generalizability of the classifier across activities.

In Study #3, we evaluated subjective and objective measures of trust in human-automation interactions. We compared real-time physiological responses and trust levels (reported via joystick) to the state-of-the-art method of post-trial Likert surveys. Our results suggest that eye gaze features and heart rate are effective, nonintrusive metrics for real-time monitoring of human trust in automated systems.

In summary, we developed machine learning-based action recognition methods using novel 3D gaze-related features, and we related uncertainty in robot competence to real-time measures of trust variation. Our work establishes a foundation for enhancing human-robot collaborative systems by leveraging eye tracking for intention recognition and trust calibration.

The dissertation of Alireza Haji Fathaliyan is approved.

Tsu-Chin Tsao

Robert M'Closkey

Jonathan Kao

Veronica J. Santos, Committee Chair

University of California, Los Angeles

2022

To my beloved family . . .

Zahra, the source of inspiration, courage, and hope in my life

Nadia, who is capable of putting the biggest smile on my face for no reason

My parents, who dedicated their entire life to me

My parents-in-law who have given me the best gift of my life

My siblings who have supported me in countless ways

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I will forever be thankful to my advisor Professor Veronica J. Santos, for her incredible support, valuable guidance, and friendly discussions. I have learned something insightful through every interaction with her during the entire course of working under her supervision.

I am very grateful to my committee members, Professor Tsu-Chin Tsao, Professor Robert M'Closkey, and Professor Jonathan Kao. I have greatly benefited from their advice, discussions, and invaluable feedback on my proposed research and final dissertation. I have also developed many academic skills through the courses, seminars, or research publications led by all of my committee members.

A special thanks to Dr. Xiaoyu Wang, my incredible collaborator who has been my research partner during the entire course of my doctoral work. We brainstormed novel ideas, designed experiments, overcame technical challenges, and produced deliverable outcomes together. I would also like to thank all UCLA Biomechatronics Laboratory members for providing an excellent friendly, and collaborative research environment.

For Chapter 2, we thank Allison Walters and Manuel Cisneros for contributions to the initial experimental protocol and equipment setup. We thank Sahm Bazargan, Kevin Hsu, John-Pierre Sawma, and Sarah Anaya for assistance with data collection and analysis. Finally, we thank Dr. Eunsuk Chong, Cheng-Ju Wu, and Eric Peltola for discussions on early drafts. For Chapter 3, we thank Daniela Zokaeim, Aarranon Bharathan, Kevin Hsu, and Emma Suh for assistance with data analysis. We thank Dr. Eunsuk Chong, Yi Zheng, and Eric Peltola for discussions on early drafts. For Chapter 4, we thank Annalise Wulf and Eric Peltola for assistance with the experimental setup, Aileen Villalpando and Lionel Zhang for assistance with early stage testing, Jonathan Bopp for assistance with photos, and Huajing Zhao and Dr. Alexis E. Block for discussions on data analysis and early drafts.

Chapter 2 was supported in part by National Science Foundation Awards #1461547 and #1463960. Chapters 2, 3, and 4 were supported in part by Office of Naval Research Award #N00014-16-1-2468. Any opinions, findings, conclusions, or recommendations are those of the authors and do not necessarily represent the official views, opinions, or policies of the funding agencies.

VITA

2017–Pres. Graduate Student Researcher, UCLA Biomechatronics Lab, Mechanical and Aerospace Engineering Department, UCLA, Los Angeles, California

2019–Pres. Data Scientist, Research and Development Team, Meredith/Dotdash Corporation, Los Angeles, California

2017 M.S. Mechanical Engineering, UCLA, Los Angeles, California

2015 B.S. Mechanical and Aerospace Engineering, UCLA, Los Angeles, California

PUBLICATIONS

Haji Fathaliyan, A., Wang, X., and Santos, V.J. ”Exploiting 3D gaze tracking for action recognition during bimanual manipulation to enhance human-robot collaboration.” in Frontiers in Robotics and AI, vol. 5, Article 25, Apr 2018.

Wang, X., **Haji Fathaliyan, A.**, and Santos, V.J. ”Toward shared autonomy control schemes for human-robot systems: Action primitive recognition using eye gaze features.” in Frontiers in Neurorobotics, vol. 14, Article 567571, Oct 2020.

Haji Fathaliyan, A., Wang, X., Bazargan, S., and Santos, V.J. ”Hand-object kinematics and gaze fixation during bimanual tasks.” in Proceedings of the Annual Meeting of the American Society of Biomechanics, Boulder, CO, August 9, 2017.

CHAPTER 1

Introduction

1.1 Motivation

Eye movements can encode rich information regarding human interactions with other agents and/or the environment. For instance, in scenarios where eye-hand coordination is involved, gaze movements lead hand movements [2]. Thus, there is the potential that visual feedback can assist with a robot's situational awareness [3] for the purposes of enhancing human-robot systems. Furthermore, considering the substantial growth in human-machine interactions in the past decade and the presence of multiple cameras in devices, there is a greater opportunity to implement eye tracking hardware and software in daily devices. An individual's intent, attention level, and trust could be captured to improve their daily experience when interacting with artificially intelligent systems.

1.1.1 Prior Applications of Eye Tracking

Eye gaze movements have been studied in different fields to advance various applications. For example, in the field of robotics, eye tracking techniques have been leveraged as teleoperation interfaces for controlling wheelchairs [4], drones [5, 6, 7], and medical robots [8]. In a study by Raymond et al., gaze signals were collected as operators used a joystick to control a wheelchair [4]. A fitting function was trained to predict joystick signals from inputs consisting of gaze point positions. Using the trained fitting function, the operators then controlled the linear and angular velocity of the wheelchair using gaze commands. Yu et al. introduced "gaze gestures" and assigned each gesture to the direct control of each degree of freedom of a drone, including speed, rotation, translation, and altitude [9]. For medical applications, Li et al. developed a robotic laparoscopic system that could automatically steer the laparoscope to focus on the operator's gazed target points [8]. Such a gaze-based targeting system has the potential to make the execution of surgeries smoother and more efficient.

Gaze patterns have also been tracked for the purpose of marketing and advertising enhancements. Researchers have found a strong correlation between consumers' gaze fixation duration and gaze counts with items purchased while observing webpages or product packages [10]. Retailers can take advantage of customers' gaze patterns and attention to optimize packaging and shelf designs in order to increase sales [11]. With eye tracking-equipped mobile phones, online shoppers could also be monitored such that web interface enhancements can be implemented that are based on gaze patterns of potential buyers.

Eye gaze patterns can be affected by skill level. For the task of observing laparoscopic operation videos, Khan et al. found a significant difference in gaze patterns between novice and expert surgeons [12]. While novice surgeons' eyes often wandered from key areas, expert surgeons' eyes were more focused on key target areas of the operative field. For training purposes, novice agents could learn from expert agents' key areas of focus in order to expedite learning.

In cognitive psychology, pupil dilation was found to reflect the intensity of cognitive load [13]. Researchers have monitored pupil size while performing arithmetic operations, digit sorting, and reading comprehension [14]. Larger pupil size was observed when participants were loaded with tasks with more challenging conditions. In another study, Bradley et al. observed pupil diameter changes when subjects looked at pleasant, unpleasant, and neutral pictures [15]. Their findings suggest that emotional arousal can affect pupil size for both pleasant and unpleasant pictures.

Eye tracking has been studied as a non-intrusive measure of trust by continuously monitoring eye movements during automation. Lu and Sarter statistically compared gaze metrics in simulated high/low-reliability modes [16]. Gaze data were collected while participants viewed a series of static images on a screen. As will be discussed further in Chapter 4, we designed a study that was influenced by the lack of a comparative evaluation of different objective and subjective measures of trust in real-time, in a real robot automation setting.

1.2 Contributions

This dissertation investigates and utilizes novel 3D gaze-related features to train action recognition classifiers and presents a comparative evaluation of subjective and objective measures of human trust in automation. First, we identified gaze features that are particularly useful for action recognition, including gaze object, gaze object sequence, gaze object angle, and gaze object angular speed. Using traditional dynamic time warping, we were able to recognize actions at the subtask level, and through recurrent neural networks machine learning approach, we were able to recognize actions at the action primitive level. Finally, we evaluated the efficacy of eye gaze features as a measure of human trust in automation as compared to other physiological objective signals such as heart rate and galvanic skin response.

Chapter 2 presents an investigation of human gaze behavior and gaze-object interactions in 3D during the performance of a bimanual, instrumental activity of daily living. We identify useful features that can be extracted from 3D gaze behaviors and used as inputs to machine learning algorithms for human action recognition. Using dynamic time warping, we created a population-based set of characteristic gaze object sequences and demonstrated action recognition at the subtask level.

Chapter 3 presents a gaze-based action primitive classifier that could be used for human intent recognition in shared autonomy, human-robot systems. In this work, we defined an action primitive as a triplet comprised of a verb, “target object,” and “hand object.” We used a long short-term memory recurrent neural network to recognize participants’ intended verb and target object as the output of the machine learning model. We found that the use of novel gaze-related features, such as gaze object angle and gaze object angular speed, are especially useful for enhancing the recognition accuracy and reducing the observational latency of the model.

Chapter 4 presents an investigation of objective and subjective measures of trust variation in human-automation. Our experiment consisted of an automated assistive robot performing two common tasks for activities of daily living. We compared real-time physiological responses and trust levels (reported via joystick) to the state-of-the-art method of post-trial Likert surveys. Our results suggest that eye gaze features and heart rate are effective, nonintrusive metrics for real-time monitoring of human trust in automated systems. Such real-time measures of human-automation trust could be used to train supervised machine learning models for the real-time prediction of human trust variation during human-robot interactions. By monitoring the human’s mental state, the robot could modify its behavior in order to enhance the human experience in shared autonomy systems.

Chapter 5 summarizes the dissertation and presents potential applications and future enhancements for the presented research.

CHAPTER 2

Exploiting 3D Gaze Tracking for Action Recognition During Bimanual Manipulation to Enhance Human-robot Collaboration

This chapter was based on work published in the journal Frontiers in Robotics and AI [17].

2.1 Abstract

Human-robot collaboration could be advanced by facilitating the intuitive, gaze-based control of robots, and enabling robots to recognize human actions, infer human intent, and plan actions that support human goals. Traditionally, gaze tracking approaches to action recognition have relied upon computer vision-based analyses of 2D egocentric camera videos. The objective of this study was to identify useful features that can be extracted from 3D gaze behavior and used as inputs to machine learning algorithms for human action recognition. We investigated human gaze behavior and gaze-object interactions in 3D during the performance of a bimanual, instrumental activity of daily living: the preparation of a powdered drink. A marker-based motion capture system and binocular eye tracker were used to reconstruct 3D gaze vectors and their intersection with 3D point clouds of objects being manipulated. Statistical analyses of gaze fixation duration and saccade size suggested that some actions (pouring, stirring) may require more visual attention than other actions (reach, pick up, set down, move). Three-dimensional gaze saliency maps, generated with high spatial resolution for six subtasks, appeared to encode action-relevant information. The ”gaze object sequence” was used to capture information about the identity of objects in concert with the temporal sequence in which the objects were visually regarded. Dynamic time warping barycentric averaging was used to create a population-based set of characteristic gaze object sequences that accounted for intra- and inter-subject variability. The gaze object sequence was used to demonstrate the feasibility of a simple action recognition algorithm that utilized a dynamic time warping Euclidean distance metric. Recognition accuracy results of 91.5%, averaged over the six subtasks, suggest that the gaze object sequence is a promising feature for action recognition whose impact could be enhanced through the use of sophisticated machine learning classifiers and algorithmic improvements for real-time implementation. Robots capable of robust, real-time recognition of human actions during manipulation tasks could be used to improve quality of life in the home as well as quality of work in industrial environments.

2.2 Introduction

Recognition of human motion has the potential to greatly impact a number of fields, including assistive robotics, human-robot interaction, and autonomous monitoring systems. In the home, recognition of instrumental activities of daily living (iADLs) could enable an assistive robot to infer human intent and collaborate more seamlessly with humans while also reducing the cognitive burden on the user. A wheelchair-mounted robot with such capabilities could enhance the functional independence of wheelchair users with upper limb impairments [18]. During bimanual iADLs, humans rely heavily on vision to proactively gather task-relevant visual information for planning [2]. For example, task-relevant information for manipulation could include the three-dimensional (3D) location of an object as well as its structure-related and substance-related properties, such as shape and weight, respectively [19]. Saccades typically precede body movement [20] and reflect one's stratey for successful completion of a task.

The relationships between human vision, planning, and intent have inspired roboticists to adopt similar vision-based principles for planning robot movements and to use human gaze tracking for the intuitive control of robot systems. For instance, gaze fixation data collected during the human navigation of rocky terrain have been used to inspire the control of bipedal robots, specifically for the identification and selection of foot placement locations during traversal of rough terrain [21]. Human eye tracking data have also been used in the closed loop control of robotic arms. Recently, [22] demonstrated how 3D gaze tracking could be used to enable individuals with impaired mobility to control a robotic arm in an intuitive manner. Diverging from traditional gaze tracking approaches that leverage two-dimensional (2D) egocentric camera videos, Li et al. presented methods for estimating object location and pose from gaze points reconstructed in 3D. A visuomotor grasping model was trained on gaze locations in 3D along with grasp configurations demonstrated by unimpaired subjects. The model was then used for robot grasp planning driven by human 3D gaze.

In this work, we consider how human eye movements and gaze behavior may encode intent and could be used to inform or control a robotic system for the performance of bimanual tasks. Unlike repetitive, whole-body motions such as walking and running, iADLs can be challenging for autonomous recognition systems for multiple reasons. For instance, human motion associated with iADLs is not always repetitive, often occurs in an unstructured environment, and can be subject to numerous visual occlusions by objects being manipulated as well as parts of the human body. Prior studies on recognition of iADLs often applied computer vision-based approaches to images and videos captured via egocentric cameras worn by human subjects. Video preprocessing methods typically consist of first subtracting the foreground and then detecting human hands, regions of visual interest, and objects being manipulated [23, 24, 25, 26].

A variety of methods have been presented for feature extraction for use in machine learning classifiers. In some studies, hand-hand, hand-object, and/or object-object relationships have been leveraged [27, 28, 29]. The state of an object (e.g., open vs. closed) has been used as a feature of interest [30].Another study leveraged a saliency-based method to estimate gaze position, identify the "gaze object" (the object of visual regard), and recognize an action [31].Other studies have employed eye trackers in addition to egocentric cameras; researchers have reported significant improvements in action recognition accuracy as a result of the additional gaze point information [27, 24].

In the literature, the phrase "saliency map" has been used to reference a topographically arranged map that represents visual saliency of a corresponding visual scene [32]. In this work, we will refer to "gaze saliency maps" as heat maps that represent gaze fixation behaviors. 2D gaze saliency maps have been effectively employed for the study of gaze behavior while viewing and mimicking the grasp of objects on a computer screen [33]. Belardinelli et al. showed that gaze fixations are distributed across objects during action planning and can be used to anticipate a user's intent with the object (e.g., opening vs. lifting a teapot). While images of real world objects were presented, subjects were only instructed to mimic actions. In addition, since such 2D gaze saliency maps were constructed from a specific camera perspective, they cannot be easily generalized to other views of the same object. One of the objectives of this work was to construct gaze saliency maps in 3D that could enable gaze behavior analyses from a variety of perspectives. Such 3D gaze saliency maps could be mapped to 3D point clouds trivially obtained using low-cost RGB-D computer vision hardware, as is common in robotics applications. Furthermore, given that all manipulation tasks occur in three dimensions, 3D gaze saliency maps could enable additional insights into action-driven gaze behaviors. Although our experiments were conducted in an artificial lab setting using an uncluttered object scene, the experiment enabled subjects to perform actual physical manipulations of the object as opposed to only imagining or mimicking the manipulations, as in [33].

The primary objective of this study was to extract and rigorously evaluate a variety of 3D gaze behavior features that could be used for human action recognition to benefit human–robot collaborations. Despite the increasing use of deep learning techniques for end-to-end learning and autonomous feature selection, in this work, we have elected to consider the potential value of independent features that could be used to design action recognition algorithms in the future. In this way, we can consider the physical meaning, computational expense, and value added on a feature-by-feature basis. In Section “Materials and Methods,” we describe the experimental protocol, methods for segmenting actions, analyzing eye tracker data, and constructing 3D gaze vectors and gaze saliency maps. In Section “Results,” we report trends in eye movement characteristics and define the “gaze object sequence.” In Section “Discussion,” we discuss observed gaze behaviors and the potential and practicalities of using gaze saliency maps and gaze object sequences for action recognition. Finally, in Section “Conclusion,” we summarize our contributions and suggest future directions.

2.3 Materials and Methods

2.3.1 Experimental Protocol

This study was carried out in accordance with the recommendations of the UCLA Institutional Review Board with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the UCLA Institutional Review Board. A total of 11 subjects (nine males, two females; aged 18–28 years) participated in the study, whose preliminary results were first reported in [34]. According to a handedness assessment [35], two subjects were "pure right handers," seven subjects were "mixed right handers," and two subjects were "neutral."

Subjects were instructed to perform a bimanual tasks involving everyday objects and actions. In this work, we focus on one bimanual task that features numerous objects and subtasks: the preparation of a powdered drink. To investigate how the findings of this study may generalize to other iADL tasks, we plan to apply similar analyses to other bimanual tasks in the future. The objects for the drink preparation task were selected from the benchmark Yale-CMU-Berkeley (YCB) Object Set [36]: mug, spoon, pitcher, and pitcher lid. The actions associated with these objects were reach for, pick up, set down, move, stir, scoop, drop, insert, and pour.

Subjects were instructed to repeat the task four times with a 1 min break between each trial. The YCB objects were laid out and aligned on a table (adjusted to an ergonomic height for each subject) as shown in Figure 1. The experimental setup was reset prior to each new trial. Subjects were instructed to remove a pitcher lid, stir the contents of the pitcher, which contained water only (the powdered drink was imagined), and transfer the drink from the pitcher to the mug in two different ways. First, three spoonfuls of the drink were to be transferred from the pitcher to the mug using a spoon. Second, the pitcher lid was to be closed to enable to pouring of the drink from the pitcher to the mug until the mug was filled to two-third of its capacity. In order to standardize the instructions provided to subjects, the experimental procedure was demonstrated via a prerecorded video.

Subjects wore an ETL-500 binocular, infrared, head-mounted eye tracker (ISCAN, Inc., Woburn, MA, USA) that tracked their visual point of regard, with respect to a head-mounted egocentric scene camera, at a 60 Hz sampling frequency. Calibration data suggest that the accuracy and precision of the eye tracker are approximately 1.43° and 0.11°, respectively. Six T-Series cameras sampled at 100 Hz and a Basler/Vue video camera (Vicon, Culver City, CA, USA) were used to track the motion of the subjects and YCB objects (Figure 1). Retroreflective markers were attached to the YCB objects, eye tracker, and subjects' shoulders, upper arms, forearms, and hands (dorsal aspects). Visual distractions were minimized through the use of a blackout curtain that surrounded the subject's field of view.

2.3.2 Action Segmentation: Task, Subtask, and Action Unit Hierarchy

[20] reported on gaze fixation during a tea-making task. In that work, a hierarchy of four activity levels was considered: "make the tea" (level 1), "prepare the cups" (level 2), "fill the kettle" (level 3), and "remove the lid" (level 4). [37] reported on a brownie-making task and divided the task into 29 actions, such as "break one egg" and "pour oil in cup." Adopting a similar approach as these prior works, we defined an action hierarchy using a task–subtask–action unit format (Table 1). Subtasks were defined similar to Land et al.'s "4th level activities" while the action units were defined according to hand and object kinematics. All subjects performed all six subtasks listed in Table 1, but not all subjects performed all action units. For example, a couple of subjects did not reach for the pitcher

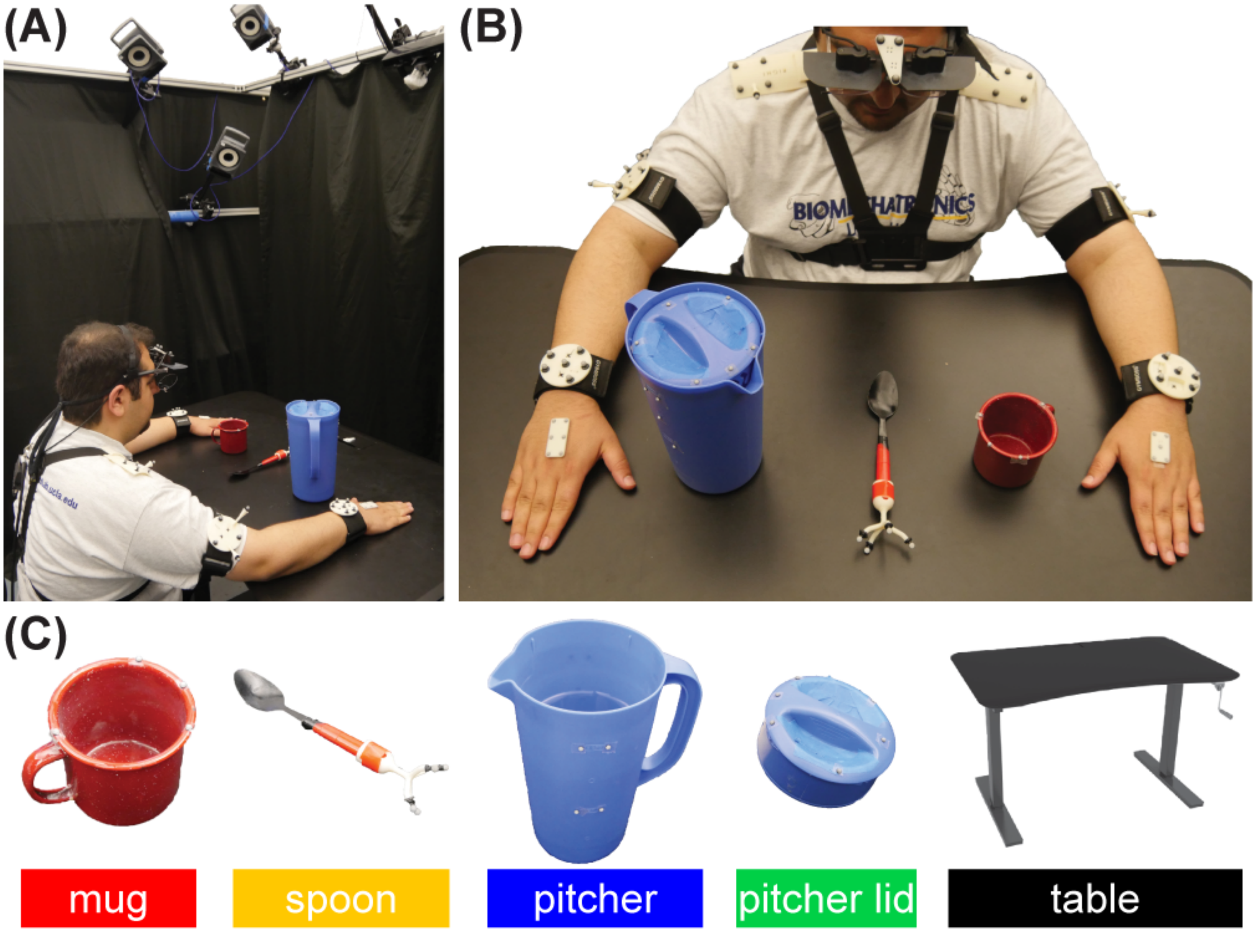


Figure 2.1: (A) Each subject was seated in the motion capture area. A blackout curtain was used to minimize visual distractions. (B) The subject wore a head-mounted eye tracker. Motion capture markers were attached to the Yale-CMU-Berkeley objects, the eye tracker, and subjects' upper limbs. Each trial used the object layout shown. (C) Retroreflective markers were placed on a mug, spoon, pitcher, pitcher lid, and table. These objects will be referenced using the indicated color code throughout this manuscript. The subject shown in panels (A,B) has approved of the publication of these images.

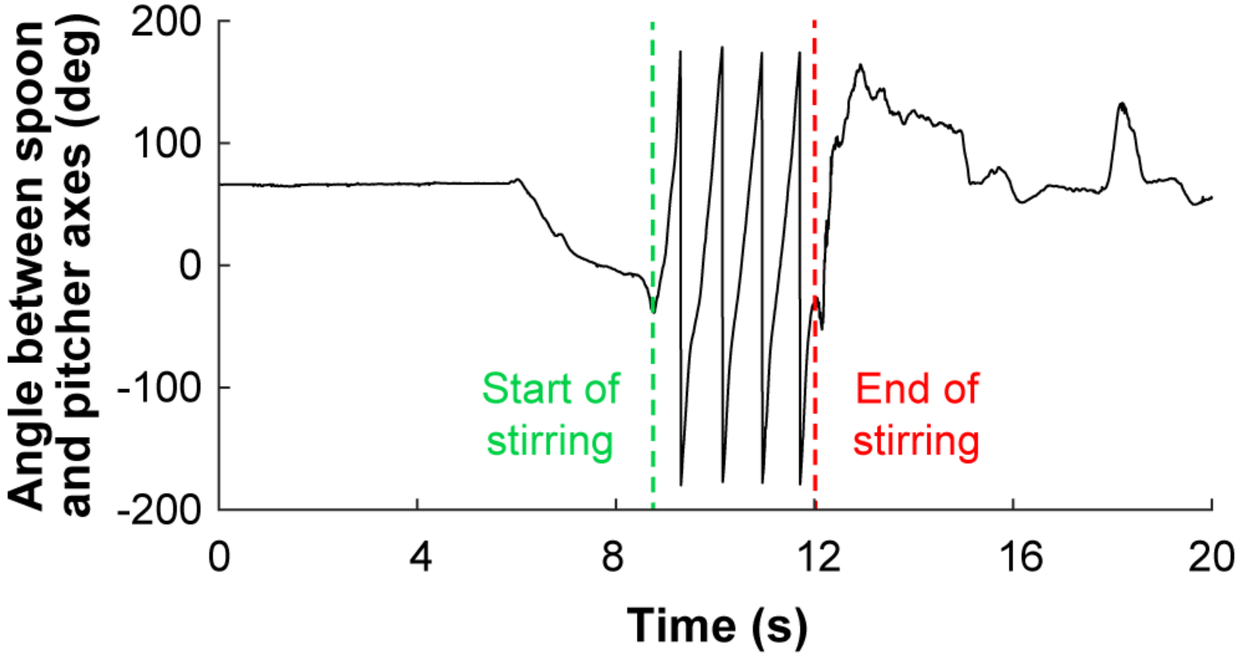


Figure 2.2: The repetitive nature of the spoon's kinematics with respect to the pitcher was used to identify the start and end of the action unit "stir inside pitcher." Although the spoon was not manipulated until approximately 6 s had elapsed in the representative trial shown, the full trial is provided for completeness.

during Subtask 2 ("move spoon into pitcher").

The start and end time of each action unit were identified according to hand and object kinematics and were verified by observing the egocentric video recorded from the eye tracker. For example, the angle of the spoon's long axis with respect to the pitcher's long axis and the repetitive pattern of the angle were used to identify the beginning and end of the action unit "stir inside pitcher" (Figure 2).

2.3.3 Gaze Fixation and Saccade Labeling

Saccadic movements of the eye were discovered by Edwin Landott in 1890 while studying eye movements during reading [38]. According to Kandel et al., saccadic eye movements are characterized by “jerky movements followed by a short pause” or “rapid movements between fixation points.” In our study, saccades were detected using the angular velocity of the reconstructed gaze vector (see 3D Gaze Vector and Gaze Saliency Map Construction) and intervals between saccades that exceeded 200 ms were labeled as gaze fixations, as in [39]. As described previously, the beginning and end of action units were defined based on hand and object kinematics. A heuristic approach, as outlined in Figure 3, was used to associate gaze fixation periods and saccades in the eye tracker data with action units. A given gaze fixation period was associated with a specific action unit if the gaze fixation period overlapped with the action unit period ranging from 0.3 to 0.7 T, where T was the duration of the specific action unit. A given saccade was associated with a specific action unit if the saccade occurred during the action unit period ranging from -0.2 to 0.8 T. Saccade to action unit associations were allowed prior to the start of the action unit (defined from hand and object kinematics) based on reports in the literature that saccades typically precede related motions of the hand [20, 2]. The results of the approach presented in Figure 3 were verified through careful comparison with egocentric scene camera videos recorded by the eye tracker.

2.3.4 3D Gaze Vector and Gaze Saliency Map Construction

The eye tracker provided the 2D pixel coordinates of the gaze point with respect to the image plane of the egocentric scene camera. The MATLAB Camera Calibration Toolbox [40, 41] and a four-step calibration procedure were used to estimate the camera’s intrinsic and

Subtasks	**Action units**
1: remove pitcher lid	Reach for pitcher lid Reach for pitcher Pick up pitcher lid Set down pitcher lid
2: move spoon into pitcher	Reach for pitcher Reach for spoon Pick up spoon Move spoon
3: stir inside pitcher	Stir
4: transfer liquid from pitcher to mug using spoon	Scoop inside pitcher Reach for mug Move mug to pitcher Move spoon to mug Drop liquid into mug using spoon Set down mug Set down spoon
5: replace pitcher lid	Reach for pitcher lid Reach for pitcher Pick up pitcher lid Move pitcher lid to pitcher Insert pitcher lid into pitcher
6: pour liquid into mug	Reach for mug Pick up mug Move mug to pitcher Reach for pitcher handle Pick up pitcher Pour liquid Set down pitcher

Table 2.1: Six subtasks were defined for the task of making a powdered drink; action units were defined for each subtask according to hand and object kinematics.

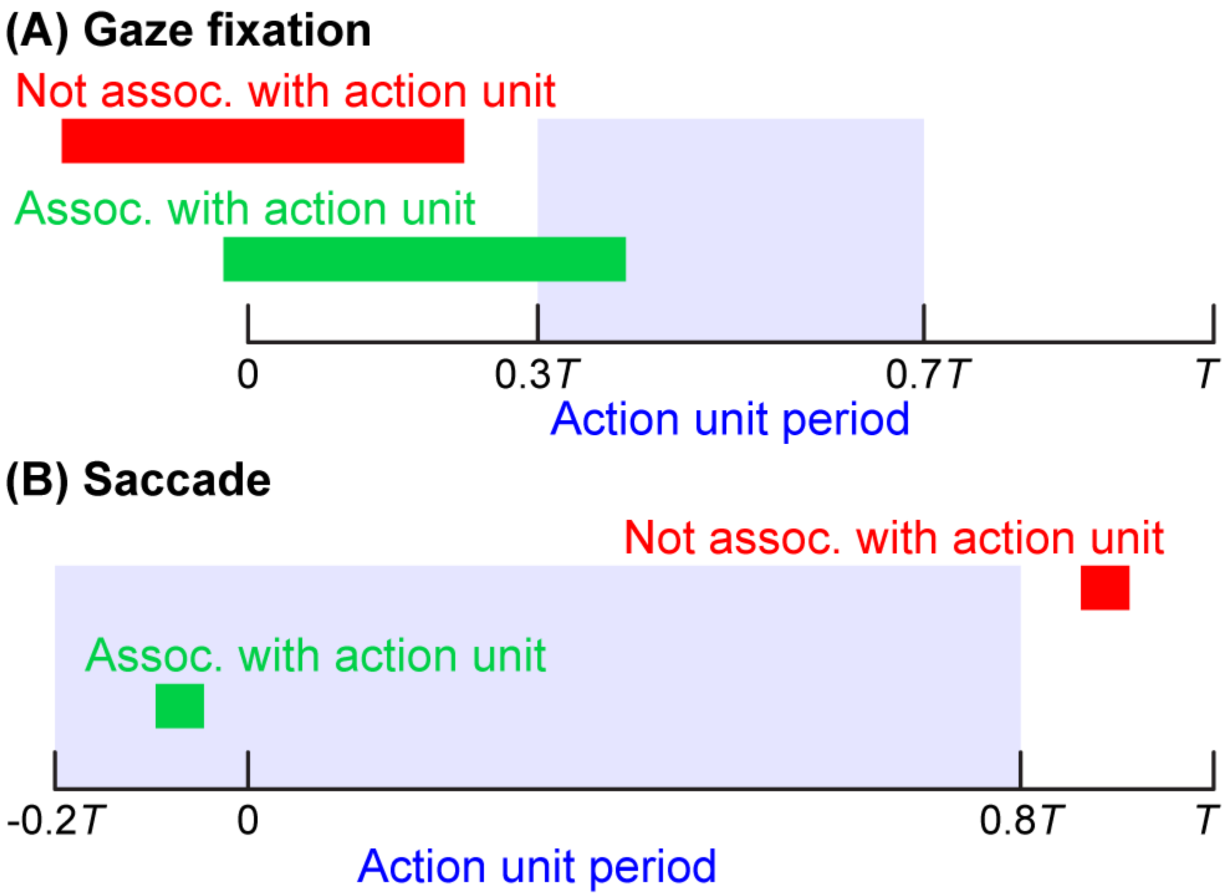


Figure 2.3: (A) A given gaze fixation period was associated with a specific action unit if the gaze fixation period overlapped with the action unit period ranging from 0.3 to 0.7 T (blue shaded region), where T was the duration of the specific action unit. (B) A given saccade was associated with a specific action unit if the saccade occurred during the action unit period ranging from -0.2 to 0.8 T.

extrinsic parameters. These parameters enabled the calculation of the pose of the 2D image plane in the 3D global reference frame. The origin of the camera frame was located using motion capture markers attached to the eye tracker. The 3D gaze vector was reconstructed by connecting the origin of the camera frame with the gaze point's perspective projection onto the image plane.

Using the reconstructed 3D gaze vector, we created 3D gaze saliency maps by assigning RGB colors to the point clouds obtained from 3D scans of the YCB objects. The point cloud for the mug was obtained from [36]. The point clouds for the pitcher, pitcher lid, and spoon were scanned with a structured-light 3D scanner (Structure Sensor, Occipital, Inc., CA, USA) and custom turntable apparatus. This was necessary because the YCB point cloud database only provides point clouds for the pitcher lid assembly and because the proximal end of the spoon was modified for the application of motion capture markers (Figure 1C). Colors were assigned to points based on the duration of their intersection with the subject's 3D gaze vector. In order to account for eye tracker uncertainty, colors were assigned to a 5 mm-radius spherical neighborhood of points, with points at the center of the sphere (intersected by the 3D gaze vector) being most intense. Color intensity for points within the sphere decreased linearly as the distance from the center of the sphere increased. Both gaze fixation and saccades were included during RGB color assignment. For each subtask, the RGB color intensity maps were summed across subjects and then normalized to the [0, 1] range, with 0 as black and 1 as red. The normalization was performed with all task-relevant objects considered simultaneously and not on an object-specific basis. This enabled the investigation of the relative visual importance of each object for each subtask.

2.4 Results

2.4.1 Eye Movements: Gaze Fixation Duration and Saccade Size

Gaze fixation duration and saccade size have previously been identified as important features for gaze behaviors during iADLs. As in [42], we use "saccade size" to refer to the angle spanned by a single saccade. [20] reported overall trends and statistics for the entire duration of a tea-making task. However, information about dynamic changes in gaze behavior is difficult to extract and analyze when eye tracker data are convolved over a large period of time. In order to address eye movements at a finer level of detail, we investigated trends in gaze fixation duration and saccade size at the action unit level. Gaze fixation duration data were normalized by summing the durations of gaze fixation periods that belonged to the same action unit and then dividing by the total duration of that action unit. This normalization was performed to minimize the effect of action unit type, such as reaching vs. stirring, on gaze fixation duration results. Gaze fixation duration and saccade size were analyzed according to groupings based on six common action unit verbs: "reach," "pick up," "set down," "move," "pour," and "stir" (Figure 4). "Drop" and "insert" were excluded, as they occurred infrequently and their inclusion would have further reduced the power of the statistical tests.

We conducted two ANOVA tests with a significance level of $\alpha = 0.05$. One test compared the distributions of gaze fixation duration across the six action unit verb groups while the other test compared the distributions of saccade size. In both cases, the ANOVA resulted in p ¡ 0.001. Thus, post hoc pairwise t-tests were conducted to identify which verb groups were significantly different (Table 2). A Bonferroni correction was additionally applied (α = 0.05/k, where k = 15, the total number of pairwise comparisons) to avoid type I errors

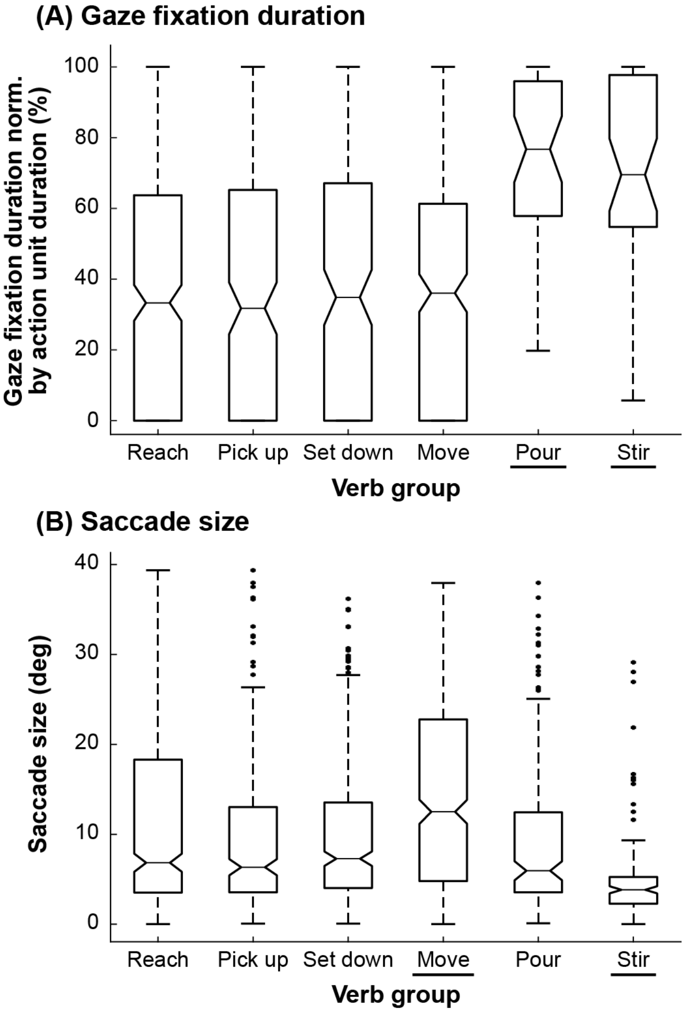


Figure 2.4: Box and whisker plots are shown for each of the six action unit verb groups for (A) normalized gaze fixation duration and (B) saccade size. The tapered neck of each box marks the median while the top and bottom edges mark the first and third quantiles. The whiskers extend to the most extreme data points that are not considered outliers (black dots). For normalized gaze fixation duration, both "pour" and "stir" were statistically significantly different from the other action unit verb groups, as indicated by underlines. For saccade size, both "move" and "stir" were statistically significantly different from the other action unit verb groups.

Fixation Saccade	Reach	Pick up	Set down	Move	Pour	Stir
Reach		0.012	0.050	3e-6*	0.030	2e-13*
Pick up	0.707		0.450	5e-10*	0.462	3e-12*
Set down	0.242	0.496		3e-10*	0.938	2e-9*
Move	0.666	0.992	0.432		9e-8*	9e-23*
Pour	1e-10*	6e-9*	2e-8*	4e-10*		3e-8*
Stir	3e-9*	1e-7*	4e-7*	1e-8*	0.512	

Table 2.2: The lower left triangle of the table (shaded in gray) summarizes p-values for t-tests of average normalized gaze fixation duration for different pairs of action unit verbs while the upper right triangle represents p-values for t-tests with regards to saccade size. Asterisks indicate the t-tests that were statistically significant for a Bonferroni-corrected $\alpha = 0.003$.

when performing the post hoc pairwise comparisons. It was found that the average gaze fixation durations for “pour” and “stir” were significantly greater than those of other verbs (Figure 4A). Saccade sizes for “move” and “stir” were significantly different from those of other verbs (Figure 4B). Saccade sizes for “move” were significantly larger than those of other verbs while those for “stir” were significantly smaller (Figure 5).

2.4.2 3D Gaze Saliency Maps and Gaze Object Percentages

The 3D gaze saliency map for each object is shown for each of the six subtasks in Figure 5. We use “gaze object” to refer to the object that is intersected by the reconstructed 3D gaze vector. This 3D approach is analogous to the use of 2D egocentric camera videos to identify the gaze object defined as the “object being fixated by eyes” or the “visually attended object” [23]. In the case that multiple objects were intersected by the same gaze vector, we selected the closest object to the subject as the gaze object. We defined the gaze object percentage as the amount of time, expressed as a percent of a subtask, that an object was intersected by a gaze vector. Gaze object percentages, averaged across all 11 subjects, are presented for each of the six subtasks in pie chart form (Figure 5). Although the table in the experiment setup was never manipulated, during some subtasks, the gaze object percentage for the table exceeded 20% for subtasks that included action units related to “set down.”

2.4.3 Recognition of Subtasks Based on Gaze Object Sequences

2.4.3.1 The Gaze Object Sequence

In order to leverage information about the identity of gaze objects in concert with the sequence in which gaze objects were visually regarded, we quantified the gaze object sequence

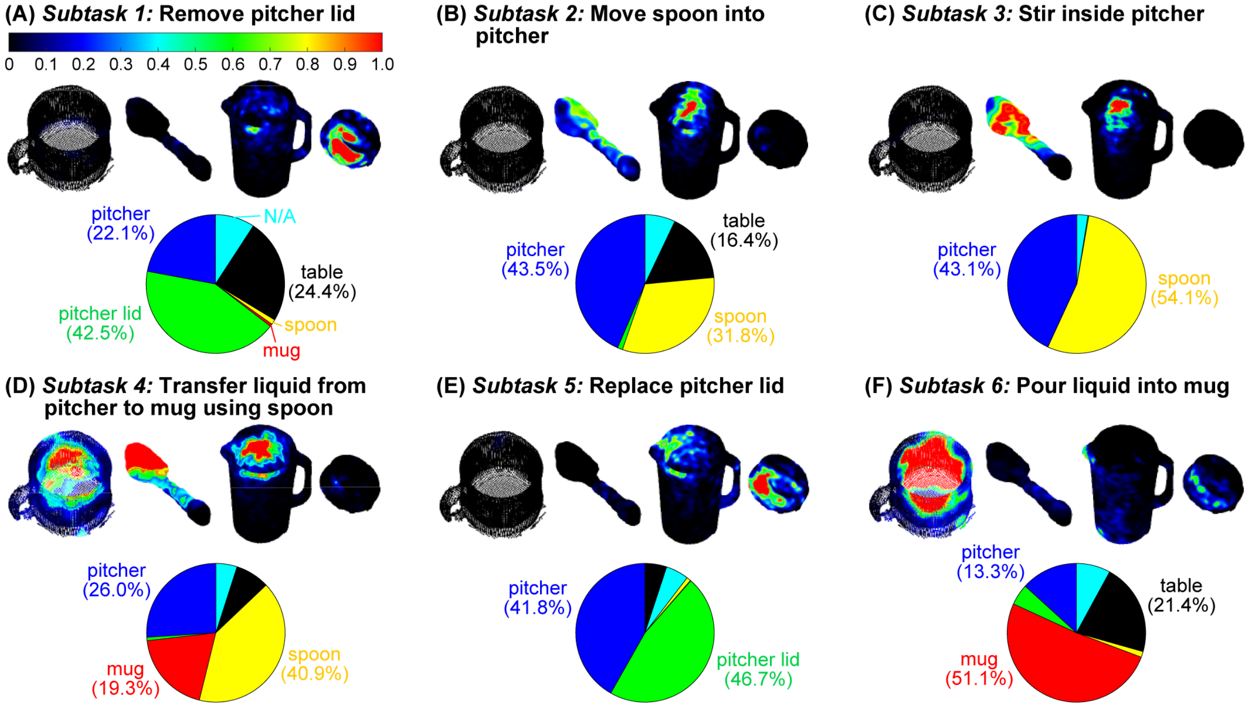


Figure 2.5: Three-dimensional gaze saliency maps of the task-related objects (mug, spoon, pitcher, and pitcher lid) are shown for each of the six subtasks (A–F). The RGB color maps were summed across subjects and then normalized to the [0, 1] range for each subtask. The RGB color scale for all gaze saliency maps is shown in panel (A). Gaze object percentages are reported via pie charts. The colors in the pie charts correspond to the color-coded objects in Figure 1C.

for use in the automated recognition of subtasks. The concept of a gaze object sequence has been implemented previously for human action recognition, but in a different way. [23] performed action recognition with a dynamic Bayesian network having four hidden nodes and four observation nodes. One of the hidden nodes was the true gaze object and one of the observation nodes was the estimated gaze object extracted from 2D egocentric camera videos. In this work, we define the gaze object sequence as being comprised of an (M × N) matrix, where M is the number of objects involved in the manipulation task and N is the total number of instances (frames sampled at 60 Hz) that at least one of the M objects was visually regarded, whether through gaze fixation or saccade (Figure 6C). Each of the $M = 5$ rows corresponds to a specific object. Each of the N columns indicates the number of times each object was visually regarded within a sliding window consisting of 10 frames (Figures 6A,B).

A sliding window was used to filter the raw gaze object sequence to alleviate abrupt changes of values in the matrix. The size of the sliding window was heuristically selected to be large enough to smooth abrupt changes in the object sequence that could be considered as noise, but also small enough so as not to disregard major events within its duration. In preliminary analyses, this sliding window filtration step was observed to improve recognition accuracy.

2.4.3.2 Creating a Library of Characteristic Gaze Object Sequences

Intra- and inter-subject variability necessitate analyses of human subject data that account for variations in movement speed and style. In particular, for pairs of gaze object sequences having different lengths, the data must be optimally time-shifted and stretched prior to comparative analyses. For this task, we used dynamic time warping (DTW), a

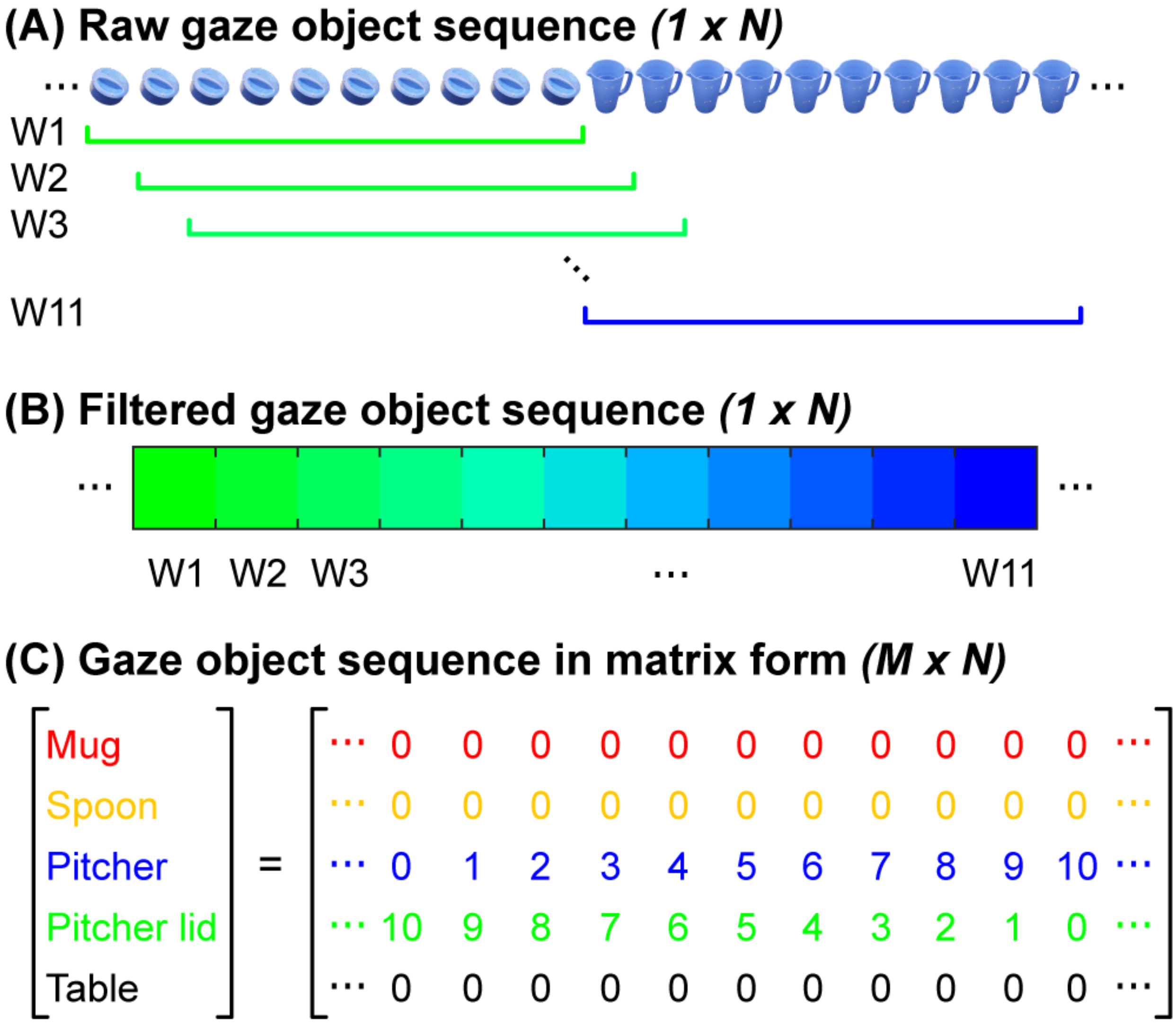


Figure 2.6: (A) Each raw gaze object sequence was represented by a (1 × N) set of frames. In this example, the gaze object transitioned from the pitcher lid to the pitcher. The colors in the figure correspond to the color-coded objects in Figure 1C. (B) The raw sequence of gaze objects was filtered using a rolling window of 10 frames. (C) The gaze object sequence was represented by an (M × N) matrix for M task-relevant objects.

technique that has been widely used for pattern recognition of human motion, such as gait recognition [43] and gesture recognition [44].

Dynamic time warping compares two time-dependent sequences X and Y, where $X \in \mathbb{R}^{S\times U}$ and $Y \in \mathbb{R}^{S\times V}$. A warping path $W_i = [p_{i1}, p_{i2}, ..., p_{ij}, ..., p_{iK_i}]$ defines an alignment between pairs of elements in X and Y by matching element(s) of X to element(s) of Y. For example, $p_{ij} = (u, v)$ represents the matched pair of x_u and y_v. If the warping path is optimized to yield the lowest sum of Euclidean distances between the two sequences, the DTW distance between the two sequences X and Y can be defined as the following:

$$DTW(X, Y) = \min_{W_i}\{d(W_i) \,|\, W_i \in \langle W_1, W_2, ..., W_L\rangle\}, \tag{2.1}$$

where $d(W_i) = \sum_{j=1}^{K_i}\langle p_{ij}\rangle$ and $\langle p_{ij}\rangle = \|x_u - y_v\|_2$.

In order to identify a characteristic gaze object sequence for each subtask, we employed a global averaging method called dynamic time warping barycenter averaging (DBA), which performs the DTW and averaging processes simultaneously. This method uses optimization to iteratively refine a DBA (average) sequence until it yields the smallest DTW Euclidean distance (see Recognition of Subtasks Using DTW Euclidean Distances) with respect to each of the input sequences being averaged ([45]). The gaze object sequences were averaged across all trials for all subjects for each subtask using an open source MATLAB function provided by the creators of the DBA process ([45]). A total of 43 trials (4 repetitions per each of 11 subjects, less 1 incomplete trial) were available for each subtask. Figure 7 shows visual representations of the DBA gaze object sequence for each of the six subtasks.

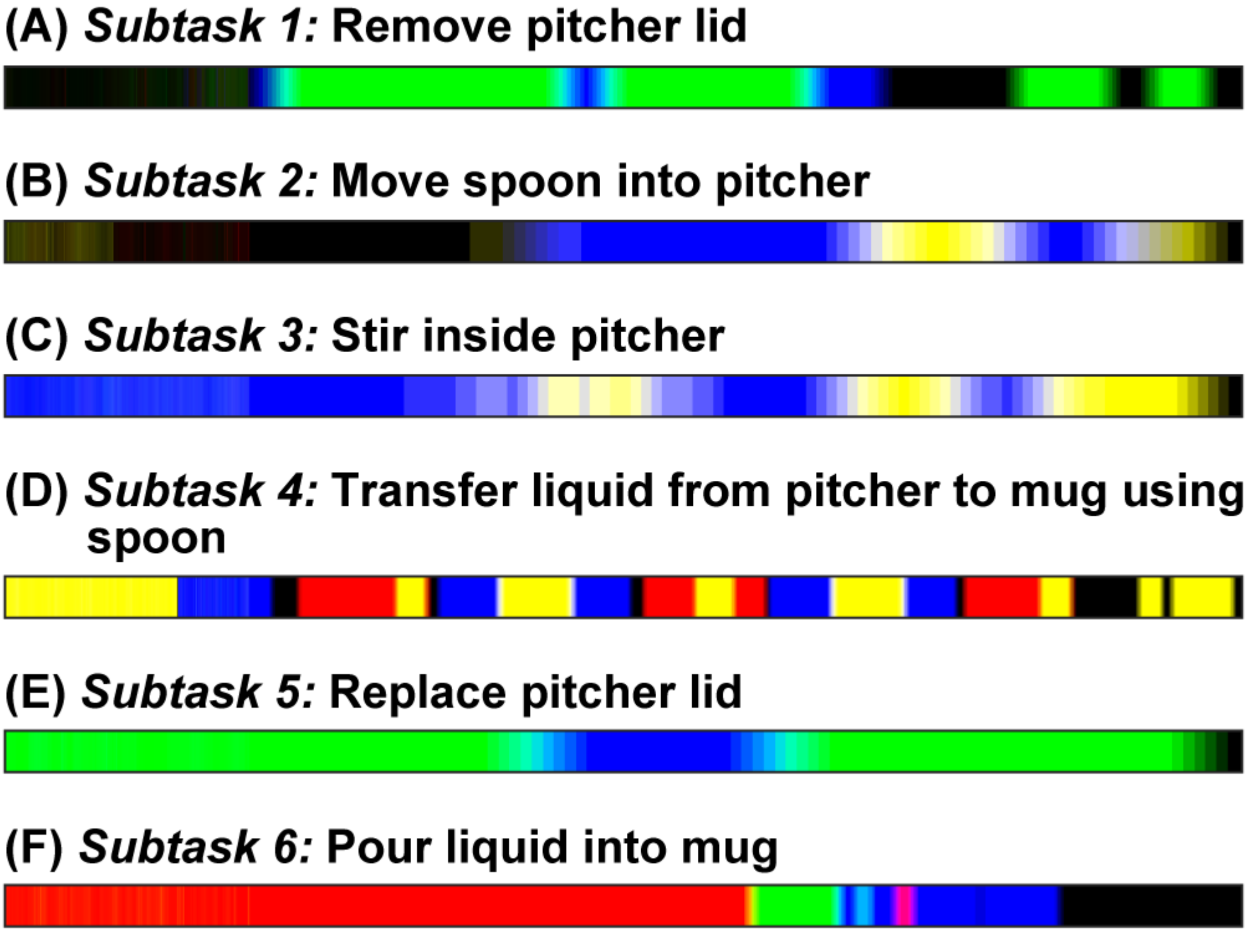


Figure 2.7: Characteristic gaze object sequences were produced using dynamic time warping barycenter averaging over data from 11 subjects for each of six subtasks (A–F). The colors in the figure correspond to the color-coded objects in Figure 1C. The lengths of the sequences were normalized for visualization.

2.4.3.3 Recognition of Subtasks Using DTW Euclidean Distances

Traditionally, the Euclidean distance is used as a metric for similarity between two vectors. However, the Euclidean distance alone is not an accurate measure of similarity for time series data ([45]). Here, we use the "DTW Euclidean distance," which is calculated as the sum of the Euclidean distances between corresponding points of two sequences. The DTW process minimizes the sum of the Euclidean distances, which enables a fair comparison of two sequences. The smaller the DTW Euclidean distance, the greater the similarity between the two sequences. A simple way to associate a novel gaze object sequence with a specific subtask is to first calculate the DTW Euclidean distance between the novel sequence and a characteristic sequence (generated using the DBA process) for each of the six candidate subtasks and to then select the subtask label that results in the smallest DTW Euclidean distance.

Figure 8 shows a novel gaze object sequence and its DTW Euclidean distance with respect to each of the candidate DBA sequences (one for each of six subtasks). The DTW Euclidean distance is reported as a function of the (equal) elapsed times for the novel and DBA gaze object sequences. This enables us to relate recognition accuracy to the percent of a subtask that has elapsed and to comment on the feasibility of real-time action recognition. For instance, for Subtask 4 ("transfer water from pitcher to mug using spoon"), the DTW Euclidean distance between the novel gaze object sequence and the correct candidate DBA sequence does not clearly separate itself from the other five DTW distances until 30% of the novel gaze object sequence has elapsed for the specific case shown (Figure 8). Subtask recognition accuracy generally increases as the elapsed sequence time increases. Figure 8 illustrates how a primitive action recognition approach could be used to label a subtask based on a gaze object sequence alone. However, only one representative novel gaze object sequence was shown as an example.

In order to address the accuracy of the approach as applied to all 43 gaze object sequences, we used a leave-one-out approach. First, one gaze object sequence was treated as an unlabeled, novel sequence. Dynamic time warping barycenter averaging was applied to the remaining sequences. The DTW Euclidean distance was calculated between the novel and candidate DBA sequences, and the pair with the smallest DTW distance was used to label the novel sequence. This process was repeated for each of the gaze object sequences. The DTW distance was calculated using equal elapsed times for the novel and DBA sequences.

The resulting recognition accuracy, precision, and recall for each subtask are reported in Figure 9 as a function of the percent of the subtask that has elapsed. Accuracy represents the fraction of sequences that are correctly labeled. Precision represents the fraction of identified sequences that are relevant to Subtask i. Recall represents the fraction of relevant sequences that are identified ([46])

$$accuracy_i = \frac{TP_i + TNi}{TPi + TNi + FPi + FNi}, \tag{2.2}$$

$$precision_i = \frac{TP_i}{TPi + FPi}, \tag{2.3}$$

$$recall_i = \frac{TP_i}{TPi + FNi}. \tag{2.4}$$

TP_i, TN_i, FP_i, and FN_i represent the number of true positive, true negative, false positive, and false negative sequences when attempting to identify all sequences associated with Subtask i. For example, consider the task of identifying the 43 sequences relevant to Subtask 1 out of the total of (43*6) unlabeled sequences. Using all sequence data, at 100% elapsed

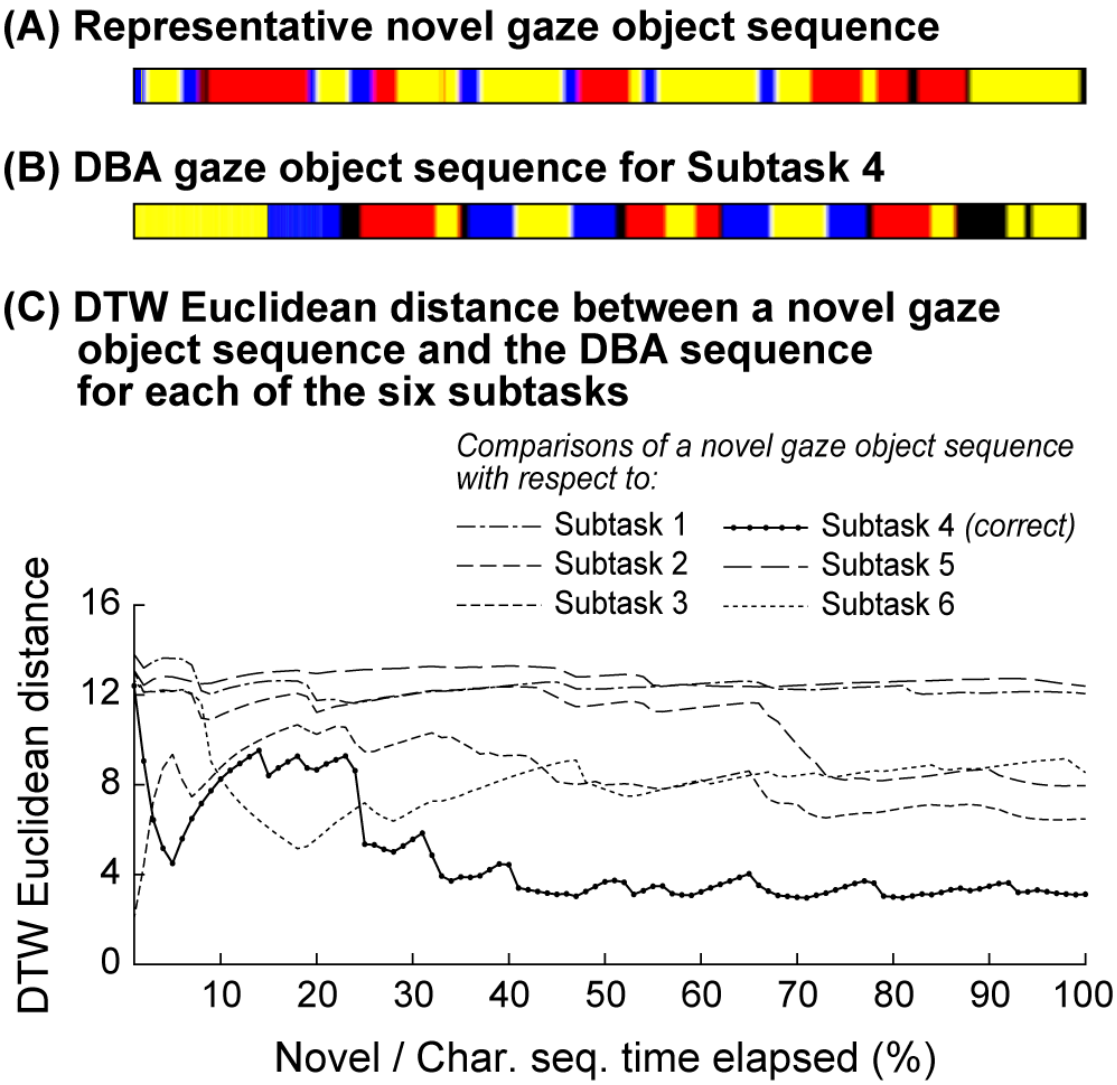


Figure 2.8: (A) A representative novel gaze object sequence is shown. The colors in the figure correspond to the color-coded objects in Figure 1C. (B) A DBA gaze object sequence is shown for Subtask 4, which is the correct subtask label for the novel gaze object sequence shown in panel (A). (C) The DTW Euclidean distance is shown for the comparisons of a novel gaze object sequence and the DBA sequence for each of the six subtasks. The DTW distance was calculated using equal elapsed times for the novel and DBA sequences. The lowest DTW distance would be used to apply a subtask label. Subtask recognition accuracy generally increases as the elapsed sequence time increases.

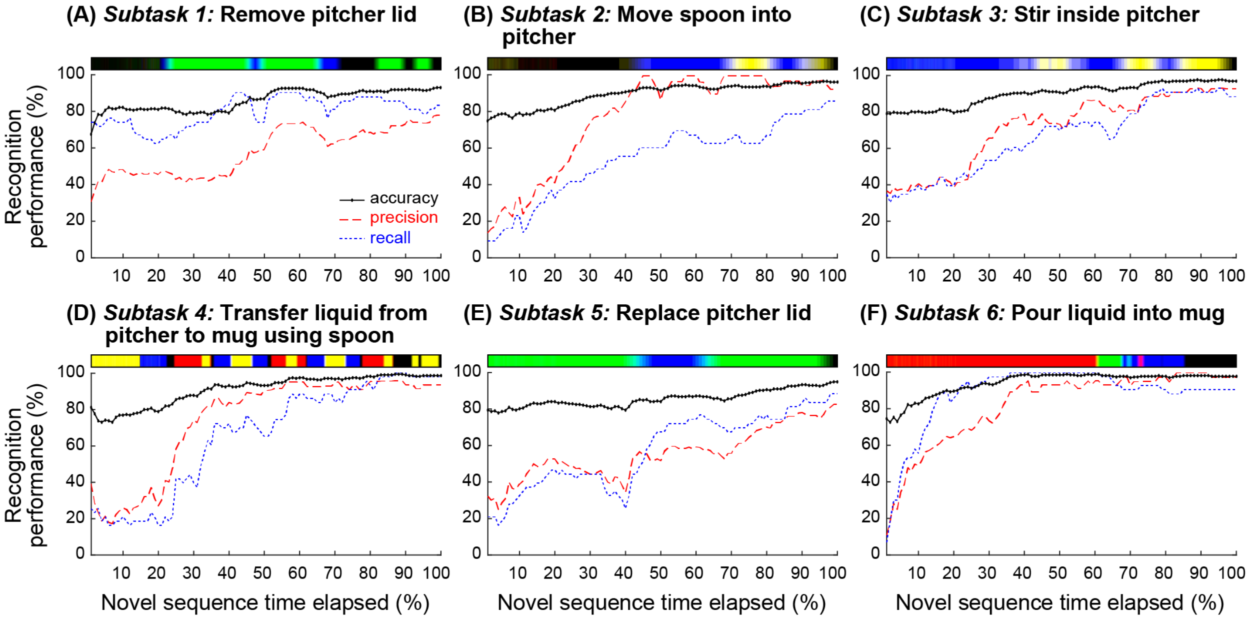


Figure 2.9: Using a leave-one-out approach, the performance of the action recognition algorithm is reported as a function of the elapsed time of a novel gaze object sequence for each subtask. Accuracy (black solid line), precision (red dashed line), and recall (blue dotted line) are shown for each of the six subtasks (A–F). The characteristic gaze object sequence is shown above each subplot. The colors in the sequence correspond to the objects shown in Figure 1C.

time of a novel gaze object sequence, the classifier correctly labeled 36 of the 43 relevant sequences as Subtask 1, but also labeled 10 of the (43*5) irrelevant sequences as Subtask 1. In this case, $TP_1 = 36, TN_1 = 205, FP_1 = 10$, and $FN_1 = 7$. Using Eqs 2–4, this results in an accuracy of 93.4%, precision of 78.2%, and recall of 83.7% for Subtask 1, as shown in Figure 9A.

Figure 10 shows a confusion matrix that summarizes the subtask labeling performance of our simple action recognition algorithm at 100% of the elapsed time for the novel and DBA gaze object sequences. Predictions of subtask labels (columns) are compared to the true subtask labels (rows). Consider again the task of identifying the 43 sequences relevant to Subtask 1. TP_1 is shown as the first diagonal element in the confusion matrix (row 1, column 1). FP_1 and FN_1 are the sum of off-diagonal elements in the first column and first row, respectively.

2.4.4 Discussion

2.4.4.1 Gaze Fixation Duration and Saccade Size May Reflect Differences in Visual Attention

Eye movements were investigated at the action unit level through gaze fixation duration and saccade size. For gaze fixation duration, both “pour” and “stir” were statistically significantly different from the other action unit verb groups (Figure 4A). The median normalized gaze fixation duration values for “pour” and “stir” were, respectively, 41 and 33% greater than the largest median duration value of the “reach,” “pick up,” “set down,” and “move” verb groups (36% for “move”). The lengthier gaze fixation durations could be due to the fact that pouring and stirring simply took longer than the other movements. The trends could

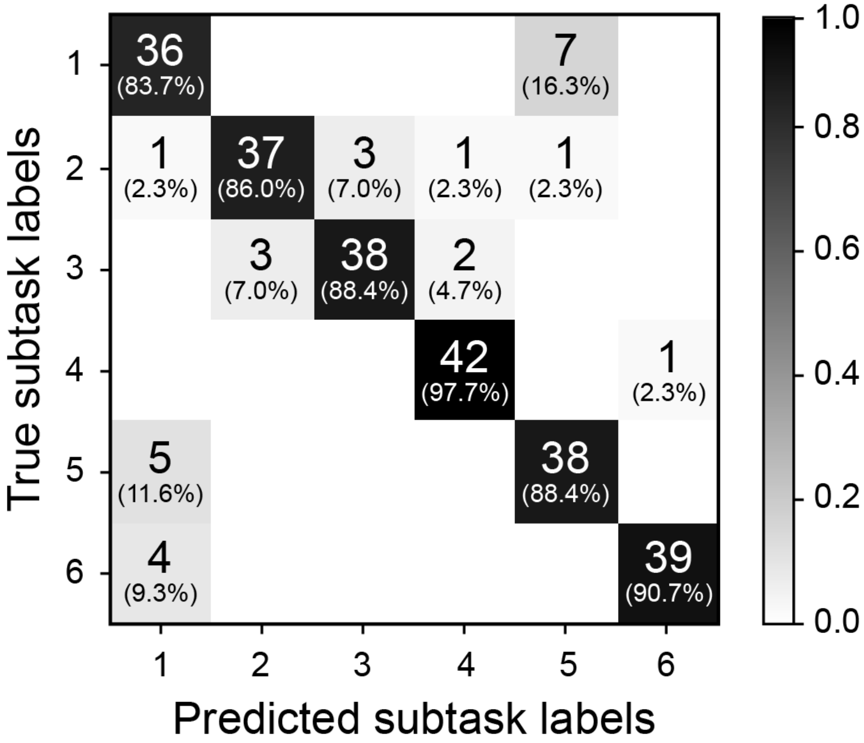


Figure 2.10: The confusion matrix is shown for 100% of the elapsed time of a novel gaze object sequence for each subtask. Predicted subtask labels (columns) are compared to the true subtask labels (rows). Each subtask has a total of 43 relevant sequences and (43*5) irrelevant sequences. Each shaded box lists the number of label instances and parenthetically lists the percentage of those instances out of 43 relevant subtasks.

also indicate that more visual attention is required for successful performance of pouring and stirring. For instance, pouring without spilling and stirring without splashing might require greater manipulation accuracy than reaching, picking up, setting down, or moving an object. However, based on the data collected, it is unknown whether subjects were actively processing visual information during these fixation periods. Gaze fixation durations could also be affected by object properties, such as size, geometry, color, novelty, etc. For instance, fixation durations might be longer for objects that are fragile, expensive, or sharp as compared to those for objects that are durable, cheap, or blunt. The effects of object properties on gaze fixation duration and saccade size require further investigation.

For saccade size, both “move” and “stir” were statistically significantly different from the other action unit verb groups (Figure 4B). The relatively large saccade size for “move” was likely a function of the distance by which the manipulated objects were moved during the experimental task. The relatively small saccade size for “stir” (4.7° ± 2.7°) could be due to the small region associated with the act of stirring within a pitcher and the fact that subjects did not follow the cyclic movements of the spoon with their gaze during stirring.

The concept of “quiet eye,” originally introduced in the literature with regards to the cognitive behaviors of elite athletes, has been used to differentiate between expert and novice surgeons [47]. Quiet eye has been defined as “the final fixation or tracking gaze that is located on a specific location or object in the visuomotor workspace within 3° of the visual angle for >100 ms” [47]. It has been hypothesized that quiet eye is a reflection of a “slowing down” in cognitive planning (not body movement speed) that occurs when additional attention is paid to a challenging task [48]. Based on the gaze fixation duration trends (Figure 4A), one might hypothesize that pouring and stirring require additional attention. Yet, “stir” was the only verb group that exhibited a small saccade size in the range reported for quiet eye.

We are not suggesting that stirring is a special skill that can only be performed by experts; we would not expect a wide range of skill sets to be exhibited in our subject pool for iADL. Nonetheless, it could be reasoned that certain action units may require more visual attention than others and that gaze fixation and saccade size could assist in recognition of such action units employed during everyday tasks.

2.4.4.2 Gaze Saliency Maps Encode Action-Relevant Information at the Subtask and Action Unit Levels

Gaze saliency maps at the subtask level can be used to represent gaze fixation distribution across multiple objects. The gaze saliency maps for the six subtasks (Figure 5) supported Hayhoe and Ballard's finding that gaze fixation during task completion is rarely directed outside of the objects required for the task [49]. Considering Subtask 4, ("transfer water from pitcher to mug using spoon"), the objects comprising the majority of the gaze object percentage pie chart (Figure 5D) were grasped and manipulated (spoon) or were directly affected by an action being performed by a manipulated object (pitcher and mug). While the table was not manipulated, it was often affected by action units that required the picking up or setting down of an object, as for the pitcher lid, spoon, and pitcher in Subtasks 1, 2, and 6 (Figures 5A,B,F), respectively. The gaze fixation percentage for the table was dwarfed by the importance of other objects in Subtasks 4 and 5 (Figures 5D,E).

In some cases, a gaze saliency map could be easily associated with a subtask. For instance, gaze saliency was uniquely, simultaneously intense on the spoon bowl and tip, inner wall of the mug, and inner wall of the pitcher for Subtask 4 ("transfer water from pitcher to mug using spoon") (Figure 5D). In other cases, differences between gaze saliency maps were subtle. For example, the gaze saliency maps were quite similar for the inverse subtasks “remove pitcher lid” and “replace pitcher lid” (Figures 5A,E). In both cases, gaze saliency was focused near the handle of the pitcher lid and the upper rim of the pitcher. However, gaze fixation was slightly more intense near the pitcher spout for Subtask 5 (“replace pitcher lid”) because subjects spent time to carefully align the slots in the pitcher lid with the spout for the “pour liquid into mug” Subtask 6 that was to immediately follow.

Likewise, the gaze saliency maps for Subtask 2 (“move spoon into pitcher”) and Subtask 3 (“stir inside pitcher”) were distinguished only by the subtle difference in gaze fixation distribution on the spoon (Figures 5B,C). The diffuse and homogeneous distribution across the entirety of the spoon for Subtask 2 was contrasted by a focused intensity on the bowl of the spoon for stirring. This was because the “reach for,” “pick up,” and “move” action units performed with the spoon were summed over time to produce the gaze saliency map at the subtask level. Given that the details of each action unit’s unique contribution to the saliency map becomes blurred by temporal summation, it is worth considering gaze saliency maps at a finer temporal resolution, at the action unit level. Due to the short duration of action units (approximately 1 s long), the gaze saliency maps at the action unit level only involve one object at a time. A few representative gaze saliency maps for different action units are shown in Figure 11. The RGB color intensity maps were summed across subjects and then normalized to the [0, 1] range, with 0 as black and 1 as red, according to the duration of the action unit.

Some gaze saliency maps could also be easily associated with specific action units. For instance, gaze saliency intensity was greatest at the top of the pitcher for the action unit “reach for pitcher,” but greatest at the bottom for “set down pitcher” (Figure 11C). By contrast, the gaze saliency maps for the pitcher lid were similar for action units “pick up pitcher lid” and “insert pitcher lid into pitcher.” Subtle differences were observed, such as

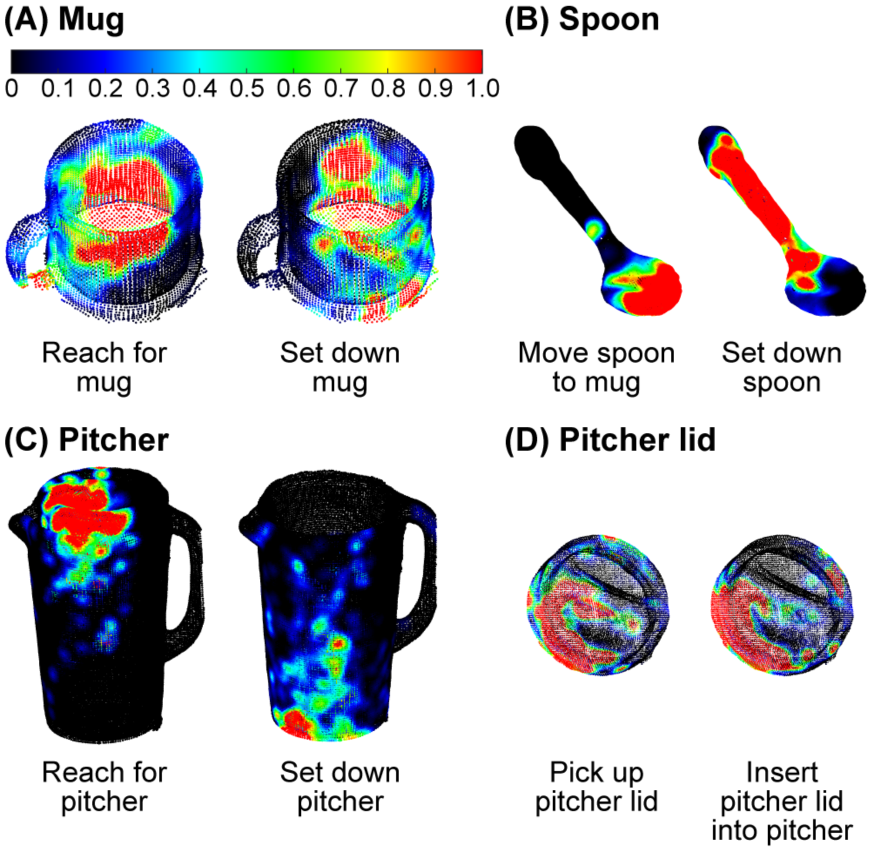


Figure 2.11: Three-dimensional gaze saliency maps of the task-related objects [mug (A), spoon (B), pitcher (C), and pitcher lid (D)] are shown for a subset of action units. The RGB color scale for all gaze saliency maps is shown in panel (A).

more focused gaze intensity near the slots in the lid, in preparation for the "pour liquid into mug" Subtask 6 that was to immediately follow. Gaze saliency maps for different action units were also similar for the mug (Figure 11A), possibly due to its aspect ratio. Not only is the mug a relatively small object but also its aspect ratio from the subject's viewpoint is nearly one. During both "reach for mug" and "set down mug," gaze fixation was spread around the mug's centroid. This was surprising, as we had expected increased intensity near the mug's handle or base for the "reach" and "set down" action units, respectively, based on the findings of [33]. There are a couple of possible explanations for this. First, the Belardinelli et al. study was conducted with a 2D computer display and subjects were instructed to mimic manipulative actions. In this work, subjects physically interacted with and manipulated 3D objects. It is also possible that subjects grasped the mug with varying levels of precision based on task requirements (or lack thereof). For instance, a mug can be held by grasping its handle or its cylindrical body. Had the task involved a hot liquid, for example, perhaps subjects would have grasped and fixated their gaze on the handle of the mug for a longer period.

Although 3D gaze saliency maps are not necessarily unique for all subtasks and action units, it is likely that a combination of the gaze saliency maps for a subtask and its constituent action units could provide additional temporal information that would enable recognition of a subtask. While beyond the scope of this work, we propose that a sequence of gaze saliency maps over time could be used for action recognition. The time series approaches presented for the analysis of gaze object sequences could similarly be applied to gaze saliency map sequences.

2.4.4.3 Practical Considerations and Limitations of Gaze Saliency Maps

If the dynamic tracking of 3D gaze saliency maps is to be practically implemented, one must address the high computational expense associated with tracking, accessing, and analyzing dense 3D point clouds. In this work, the 3D point clouds for the spoon and pitcher were comprised of approximately 3,000 and 20,000 points, respectively. At least two practical modifications could be made to the gaze saliency map representation. First, parametric geometric shapes could be substituted for highly detailed point clouds of rigid objects, especially if fine spatial resolution is not critical for action recognition. The use of a geometric shapes could also enable one to analytically solve for the intersection point(s) between the object and gaze vector. Second, gaze fixation can be tracked for a select subset of regions or segments, such as those associated with “object affordances,” which describe actions that can be taken with an object [50], or ”grasp affordances,” which are defined as “object-gripper relative configurations that lead to successful grasps” [51]. Computational effort could then be focused on regions that are most likely to be task-relevant, such as the spout, rim, handle, and base of a pitcher. Additionally, techniques can be leveraged from computer-based 3D geometric modeling. For example, triangle meshes and implicit surfaces have been used for real-time rendering of animated characters[52]. A similar approach could be used to simplify the 3D point clouds. In addition to tracking the shape and movement of an object, one could track the homogeneous properties (e.g., RGB color associated with gaze fixation duration) of patch elements of surfaces. The spatial resolution of each gaze saliency map could be tuned according to the task-relevant features of the object and reduced to the minimal needs for reliable action recognition.

One limitation of this work is that we cannot comment on the subject’s true focal point or whether subjects were actively processing visual information. A gaze vector may pass through multiple objects, or even through materials that are not rigid objects (e.g., a stream of flowing water). We calculated the intersection points between a gaze vector and objects in its path and then treated the closest intersection point to the user as a gaze fixation point. This approach may not work if some of the task-relevant objects are transparent and subjects look through one object to visually attend to a more distant object. In this work, objects sometimes passed through the path of a stationary gaze vector, but may not have been the focus of active visual attention. For example, the gaze saliency map for Subtask 3 ("stir inside pitcher") displayed regions of greater intensity on both the bowl of the spoon and the inner wall of the pitcher (Figure 5C). However, the egocentric camera attached to the eye tracker revealed that the gaze fixation point remained near the water level line in the pitcher. Since the spoon was moved cyclically near the inner wall of the pitcher, in the same region as the surface of the water, the gaze fixation point alternated between the spoon and the pitcher. As a result, both the spoon and pitcher gaze saliency maps were affected. In one case, a subject's gaze fixation point was calculated as being located on the outer wall of the pitcher during stirring. This interesting case highlights the fact that a direct line of sight (e.g., to the spoon, water, or inner pitcher surface) may not be necessary for subtask completion, and mental imagery ("seeing with the mind's eye") may be sufficient [53].

Future work should address methods for enhancing the robustness of action recognition algorithms to occlusions. For example, if a gaze object is briefly occluded by a moving object that passes through the subject's otherwise fixed field of view, an algorithm could be designed to automatically disregard the object as noise to be filtered out. In addition, a more advanced eye tracker and/or calibration process could be leveraged to estimate focal length. Focal length could be combined with 3D gaze vector direction to increase the accuracy of gaze object identification in cases, where the 3D gaze vector intersects multiple objects.

Human gaze behavior “in the wild” will differ to some (as yet unknown) extent as compared to the gaze behavior observed in our laboratory setting. Our use of black curtains and the provision of only task-relevant objects enabled the standardization of the experimental setup across subjects. However, this protocol also unrealistically minimized visual clutter, the presence of novel objects, and distractions to the subject. In a more natural setting, one’s gaze vector could intersect with task-irrelevant objects in the scene. This would result in the injection of noise into the gaze object sequence, for example, and could decrease the speed and/or accuracy of action recognition. Probabilistic modeling of the noise could alleviate this challenge.

2.4.4.4 The Gaze Object Sequence Can Be Leveraged for Action Recognition to Advance Human–Robot Collaborations

During everyday activities, eye movements are primarily associated with task-relevant objects [54]. Thus, identification of gaze objects can help to establish a context for specific actions. [24] showed that knowledge of gaze location significantly improves action recognition. However, action recognition accuracy was limited by errors in the extraction of gaze objects from egocentric camera video data (e.g., failing to detect objects or detecting irrelevant objects in the background), and gaze objects were not treated explicitly as features for action recognition. Moreover, model development for gaze-based action recognition is challenging due to the stochastic nature of gaze behavior [55]. Using objects tagged with fiducial markers and gaze data from 2D egocentric cameras, Admoni and Srinivasa presented a probabilistic model for the detection of a goal object based on object distance from the center of gaze fixation. In this work, we propose to leverage 3D gaze tracking information about the identity of gaze objects in concert with the temporal sequence in which gaze objects were visually regarded to improve the speed and accuracy of automated action recognition.

In the context of human-robot collaboration, the gaze object sequence could be used as an intuitive, non-verbal control signal by a human operator. Alternatively, the gaze object sequence could be provided passively to a robot assistant that continuously monitors the state of the human operator and intervenes when the human requires assistance. A robot that could infer human intent could enable more seamless physical interactions and collaborations with human operators. For example, a robot assistant in a space shuttle could hand an astronaut a tool during a repair mission, just as a surgical assistant might provide support during a complicated operation. [56] introduced a probabilistic framework for collaboration between a semi-autonomous robot and human co-worker. For a box assembly task, the robot decided whether to hold a box or to hand over a screwdriver based on the movements of the human worker. As there were multiple objects involved in the task, the integration of the gaze object sequence into the probabilistic model could potentially improve action recognition accuracy and speed.

The practical demonstration of the usefulness of gaze object sequence is most likely to occur first in a relatively structured environment, such as that of a factory setting. Despite the unpredictability of human behavior, there are consistencies on a manufacturing line that suggest the feasibility of the gaze object sequence approach. The number of parts and tools used during manual manufacturing operations are uniform in their size and shape and are also limited in number. Although the speed with which a task is completed may vary, the task itself is repetitive. [57] have demonstrated human–robot collaboration for industrial manipulation tasks for which human reaching motions were predicted to enable robot collaboration without collision in a small-shared workspace. In that work, the robot had access to real-time information about the human collaborator's upper limb kinematics, such as palm and arm joint center positions. Focusing on the safety of human–robot collaboration, [58] developed a framework that uses a collision avoidance strategy to assist human workers performing an assembly task in close proximity with a robot arm. Numerous RGB-D cameras were used to track the location and configuration of humans within the collaborative workspace. The common theme of such approaches is to track human kinematics and infer intent from kinematic data alone. The additional use of the gaze object sequence could infer human intent at an earlier stage and further advance safety and efficiency for similar types of human–robot collaboration tasks.

The gaze object sequence could also be demonstrated in the familiar environment of someone's home if a recognition system were properly trained on commonly used objects, where the objects are typically located (e.g., kitchen vs. bathroom), and how they are used. The performance of household robots will largely depend on their ability to recognize and localize objects, especially in complex scenes [59]. Recognition robustness and latency will be hampered by large quantities of objects, the degree of clutter, and the inclusion of novel objects in the scene. The gaze object sequence could be used to address challenges posed by the presence of numerous objects in the scene. While the combinatorial set of objects and actions could be large, characteristic gaze object sequences for frequently used subject-specific iADLs could be utilized to quickly prune the combinatorial set.

Up to now, we have focused primarily on the task-based aspects of gaze tracking for human–robot collaboration. However, gaze tracking could also provide much needed insight into intangible aspects such as human trust in robot collaborators [60]. Our proposed methods could be used to quantify differences in human gaze behavior with and without robot intervention and could enhance studies on the effects of user familiarity with the robot, human vs. non-human movements, perceived risk of robot failure, etc. Consider, for example, a robot arm that is being used to feed oneself [18]. Such a complicated task requires the safe control of a robot near sensitive areas such as the face and mouth and may also be associated with a sense of urgency on the part of the user. A gaze object sequence could reveal high-frequency transitions between task-relevant objects and the robot arm itself, which could indicate a user's impatience with the robot's movements or possibly a lack of trust in the robot and concerns about safety. As the human–robot collaboration becomes more seamless and safe, the frequency with which the user visually checks the robot arm may decrease. Thus, action recognition algorithms may need to be tuned to inter-subject variability and adapted to intra-subject variability as the beliefs and capabilities of the human operator change over time.

Other potential applications of the gaze object sequence include training and skill assessment. For instance, [61] developed a framework that combines Augmented Reality with an Intelligent Tutoring System to train novices on computer motherboard assembly. Via a head-mounted display, trainees were provided real-time feedback on their performance based on the relative position and orientation of tools and parts during the assembly process. Such a system could be further enhanced by, for example, using an expert's gaze object sequence to cue trainees via augmented reality and draw attention to critical steps in the assembly process or critical regions of interest during an inspection process. Gaze object sequences could also be used to establish a continuum of expertise with which skill level can be quantified and certified. [47] described the concepts of "quiet eye" and "slowing down" observed with surgeons performing thyroid lobectomy surgeries. Interestingly, expert surgeons fixated their gaze on the patient's delicate laryngeal nerve for longer periods than novices when performing "effortful" surgical tasks that required increased attention and cognition. Gaze behavior has also been linked with sight reading expertise in pianists [62]. Gaze fixation duration on single-line melodies was shorter for more skilled sight-readers than less skilled sight-readers.

In short, the gaze object sequence generated from 3D gaze tracking data has been demonstrated as a potentially powerful feature for action recognition. By itself, the gaze object sequence captures high-level spatial and temporal gaze behavior information. Moreover, additional features can be generated from the gaze object sequence. For instance, gaze object percentage can be extracted by counting instances of objects in the gaze object sequence. Gaze fixation duration and saccades from one object to another can be extracted from the gaze object sequence. Even saccades to different regions of the same object could potentially be identified if the resolution of the gaze object sequence were made finer through the use of segmented regions of interest for each object (e.g., spout, handle, top, and base of a pitcher).

2.4.4.5 Practical Considerations and Limitations of Gaze Object Sequences

In this work, we have presented a simple proof-of-concept methods for action recognition using a DTW Euclidean distance metric drawn from comparisons between novel and characteristic gaze object sequences. In the current instantiation, novel and characteristic sequences were compared using the same elapsed time (percentage of the entire sequence) (Figure 8). This approach was convenient for a post hoc study of recognition accuracy as a function of time elapsed. However, in practice, the novel gaze object sequence will roll out in real-time and we will not know a priori what percent of the subtask has elapsed. To address this, we propose the use of parallel threads that calculate the DTW Euclidean distance metric for comparisons of the novel sequence with different portions of each characteristic sequence. For instance, one thread runs a comparison with the first 10% of one characteristic gaze object sequence; another thread runs a comparison for the first 20% of the same characteristic gaze object sequence, etc. Such an approach would also address scenarios in which an individual happens to be performing a subtask faster than the population, whose collective behavior is reflected in each characteristic gaze object sequence. For example, it can be seen that the novel gaze object sequence in Figure 8A has a similar pattern as the characteristic gaze object sequence in Figure 8B. However, the individual subject is initially performing the subtask at a faster rate than the population average. The (yellow, blue, black, red, etc.) pattern occurs within the first 10% of the novel sequence, but does not occur until 30% of the characteristic sequence has elapsed. The delayed recognition of the subtask could be addressed using the multi-thread approach described above Figure 8. To further address the computational expense commonly associated with DTW algorithms, one could implement an "unbounded" version of DTW that improves the method for finding matching sequences, which occur arbitrarily within other sequences [63].

For human-robot collaborations, the earlier that a robot can recognize the intent of the human, the more time the robot will have to plan and correct its actions for safety and efficacy. Thus, practical limitations associated with the computational expense of real-time gaze object sequence recognition must be addressed. At the least, comparisons of a novel sequence unfolding in real-time could be made with a library of characteristic subtask sequences using GPUs and parallel computational threads (one thread for each distinct comparison). The early recognition of a novel subtask is not just advantageous for robot planning and control. The computational expense of DTW increases for longer sequences. Thus, the sooner a novel sequence can be recognized, the less time is spent on calculating the proposed DTW Euclidean metric. Since DTW uses dynamic programming to find the best warping paths, a quadratic computational complexity results. While not implemented in this work, the computational expense of the DTW process could be further reduced by leveraging a generalized time warping technique that temporally aligns multimodal sequences of human motion data while maintaining linear complexity [37].

2.4.4.6 Potential Advancements for a Gaze Object Sequence-Based Action Recognition System

As expected, recognition accuracy increased as more of the novel gaze object sequence was compared with each characteristic gaze object sequence (Figure 9). However, the simple recognition approach presented here is not perfect. Even when an entire novel gaze object sequence is compared with each characteristic gaze object sequence, the approach only achieves an accuracy of 96.4%, precision of 89.5%, and recall of 89.2% averaged across the six subtasks. The confusion matrix (Figure 10) shows which subtasks were confused with one another even after 100% elapsed time. Although the percentage of incorrect subtask label predictions is low, the subtasks that share the same gaze objects have been confused the most. For instance, the Subtask 1 (“remove pitcher lid”) and Subtask 5 (“replace pitcher lid”) were occasionally confused with one another. It is hypothesized that the training of a sophisticated machine learning classifier could improve the overall accuracy of the recognition results, especially if additional features were provided to the classifier. Potential additional features include quantities extracted from upper limb kinematics and other eye tracker data, such as 3D gaze saliency maps.

As with the processing of any sensor data, there are trade-offs with speed and accuracy in both the spatial and temporal domains. In its current instantiation, the gaze object sequence contains rich temporal information, but at the loss of spatial resolution; entire objects are considered rather than particular regions of objects. By contrast, the 3D gaze saliency map and gaze object percentage contain rich spatial information, but at the loss of temporal resolution due to the convolution of eye tracker data over a lengthy period of time. For practical purposes, we are not suggesting that spatial and temporal resolution should be maximized. In practice, an action recognition system need not be computationally burdened with the processing of individual points in a 3D point cloud or unnecessarily high sampling frequencies. However, one could increase spatial resolution by segmenting objects into affordance-based regions [64], or increase temporal resolution by considering the temporal dynamics of action units rather than subtasks.

While object recognition from 2D egocentric cameras is an important problem, solving this problem was not the focus of the present study. As such, we bypassed challenges of 2D image analysis such as scene segmentation and object recognition, and used a marker-based motion capture system to track each known object in 3D. Data collection was performed in a laboratory setting with expensive eye tracker and motion capture equipment. Nonetheless, the core concepts presented in this work could be applied in non-laboratory settings using low-cost equipment such as consumer-grade eye trackers, Kinect RGB-D cameras, and fiducial markers (e.g., AprilTags and RFID tags).

2.4.5 Conclusion

The long-term objective of the work is to advance human-robot collaboration by (i) facilitating the intuitive, gaze-based control of robots and (ii) enabling robots to recognize human actions, infer human intent, and plan actions that support human goals. To this end, the objective of this study was to identify useful features that can be extracted from 3D gaze behavior and used as inputs to machine learning algorithms for human action recognition. We investigated human gaze behavior and gaze-object interactions in 3D during the performance of a bimanual, iADL: the preparation of a powdered drink. Gaze fixation duration was statistically significantly larger for some action verbs, suggesting that some actions such as pouring and stirring may require increased visual attention for task completion. 3D gaze saliency maps, generated with high spatial resolution for six subtasks, appeared to encode action-relevant information at the subtask and action unit levels. Dynamic time warping barycentric averaging was used to create a population-based set of characteristic gaze object sequences that accounted for intra- and inter-subject variability. The gaze object sequence was then used to demonstrate the feasibility of a simple action recognition algorithm that utilized a DTW Euclidean distance metric. Action recognition results (96.4% accuracy, 89.5% precision, and 89.2% recall averaged over the six subtasks), suggest that the gaze object sequence is a promising feature for action recognition whose impact could be enhanced through the use of sophisticated machine learning classifiers and algorithmic improvements for real-time implementation. Future work includes the development of a comprehensive action recognition algorithm that simultaneously leverages features from 3D gaze–object interactions, upper limb kinematics, and hand–object spatial relationships. Robots capable of robust, real-time recognition of human actions during manipulation tasks could be used to improve quality of life in the home as well as quality of work in industrial environments.

CHAPTER 3

Toward Shared Autonomy Control Schemes for Human-Robot Systems: Action Primitive Recognition Using Eye Gaze Features

This chapter was based on work published in the journal Frontiers in Neurorobotics [65].

3.1 Abstract

The functional independence of individuals with upper limb impairment could be enhanced by teleoperated robots that can assist with activities of daily living. However, robot control is not always intuitive for the operator. In this work, eye gaze was leveraged as a natural way to infer human intent and advance action recognition for shared autonomy control schemes. We introduced a classifier structure for recognizing low-level action primitives that incorporates novel three-dimensional gaze-related features. We defined an action primitive as a triplet comprised of a verb, target object, and hand object. A recurrent neural network was trained to recognize a verb and target object, and was tested on three different activities. For a representative activity (making a powdered drink), the average recognition accuracy was 77% for the verb and 83% for the target object. Using a non-specific approach to classifying and indexing objects in the workspace, we observed a modest level of

generalizability of the action primitive classifier across activities, including those for which the classifier was not trained. The novel input features of gaze object angle and its rate of change were especially useful for accurately recognizing action primitives and reducing the observational latency of the classifier.

3.2 Introduction

Activities of daily living (ADLs) can be challenging for individuals with upper limb impairment. The use of assistive robotic arms is an active area of research, with the aim of increasing an individual's functional independence [66]. However, current assistive robotic arms, such as the Kinova arm and Manus arm, are controlled by joysticks that require operators to frequently switch between several modes for the gripper, including a position mode, an orientation mode, and an open/close mode [67, 68].Users need to operate the arm from the gripper's perspective, in an unintuitive Cartesian coordinate space. Operators would greatly benefit from a control interface with a lower cognitive burden that can accurately and robustly inference human intent.

The long-term objective of this work is to advance shared autonomy control schemes so that individuals with upper limb impairment can more naturally control robots that assist with activities of daily living. Toward this end, the short-term goal of this study is to advance the use of eye gaze for action recognition. Our approach is to develop a neural-network based algorithm that exploits eye gaze-based information to recognize action primitives that could be used as modular, generalizable building blocks for more complex behaviors. We define new gaze-based features and show that they increase recognition accuracy and decrease the observational latency [69] of the classifier.

This article is organized as follows. Section Related Work outlines related work with respect to user interfaces for assistive robot arms and action recognition methods. Section Materials and Methods introduces the experimental protocol and proposed structure of an action primitive recognition model, whose performance is detailed in section Results. Section Discussion addresses the effects of input features on classifier performance and considerations for future real-time implementation. Contributions are summarized in section Conclusion.

3.3 Related Work

3.3.1 User Interfaces for Assistive Robot Arms

Many types of non-verbal user interfaces have been developed for controlling assistive robot arms that rely on a variety of input signals, such as electrocorticographic (ECoG) [70], gestures ([71]), electromyography (EMG) ([72]), and electroencephalography (EEG) [73, 74]. Although ECoG has been mapped to continuous, high-DOF hand and arm motion [75, 76], a disadvantage is that an invasive surgical procedure is required. Gesture-based interfaces often require that operators memorize mappings from specific hand postures to robot behaviors [71, 77, 78], which is not natural. EMG and EEG-based interfaces, although non-invasive and intuitive, require users to don and doff EMG electrodes or an EEG cap, which may be inconvenient and require a daily recalibration.

In this work, we consider eye gaze-based interfaces, which offer a number of advantages. Eye gaze is relatively easy to measure and can be incorporated into a user interface that is non-verbal, non-invasive, and intuitive. In addition, with this type of interface, it may be possible to recognize an operator's intent in advance, as gaze typically precedes hand motions [79].

Numerous studies have reported on the use of eye gaze for robot control. In the early 2000's, the eyetracker was used as a direct substitute for a handheld mouse such that the gaze point on a computer display designates the cursor's position, and blinks function as button clicks ([80, 81]). Since 2015, eye gaze has been used to communicate a 3D target position ([8, 82, 83, 84, 85, 86]) for directing the movement of the robotic end effector. No action recognition was required, as these methods assumed specific actions in advance, such as reach and grasp ([22]), write and draw ([83]), and pick and place ([85]). Recently, eye gaze has been used to recognize an action from an a priori list. For instance, Shafti et al. developed an assistive robotic system that recognized subjects' intended actions (including reach to grasp, reach to drop, and reach to pour) using a finite state machine ([87]).

In this work, we advance the use of eye gaze for action recognition. We believe that eye gaze control of robots is promising due to the non-verbal nature of the interface, the rich information that can be extracted from eye gaze, and the low cognitive burden on the operator during tracking of natural eye movements.

3.3.2 Action Representation and Recognition

Moeslund et al. described human behaviors as a composition of three hierarchical levels: (i) activities, (ii) actions, and (iii) action primitives [88]. At the highest level, activities involve a number of actions and interactions with objects. In turn, each action is comprised of a set of action primitives. For example, the activity "making a cup of tea" is comprised of a series of actions, such as "move the kettle to the stove." This specific action can be further divided into three action primitives: "dominant hand reaches for the kettle," "dominant hand moves the kettle to the stove," and "dominant hand sets down the kettle onto the stove."

A great body of computer vision-based studies has already contributed to the recognition of activities of daily living such as walk, run, wave, eat, and drink [89, 90, 91, 92]. These studies detected joint locations and joint angles as input features from external RGB-D cameras and classified ADLs using algorithms such as hidden Markov models (HMMs) and recurrent neural networks (RNNs).

Other studies leveraged egocentric videos taken by head-mounted cameras or eyetrackers ([27, 23, 28, 24, 29, 30, 31, 93, 94]). Video preprocessing methods necessitated first subtracting the foreground and then detecting human hands and activity-relevant objects. Multiple features related to hands, objects, and gaze were then used as inputs for the action recognition using approaches such as HMMs, neural networks, and support vector machines (SVMs). Hand-related features included hand pose, hand location, relationship between left and right hand, and the optical flow field associated with the hand ([28, 94]). Object-related features included pairwise spatial relationships between objects ([29]), state changes of an object (open vs. closed) ([30]), and the optical flow field associated with objects ([28]). The "visually regarded object," defined by [23] as the object being fixated by the eyes, was widely used as the gaze-related feature ([27, 23, 31]). Some studies additionally extracted features such as color and texture near the visually regarded object ([24, 93]).

Due to several limitations, state-of-the-art action recognition methods cannot be directly applied to the intuitive control of an assistive robot via eye gaze. First, computer vision-based approaches to the automated recognition of ADLs have focused on the activity and action levels according to Moeslund's description of action hierarchy ([88]). Yet, state-of-the-art robots are not sophisticated enough to autonomously plan and perform these high-level behaviors. Second, eye movements are traditionally used to estimate gaze point or gaze object alone ([27, 23, 31]). More work could be done to extract other useful features from spatiotemporal eye gaze data, such as time histories of gaze object angle and gaze object angular speed, which are further described in section Gaze-Related Quantities.

3.4 Materials and Methods

3.4.1 Experimental Set-Up

This study was approved by the UCLA Institutional Review Board. The experimental setup and protocol were previously reported in our prior paper ([17]). Data from 10 subjects are reported [nine males, one female; aged 18–28 years; two pure right-handers, six mixed right-handers, two neutral, per a handedness assessment [95] based on the Edinburgh Handedness Inventory [35].Subjects were instructed to perform three bimanual activities involving everyday objects and actions: make instant coffee, make a powdered drink, and prepare a cleaning sponge (Figure 1). The objects involved in these three activities were selected from the benchmark Yale-CMU-Berkeley (YCB) Object Set [36]. We refer to these objects as activity-relevant objects since they would be grasped and manipulated as subjects performed specific activities.

For Activity 1, subjects removed a pitcher lid, stirred the water in the pitcher, and transferred the water to a mug using two different methods (scooping with a spoon and pouring). For Activity 2, subjects were instructed to remove a coffee can lid, scoop instant coffee mix into a mug, and pour water from a pitcher into the mug. For Activity 3, subjects unscrewed a spray bottle cap, poured water from the bottle into a mug, sprayed the water onto a sponge, and screwed the cap back onto the bottle. In order to standardize the instructions provided to subjects, the experimental procedures were demonstrated via a prerecorded video. Each activity was repeated by the subject four times; the experimental setup was reset prior to each new trial.

A head-mounted eyetracker (ETL-500, ISCAN, Inc., Woburn, MA, USA) was used to track the subject's gaze point at 60 Hz with respect to a built-in egocentric scene camera. Per calibration data, the accuracy and precision of the eyetracker were 1.4 deg and 0.1 deg, respectively. The motion of the YCB objects, eyetracker, and each subject's upper limb were tracked at 100 Hz by six motion capture cameras (T-Series, Vicon, Culver City, CA, USA). A blackout curtain surrounded the subject's field of view in order to minimize visual distractions. A representative experimental trial is shown in Supplementary Video 1.

3.4.2 Gaze-Related Quantities

We extract four types of gaze-related quantities from natural eye movements as subjects performed Activities 1–3. The quantities include the gaze object (GO) ([27, 23, 31]) and gaze object sequence (GOS) ([17]). This section describes how these quantities are defined and constructed. As described in section Input Features for the Action Primitive Recognition Model, these gaze-related quantities are used as inputs to a long-short term memory (LSTM) recurrent neural network in order to recognize action primitives.

The raw data we obtain from the eyetracker is a set of 2D pixel coordinates. The coordinates represent the perspective projection of a subject's gaze point onto the image plane of the eyetracker's egocentric scene camera. In order to convert the 2D pixel coordinate into a 3D gaze vector, we use camera calibration parameters determined using a traditional chessboard calibration procedure ([96]) and the MATLAB Camera Calibration Toolbox ([40]). The 3D gaze vector is constructed by connecting the origin of the egocentric camera frame with the gaze point location in the 2D image plane that is now expressed in the 3D global reference frame.

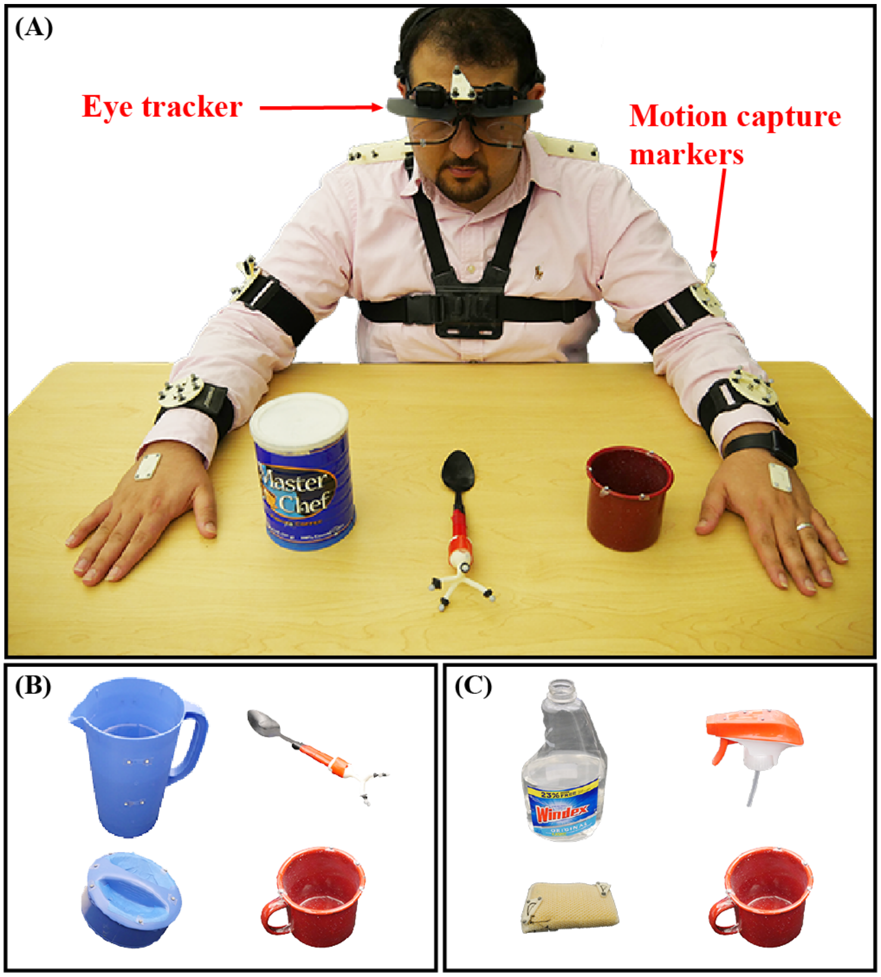


Figure 3.1: (A) A subject prepares to perform Activity 2 (make instant coffee) while eye gaze and kinematics are tracked with a head-mounted eyetracker and motion capture system (not shown). Activity 2 involves a coffee can, spoon and mug. (B) Activity 1 (make a powdered drink) involves a coffee can, spoon and mug. (C) Activity 3 (prepare a cleaning sponge) involves a spray bottle and cap, sponge, and mug. The subject shown in panel (A) has approved of the publication of this image.

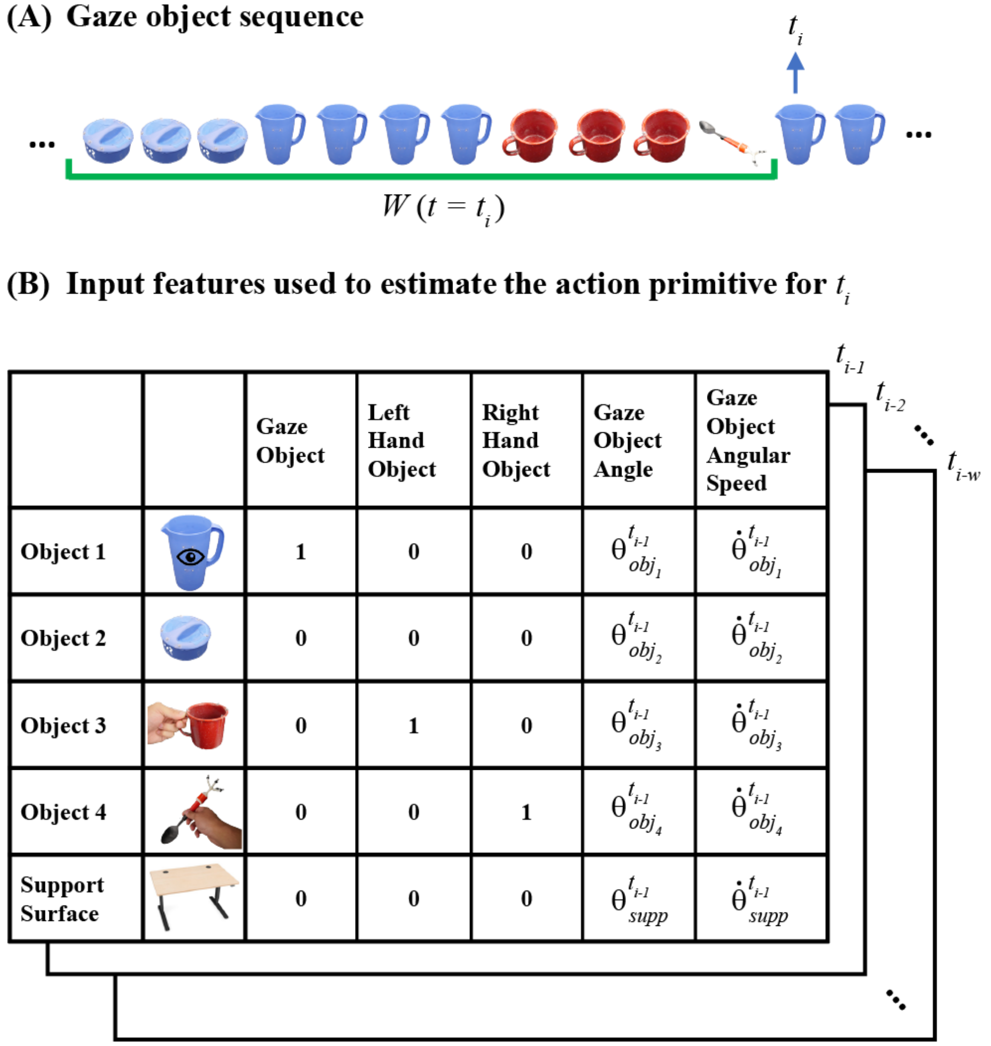
The gaze object (GO) is defined as the first object to be intersected by the 3D gaze vector, as the gaze vector emanates from the subject. Thus, if the gaze vector pierces numerous objects, then the object that is closest to the origin of the 3D gaze vector (within the head-mounted eyetracker) is labeled as the gaze object.

As defined in our prior paper, the gaze object sequence (GOS) refers to the identity of the gaze objects in concert with the sequence in which the gaze objects are visually regarded ([17]). Specifically, the gaze object sequence time history $GOS(t_i)$ is comprised of a sequence of gaze objects sampled at 60 Hz within a given window of time $W(t_i)$ (Figure 2). The time window $W(t_i)$ contains w time steps from t_{i-w} to t_{i-1}.

In this work, we use a value of w = 75 time steps, equivalent to 1.25 s. This time window size was determined from a pilot study whose results are presented in section Effect of Time Window Size on Recognition Accuracy. The pilot study was motivated by the work of Haseeb et al. in which the accuracy of an LSTM RNN was affected by time window size ([97]).

The gaze object angle (GOA) describes the spatial relationship between the gaze vector and each gaze object. The GOA is defined as the angle between the gaze vector and the eye-object vector (Figure 3). The eye-object vector shares the same origin as the gaze vector but ends at an object's center of mass. Each object's center of mass was estimated by averaging the 3D coordinates of the points in the object's point cloud. Each object's point cloud was scanned with a structured-light 3D scanner (Structure Sensor, Occipital, Inc., CA, USA) and custom turntable apparatus. Containers, such as the pitcher and mug, are assumed to be empty for center of mass estimation.

The gaze object angular speed (GOAS) is calculated by taking the time derivative of the GOA. We use the GOAS to measure how the gaze vector moves with respect to other activity-relevant objects. Previously, the gaze object and gaze object sequence have been



		Gaze Object	Left Hand Object	Right Hand Object	Gaze Object Angle	Gaze Object Angular Speed
Object 1		1	0	0	$\theta^{t_{i-1}}_{obj_1}$	$\dot{\theta}^{t_{i-1}}_{obj_1}$
Object 2		0	0	0	$\theta^{t_{i-1}}_{obj_2}$	$\dot{\theta}^{t_{i-1}}_{obj_2}$
Object 3		0	1	0	$\theta^{t_{i-1}}_{obj_3}$	$\dot{\theta}^{t_{i-1}}_{obj_3}$
Object 4		0	0	1	$\theta^{t_{i-1}}_{obj_4}$	$\dot{\theta}^{t_{i-1}}_{obj_4}$
Support Surface		0	0	0	$\theta^{t_{i-1}}_{supp}$	$\dot{\theta}^{t_{i-1}}_{supp}$

Figure 3.2: (A) The gaze object sequence time history $GOS(t_i)$ within a window of time $W(t_i)$ (green bracket) is shown for Activity 1 (make a powdered drink). (B) To predict the action primitive at time step t_i, input feature vectors (shown as 5×5 matrices for clarity) are created for each of the times from t_{i-w} to t_{i-1}. Activity-relevant objects are sorted according to their frequency of occurrence in the $GOS(t_i)$.

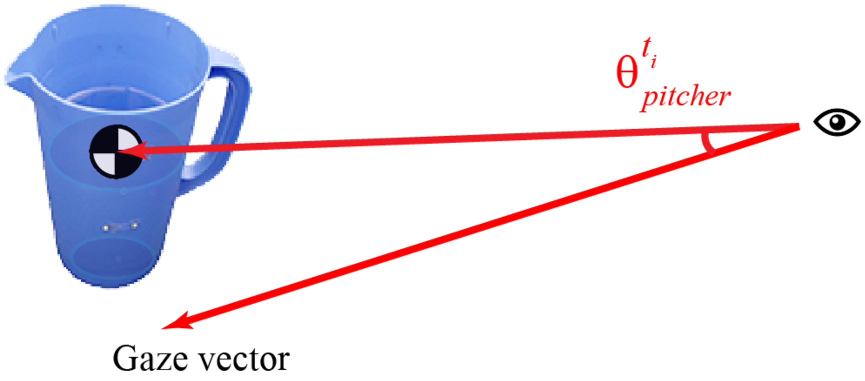


Figure 3.3: Gaze object angle is defined as the angle between the gaze vector and the eye-object vector (ending at the object's center of mass).

used to recognize actions ([23, 31]). To our knowledge, this is the first work to leverage the gaze object angle and gaze object angular speed for action primitive recognition.

3.4.3 Action Primitive Recognition Model

3.4.3.1 Action Primitive Representation

We represent each action primitive as a triplet comprised of a verb, target object (TO), and hand object (HO). Each action primitive can be performed by either the dominant hand or non-dominant hand. When both hands are active at the same time, hand-specific action primitives can occur concurrently.

The verb can be one of four classes: Reach, Move, Set down, or Manipulate. The classes Reach, Move, and Set down describe hand movements toward an object or support surface, with or without an object in the hand. Notably, these verbs are not related to or dependent upon object identity. In contrast, the class Manipulate includes a list of verbs that are highly related to object-specific affordances ([50]). For instance, in Activity 1, the verb "scoop" and “stir” are closely associated with the object “spoon” (Table 1). We refer to these verbs as manipulate-type verbs.

In addition to a verb, the action primitive triplet includes the identity of two objects. The target object TO refers to the object that will be directly affected by verbs such as Reach, Move, Set down, and Manipulate. The hand object HO refers to the object that is currently grasped. For instance, when the dominant hand grasps a spoon and stirs inside a mug, the triplet of the action primitive for the dominant hand is: manipulate (verb), mug (TO), and spoon (HO). A hierarchical description of activities, actions, and action primitives for Activities 1–3 are presented in Table 1.

In order to develop a supervised machine learning model for action primitive recognition, we manually label each time step with the action primitive triplet for either the dominant or non-dominant hand. The label is annotated using video recorded by an egocentric scene camera mounted on the head-worn eyetracker. We annotate each time step with the triplet of a subject’s dominant hand as it is more likely the target of the subject’s attention. For instance, when the dominant hand (holding a spoon) and the non-dominant hand (holding a mug) move toward each other simultaneously, we label the action primitive as “move the spoon to the mug,” where the verb is “move” and the target object is “mug.” However, when the dominant hand is not performing any action primitive, we refer to the non-dominant hand instead. If neither hand is moving or manipulating an object, we exclude that time step from the RNN training process.

3.4.3.2 Input Features for the Action Primitive Recognition Model

Given that the identity of gaze objects will vary across activities, we substitute the specific identities of gaze objects with numerical indices. This is intended to improve the

Activities		Activity 1: make a powdered drink	Activity 2: make instant coffee	Activity 3: prepare a cleaning sponge
Actions		Remove pitcher lid Stir liquid inside pitcher Scoop liquid into mug Pour liquid into mug	Remove coffee can lid Scoop coffee insider can Transfer coffee into mug Stir liquid inside mug	Remove spray bottle cap Transfer cleanser into mug Close spray bottle cap Spray cleanser onto sponge
Action primitives	**Verb**	Reach, Move, Set down, Manipulate (open, close, stir, scoop, drop pour)	Reach, Move, Set down, Manipulate (open, close, stir, scoop, drop, pour)	Reach, Move, Set down, Manipulate (screw, unscrew, lift, pour, insert, spray)
	TO	Pitcher, pitcher lid, mug, spoon, table	Coffee can, coffee lid, mug, spoon, table	Spray bottle, spray cap, mug, sponge, table
	HO	Pitcher, pitcher lid, mug, spoon	Coffee can, coffee lid, mug, spoon,	Spray bottle, spray cap, cap, mug, sponge

Table 3.1: Each of three activities is divided into actions that are further decomposed into action primitives. Each action primitive is defined as a triplet comprised of a verb, target object (TO), and hand object (HO).

generalizability of our action primitive recognition algorithm across different activities. For each time step t_i, the n activity-relevant objects are sorted in descending order according to their frequency of occurrence in $GOS(t_i)$. Once sorted, the objects are indexed as Object 1 to Object n, such that Object 1 is the object that most frequently appears in the gaze object sequence at t_i. If two or more objects appear in the gaze object sequence with the same frequency, the object with the smaller gaze object angle is assigned the smaller numerical index, as it is aligned most closely to the gaze vector and will be treated preferentially.

Figure 2 exemplifies how activity-relevant objects in a gaze object sequence would be assigned indices at a specific time step t_i. The activity-relevant objects (n = 4) in Activity 1 were sorted according to their frequency of occurrence in $GOS(t_i)$, which is underlined by a green bracket in Figure 2A. Based on frequency of occurrence, the activity-relevant objects were indexed as follows: pitcher (Object 1), pitcher lid (Object 2), mug (Object 3), and spoon (Object 4).

We introduce here the idea of a "support surface," which could be a table, cupboard shelf, etc. In this work, we do not consider the support surface (experiment table) as an activity-relevant object, as it cannot be moved or manipulated and does not directly affect the performance of the activity. Nonetheless, the support surface still plays a key role in the action primitive recognition algorithm due to the strong connection with the verb Set down. In addition, the support surface frequently appears in the GOS.

To predict the action primitive at time step t_i, input feature vectors are created for each of the time steps from time t_{i-w} to t_{i-1}, as shown in Figure 2B. For Activity 1, each input feature vector consists of five features for each of four activity-relevant objects and a support surface. For clarity, each resulting 25×1 feature vector is shown as a five-by-five matrix in Figure 2B. Gaze object, left-hand object, and right-hand object are encoded in the form of one-hot vectors while gaze object angle and angular speed are scalar values.

Gaze object identity was included as an input feature because it supported action recognition in prior studies [27, 23, 31]. We included the hand object as an input feature although it is a component of the action primitive triplet that we seek to recognize. Considering the application of controlling a robotic arm through eye gaze, we expect the robotic system to determine an object's identity before it plans any movements with respect to the object. As a result, we assume that the hand object's identity is always accessible to the classification algorithm. We included the GOA and GOAS as input features because we hypothesized that spatiotemporal relationships between eye gaze and objects would be useful for action primitive recognition. The preprocessing pipeline for the input features is shown in Supplementary Video 1.

3.4.3.3 Action Primitive Recognition Model Architecture

We train a long short-term memory (LSTM) recurrent neural network to recognize the verb and the target object TO for each time step ti. With this supervised learning method, we take as inputs the feature vectors described in section Input Features for the Action Primitive Recognition Model. For the RNN output, we label each time step ti with a pair of elements from a discrete set of verbs and generic, indexed target objects:

$$Verb(t_i) \in V = \{Reach, Move, Setdown, Manipulate\} \tag{3.1}$$

$$TO(t_i) \in O = \{Object_1, Object_2, Object_3, Support\ surface\} \tag{3.2}$$

The target object class Object 4 was excluded from the model output since its usage accounted for ¡1% of the entire dataset. The four verb labels and four TO labels are combined as 16 distinct verb-TO pairs, which are then taken as output classes when we train the RNN.

$$(Verb(t_i, TO(t_i))) \in O \times V = (Reach, Object_1), ..., (Manipulate, Supportsurface) \quad (3.3)$$

As a result, verb-TO pairs that never occur during the training process, such as (Manipulate, Support surface), can be easily eliminated.

In order to evaluate the RNN's performance on the verb and target object individually, we split the verb-TO pairs after recognition. A softmax layer was used as the final layer of the RNN.

$$Verb(t_i) = argmax_{v \in V}(\sum_{o \in O} softmax(Verb(t_i = v, TO(t_i = o))) \quad (3.4)$$

$$TO(t_i) = argmax_{o \in O}(\sum_{v \in V} softmax(Verb(t_i = v, TO(t_i = o))) \quad (3.5)$$

The RNN was comprised of one LSTM layer, three dense layers, and one softmax layer. The LSTM contained 64 neurons and each of the three dense layers contained 30 neurons. The RNN was trained with an Adaptive Momentum Estimation Optimization (Adam), which was used to adapt the parameter learning rate [98]. A dropout rate of 0.3 was applied in order to reduce overfitting and improve model performance. The batch size and epoch number were set as 128 and 20, respectively. The RNN was built using the Keras API in Python with a TensorFlow (version 1.14) backend, and in the development environment of Jupyter Notebook.

Class imbalance is a well-known problem that can result in a classification bias toward the majority class [99]. Since our dataset was drawn from participants naturally performing activities, the training set of samples was not balanced among various verb and TO classes (see sample sizes in Figure 5). An imbalance in TO classes might also result from sorting and indexing the objects as described in section Input Features for the Action Primitive Recognition Model. For instance, Object 1 occurs most frequently in the GOS by definition. Thus, Object 1 is more likely to be the target object than Objects 2 or 3. In order to compensate for the class imbalance, each class' contribution in the cross-entropy loss function was weighted by its corresponding number of samples [100].

The temporal sequence of the target object and verb recognized by the RNN can contain abrupt changes, as shown in the top rows of Figures 5A,B. These abrupt changes occur for limited time instances and make the continuous model prediction unsmooth. Such unstable classifier results might cause an assistive robot to respond unexpectedly. Thus, we implemented a one-dimensional mode filter with an order of m (in our work, m = 12 time steps, equivalent to 0.2 s) to smooth out these sequences [101]:

$$verb(t_i) = mode(\{verb(t_{i-m}), verb(t_{i-m+1}), ..., verb(t_{i-1})\}) \tag{3.6}$$

$$TO(t_i) = mode(\{TO(t_{i-m}), TO(t_{i-m+1}), ..., TO(t_{i-1})\}) \tag{3.7}$$

The sequences after filtering are shown in the middle rows of Figures 5A,B.

Considering that 10 subjects participated in our study, we adopted a leave-one-out cross-validation method. That is, when one subject's data were reserved for testing, the other nine subjects' data were used for training.

3.4.3.4 Performance Metrics for Action Recognition

In order to evaluate the performance of the action primitive classification, we assessed overall accuracy, precision, recall, and the F1-score. Overall accuracy is the number of correctly classified samples divided by the total size of the dataset. For each class of verb or target object, precision represents the fraction of correctly recognized time steps that actually belong to the given class, and recall represents the fraction of the class that are successfully recognized. We use TP, TN, and FP to represent the number of true positives, true negatives, and false positives when classifying a verb or target object class.

$$overall\ accuracy = \frac{\sum TP}{total\ size\ of\ dataset} \tag{3.8}$$

$$precision = \frac{TP}{TP + FP} \tag{3.9}$$

$$recall = \frac{TP}{TP + TN} \tag{3.10}$$

The F1-score is the harmonic mean of precision and recall.

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{3.11}$$

We also used performance metrics that were related to the temporal nature of the data. In order to evaluate how early an action primitive was successfully recognized, we adopted the terminology "observational latency," as defined in [69]. The term was defined as "the difference between the time a subject begins the action and the time the classifier classifies the action," which translates to the amount of time that a correct prediction lags behind the start of an action primitive. It should be noted that the observational latency does not include the computation time that the recognition algorithm requires to preprocess the input data and recognize the actions by the model.

We conservatively judged the success of an action primitive's classification by checking whether more than 75% of its time period was predicted correctly. Summary statistics for observational latency are reported for action primitives that were deemed correct according to this 75% threshold. Observational latency is negative if the action primitive is predicted before it actually begins.

3.5 Results

Recall our aim of specifying the three components of the action primitive triplet: verb, target object, and hand object. Given that the hand object is already known, as described in section Input Features for the Action Primitive Recognition Model, we report on the ability of the RNN to recognize the verb and target object. A demonstration of the trained RNN is included in Supplementary Video 1.

3.5.1 Effect of Time Window Size on Recognition Accuracy

In order to set the time window size, we conducted a pilot study inspired by [97]. We tested how the F1-scores of the verb and TO classes varied as the time window size was increased from five time steps (equivalent to 83 ms) to 2 s in increments of five time steps (Figure 4). Considering the average duration of an action primitive was only 1.2 s, we did not consider time window sizes beyond 2 s.

As seen in Figure 4A, time window size had a more substantial effect on the recognition

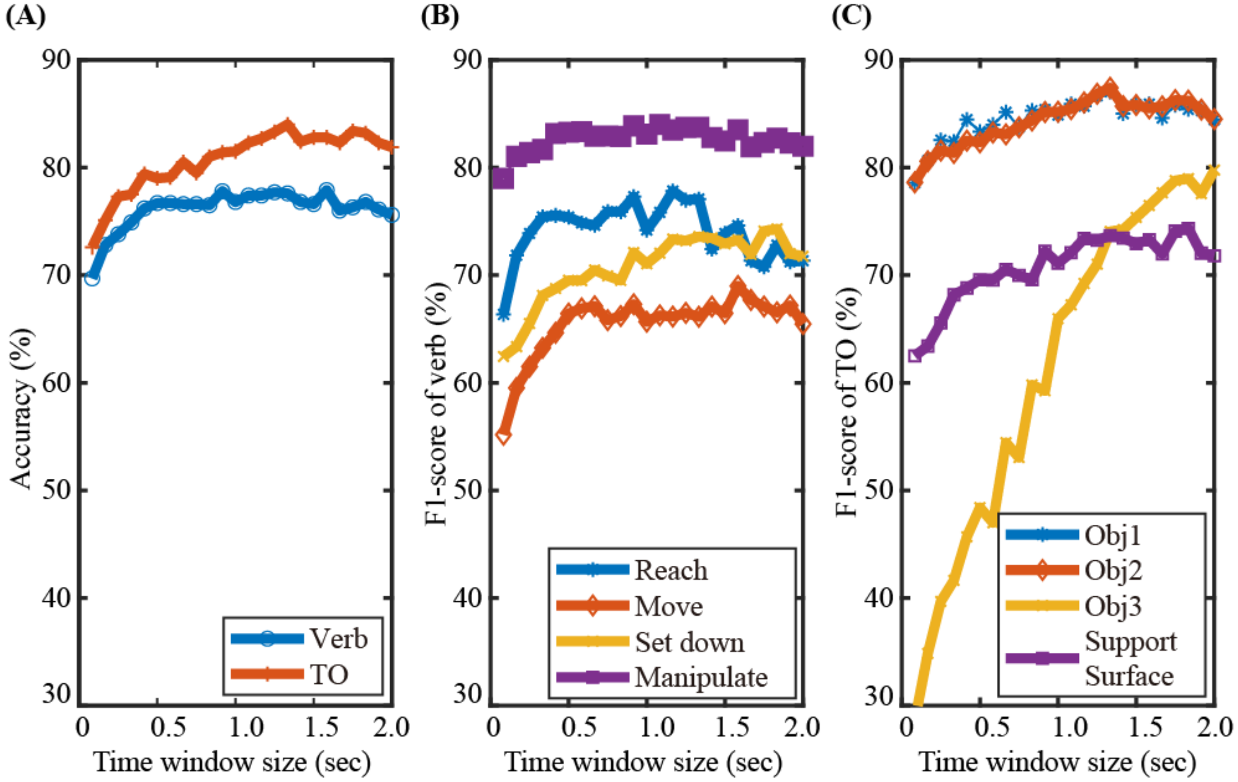
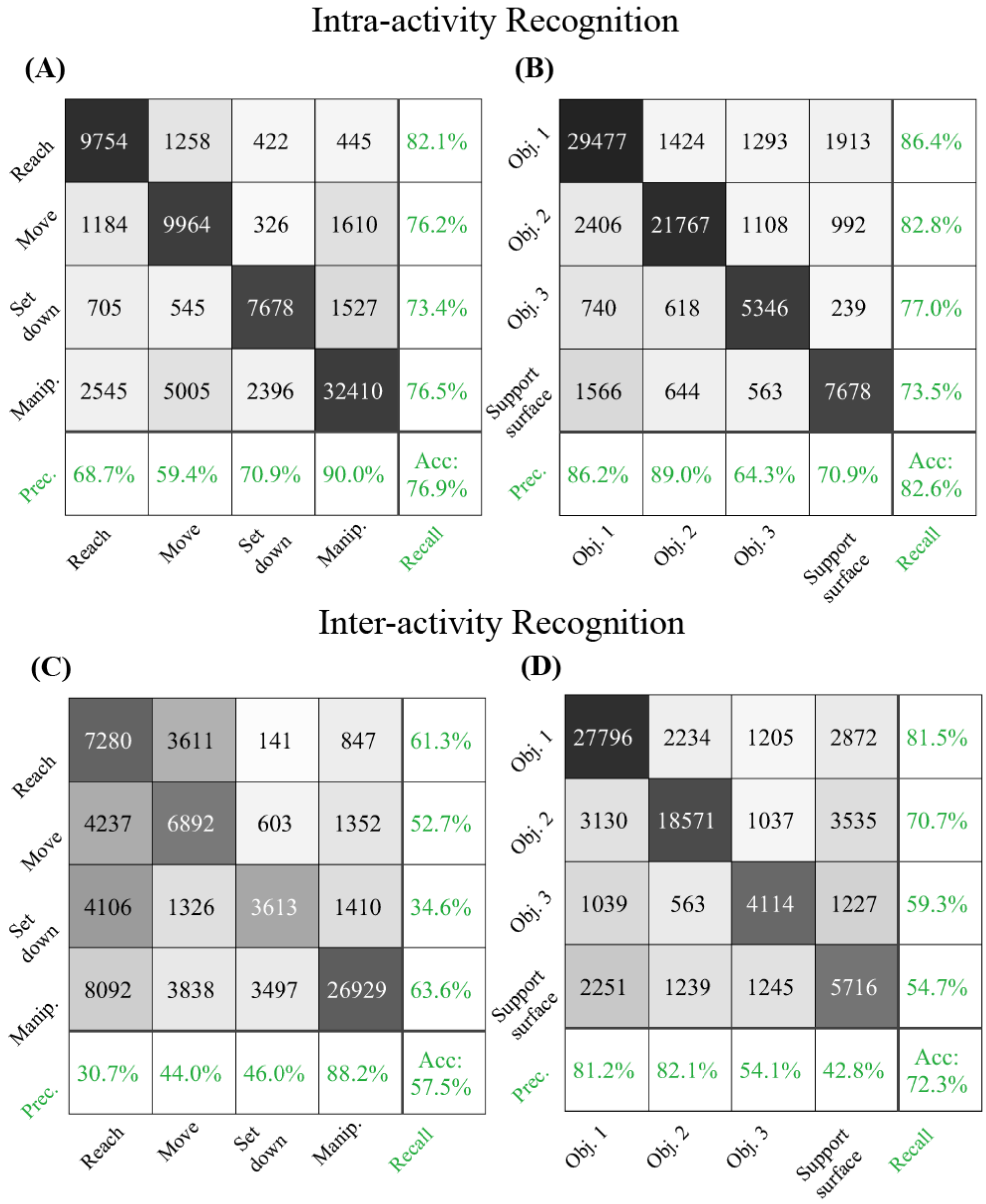


Figure 3.4: The effect of time window size (ranging from 83ms to 2 s) on recognition performance is shown for Activity 1. The overall recognition accuracy for verb and target object are shown in (A). F1-scores for the verb and target object classes are shown in (B,C), respectively.

of TO than that of verb. This is due to the fact that time window size can greatly affect the data sample distributions among target object classes as a result of sorting and indexing the activity-relevant objects. Figure 4C shows that the TO class Object 3 was especially sensitive to the window size. The corresponding F1-score continuously increased from 30% to 80% until the window size reached 1.8 s. Recognition performance of the other three TO classes Object 1, Object 2, and Support surface were also improved as the time-window size was increased from 80 ms to 1.25 s. The increased F1-scores of the TO classes can be partly attributed to alleviated class imbalance problem as the time window was lengthened, especially for the class Object 3. The number of data samples of Object 3 greatly increased due to the nature of sorting and indexing objects according to their frequency of occurrence in gaze object sequence.

As seen in Figure 4B, the F1-scores of the verb classes Reach, Move, and Manipulate increased as the time-window size increased from 80 ms to 0.5 s. Little improvement in the F1-scores was observed for time window sizes > 0.5 s, except for Set down. This suggested that a memory buffer of 0.5 s might be sufficient for predicting the verb class based on eye gaze. Gaze-related information collected long before the start of an action primitive was very likely to be irrelevant to the verb.

Considering the effect of the time window size on the classification accuracy of both the verb and target object (Figure 4), we decided to use a time window size of 1.25 s. A time window longer than 1.25 s might slightly improve recognition performance, but with additional computational cost.



Intra-activity Recognition

(A)

	Reach	Move	Set down	Manip.	Recall
Reach	9754	1258	422	445	82.1%
Move	1184	9964	326	1610	76.2%
Set down	705	545	7678	1527	73.4%
Manip.	2545	5005	2396	32410	76.5%
Prec.	68.7%	59.4%	70.9%	90.0%	Acc: 76.9%

(B)

	Obj. 1	Obj. 2	Obj. 3	Support surface	Recall
Obj. 1	29477	1424	1293	1913	86.4%
Obj. 2	2406	21767	1108	992	82.8%
Obj. 3	740	618	5346	239	77.0%
Support surface	1566	644	563	7678	73.5%
Prec.	86.2%	89.0%	64.3%	70.9%	Acc: 82.6%

Inter-activity Recognition

(C)

	Reach	Move	Set down	Manip.	Recall
Reach	7280	3611	141	847	61.3%
Move	4237	6892	603	1352	52.7%
Set down	4106	1326	3613	1410	34.6%
Manip.	8092	3838	3497	26929	63.6%
Prec.	30.7%	44.0%	46.0%	88.2%	Acc: 57.5%

(D)

	Obj. 1	Obj. 2	Obj. 3	Support surface	Recall
Obj. 1	27796	2234	1205	2872	81.5%
Obj. 2	3130	18571	1037	3535	70.7%
Obj. 3	1039	563	4114	1227	59.3%
Support surface	2251	1239	1245	5716	54.7%
Prec.	81.2%	82.1%	54.1%	42.8%	Acc: 72.3%

Figure 3.5: Intra-activity recognition results for Activity 1 are shown in confusion matrix form for (A) verb and (B) target object. Inter-activity recognition results for an RNN trained on Activity 2 and tested on Activity 1 are shown for (C) verb and (D) target object. Integers in the confusion matrices represent numbers of samples. The confusion matrices are augmented with precision, recall, and accuracy results (green).

3.5.2 Intra-Activity Recognition

We report results for intra-activity recognition, in which we trained and tested the recurrent neural network on the same activity. These results describe how well the RNN recognized novel instances of each activity despite variability inherent to activity repetition. Intra-activity recognition results for Activity 1 are shown in Figure 5 in the traditional form of confusion matrices. The rows correspond to the true class and the columns correspond to the predicted class. For brevity, intra-activity recognition results for Activities 1 and 2 are also shown in Table 2 in the form of F1-scores. The weighted averages of F1-scores for verb and target object were each calculated by taking into account the number of data samples for each class. The RNN was not trained on Activity 3 due to its smaller dataset as compared to Activities 1 and 2. Thus, no intra-activity recognition results were reported for Activity 3.

We augmented the traditional confusion matrix used to report results according to true and predicted classes with additional metrics of precision and recall (Figure 5). Precision and recall were reported as percentages (in green) in the far right column and bottom-most row, respectively. The cell in the lower-right corner represented the overall recognition accuracy.

The data samples were not balanced among various verb and TO classes since our dataset was drawn from participants naturally performing activities. The proportion of each verb and TO class in Activity 1 was the sum of the corresponding row in Figures 5A,B divided by the total size of the dataset (77,774 time step samples). The proportions for the verb classes were 15% for Reach, 17% for Move, 13% for Set down, and 55% for Manipulate. The proportions for the target object classes were 44% for Object 1, 34% for Object 2, 9% for Object 3, and 13% for Support surface.

Intra- or Inter-activity recognition	**Intra**	**Inter**	**Inter**	**Intra**	**Inter**	**Inter**
Activity # (training)	1	1	1	2	2	2
Activity # (testing)	1	2	3	2	1	3
F1-scores for verb recognition (%)						
Reach	74.8	52.9	54.8	56.5	40.9	55.6
Move	66.8	36.6	61.1	59.5	48.0	60.5
Set down	72.1	49.3	45.3	59.6	39.5	44.4
Manipulate	82.7	73.7	72.7	81.4	73.9	71.8
Verb Average	77.4	60.3	63.6	68.6	59.9	63.1
F1-scores for target object recognition (%)						
Object 1	86.3	72.1	78.0	80.2	81.3	77.4
Object 2	85.8	80.7	83.6	87.2	76.0	80.8
Object 3	70.1	41.7	52.5	55.2	56.6	56.8
Support surface	72.2	56.9	49.8	69.3	48.0	46.6
TO Average	82.8	73.0	74.9	81.1	72.8	73.4

Table 3.2: The RNN performance for intra- and inter-activity recognition is reported via F1-scores (%). Weighted averages of F1-scores that account for the number of data samples in each class are reported for both verb and target object (TO).

The RNN achieved a good performance in recognizing the majority verb class Manipulate (precision: 90%, recall: 77%) and the TO class Object 1 (precision: 86%, recall: 86%), which laid a solid foundation for its overall accuracy (verb: 77%, TO: 83%).

3.5.3 Inter-Activity Recognition

We report results for inter-activity recognition, in which we trained and tested the recurrent neural network on different activities. These results describe how well the RNN can recognize verbs and target objects despite variability across different activities. To evaluate the algorithm's cross-activity generalizability, an RNN trained on Activity 2 (make instant coffee) was tested on Activity 1 (make a powdered drink), and vice versa. RNNs trained on Activity 1 and Activity 2 were additionally tested on Activity 3 (prepare a cleaning sponge). The confusion matrices of an RNN trained on Activity 2 and tested on Activity 1 are shown in Figures 5C,D for verb and target object estimation, respectively. For brevity, additional inter-activity recognition results are presented in Table 2 in the form of F1 scores.

We also compared intra-activity and inter-activity performance of RNN models tested on the same activity. For this, we subtracted the average F1-scores for inter-activity recognition from those of the appropriate intra-activity recognition for RNNs tested on Activity 1 and Activity 2. As expected, when testing with an activity that differed from the activity on which the RNN was trained, the classification performance decreased. The average F1-scores of verb and target object each dropped by 8% when the RNN was trained on Activity 1 and tested on Activity 2. The average F1-scores of verb and target object dropped by 18 and 10%, respectively, when the RNN was trained on Activity 2 and tested on Activity 1. The average F1-score decreases were no larger than 20%, which suggested that the classification algorithm was able to generalize across activities to some degree. In addition, despite the

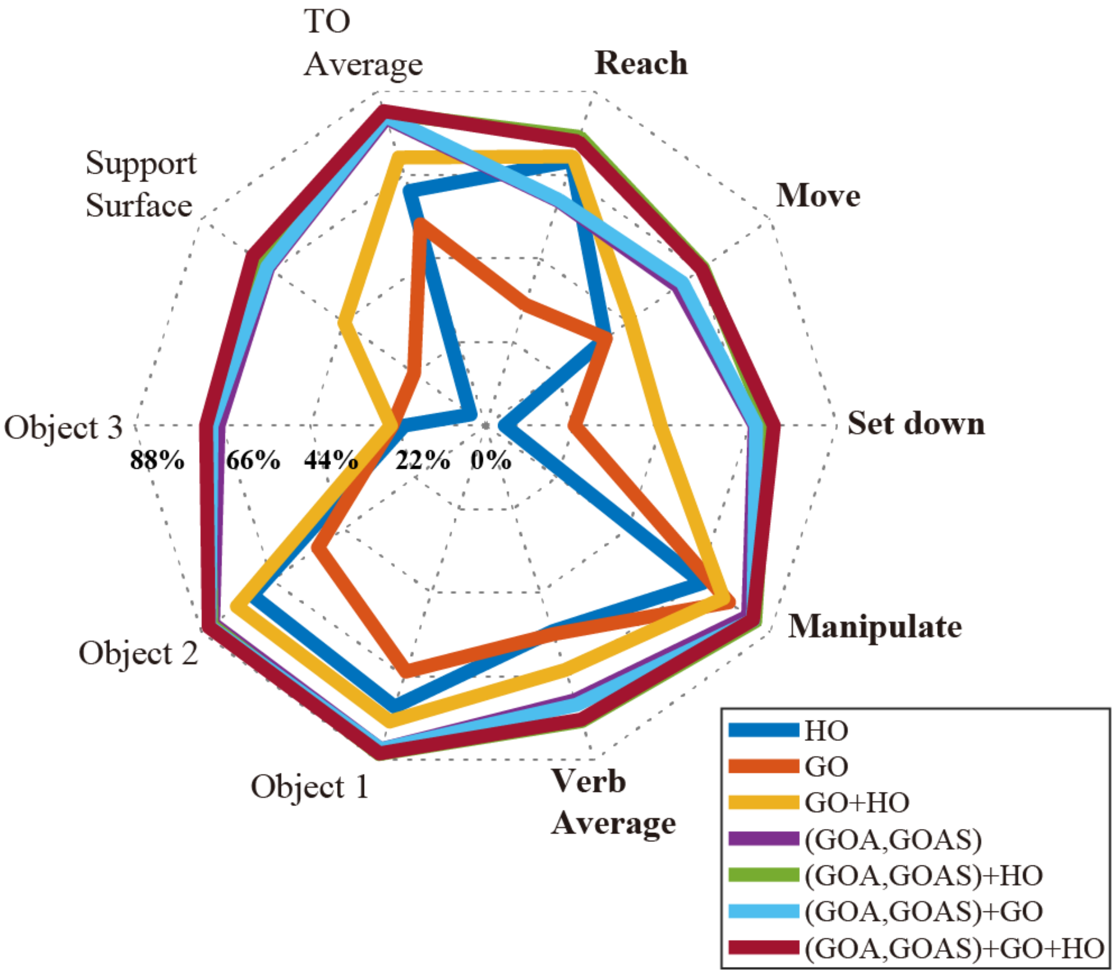


Figure 3.6: For Activity 1, RNN performance is reported by F1-scores for different combinations of input features (HO, GO, GOA, GOAS) using a radar chart. Axes represent the verb (bold) and target object classes. F1-score gridlines are offset by 22%. Each of the polygons corresponds to one combination of input features. The combined use of HO, GO, GOA, and GOAS features resulted in the best performance; HO alone performed the worst.

fact that Activity 3 shared only one common activity-relevant object (mug) with the other two activities, the average F1-scores of verb and TO achieved for Activity 3 were slightly higher than those of the other inter-activity recognition tests (Table 2).

3.5.4 Effect of Input Features on Recognition Accuracy

In order to evaluate feature importance, we compared the classification performance achieved in Activity 1 with various combinations of input features using a radar chart (Figure 6). Axes represented the verb and target object classes. Gridlines marked F1-scores in increments of 22%. Classification using HO alone was poor, with F1-scores for "Set down" and "Object 3" being ¡10%. Only slightly better, classification using GO alone was still not effective, with F1-scores of the "Set down," "Object 3," and "Support surface" only reaching values near 22%. In contrast, GOA-based features (GOA, GOAS) alone outperformed both HO and GO on their own in every verb and target object class. With the exception of "Reach," GOA-based features alone also outperformed the use of HO and GO together.

Although the feature HO alone did not provide good recognition result, it could substantially improve the classification performance when used in concert with GOA-based features. For every class, the F1-scores achieved with the combination of GOA-based feature and HO were equal to or higher than with the GOA-based feature alone.

3.5.5 Effect of Input Features on Observational Latency

The time histories of the verb and target object recognition for a representative Activity 1 trial are shown in Figures 7A,B. In each of Figures 7A,B, the top colorbar represents a time history of raw prediction results. The middle colorbar shows the output of the mode filter that smooths the raw prediction results. The bottom colorbar represents the ground truth. White gaps in the ground truth correspond to instances when neither hand was moving or manipulating an object. The observational latency is obtained by comparing the middle and bottom colorbars.

While Figure 7 shows the observational latency for a single representative trial, the observational latencies for all trials and participants are presented in Figure 8. Specifically, Figures 8A,B, summarize results for the recognition of verb and target object, respectively, for an RNN trained and tested on Activity 1. Figure 8 illustrates the effect of input features on observational latency by comparing the results of an RNN that only used GO and HO as input features to those of an RNN that additionally used GOA, and GOAS as input features.

We hypothesized that the incorporation of GOA-based input features could significantly decrease observational latency. To test this, we conducted a Wilcoxon signed-rank test (following a Lilliefors test for normality) with a total of 714 action primitives. The one-tailed p-values for the verbs and target objects were all less than the α level of 0.05 except for the target object of pitcher lid. Thus, we concluded that the use of GOA and GOAS as input features in addition to GO and HO resulted in a reduction in observational latency (Figure 8).

3.6 Discussion

3.6.1 Features Based on Gaze Object Angle Improve Action Primitive Recognition Accuracy

The long-term objective of this work is to advance shared autonomy control schemes so that individuals with upper limb impairment can more naturally control robots that assist with activities of daily living. One embodiment of such a teleoperated system could include both a joystick and eyetracker as user input devices. The short-term goal of this study was to improve action primitive recognition accuracy and observational latency. We pursued this goal by (i) focusing on the recognition of low-level action primitives, and (ii) defining eye gaze-based input features that improve action primitive recognition.

Previous studies leveraged egocentric videos to recognize actions when a subject was naturally performing ADLs. The features reported in these studies can be divided into three categories: features based on human hands, objects, or human gaze. Examples of hand-based features include hand location, hand pose, and relative location between left and right hands [28, 94]. Fathi et al. relied on changes in the state of objects, such as the state of the "coffee jar" (open vs. closed) [30], to recognize actions. Behera et al. used spatiotemporal relationships between objects as classifier inputs ([29]). Features related to human gaze included the gaze-object, which was widely used to classify actions ([23, 31]). The use of object appearance (histogram of color and texture) in the neighborhood of the gaze point was also effective in improving recognition accuracy ([24, 93]).

Considering the long-term objective of this work, we elected not to rely solely on features based on human hands or objects for action primitive recognition. Features based on human hands are only available when subjects use their own hands to directly grasp and manipulate objects. For the assistive robot application we envision, features of human hands such as hand location, hand pose, and relative location between left and right hands ([28, 94]) will not be available. Features based on objects are consequence of hand motions, such as changes in the states of objects or spatiotemporal relationships between objects. Such object-based features would only be available in hindsight and cannot be collected early enough to be useful for the proposed assistive robot application.

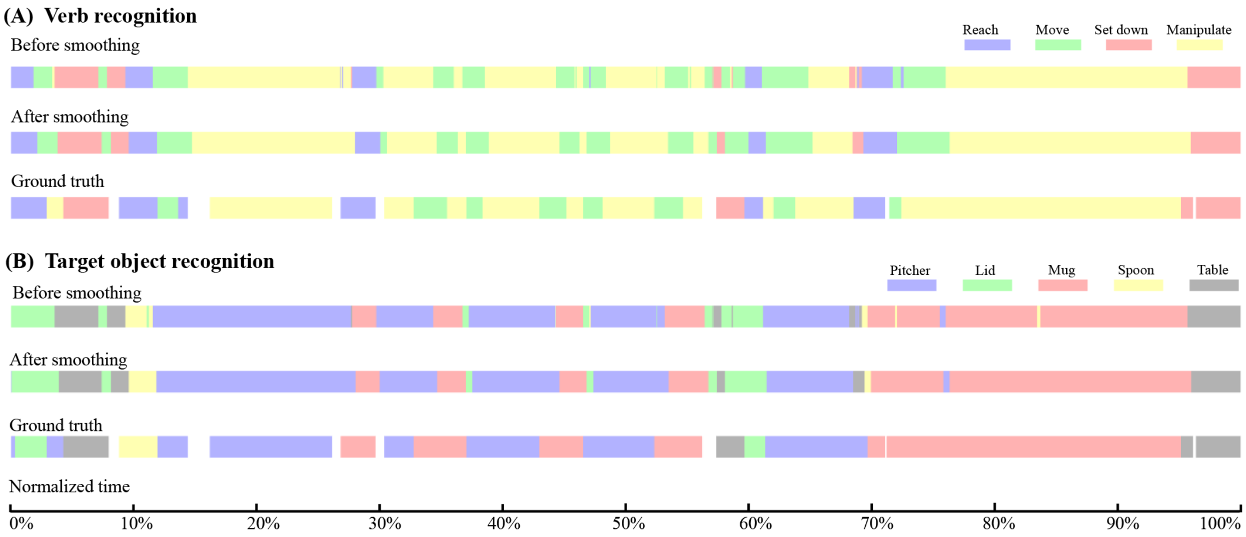
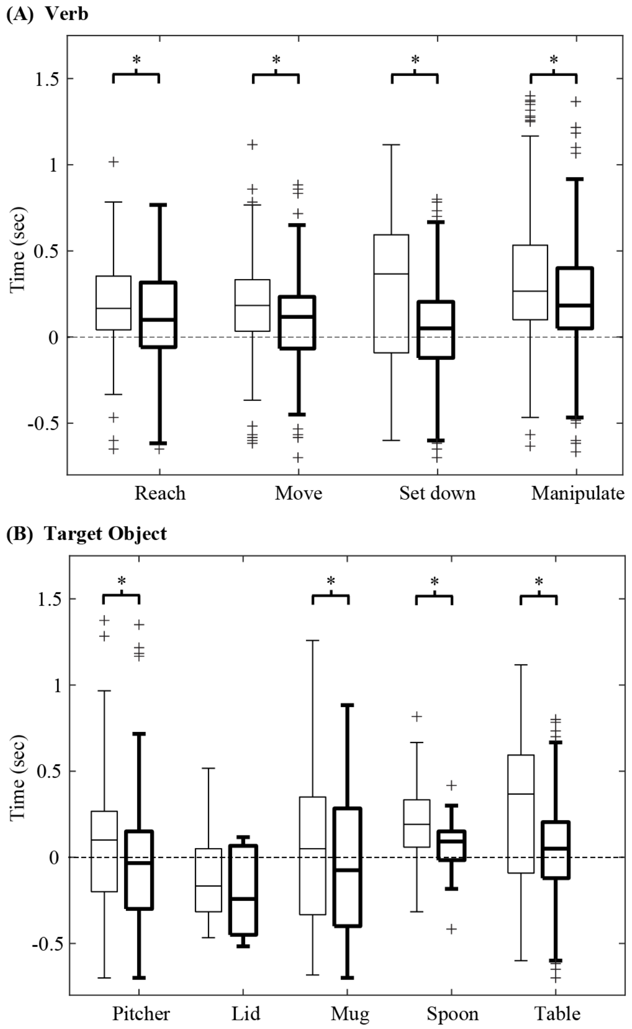


Figure 3.7: For a representative trial of Activity 1, temporal sequences of recognition results and ground truth are presented for (A) verb and (B) target object. In both (A,B), the top, middle, and bottom color bars represent the raw RNN output, RNN output smoothed by a mode filter, and hand-labeled ground truth, respectively. The total duration of this trial is 36 s.



(A) Verb

(B) Target Object

Figure 3.8: For Activity 1, the observational latency for recognition of (A) verb and (B) target object are shown using box and whisker plots. A negative latency value indicates that a verb or target object is identified before the start of the action primitive. For each boxplot pair, the observational latency without using GOA and GOAS (thin lines) is compared with that using GOA and GOAS (thick lines). Each boxplot indicates the 25, 50, and 75th percentiles. The whiskers extend to the most extreme data points that are not considered outliers ("+") having values of more than 1.5 times the interquartile range from the top or bottom of the box. Asterisks indicate $p < \alpha = 0.05$.

We aim to exploit observations that gaze behavior is a critical component of sighted grasp and manipulation activities, and that eye movements precede hand movements ([2, 102]). As such, we adopted the gaze-based feature GO from the literature (e.g., [23]) and supplemented it with two new features that we defined: GOA and GOAS.

As reported in section Effect of Input Features on Recognition Accuracy, models that included GOA and GOAS as input features outperformed models that relied primarily on GO or HO for every verb and target object class. The addition of GOA and GOAS substantially improved the average F1-score from 64% to 77% for verb and from 71 to 83% for target object (Figure 6).

The advantages of using features based on gaze object angle for action primitive recognition are 2-fold. First, the gaze object angle quantifies the spatiotemporal relationship between the gaze vector and every object in the workspace, including objects upon which the subject is not currently gazing. In contrast, the gaze object only captures the identity of the object upon which the subject is gazing at that particular instant. Considering that daily activities generally involve a variety of objects, it is vital for the classifier to collect sufficient information related to gaze-object interactions. The feature GOA could indirectly provide information similar to that of GO. For example, a GOA value that is close to zero would result if the gaze vector is essentially pointing at the gaze object. When GOA, GOAS, and HO have already been included as input features, the addition of GO as an input feature has little to no impact on classification accuracy (Figure 6). Also, classifier performance improves when using GOA and GOAS as input features as compared to using GO, HO, or their combination (Figure 6).

Second, the input feature GOAS contains GOA rate information. To some extent, GOAS also captures directional information, as positive and negative GOAS values reflect whether the gaze vector is approaching or departing from each object in the workspace, respectively. We believe that approach/departure information can be leveraged to predict the target object for a given action primitive because gaze is used to gather visual information for planning before and during manual activities ([102]). An object being approached by the gaze vector is not necessarily the target object, as the object could simply be in the path of the gaze vector during its movement. However, objects are less likely to be labeled as the "target object" when the gaze vector moves away from them.

3.6.2 Features Based on Gaze Object Angle Improve Observational Latency

While recognition accuracy is important, human-robot systems also require low observational latency ([69]). Even an action primitive that is correctly recognized 100% of the time will cease to be useful if the delay in recognition prohibits an effective response or adds to the cognitive burden of the operator. The earlier that a robotic system can infer the intent of the human operator or collaborator, the more time will be available for computation and the planning of appropriate robot movements.

Previous studies have focused on classifying actions in videos that have already been segmented in time (e.g., [24]. However, these methods that were designed to recognize actions in hindsight would be less effective for real-time use. We desire the intended action primitive to be predicted in advance of robot movement and with as low an observational latency as possible.

Hoffman proposed several metrics to evaluate fluency in human-robot collaborative tasks. For instance, the robot's functional delay was defined as the amount of time that the human spent waiting for the robot ([103]). This concept of fluency reflects how promptly a robot can respond correctly to an operator's commands. A high observational latency will degrade the fluency of a human-robot system and increase the operator's cognitive burden, effort, and frustration levels. A user interface that requires operators to intentionally gaze at specific objects or regions for a fixed period of time may be less natural and have lower fluency than a user interface that leverages natural eye gaze behaviors ([84, 85]).

In this work, the use of gaze-related features enabled the recognition of action primitives at an early stage. The average observational latency for verb recognition was 120 ms, 10% of the average duration of an action primitive (1.2 s). The average observational latency for target object was -50 ms; the negative latency value indicates that the target object was sometimes identified before the start of the action primitive. Unfortunately, pooled across all classes, the observational latency for the target object was not statistically significantly less than zero (p = 0.075; $\alpha = 0.05$). Nonetheless, the fact that some of the trials resulted in negative observational latency values was surprising and encouraging.

Among gaze-related input features, the use of GOA and GOAS decreased the observational latency as compared with using GO alone (Figure 8). Per a Wilcoxon signed rank test, observational latency was statistically significantly smaller when GOA and GOAS were used as input features than when they were excluded ($p < \alpha = 0.05$). This was true for all verb classes and all target object classes, with the exception of lid. For the verb and target object, the observational latency dropped by an average of 108 and 112 ms, respectively. One reason for this could be that GOA-based features may encode the tendency of the gaze vector to approach an object once the eyes start to move. In contrast, the GO feature does not capture the identity of any object until the gaze vector reaches the object.

The sub-second observational latency values that we report likely resulted from the fact that eye movement generally precedes hand movement for manual activities ([2, 102]). Land et al. reported that the gaze vector typically reached the next target object before any visible signs of hand movement during the activity of making tea ([54]). The small observational latency values may also result from the fact that our classifier was designed to recognize action primitives, which are much simpler than actions or activities ([88]). Action primitives often involve a single object, a single hand, and occur over a shorter period of time than actions and activities. The recognition of actions and activities for ADLs would require observations over a longer period of time and would necessarily involve more complex eye behaviors, more complex body movements, and gaze interactions with multiple objects.

Ryoo predicted activities of daily living and defined the “observation ratio” as the ratio between the observational latency and the activity duration ([104]). Ryoo reported that a minimum observation ratio of 45% was needed to classify activities with at least 60% accuracy. In this work, we found that minimum observation ratios of 18 and 5% were needed to achieve an accuracy of 60% for each the verb and the target object, respectively. This suggests that recognition of low-level action primitives can be achieved at lower observation ratios and within shorter time periods than high-level activities, which require the passage of more time and collection of more information for similar levels of accuracy.

One limitation of this work is that the action primitive recognition algorithm has not yet been tested in real-time. This is an area of future work and considerations for real-time implementation are discussed in section Comparisons to State-of-the-Art Recognition Algorithms. Based on our experience, we expect that the overall latency will be dominated by observational latency and less affected by computational latency. This is due to the relatively simple structure of the proposed RNN architecture and the fact that the RNN model would be trained offline a priori.

3.6.3 Segmenting Objects Into Regions According to Affordance Could Improve Recognition Performance

The distribution of gaze fixations can be concentrated on certain regions of an object, such as those associated with "object affordances." An object affordance describes actions that could be performed on an object ([50]). For example, Belardinelli et al. showed human subjects a 2D image of a teapot and instructed them to consider lifting, opening, or classifying the teapot as an object that could or could not hold fluid ([33]). It was observed that subjects' gaze fixations were focused on the teapot handle, lid, and spout for lifting, opening, and classifying, respectively. In addition, in a prior study, we reported 3D gaze heat maps for the activity "make a powdered drink" ([17]). We observed that gaze fixations were focused on the top and bottom of pitcher during the action unit "reach for pitcher" and "set down pitcher."

Inspired by these findings, we hypothesized that information about the action primitive can, in theory, be encoded by gaze behavior with respect to specific regions of objects. This would provide a classification algorithm with information at a finer spatial resolution than when considering each object as a whole. In a post hoc study, we segmented the point clouds of each of the four activity-relevant objects in Activity 1 (make a powdered drink) into several regions according to object affordances (Figure 9). For instance, the spoon was segmented into the upper and bottom faces for the bowl, the handle, and the tip of the handle. Notably, the inner and outer wall of containers (pitcher and mug) were treated as different regions since the inner and outer walls were often fixated upon differently depending on the action primitive.

After the segmentation, we augmented the gaze-related features (GO, GOA, GOAS) by treating each region as an independent object while keeping the features left-hand object and right-hand object unchanged. We then retrained the RNN with the new augmented features. The recognition accuracy for verb increased slightly from 77 to 79% and accuracy for the target object increased from 83 to 86%. By increasing the total number of object regions from 4 to 20, the time taken for the trained RNN to produce one classifier output increased by 26%. Depending on the consequences of an incorrect classification and the minimum acceptable accuracy level, one could decide which objects to segment and how finely the objects should be segmented. For instance, one may still be able to improve recognition performance if the mug were segmented into inner wall, outer wall, and handle, as opposed to the five segments that we tested.

3.6.4 Comparison to State-of-the-Art Recognition Algorithms

In the evaluation of our proposed gaze-based action primitive recognition method, we were unable to identify suitable benchmarks for a direct quantitative comparison. First, our approach is designed to recognize low-level action primitives that could be used as modular, generalizable building blocks for more complex levels of the action hierarchy ([88]). The literature on action recognition provides methods for recognition at the level of actions and activities, but not at the level of action primitives that are investigated in our work. For instance, the public dataset "GTEA+" and "EGTEA Gaze+" provided by [24, 105] involve actions such as "take bread." This action would need to be split into two separate action primitives: "reach bread," and "set down bread onto table." Likewise, the public dataset "CMU-MMAC" provided by [37] involves actions such as "stir egg." This action would need to be split into three action primitives: "reach fork," "move fork into bowl," and "stir egg in the bowl using fork." Many state-of-the-art recognition methods for ADLs (whether

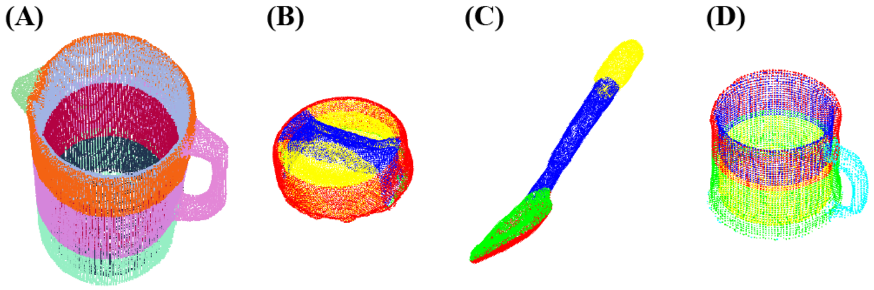


Figure 3.9: Point clouds of the four activity-relevant objects involved in Activity 1 were segmented into multiple regions for finer spatial resolution: (A) pitcher, (B) pitcher lid, (C) spoon, and (D) mug.

leveraging gaze behavior or not) are based on these publicly available datasets at the action level.

Second, action recognition models in the literature rely on computer-vision based approaches to analyze 2D videos recorded by an egocentric camera, e.g., ([24, 30, 31, 106, 94, 105, 107, 108, 109]). Whether using hand-crafted features ([28, 24, 30, 31, 106, 94, 107]) or learning end-to-end models ([105, 108, 109]), the computer vision-based approaches to action recognition must also address the challenges of identifying and tracking activity-relevant objects. In contrast, we bypassed the challenges inherent in 2D image analysis by combining an eyetracker with a marker-based motion capture system. This experimental set-up enabled the direct collection of 3D gaze-based features and object identity and pose information so that we could focus on the utility of 3D gaze features, which are unattainable from 2D camera images. Our method could be introduced into non-lab environments by combining an eyetracker with 2D cameras and ArUco markers, for example, in place of a marker-based motion capture system.

3.6.5 Considerations for Real-Time Implementation of an Action Primitive Recognition Algorithm in Human-Robot Systems

As an example of how our action primitive recognition model could be applied in a human-robot shared autonomy scenario, consider the action “stir contents inside a mug.” First, as a subject’s eye gaze vector moves toward the spoon, the probability of the potential action primitive “reach spoon” increases until it exceeds a custom threshold. The crossing of the threshold triggers the robotic end effector to move autonomously toward the spoon handle in order to grasp the spoon. The robot would use its real-time 3D model of the scene to plan its low-level movements in order to reduce the cognitive burden on the human operator. Second, as the subject’s eye gaze switches to the mug after a successful grasp of the spoon, the model would recognize the highest probability action primitive as “move spoon to mug.” Again the crossing of a probability threshold, or confidence level, would trigger the autonomous placement of the grasped spoon within the mug for a subsequent, allowable manipulate-type action primitive, which would be limited to a set of allowable manipulate-type action primitives based on the gaze object and hand object. Third, as the subject fixates their gaze on the mug, the model would recognize the highest probability action primitive as “stir inside mug” and autonomous stirring would begin. The stirring trajectory could be generated using parametric dynamic motion primitives [110], for example. Lastly, as the subject’s gaze saccades to a support surface and the action primitive is recognized as “set down spoon,” the system would proceed to determine a location on the table at which to place the spoon. This exact location could be extracted from filtered eye gaze signals as introduced in [8].

As described in the above example, we envision that our model could be used to recognize subjects’ intended action primitives through their natural eye gaze movements while the robot handles the planning and control details necessary for implementation. In contrast to some state-of-the-art approaches to commanding robot movements [22, 85, 87, 86], subjects would not be forced to unnaturally, intentionally fixate their gaze at target objects in order to trigger pre-programmed actions. Of course, much work is necessary to implement the proposed shared autonomy control scheme and this is the subject of future work.

Concerning the practical implementation of the proposed action primitive recognition method, several limitations must be addressed.

3.6.6 Specificity of the Action Primitive

The proposed recognition method is intended to assign generalized labels to each time step as one of the four verb classes (reach, move, set down, and manipulate). The current method does not distinguish between subclasses of manipulate-type verbs, such as “pour” and “stir.” Recognition of subclasses of a verb could enable assistive robots to provide even more specific assistance than that demonstrated in this work.

Recognition specificity could be advanced by incorporating additional steps. One idea is to create a lookup table based on the affordances of the objects involved in the activities. For example, the action primitive triplet of (verb = manipulate, TO = mug, HO = pitcher) is associated with the verb subclass “pour.” However, the triplet (verb = manipulate, TO = pitcher, HO = spoon) is associated with both verb subclasses “stir” and “scoop.” As an alternative, we suggest the use of gaze heat maps to facilitate the classification of verb subclasses since action primitives are activity-driven and the distribution of gaze fixations can be considerably affected by object affordance ([33, 17]).

3.6.7 Distracted or Idle Eye Gaze States

The proposed recognition method does not recognize human subjects' distracted or idle states. For example, a subject's visual attention can be distracted by environmental stimuli. In this study, we minimized visual distractions through the use of black curtains and by limiting the objects in the workspace to those required for the instructed activity. The incorporation of distractions (audio, visual, cognitive, etc.) is beyond the scope of this work, but would need to be addressed before transitioning the proposed recognition method to natural, unstructured environments.

Idle states are not currently addressed in this work. Hands are not used for every activity and subjects may also wish to rest. If the gaze vector of a daydreaming or resting subject happens to intersect with an activity-relevant object, an assistive robot may incorrectly recognize an unintended action primitive and perform unintended movements. This is similar to the "Midas touch" problem in the field of human-computer interaction, which faces a similar challenge of "how to differentiate 'attentive' saccades with intended goal of communication from the lower level eye movements that are just random" ([111]). This problem can be addressed by incorporating additional human input mechanisms, such as a joystick, which can be programmed to reflect the operator's agreement or disagreement with the robot's movements. The inclusion of "distracted" and "idle" verb classes would be an interesting area for future advancement.

3.6.8 Integration With Active Perception Approaches

The proposed recognition method could be combined with active perception approaches that could benefit a closed-loop human-robot system that leverages the active gaze of both humans and robots. In this work, the 3rd person cameras comprising the motion capture system passively observed the scene. However, by leveraging the concept of "joint attention" [112], one could use an external and/or robot-mounted camera set-up to actively explore a scene and track objects of interest, which could be used to improve the control of a robot in a human-robot system.

As discussed in section Comparisons to State-of-the-Art Recognition Algorithms, for the purposes of this work, we bypassed the process of identifying and locating activity-relevant objects by implementing a marker-based motion capture system in our experiment. Nonetheless, the perception of activity-relevant objects in non-laboratory environments remains a challenge due to object occlusions and limited field of view. Active perception-based approaches could be leveraged in such situations. In multi-object settings, such as a kitchen table cluttered with numerous objects, physical camera configurations could be actively controlled to change 3rd person perspectives and more accurately identify objects and estimate their poses [113]. Once multiple objects' poses are determined, a camera's viewpoint could then be guided by a human subject's gaze vector to reflect the subject's localized visual attention. Since humans tend to align visual targets with the centers of their visual fields ([114]), one could use natural human gaze behaviors to control camera perspectives (external or robot-mounted) in order to keep a target object, such as one recognized by our proposed recognition method, in the center of the image plane for more stable computer vision-based analysis and robotic intervention (Li et al., 2015a). When realized by a visible robot-mounted camera, the resulting bio-inspired centering of a target object may also serve as an implicit communication channel that provides feedback to a human collaborator. Going further, the camera's perspective could be controlled actively and autonomously to focus on the affordances of a target object after a verb-TO pair is identified using our proposed recognition method. Rather than changing the physical configuration of a camera to center an affordance in the image plane, one could instead focus a robot's attention on an affordance at the image processing stage ([115]). For instance, the camera's foveal vision could be moved to a pitcher's handle in order to guide a robot's reach-to-grasp movement. Such focused robot attention, whether via physical changes in camera configuration or via digital image processing methods, could be an effective way to maximize limited computational resources. The resulting enhanced autonomy of the robot could help to reduce the cognitive burden on the human in a shared autonomy system.

Considering the goal of our work to infer human intent and advance action recognition for shared autonomy control schemes, one could also integrate our proposed methods with the concept of "active event recognition," which uses active camera configurations to simultaneously explore a scene and infer human intent [116]. Ognibene and Demiris developed a simulated humanoid robot that actively controlled its gaze to identify human intent while observing a human executing a goal-oriented reaching action. Using an optimization-based camera control policy, the robot adjusted its gaze in order to minimize the expected uncertainty over numerous prospective target objects. It was observed that the resulting robot gaze gradually transitioned from the human subject's hand to the true target object before the subject's hand reached the object. As future work, it would be interesting to investigate whether and how the integration of 1st person human gaze information, such as that collected from an ego-centric camera, could enhance the control of robot gaze for action recognition. For instance, the outputs of our proposed action primitive recognition method (verb-TO pairs) could be used as additional inputs to an active event recognition scheme in order to improve recognition accuracy and reduce observational latency.

3.6.9 Effects of the Actor on Eye Gaze Behavior

The proposed recognition model was trained using data in which non-disabled subjects were performing activities with their own hands instead of subjects with upper-limb impairment who were observing a robot that was performing activities. In our envisioned human-robot system, we seek to identify operator intent via their natural gaze behaviors before any robotic movements occur. It is known that gaze behaviors precede and guide hand motions during natural hand-eye coordination ([79]). In contrast, we hypothesize that the eye gaze behaviors of subjects observing robots may be reactive in nature. Aronsen et al. have shown that subjects' gaze behaviors are different in human-only manipulation tasks and human-robot shared manipulation tasks ([117]). The further investigation of the effect of a robot on human eye gaze is warranted, but is beyond the scope of this work. We propose that the eye gaze behaviors reported in this work could be used as a benchmark for future studies of human-robot systems that seek to recreate the seamlessness of human behaviors.

The direct translation of the model to a human-robot system may not be possible. For one, the robot itself would need to be considered as an object in the shared workspace, as it is likely to receive some of the operator's visual attention. Fortunately, as suggested by Dragan and Srinivasa in [118], the action primitive prediction does not need to be perfect since the recognition model can be implemented with a human in the loop. The robotic system could be designed to wait until a specific confidence level for its prediction of human intent has been achieved before moving.

Another important consideration is that the recognition of action primitives via human eye gaze will necessarily be affected by how the robot is programmed to perform activities. For example, eye gaze behaviors will depend on experimental variables such as manual teleoperation vs. preprogrammed movements, lag in the robot control system and processing for semi-autonomous behaviors (e.g., object recognition), etc. Recognizing that there are innumerable ways in which shared autonomy could be implemented in a human-robot system, we purposely elected to eliminate the confounding factor of robot control from this foundational work on human eye-hand coordination.

3.6.10 Integration of Low-Level Action Primitive Recognition Models With Higher Level Recognition Models

This work focused on the recognition of low-level action primitives. However, the envisioned application to assistive robots in a shared autonomy schema would require recognition at all three hierarchical levels of human behavior (action primitives, actions, activities) [88] in order to customize the degree of autonomy to the operator [119, 120]. For instance, the outputs of the low-level action primitive recognition models (such as in this work) could be used as input features for the mid-level action recognition models (e.g., [17], that would then feed into the high-level activity recognition models ([23]). Simultaneously, knowledge of the activity or action can be leveraged to predict lower level actions or action primitives, respectively.

3.7 Conclusion

The long-term objective of this work is to advance shared autonomy by developing a user-interface that can recognize operator intent during activities of daily living via natural eye movements. To this end, we introduced a classifier structure for recognizing low-level action primitives that incorporates novel gaze-related features. We defined an action primitive as a triplet comprised of a verb, target object, and hand object. Using a non-specific approach to classifying and indexing objects, we observed a modest level of generalizability of the action primitive classifier across activities, including those for which the classifier was not trained. We found that the gaze object angle and its rate of change were especially useful for accurately recognizing action primitives and reducing the observational latency of the classifier. In summary, we provide a gaze-based approach for recognizing action primitives that can be used to infer the intent of a human operator for intuitive control of a robotic system. The method can be further advanced by combining classifiers across multiple levels of the action hierarchy (action primitives, actions, activities) [88] and finessing the approach for real-time use. We highlighted the application of assistive robots to motivate and design this study. However, our methods could be applied to other human-robot applications, such as collaborative manufacturing.

CHAPTER 4

Effects of Uncertainty in Robot Competence on Human Trust in Automated Systems

4.1 Introduction

Over the past three decades, human trust in autonomous systems has received a great deal of attention. Trust in automation can determine the level of an individual's overreliance or underreliance on an automated system [121]. Trust miscalibration, defined as "a mismatch between the perceived and the actual system performance and capabilities," is the leading cause of suboptimal performance in human-machine teams [122].

Consider a "human-automation interaction" (HAI) defined as "the way a human is affected by, controls and receives information from automation while performing a task" [123]. Here, we address two key challenges in HAI: (i) the identification of reliable measures of human trust during human-automation interaction, and (ii) "trust calibration," which is defined as the "process of adjusting trust to correspond to an objective measure of trustworthiness [124]." Trust calibration refers to the "correspondence between a person's trust in the automation and the automation's capabilities" [125]. To this end, we focused on variation of "robot competence" to study its effects on human trust in real-time. Robot competence is defined as "traits that are related to perceived ability" of the robot [126].

The state-of-the-art practice is to collect subjective measures of trust at the end of experimental sessions [127, 128, 129]. As a result, trust is only evaluated at a single, post-experiment timepoint and is not monitored continuously, in real-time during human-automation interactions. In addition, when trust is measured through post-trial questionnaires, experimenters are unable to observe and model the dynamics of "trust variation," or how trust changes in real-time.

The ability to measure trust variation continuously is critical for designing automated systems that center the human experience by adjusting to the human in real-time. Typically, robot accommodations of humans are considered from the perspective of physical safety. Our work aims to address dynamic accommodations of robots for the human mental state and, in particular, human trust in the robotic system. For comparison with benchmark, post-trial Likert scale questionnaires, we introduced a novel joystick-based method for collecting subjective measures of trust throughout each experimental trial.

It should be noted the collection of subjective measures of trust throughout experimental trials has been done previously [130, 131, 132, 133]. However, for shared autonomy systems, in which the human is not simply a passive observer of an automated system and must provide intermittent feedback or input, the overall performance of the human-robot collaboration is negatively affected by an increased cognitive burden on the human. Therefore, in our work, we also collected three types of physiological responses and hypothesized that these responses could serve as real-time, objective measures of human trust without increasing cognitive load.

Here, *we evaluate the effects of uncertainty in robot competence on objective and subjective measures of human-automation trust in real-time.* The contributions of our work include:

1. A comparative evaluation of numerous real-time physiological responses to uncertainty

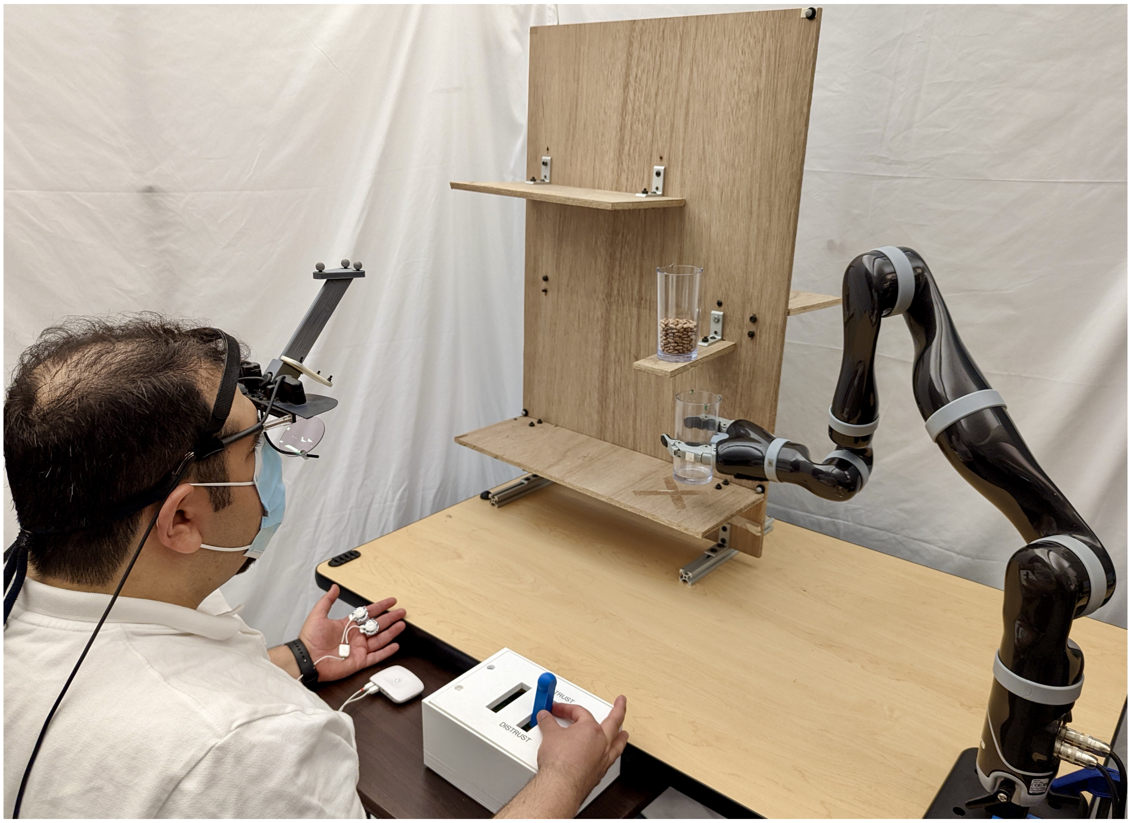


Figure 4.1: Experimental setup consisting of a robot arm, eye tracker, electrocardiogram sensors, galvanic skin response sensors, joystick, a custom shelf, and two cups containing beans.

in robot competence: (i) eye gaze features, (ii) heart rate, (iii) galvanic skin response (GSR).

2. The use of eye tracking to measure trust in automation with a real robot, and not via simulation or videos.

3. The introduction of a joystick-based method for capturing a real-time subjective measure of trust that can be used to train a supervised machine learning model for recognizing and calibrating human-automation trust in real-time.

Our work is organized as follows: Section 4.2 highlights related work on human-automation trust. Section 4.3 describes the experiment with a real robot. Section 4.4 discusses key experimental results, and Section 4.5 concludes with a discussion of limitations and future work.

4.2 Related Work

Different definitions of trust can be found across a wide range of research domains, including but not limited to psychology, philosophy, economics, human-automation interfaces, etc. However, three necessary elements are common across all fields: a "truster" who gives trust, a "trustee" who accepts trust, and something (e.g. human safety) that is at stake [121]. When it comes to humans interacting with technology (e.g., computers, robots), the efficacy of the human-automation interaction can be improved if trust variation can be observed by or communicated to the automated agent. We adopt the definition stated by Lee and See that trust in the context of human-automation is the "attitude that an agent will help achieve an individual's goals in situations characterized by uncertainty and vulnerability" [125]. In particular, we focused on the real-time measurement of trust variation in response to uncertainty about robot competence.

In the context of HAI, Hoff and Bashir described trust variation as being comprised of three layers of trust: dispositional, situational, and learned [121]. Dispositional trust refers to an "individual's enduring tendency to trust automation." Situational trust, depends on the "specific context of an interaction." Learned trust is based on "past experiences relevant to a specific automated system." In our study, we focus on the situational layer of trust that depends on time-varying factors, such as task difficulty and risk of automation failures.

Since the time-varying factors can vary throughout a human-automation interaction, real-time measures of human-automation trust are especially useful. Our emphasis on real-time measures of trust distinguishes our study from the literature, which has mainly focused on post-trial measurements via surveys.

To study the internal state of human trust during robotic automation, numerous works have focused on different modalities such as gaze and physiological measures [134]. A key advantage of physiological responses (e.g., heart rate, galvanic skin response (GSR), and electrocardiogram (ECG)) is that they are not subjective, although they could be influenced by factors independent of experimental conditions. While variations in such vital signs can be observed during human-robot interactions, the measurement of these signals can be subject to delays on the order of 1-4 sec [135].

In order to assess real-time human-automation trust, previous studies have leveraged self-reported behaviors [136] as well as measurements of psychophysiological signals and gaze [137]. Hu et al. introduced an empirical trust sensor model using electroencephalography and GSR features [138]. In their study on driving, simulated responses of an obstacle detector sensor under two conditions (faulty and reliable) were shown to the participants on a screen. A binary classifier for trust/distrust was developed using electroencephalography and GSR features as inputs. However, classification was not performed in real-time, and the stimuli were screen-based. In another work [139], Xu and Dudek developed a probabilistic trust model using a dynamic Bayesian network based on interaction experiences with a simulated autonomous robot. Here, we used a real robot to investigate human-automation trust, and we collected all experimental measures in real-time.

Rahman and Wang introduced the necessity for bilateral trust in human-robot collaboration during a sequential assembly task [140]. Although a robust method for quantifying trust in real-time was developed, the trust measures were based primarily on speed of task performance and robot fault counts. Here, we collected data on multiple measures of trust in real-time, including eye tracking.

Eye tracking has been studied as a nonintrusive measure of trust during automation. Lu and Sarter compared gaze metrics for two different modes of simulated reliability (low, high) of an automated unmanned aerial vehicle system [16]. However, gaze data were collected while participants viewed static images on a screen. Influenced by the aforementioned studies, *we designed an experiment to evaluate multiple objective and subjective measures of trust, in real-time, and with a real robot.*

4.3 Methods

Using a protocol approved by the UCLA Institutional Review Board, we conducted a study on the effects of uncertainty in robot competence on human trust in automated systems. All 18 participants (15 male, 3 female; aged 18-35 years with a mean ± stdev of 27.2 ± 4.5 years) gave written informed consent in conformity with the Declaration of Helsinki. Six out of the 18 participants reported prior experience in interacting with robots.

4.3.1 Experimental Set-up

Participants were seated at a table where they observed automated robot movements for two tasks that are important for activities of daily living (Figures 4.1 and 4.2): (i) a pick-and-place task in which a cup was grasped from a low shelf, moved around an obstacle, and placed on a high shelf, and (ii) a pouring task in which contents from one cup were poured into a second cup.

We used a 7 degree-of-freedom (DOF) assistive robotic arm with a three-fingered end-effector (JACO2 7-DOF spherical, Kinova Robotics, Quebec, Canada). A motion capture system (T-Series, Vicon, Culver City, CA, USA) with a sampling rate of 100 Hz was used to track the pose of the cups to be grasped and manipulated by the robot arm.

Three types of objective, physiological responses were measured in real-time throughout each trial (Figure 4.1). Participants wore an eye tracker (ETL-500, ISCAN, Inc., Woburn, MA, USA) that provided pupil diameter and 2-D pixel coordinates of the gaze point at 60 Hz. Pre-gelled, self-adhesive disposable electrodes (biosignalsplux, Lisboa, Portugal) were used to collect electrocardiogram (ECG) data and galvanic skin response (GSR) data from the participant's non-dominant hand at 500 Hz. To minimize noise, subjects were instructed to rest their non-dominant hand on the table and to keep it still during experimental trials.

We used a custom joystick to collect subjective data on each participant's level of human-automation trust in real-time throughout each trial as they observed the robot. A 3D-printed enclosure was used to constrain the movement of a Logitech F710 Wireless Gamepad joystick to one degree-of-freedom.

A computer with a 3.6 GHz Intel 9700K processor and NVIDIA GeForce RTX 2070 GPU was used to control the robot arm and synchronize all data from the motion capture system, eye tracker, ECG and GSR sensors, and joystick.

4.3.2 Experimental Protocol

Priming is defined as "the exposure to a stimulus that influences the response to another subsequent stimulus and thus influences behavior later on, without the individual necessarily being aware of this effect" [141]. To standardize the priming effect [16] across all participants,

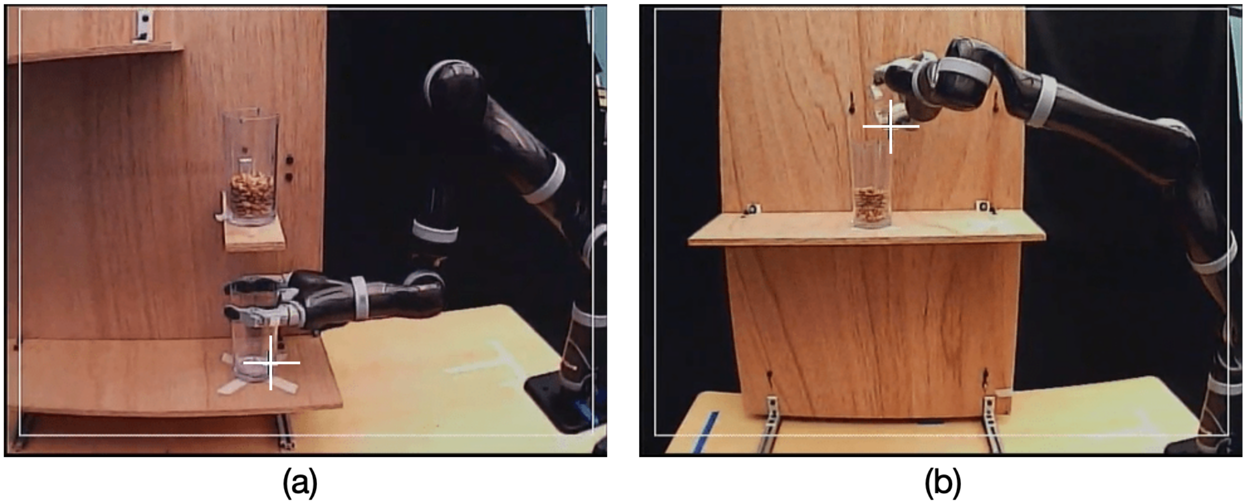


Figure 4.2: (a) Task 1 consisted of a pick-and-place action while avoiding an obstacle. (b) Task 2 consisted of a pouring action.

we began each experimental session by informing participants that they would be observing an automated robot that was in beta testing mode and that the robot might occasionally make mistakes.

The robot was preprogrammed to perform two tasks with plastic cups that were selected for their clear, glass-like appearance (Figure 4.2). Task 1 consisted of picking up an empty plastic cup from a low shelf and moving around an intermediate shelf in order to place the cup on the highest shelf. A second plastic cup containing pinto beans sat on the intermediate shelf and served as an obstacle along the path from the lowest shelf to the highest shelf.

Erroneous robot behaviors were preprogrammed for two critical phases during pick-and-place Task 1. For Critical Phase 1, the robot arm could take any one of three possible paths around the intermediate shelf: a path that collided with the shelf ("Failed"), a risky path that passed very close to but did not collide with the shelf ("Marginal"), and a conservative path that passed easily around the shelf ("Successful"). For Critical Phase 2, the robot would either place the grasped cup too close to the edge of the top shelf, in which case the cup would fall ("Failed"), or place the cup securely on the top shelf ("Successful").

Task 2 consisted of picking up a plastic cup containing pinto beans and pouring the beans into an empty cup nearby. Again, erroneous robot behaviors were preprogrammed for two critical phases. For Critical Phase 1, the robot would pour such that most of the beans would miss the empty cup ("Failed") or correctly pour all beans into the empty cup ("Successful"). As with Task 1, for Critical Phase 2, the robot would either place the grasped cup too close to the edge of the shelf, in which case the cup would fall ("Failed"), or place the cup securely on the shelf ("Successful").

Table 4.1 shows the prescribed order of experimental trials that were used for each task and for all participants. Each task was performed in blocks consisting of six trials with pre-planned combinations of successful, marginal, and failed paths.

Table 4.1: Fixed order of trials for each Task, with pre-planned combinations of Successful (green), Marginal (yellow), and Failed (red) paths for each Critical Phase. The color coding for the experimental conditions will remain consistent across all remaining tables and figures.

Task 1: Pick-and-place

Trial	Critical Phase 1	Critical Phase 2
1	Failed	Failed
2	Successful	Successful
3	Marginal	Failed
4	Failed	Successful
5	Successful	Successful
6	Marginal	Successful

Task 2: Pouring

Trial	Critical Phase 1	Critical Phase 2
1	Failed	Failed
2	Successful	Successful
3	Successful	Failed
4	Failed	Successful
5	Successful	Failed
6	Successful	Successful

4.3.2.1 Joystick

Participants were instructed to actively report their trust in the robot's competence via the 1-DOF joystick throughout each trial. They used their dominant hand to push the joystick forward or pull the joystick backward in order to report increased and decreased trust in the robot, respectively. Participants were aware that they could report their trust level over a continuous range and could also leave the joystick in a neutral position, if desired.

4.3.2.2 Post-trial Likert surveys

Upon completion of each trial, we asked participants the following questions from Muir's trust questionnaire [1]. Questions 1-4 below address robot predictability, competence, reliability, and overall human trust in the robot, respectively.

1. To what extent can the robot's behavior be predicted from moment to moment?
2. To what extent can you count on the robot to do its job?
3. What degree of faith do you have that the robot will be ale to cope with similar situations in the future?
4. Overall, how much do you trust the robot?

4.3.3 Data Processing

4.3.3.1 Eye Gaze

The following information was extracted from the eye tracker data: (i) [x,y] coordinates of the gaze point expressed in the eye tracker reference frame, (ii) reconstructed 3D gaze vector that originated from the egocentric camera reference frame, and (iii) horizontal and vertical pupil diameters and pupil area for each eye.

Custom Python code was used to extract gaze-related features. First, we used the vertical pupil diameter to detect and filter out blink periods. Then we leveraged a well-established eye movement classifier algorithm known as “robust eye-movement classification for dynamic stimulation” (REMoDNaV) [142]. With minor modifications to the REMoDNaV algorithm, we provided gaze coordinates as time series inputs and generated gaze labels (saccade, fixation, pursuit and post-saccadic oscillations) for every timestep of each trial.

Using the time series-generated gaze labels, we used a repeated measure ANOVA for within-subject comparisons, followed by paired t-tests, to analyze the eye behavior of the participants between different automation failure/success conditions ($\alpha = 0.05$). Similar to the eye gaze metrics analyzed in [16], we focused on the following eye gaze features: (i) average gaze velocity, (ii) total fixation/pursuit duration, and (iii) pupil size. Average gaze velocity (first row of Figure 4.3) was calculated using the instantaneous rate of change of gaze position in the eye tracker image plane. Gaze velocity serves as the basis for the extraction of numerous eye gaze features that can be calculated with minimal computational resources in real-time.

Periods of gaze fixation and pursuit were pooled and reported as a single measure of total fixation/pursuit duration. The fixation and pursuit behaviors were pooled because they both involve fixation on an object except that pursuit involves fixation on a moving object. Pupil size was directly outputted from the eye tracker signal. Statistical analyses were performed for vertical pupil diameter, specifically. Periods of blinking were excluded from all analyses.

4.3.3.2 Electrocardiograms

With a raw ECG signal as the input (second row of Figure 4.3), we used the BioSignalPlux ECG processing Python package to calculate the temporal heart rate (HR) sequence per trial. To analyse the HR variation in the population level, we calculated heart rate variation as the average increase above each subject minimum HR value, in one-second intervals (Figure 4.4).

4.3.3.3 Galvanic Skin Responses

Using the smoothed GSR signal (third row of Figure 4.3), we extracted key features of the signal, such as onset of physiological arousal. Considering onset instances as markers of responses to uncertainty in robot competence, we counted onset instances before and after Critical Phases of each task.

4.4 Results and Discussion

For brevity, we present and discuss the results for Task 1 only. The results for Task 2 generally align with those for Task 1. Time series data from a representative trial for a representative participant are shown in Figure 4.3. Rows 1-4 show average gaze velocity, heart rate, galvanic skin response, and joystick data, respectively. All remaining figures and tables summarize results at the population level.

4.4.1 Objective Measures

In this section, we present our analyses of the physiological response data. Each objective, physiological measure is presented in order of decreasing usefulness with respect to real-time

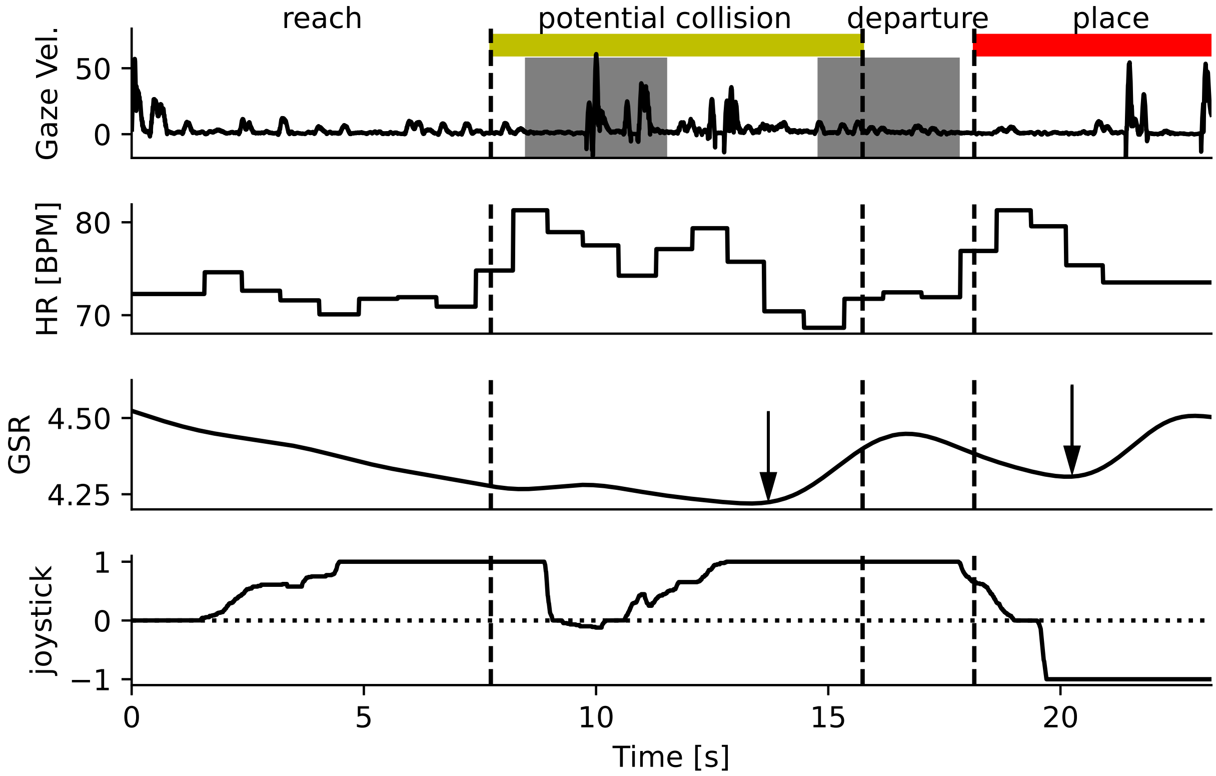


Figure 4.3: Physiological data and joystick responses for a representative participant for Task 1, Trial 3 are shown as a function of time. The “Marginal” and “Failed” conditions for the Critical Phases are indicated with yellow and red color bars, respectively. Row 1 shows the gaze velocity; Section 4.4.1.1 will address statistical tests for the periods highlighted in gray. Row 2 shows the heart rate. Row 3 shows a smoothed GSR signal; arrows indicate the onset of physiological arousal. Row 4 shows the joystick responses.

tracking of trust variation. Eye gaze features were found to be the most useful measure that reflects real-time trust variation.

4.4.1.1 Eye Gaze

To study the effect of uncertainty variation on human gaze behavior, we compared gaze patterns during 3-sec periods (gray shading in Figure 4.3) when participants would observe potential robot failures.

Table 4.2 shows the population-level results from a repeated measure ANOVA for within-subject comparisons for the three aforementioned eye gaze features during two periods of Task 1: potential collision and departure. During the potential collision period, the robot end-effector approached the obstacle with the grasped cup. In the “Marginal” condition, the cup was passed close to the edge of the obstacle edge; it was difficult to visually assess whether a collision would occur. In the “Successful” condition, the cup was moved around the obstacle with a more conservative trajectory that allowed for more distance between the cup and the obstacle. The slight difference in robot trajectories resulted in a statistically significant difference between the “marginal” and “successful” conditions for average gaze velocity and total fixation for both the potential collision and departure periods (Table 4.2). Pupil size was affected by the “Marginal” and “Successful” conditions for the departure period only.

We hypothesized that smaller gaze velocity values would be associated with increased attention levels, possibly caused by greater uncertainty in robot competence. We also hypothesised that longer fixation/pursuit duration would occur during periods of greatest risk for robot failure. For the potential collision period, the average gaze velocity was statistically significantly smaller for the “marginal” condition than for the “successful” condition.

Eye movements were slower and more time was spent on gaze fixation or pursuit for the "marginal" condition. This makes sense, as participants may be more likely to fixate on and visually track the cup when the robot is making risky movements.

For the departure period, opposite trends were observed for average gaze velocity and total fixation duration. The average gaze velocity was larger and the total fixation duration was smaller during the departure period after the participant had just observed a close call for the "marginal" conditions in the preceding potential collision period.

Table 4.2: A repeated measure ANOVA for within-subject comparisons and paired t-tests were used to analyze the effects of "Marginal"(M; yellow) and "Successful"(S; green) conditions on eye gaze features for "potential collision" and "departure" periods for Task 1 $(\alpha = 0.05)$.

Phase	**Dependent Variable**	**Comparison**	**p-value**
potential collision	Ave. Gaze Vel.	M < S	$p < 0.001$
	Total Fix. / Purs.	S < M	$p = 0.01$
	Pupil Size	S < M	$p = 0.216$
departure	Ave. Gaze Vel.	S < M	$p = 0.035$
	Total Fix. / Purs.	M < S	$p = 0.003$
	Pupil Size	M < S	$p < 0.001$

Table 4.2 suggests that eye gaze features could be one of the most practical metrics for monitoring human-automation trust in real-time. Considering that vision is the primary

sensory feedback modality during the observation of automated systems, the effects of uncertainty in robot competences are likely to be reflected sooner in eye gaze features than other physiological measures, such as heart rate.

4.4.1.2 Electrocardiograms

Figure 4.4 shows the population-level, mean increase in heart rate above the minimum heart rate in 1-sec intervals for Task 1, Trials 2 and 3. Trial 2 was "Successful" for both Critical Phases while Trial 3 was programmed for "Marginal" and "Failed" paths in Critical Phases 1 and 2, respectively.

Trends in heart rate variation for the "Successful" and "Marginal" conditions diverge near the start of Critical Phase 1 (around t=10 sec). The mean increase in heart rate rises for the "Marginal" condition, but falls for the "Successful" condition. Near the start of Critical Phase 2 (around t=20 sec), the mean increase in heart rate rises for both the "Failed" and "Successful" conditions. However, the increase in heart rate for the "Failed" condition is nearly twice that for the "Successful" condition.

Figure 4.4 suggests that increases in mean heart rate above the minimum heart rate can be affected by uncertainty in robot competence, as at the start of Critical Phase 1. However, heart rate can also increase as a result of anticipated robot failure, as at the start of Critical Phase 2. The fall in heart rate after the failure in Critical Phase 2 does not indicate that trust in the automated system has increased or been restored. This suggests that further study is needed to relate real-time heart rate variation to real-time human-automation trust.

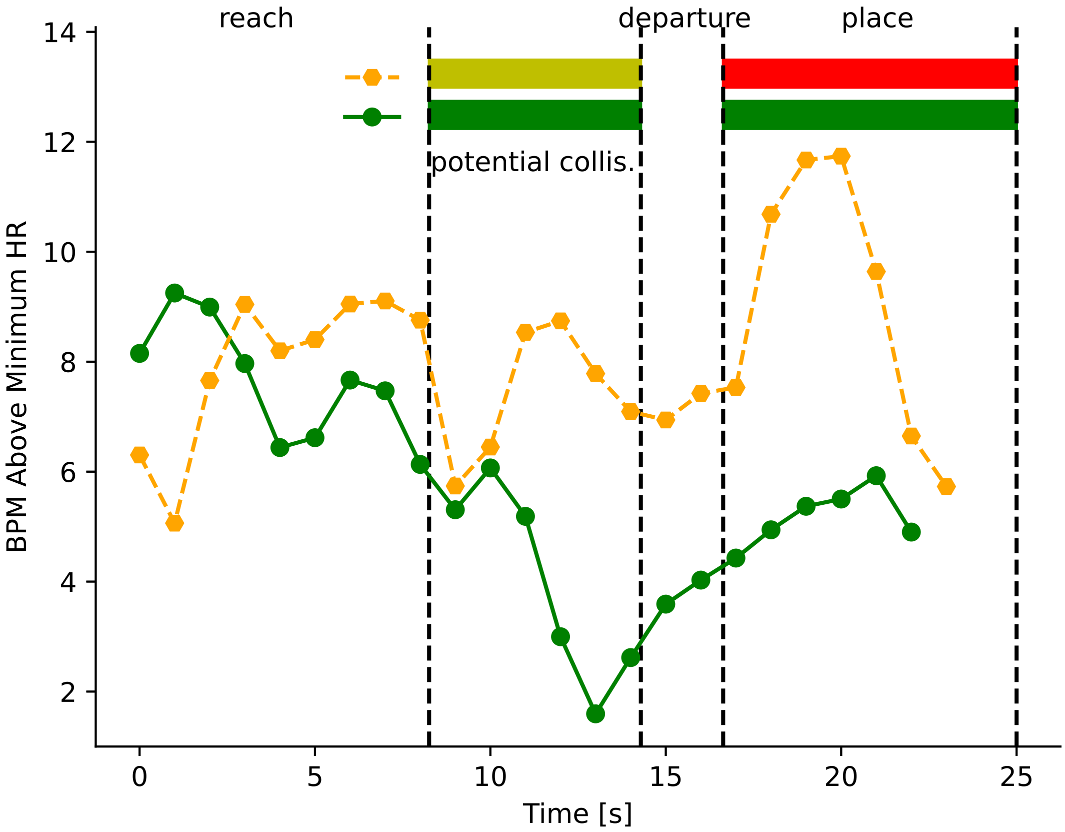


Figure 4.4: The population-level, mean increase in heart rate above the minimum heart rate in beats per minute (BPM) is shown in 1-sec intervals for Task 1, Trials 2 (green dots) and 3 (orange dots). “Successful”, “Marginal”, and “Failed” conditions for the two Critical Phases are indicated by green, yellow, and red color bars, respectively. Trends in heart rate variation for the “Successful” and “Marginal” conditions diverge near the start of Critical Phase (around t=10 sec).

4.4.1.3 Galvanic Skin Responses

Figure 4.5 shows the population-level counts of GSR onset instances for all trials for Task 1 shortly before and after Critical Phase 1. For every trial, the total GSR onset counts were greater after the potential collision period than those before the same period. In Trial 3, where the “Marginal” Critical Phase 1 and potential collision are introduced for the first time, we observed a notable increase in the total GSR onset counts as compared to the other five trials. However, the different Critical Phase 1 conditions in the subsequent trials (Trials 4, 5, and 6 with “Failed”, “Successful” and “Marginal” conditions, respectively) do not appear to effect the GSR onset counts.

Figure 4.5 suggests that total GSR onset counts may be limited in their ability to reliably reflect human-automation trust. Additionally, galvanic skin response is known to have a relatively long response time (1-3 sec) [143]. Natural delays in the galvanic skin response signal may preclude their use for the real-time observation human-automation trust for online robot planning. Based on our observations, eye gaze features and heart rate variation are recommended over galvanic skin response for real-time tracking of human-automation trust.

4.4.2 Subjective Measures

In this section, we present our analyses of two subjective measures of trust: real-time joystick responses and post-trial Likert surveys.

4.4.2.1 Joystick

Figure 4.6 shows the population-level mean (first row) and rate of change (second row) for the joystick data for Task 1, Trials 2 and 3. Trial 2 was “Successful” for both Critical Phases

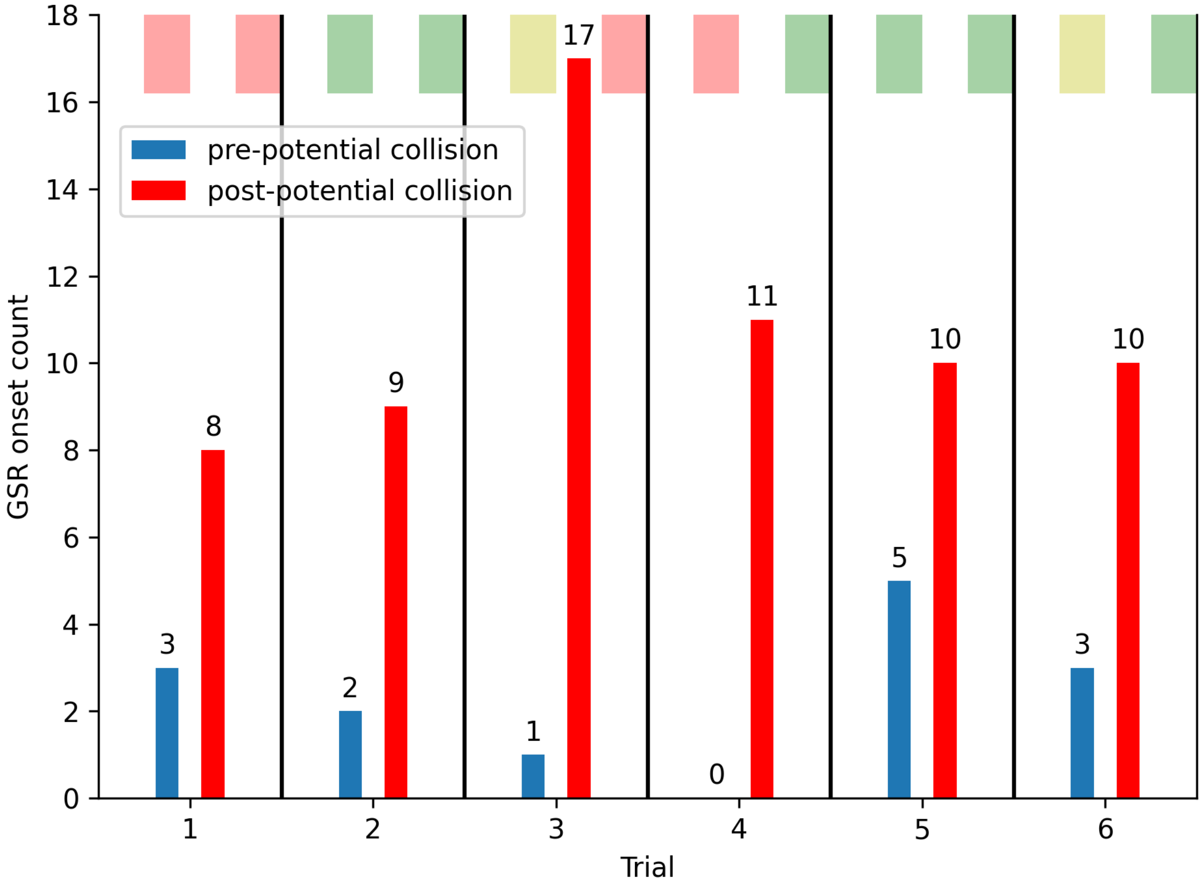


Figure 4.5: Population-level total GSR onset counts for all trials, shortly before and after Critical Phase 1 of Task 1. “Successful”, “Marginal”, and “Failed” conditions for the two Critical Phases are indicated by green, yellow, and red color bars, respectively.

while Trial 3 was programmed for “Marginal” and “Failed” paths in Critical Phases 1 and 2, respectively. Near the start of Critical Phase 1 (around t=10 sec), the rate of change for the joystick data plunges for the “Marginal” condition, but remains fairly positive and stable for the “Successful” condition. The variation of this continuous, joystick-based subjective response temporally aligns with the sudden changes in the objective, physiological measures of trust around critical phases, further supporting the results discussed in Section 4.4.1.1 and 4.4.1.2.

We introduced the 1-DOF joystick for tracking subjective responses in real-time for two main reasons. First, the 1-DOF joystick enables participants to non-verbally report their uncertainty and trust in a way that (i) does not require participants to divert their eyes from the robot in order to move the joystick, and (ii) is intuitive and, therefore, minimizes additional cognitive load. Second, the joystick signals can be used as labels for future supervised machine learning models that take objective measures as inputs and return a continuous, real-time measure of human-automation trust as an output. Real-time variations in human-automation trust could then be used for online robot planning in order to enhance the experience of the human participants.

4.4.2.2 Post-trial Likert Surveys

Figure 4.7 shows the population-level Likert scale survey results for Task 1 and Question 4 of Muir’s trust questionnaire [1]: “Overall, how much do you trust the robot?”. Results for Questions 1-3, provided in Appendix A.1, are similar to those for Question 4. The range of survey responses varied between 1 - 7, where 1 indicated “not at all,” 7 indicated “A lot,” and 4 was neutral.

A repeated measure ANOVA for within-subject comparisons and pairwise t-tests were

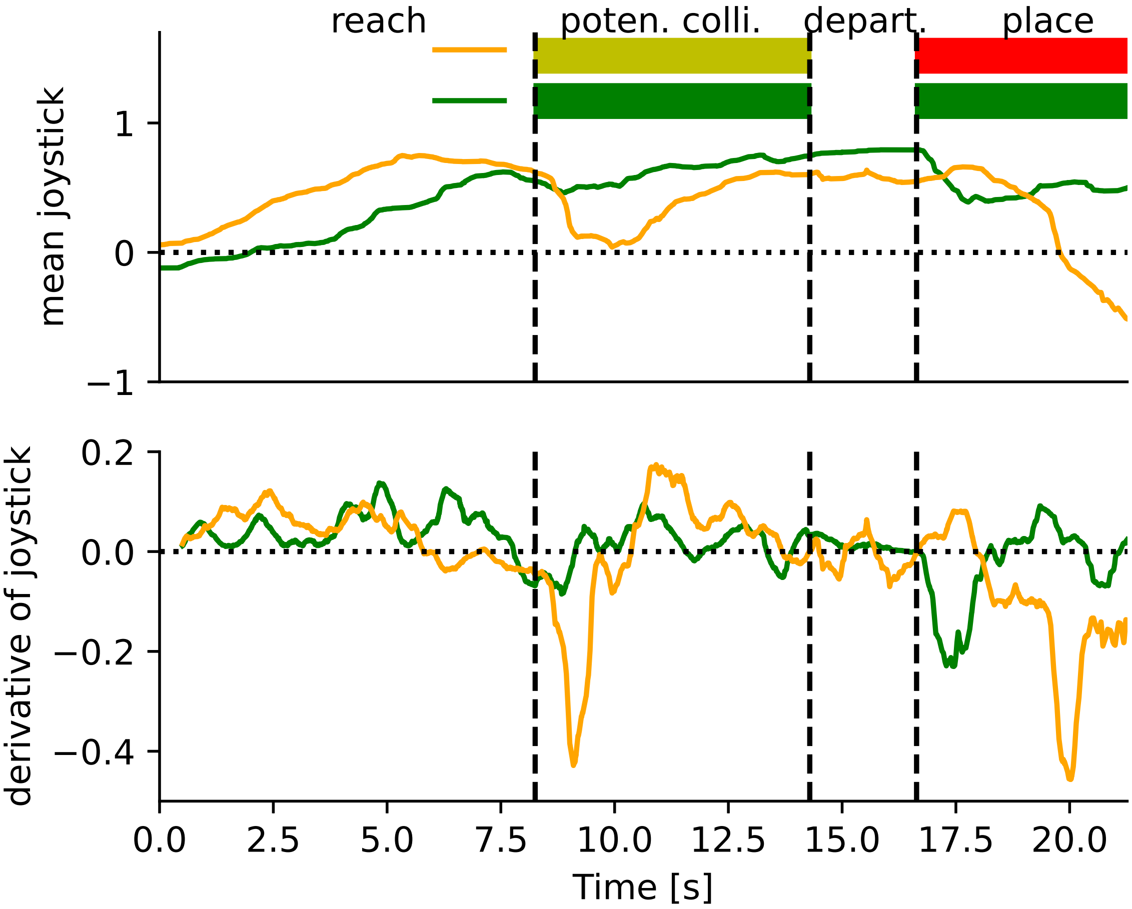


Figure 4.6: Population-level mean (first row) and rate of change (second row) for the subjective joystick data for Task 1, Trials 2 (green) and 3 (orange). “Successful”, “Marginal”, and “Failed” conditions for the two Critical Phases are indicated by green, yellow, and red color bars, respectively.

used to compare survey responses across all six trials ($\alpha = 0.05$). As expected, we observed that participant responses regarding overall trust of the robot were statistically significantly lower for trials that included “Failed” paths (Trials 1, 3, and 4).

Although the statistically significant results are aligned with our intuition based on each Trial’s Critical Phase conditions, there is a large amount of variance in survey responses within each trial. For instance, after Trial 2 (an entirely successful trial), there were still many responses in the relatively negative 1-3 range. Additionally, survey responses for Trial 5 (an entirely successful trial) were not statistically significantly different from those for Trial 6, which included a “Marginal” condition for Critical Phase 1.

We recognize the immense value of Likert scale surveys, such as Muir’s trust questionnaire [1], for capturing experimental effects on perceptions of robot predictability, competence, and reliability, as well as overall human-automation trust. Such post-trial surveys remain a valid, subjective method for evaluating human-trust automation, but they appear to be limited for the purposes of (i) real-time tracking of human-automation trust and (ii) capturing subtle effects of different experimental conditions (e.g. “Marginal” vs. “Successful” conditions). By the nature of their design, which requires reflection and an additional cognitive load, such Likert scale surveys are impractical for online robot planning in human-robot systems.

4.5 Conclusion

Here, we addressed the challenge of trust calibration in human-automation interaction. We evaluated the effects of uncertainty in robot competence on human-automation trust in real-time. We measured three objective physiological responses: (i) eye gaze features, (ii)

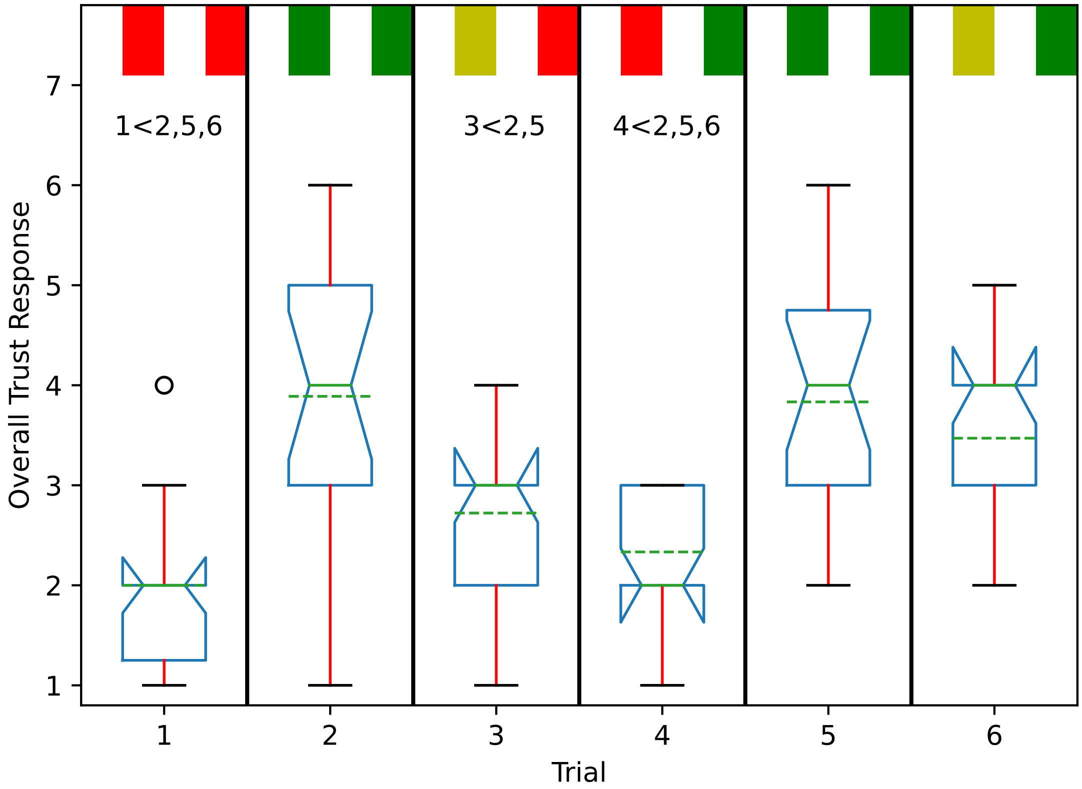


Figure 4.7: Population-level Likert scale survey results are shown for all trials for Task 1, Question 4 of Muir's trust questionnaire [1]: "Overall, how much do you trust the robot?". The scale ranges from 1 ("not at all") to 7 ("a lot"). "Successful", "Marginal", and "Failed" conditions for the two Critical Phases in each Trial are indicated by green, yellow, and red color bars, respectively. Boxplots show the 25th, 50th (narrowed) and 75th percentiles. Mean values are shown with green dashed lines. Participant responses regarding overall trust of the robot were statistically significantly lower ($\alpha = 0.05$) for trials that included "Failed" paths (Trials 1, 3, and 4).

heart rate, and (iii) galvanic skin response. We investigated significant differences in eye gaze features when participants observed a real robot that was preprogrammed for conservative, risky, and erroneous behaviors. In addition, we introduced a continuous, joystick-based subjective measure of trust that can be used to label objective measures of trust as part of a supervised machine learning model that outputs a continuous measure of human-automation trust in real-time.

One limitation of this study is that all participants experienced the same order of trials, as described in Table 4.1. As a result, the "learning effect" from preceding trials could have affected the trust level from trial to trial. However, as mentioned in Section 4.2, our focus was on the situational layer of trust as opposed to "learned trust."

While learned trust could affect the baseline values of the trust measures, we hypothesize that any effects of learned trust would occur over a relatively long, trial-to-trial timescale. In contrast, our long-term objective is to track human-automation trust in real-time at second-to-second timescales (or subsecond timescales, if possible) for online robot planning in human-robot systems. Further investigation is needed to assess whether our observations would be affected by a randomized trial order.

Another limitation of this study is that the specific participant pool may have introduced biases into our observations. It is worth noting that this experiment was conducted during the COVID-19 pandemic under restricted experimental and participant recruitment conditions. Therefore, we recommend replicating such an experiment with a gender-balanced group of subjects with a broader age range.

As future work, one could develop a model that would take objective, physiological measures as inputs and produce a continuous measure of real-time human-automation trust as the output. The real-time, subjective joystick signals could be used as labels to train a supervised machine learning model to recognize, or even predict, variations in objective, physiological responses as they relate to variations in human-automation trust. A regression model could take as inputs features with continuous values such as gaze velocity, heart rate, and pupil size as well as categorical features such as gaze label (fixation/pursuit, saccade) and task phase. The model's output could be used for online robot planning and could lead to a new action, re-planned action, speed adjustment, or even an emergency stop behavior, as needed.

CHAPTER 5

Summary and Conclusion

Attaining the optimal performance of collaborative human-machine systems remains an ever-present goal for shared autonomy systems. In this work, we focused on eye gaze features and their promise for enhancing the human experience in human-robot systems. Specifically, we investigated eye movements in the context of human intention recognition and trust for human-robot systems.

Eye tracking offers a non-verbal, nonintrusive method for human agents to naturally communicate to robots. However, it can be challenging to identify informative gaze features, recognize intent and trust level through gaze behaviors, and integrate gaze-based content into collaborative human-robot systems. In this work, we have introduced new methodologies to create 3D gaze saliency maps for gaze behavior analysis, recognize human intent at the subtask and action primitive levels. In addition, we evaluated numerous objective and subjective methods of measuring human trust in automation, and recommend gaze features as a reliable and practical method of inferring human trust in automation.

5.1 Contributions

3D gaze saliency maps: We used reconstructed 3D gaze vectors, to create high spatial resolution 3D gaze saliency maps over the manipulated objects of the experiment in Chapter

2. Our novel 3D gaze saliency maps encode characteristics of subtask and action unit levels which distingushes them from 2D saliency maps from single camera perspectives

Novel subtask recognition method using gaze object sequences: We used a gaze object sequence (GOS) to capture temporal sequence in which the objects were visually regarded. To create characteristics for the subtasks for automatic recognition, we used dynamic time warping barycentric averaging to create a population-based set of gaze object sequences. Through comparison of Euclidean distance from individual GOS of subtasks to the characteristic GOS we were able to recognize each intended subtask with high accuracy. We demonstrated that the GOS could be used with high recognition accuracy values as an important feature for action recognition.

Novel action primitive recognition method using 3D gaze-related features: We defined an action primitive as a triplet comprised of a verb, target object, and hand object. We trained a long short-term memory recurrent neural network to recognize a verb and target object, and then tested the trained network on three different activities. Using a non-specific approach to indexing objects in the workspace, we observed a modest level of generalizability of the action primitive classifier across activities, including those for which the classifier was not trained. We demonstrated that the novel input features of gaze object angle and its rate of change were especially useful for accurately recognizing action primitives and reducing the observational latency of the classifier. The classifier and the novel gaze-related features can be used to recognize intent in shared autonomy control schemes for human-robot systems.

Evaluating the effects of uncertainty in robot competence on human trust in automated systems: We investigated multiple objective and subjective measures of trust as participants observed a real robot performing everyday tasks with various levels of competence. We analyzed three types of real-time physiological responses to uncertainty in robot competence: (i) eye gaze features, (ii) heart rate, and (iii) galvanic skin response. In addition, we introduced a continuous, joystick-based subjective response measure of trust in real-time that can be used to train a supervised machine learning model for the recognition and calibration of trust in real-time.

5.2 Future Work

5.2.1 Generalizing Eye Tracking Insights from the Lab to the Field

The amount of academic research in eye tracking fields, both in hardware and software developments, has substantially increased in the past few decades. However, there is still a large gap between eye tracking in the lab and eye tracking in the field. Eye trackers now exist in modern vehicles, smartphones, Mixed Reality (MR), Augmented Reality (AR) and Virtual Reality (VR) goggles, smart displays, and smart home devices. Even so, the level of practical usage "in the wild" is very limited. The aforementioned commercially available devices capture high-level information such as attention level and gaze movements relative to objects on a screen, but not in scenarios with real robots.

In Chapters 2 and 3, we purposely focused solely on eye gaze features for intention recognition and designed experiments for specific tasks. However, insights gleaned from eye gaze behavior could be integrated with other inputs such as voice commands and/or body movement patterns. As future work, one could develop models that generalize to different tasks, interactive scenarios, and work seamlessly with additional command input modalities. Such models could integrate signals from different sensing modalities, both from the human and environment, to provide a spatiotemporal context for the human intention recognition models.

5.2.2 Closing the Human-robot Collaboration Loop by Communicating Human Intention and Trust

Our long-term objective is to close the human-robot collaboration loop and improve performance of shared autonomy systems through bidirectional communication of intent and trust. With recent advancements in augmented reality goggles, there now exist commercially available AR goggles with integrated eye tracking hardware and software, such as the Microsoft HoloLens 2 [144]. Such an AR interface could be used to track human intent and trust in order to inform the planning of robot movements. The robot's planned objective could be communicated to the human via an AR projection over the real scene. The schematic in Figure 5.1 visualizes the proposed scenario.

The eye tracker embedded in the AR goggles would continuously monitor the user's eye movements and process the gaze velocity to extract lower-level features of the eyes such as fixation, saccade, and pursuit. Such extracted data from the eyes would serve as inputs to a pipeline for a signal translator model to extract, transform and predict human intention and trust level. These outputs could then become actionable commands for robot planning. Based on real-time monitoring of human intent and trust, the robot might perform a different action, re-plan the current action, adjust its speed, or even halt its movements altogether, if warranted.

Since the proposed scenario involves a closed-loop system, we could provide the robot's current action in progress to the signal translator model in order to provide context for changes in human objective or subjective signals. This additional input will be essential in situations where variations in objective signals, such as physiological signals, are task-dependent. For example, a task involving potential collisions may require more gaze fixation

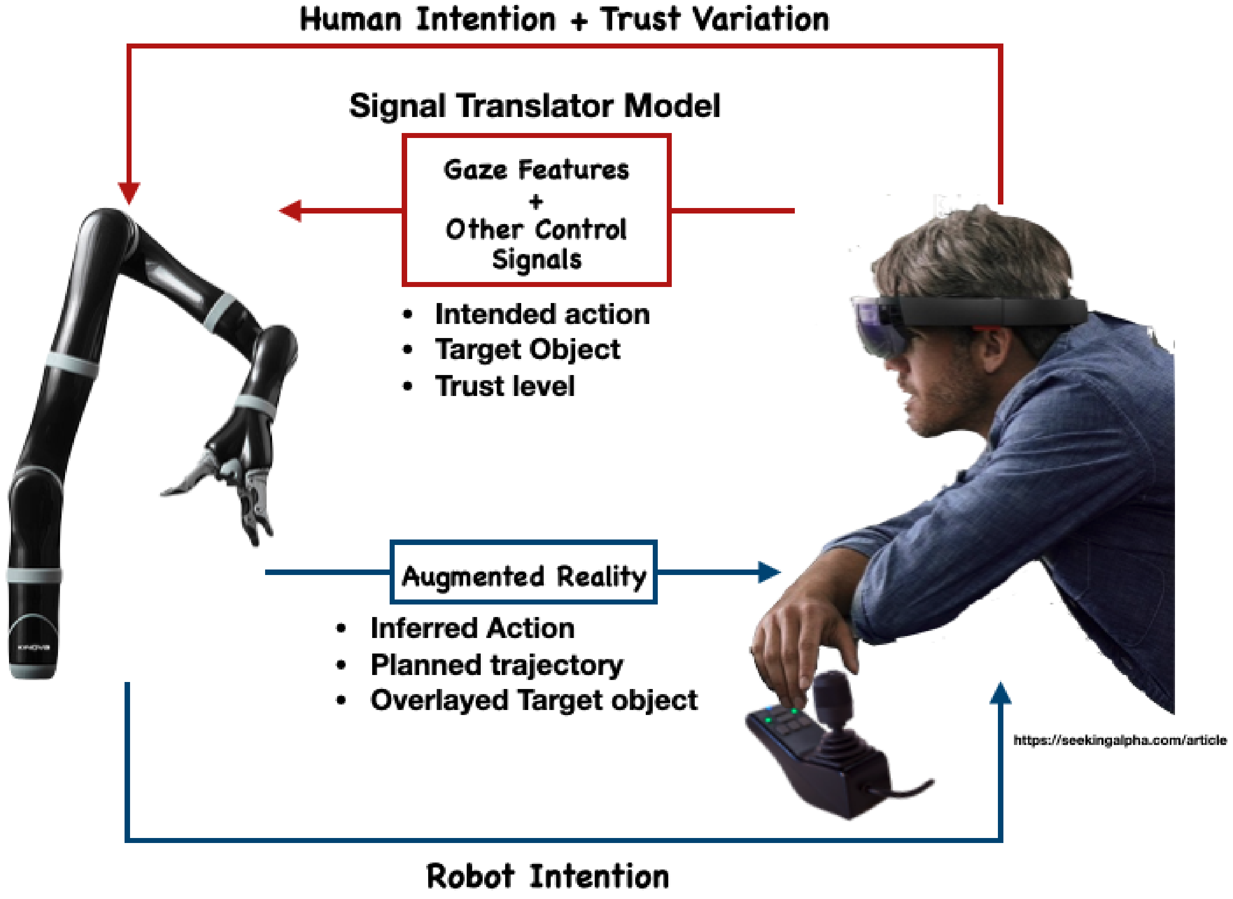


Figure 5.1: Proposed closed feedback loop system for effective human robot interaction system

than a task such as pouring, which may require more saccades between objects. Such context-dependent interpretations of eye gaze features could be provided as outputs by a signal translator model that takes the robots planned movements as inputs. Other potential inputs to the signal translator model could include voice commands, physiological signals such as heart rate and galvanic skin response, and joystick-based robot control signals.

The visual feedback provided to the human via the AR goggles could include: (i) the robot's intended action, (ii) target object, and (iii) planned trajectory in 3D space. Since no machine learning model for intention recognition can achieve 100% accuracy, the inferred human intention could be communicated via minimalistic text and/or color overlays on the target object for the planned action. The human operator could then correct the robot via voice command, a joystick input, or by adjusting their gaze focus, as needed.

In situations where the robot is constantly planning trajectories to reach or transfer objects, a visualization of the planned trajectory could help a human collaborator to foresee possible obstacle collisions. Such feedback could ultimately affect the human trust level with regards to robot competence and seamlessly inform the robot to re-plan its trajectory. Alternatively, the user could use other command methods to inform the robot of its proposed risky action.

Ultimately, human-robot, shared autonomy systems could be refined based on the application and various environmental effects. One practical example could be a repetitive and precise assembly task in a fast-paced factory setting where seamless, safe human-robot collaboration is necessary. Another example application is the use of robots to assist individuals with upper-limb impairments to regain functional independence for activities of daily living.

APPENDIX A

Supplemental Results for Chapter 4

A.1 Task 1: Post-trial Likert Surveys

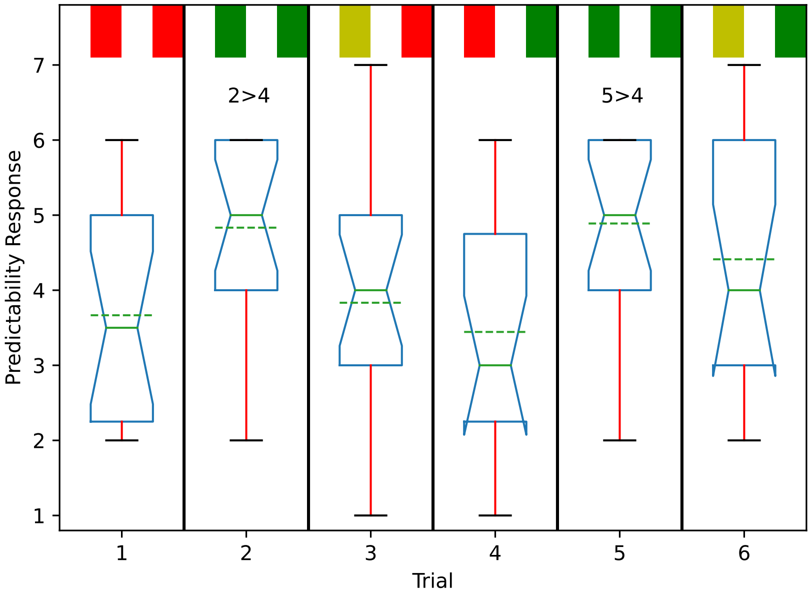


Figure A.1: Likert scale survey results are shown for Task 1 and Question 1 of Muir's trust questionnaire [1]: "To what extent can the robot's behavior be predicted from moment to moment?". The scale ranges from 1 ("not at all") to 7 ("a lot"). "Successful", "Marginal", and "Failed" conditions for the two Critical Phases in each Trial are indicated by green, yellow, and red color bars, respectively. Boxplots show the 25th, 50th (narrowed) and 75th percentiles. Mean values are shown with green dashed lines. Responses regarding robot predictability were statistically significantly higher ($\alpha = 0.05$) for trials without "Failed" paths (Trials 2, and 5).

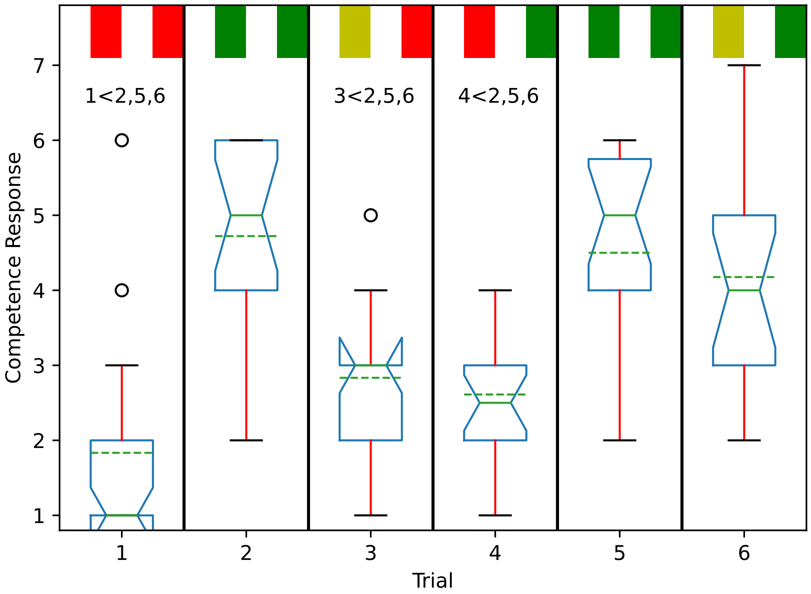


Figure A.2: Likert scale survey results are shown for Task 1 and Question 2 of Muir's trust questionnaire [1]: "To what extent can you count on the robot to do its job?". The scale ranges from 1 ("not at all") to 7 ("a lot"). "Successful", "Marginal", and "Failed" conditions for the two Critical Phases in each Trial are indicated by green, yellow, and red color bars, respectively. Boxplots show the 25th, 50th (narrowed) and 75th percentiles. Mean values are shown with green dashed lines. Responses regarding robot competence were statistically significantly lower ($\alpha = 0.05$) for trials that included "Failed" paths (Trials 1, 3, and 4).

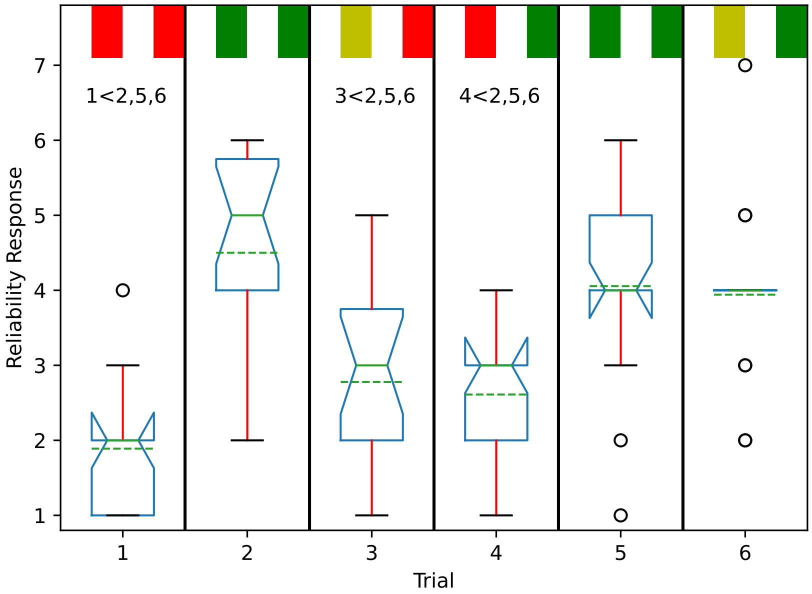


Figure A.3: Likert scale survey results are shown for Task 1 and Question 3 of Muir's trust questionnaire [1]: "What degree of faith do you have that the robot will be ale to cope with similar situations in the future?". The scale ranges from 1 ("not at all") to 7 ("a lot"). "Successful", "Marginal", and "Failed" conditions for the two Critical Phases in each Trial are indicated by green, yellow, and red color bars, respectively. Boxplots show the 25th, 50th (narrowed) and 75th percentiles. Mean values are shown with green dashed lines. Participant responses regarding robot reliability were statistically significantly lower ($\alpha = 0.05$) for trials that included "Failed" paths (Trials 1, 3, and 4).

A.2 Task 2: Eye Gaze Results

Table A.1: A repeated measure ANOVA for within-subject comparisons and paired t-tests were used to analyze the effects of "Successful" (S; green) and "Failed" (F; red) conditions on eye gaze features for Critical Phase 1 (pouring) in Task 2 ($\alpha = 0.05$).

Phase	**Dependent Variable**	**Comparison**	**p-value**
pouring	Ave. Gaze Vel.	S < F	p = 0.016
	Total Fix. / Purs.	F < S	p = 0.014
	Pupil Size	S < F	p < 0.001

REFERENCES

[1] B. M. Muir, "Operators' trust in and use of automatic controllers in a supervisory process control task." 2002.

[2] R. S. Johansson, G. Westling, A. Bäckström, and J. R. Flanagan, "Eye-Hand Coordination in Object Manipulation," *Journal of Neuroscience*, vol. 21, no. 17, pp. 6917–6932, Sep. 2001, publisher: Society for Neuroscience Section: ARTICLE. [Online]. Available: https://www.jneurosci.org/content/21/17/6917

[3] B. E. Asmar, S. Chelly, N. Azzi, L. Nassif, J. E. Asmar, and M. Färber, "Aware: A situational awareness framework for facilitating adaptive behavior of autonomous vehicles in manufacturing," in *International Semantic Web Conference*. Springer, 2020, pp. 651–666.

[4] L.-A. Raymond, M. Piccini, M. Subramanian, O. Pavel, and A. Faisal, "Natural Gaze Data Driven Wheelchair," Tech. Rep., Jan. 2018, company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article. [Online]. Available: https://www.biorxiv.org/content/10.1101/252684v1

[5] J. P. Hansen, A. Alapetite, I. S. MacKenzie, and E. Møllenbach, "The use of gaze to control drones," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, ser. ETRA '14. New York, NY, USA: Association for Computing Machinery, Mar. 2014, pp. 27–34. [Online]. Available: https://doi.org/10.1145/2578153.2578156

[6] M. Yu, Y. Lin, D. Schmidt, X. Wang, and Y. Wang, "Human-Robot Interaction Based on Gaze Gestures for the Drone Teleoperation," *Journal of Eye Movement Research*, vol. 7, pp. 1–14, Sep. 2014.

[7] L. Yuan, C. Reardon, G. Warnell, and G. Loianno, "Human Gaze-Driven Spatial Tasking of an Autonomous MAV," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1343–1350, Apr. 2019, conference Name: IEEE Robotics and Automation Letters.

[8] S. Li, X. Zhang, F. J. Kim, R. Donalisio da Silva, D. Gustafson, and W. R. Molina, "Attention-Aware Robotic Laparoscope Based on Fuzzy Interpretation of Eye-Gaze Patterns," *Journal of Medical Devices*, vol. 9, no. 4, p. 041007, Aug. 2015.

[9] M. Yu, Y. Lin, D. Schmidt, X. Wang, and Y. Wang, "Human-robot interaction based on gaze gestures for the drone teleoperation," *Journal of Eye Movement Research*, vol. 7, no. 4, pp. 1–14, 2014.

[10] H. Zamani, A. Abas, and M. K. M.Amin, "Eye Tracking Application on Emotion Analysis for Marketing Strategy," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 8, no. 11, pp. 87–91, Dec. 2016, number: 11. [Online]. Available: https://jtec.utem.edu.my/jtec/article/view/1415

[11] P. T. Huddleston, B. K. Behe, C. Driesener, and S. Minahan, "Inside-outside: Using eye-tracking to investigate search-choice processes in the retail environment," *Journal of Retailing and Consumer Services*, vol. 43, pp. 85–93, Jul. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0969698917306227

[12] R. S. A. Khan, G. Tien, M. S. Atkins, B. Zheng, O. N. M. Panton, and A. T. Meneghetti, "Analysis of eye gaze: Do novice surgeons look at the same location as expert surgeons during a laparoscopic operation?" *Surgical Endoscopy*, vol. 26, no. 12, pp. 3536–3540, Dec. 2012. [Online]. Available: https://doi.org/10.1007/s00464-012-2400-7

[13] M. K. Eckstein, B. Guerra-Carrillo, A. T. Miller Singley, and S. A. Bunge, "Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development?" *Developmental Cognitive Neuroscience*, vol. 25, pp. 69–91, Jun. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1878929316300846

[14] K. Rayner, K. H. Chace, T. J. Slattery, and J. Ashby, "Eye Movements as Reflections of Comprehension Processes in Reading," *Scientific Studies of Reading*, vol. 10, no. 3, pp. 241–255, Jul. 2006, publisher: Routledge _eprint: https://doi.org/10.1207/s1532799xssr1003_3. [Online]. Available: https://doi.org/10.1207/s1532799xssr1003_3

[15] M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang, "The pupil as a measure of emotional arousal and autonomic activation," *Psychophysiology*, vol. 45, no. 4, pp. 602–607, 2008, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.2008.00654.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8986.2008.00654.x

[16] Y. Lu and N. Sarter, "Eye Tracking: A Process-Oriented Method for Inferring Trust in Automation as a Function of Priming and System Reliability," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 6, pp. 560–568, Dec. 2019, conference Name: IEEE Transactions on Human-Machine Systems.

[17] A. Haji Fathaliyan, X. Wang, and V. J. Santos, "Exploiting Three-Dimensional Gaze Tracking for Action Recognition During Bimanual Manipulation to Enhance Human–Robot Collaboration," *Frontiers in Robotics and AI*, vol. 5, pp. 1–15, 2018.

[18] B. D. Argall, "Turning assistive machines into assistive robots," M. Razeghi, E. Tournié, and G. J. Brown, Eds., San Francisco, California, United States, Jan.

2015, p. 93701Y. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2085352

[19] S. J. Lederman and R. L. Klatzky, "Hand Movements: A Window Into Haptic Object Recognition," *Cognitive Psychology*, vol. 19, no. 3, pp. 342–368, Jul. 1987. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/3608405

[20] M. Land, N. Mennie, and J. Rusted, "The roles of vision and eye movements in the control of activities of daily living," *Perception*, vol. 28, no. 11, pp. 1311–1328, 1999.

[21] D. Kanoulas and M. Vona, "Bio-inspired rough terrain contact patch perception," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. Hong Kong, China: IEEE, 2014, pp. 1719–1724.

[22] S. Li, X. Zhang, and J. Webb, "3D-Gaze-based Robotic Grasping through Mimicking Human Visuomotor Function for People with Motion Impairments," *IEEE Transactions on Biomedical Engineering*, pp. 1–1, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7870669/

[23] W. Yi and D. Ballard, "Recognizing behavior in hand-eye coordination patterns," *International Journal of Humanoid Robotics*, vol. 06, no. 03, pp. 337–359, Sep. 2009. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/S0219843609001863

[24] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *European Conference on Computer Vision*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer, 2012, pp. 314–327.

[25] A. Behera, M. Chapman, A. G. Cohn, and D. C. Hogg, "Egocentric activity recognition using histograms of oriented pairwise relations," in *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on*, vol. 2. Lisbon, Portugal: IEEE, 2014, pp. 22–30.

[26] T.-H.-C. Nguyen, J.-C. Nebel, and F. Florez-Revuelta, "Recognition of Activities of Daily Living with Egocentric Vision: A Review," *Sensors*, vol. 16, no. 1, p. 72, Jan. 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/1/72

[27] C. Yu and D. H. Ballard, "Understanding human behaviors based on eye-head-hand coordination," in *International Workshop on Biologically Motivated Computer Vision*, H. Bülthoff, C. Wallraven, S. Lee, and T. Poggio, Eds. Springer, 2002, pp. 611–619.

[28] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding Egocentric Activities," in *Proceedings of the IEEE International Conference on Computer Vision*. Barcelona, Spain: IEEE, 2011, pp. 407–414.

[29] A. Behera, D. C. Hogg, and A. G. Cohn, "Egocentric activity monitoring and recovery," in *Asian Conference on Computer Vision*, K. Lee, Y. Matsushita, J. Rehg, and Z. Hu, Eds. Berlin, Heidelberg: Springer, 2012, pp. 519–532.

[30] A. Fathi and J. M. Rehg, "Modeling actions through state changes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Portland, OR, USA: IEEE, 2013, pp. 2579–2586.

[31] K. Matsuo, K. Yamada, S. Ueno, and S. Naito, "An Attention-based Activity Recognition for Egocentric Video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Columbus, OH, USA, 2014, pp. 551–556.

[32] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[33] A. Belardinelli, O. Herbort, and M. V. Butz, "Goal-oriented gaze strategies afforded by object interaction," *Vision Research*, vol. 106, pp. 47–57, Jan. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0042698914002818

[34] A. Haji Fathaliyan, X. Wang, S. Bazargan, and V. Santos, "Hand-object kinematics and gaze fixation during bimanual tasks," in *Proc Ann Mtg American Society of Biomechanics*, Boulder, CO, Aug. 2017.

[35] R. C. Oldfield, "The assessment and analysis of handedness: the Edinburgh inventory," *Neuropsychologia*, vol. 9, no. 1, pp. 97–113, 1971. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0028393271900674

[36] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in Manipulation Research: The YCB Object and Model Set and Benchmarking Protocols," *arXiv preprint arXiv:1502.03143*, 2015. [Online]. Available: http://arxiv.org/abs/1502.03143

[37] E. H. Spriggs, F. De La Torre, and M. Hebert, "Temporal segmentation and activity classification from first-person sensing," in *Proc of the IEEE Conf on Computer Vision and Pattern Recognition Workshops*. Miami Beach, Florida: IEEE, 2009, pp. 17–24.

[38] E. R. Kandel, J. H. Schwartz, and T. M. Jessell, Eds., *Principles of neural science*, 4th ed. New York: McGraw-Hill, Health Professions Division, 2000.

[39] M. Nyström and K. Holmqvist, "An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data," *Behavior Research Methods*, vol. 42, no. 1, pp. 188–204, Feb. 2010. [Online]. Available: https://link.springer.com/article/10.3758/BRM.42.1.188

[40] J.-Y. Bouguet, “Camera Calibration Toolbox for MATLAB,” Oct. 2015. [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/

[41] T. MathWorks, “Single Camera Calibration App,” 2017. [Online]. Available: https://www.mathworks.com/help/vision/ug/single-camera-calibrator-app.html

[42] R. E. Morrison and K. Rayner, “Saccade size in reading depends upon character spaces and not visual angle,” *Perception & Psychophysics*, vol. 30, no. 4, pp. 395–396, Jul. 1981. [Online]. Available: https://link.springer.com/article/10.3758/BF03206156

[43] N. V. Boulgouris, K. N. Plataniotis, and D. Hatzinakos, “Gait recognition using dynamic time warping,” in *IEEE Workshop on Multimedia Signal Processing*, Siena, Italy, 2004, pp. 263–266.

[44] D. M. Gavrila and L. S. Davis, “Towards 3-D model-based tracking and recognition of human movement: a multi-view approach,” in *International Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, 1995, pp. 272–277.

[45] F. Petitjean, “MATLAB function for ”DBA: Averaging time series consistently with Dynamic Time Warping”,” Nov. 2016. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/47483-fpetitjean-dba

[46] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval.* New York: Cambridge University Press, 2008, oCLC: ocn190786122.

[47] A. Harvey, J. N. Vickers, R. Snelgrove, M. F. Scott, and S. Morrison, “Expert surgeon’s quiet eye and slowing down: expertise differences in performance and quiet eye duration during identification and dissection of the recurrent laryngeal nerve,” *The American Journal of Surgery*, vol. 207, no. 2, pp. 187–193, Feb. 2014. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0002961013005679

[48] C.-a. Moulton, G. Regehr, L. Lingard, C. Merritt, and H. MacRae, “Slowing Down to Stay Out of Trouble in the Operating Room: Remaining Attentive in Automaticity:,” *Academic Medicine*, vol. 85, no. 10, pp. 1571–1577, Oct. 2010. [Online]. Available: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00001888-201010000-00013

[49] M. Hayhoe and D. Ballard, “Eye movements in natural behavior,” *Trends in Cognitive Sciences*, vol. 9, no. 4, pp. 188–194, Apr. 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364661305000598

[50] J. J. Gibson, “The theory of affordances,” in *Perceiving, Acting, and Knowing: Towards an Ecological Psychology*, R. Shaw and J. Bransford, Eds. Hoboken, NJ: John Wiley & Sons Inc., 1977, pp. 127–143.

[51] R. Detry, E. Baseski, M. Popovic, Y. Touati, N. Kruger, O. Kroemer, J. Peters, and J. Piater, "Learning object-specific grasp affordance densities," in *Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on.* Shanghai, China: IEEE, 2009, pp. 1–7.

[52] A. Leclercq, S. Akkouche, and E. Galin, "Mixing Triangle Meshes and Implicit Surfaces in Character Animation," in *Computer Animation and Simulation 2001: Proceedings of the Eurographics Workshop in Manchester, UK, September 2–3, 2001*, N. Magnenat-Thalmann and D. Thalmann, Eds. Vienna: Springer Vienna, 2001, pp. 37–47. [Online]. Available: https://doi.org/10.1007/978-3-7091-6240-8_4

[53] J. Pearson and S. M. Kosslyn, Eds., *Mental Imagery*, ser. Frontiers Research Topics. Frontiers Media SA, 2013. [Online]. Available: http://www.frontiersin.org/books/Mental_Imagery/188

[54] M. F. Land and M. Hayhoe, "In what ways do eye movements contribute to everyday activities?" *Vision Research*, vol. 41, no. 25, pp. 3559–3565, Nov. 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S004269890100102X

[55] H. Admoni and S. Srinivasa, "Predicting User Intent Through Eye Gaze for Shared Autonomy," in *Proc AAAI Fall Symposium Series: Shared Autonomy in Research and Practice*, Arlington, VA, Nov. 2016, pp. 298–303.

[56] G. Maeda, M. Ewerton, R. Lioutikov, H. B. Amor, J. Peters, and G. Neumann, "Learning interaction for collaborative tasks with probabilistic movement primitives," in *Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on.* Madrid, Spain: IEEE, 2014, pp. 527–534.

[57] R. Luo, R. Hayne, and D. Berenson, "Unsupervised early prediction of human reaching for human–robot collaboration in shared workspaces," *Autonomous Robots*, Jul. 2017. [Online]. Available: http://link.springer.com/10.1007/s10514-017-9655-8

[58] C. Morato, K. N. Kaipa, B. Zhao, and S. K. Gupta, "Toward Safe Human Robot Collaboration by Using Multiple Kinects Based Real-time Human Tracking," *Journal of Computing and Information Science in Engineering*, vol. 14, no. 1, p. 011006, Jan. 2014. [Online]. Available: http://computingengineering.asmedigitalcollection.asme.org/article.aspx?doi=10.1115/1.4025810

[59] S. S. Srinivasa, D. Berenson, M. Cakmak, A. Collet, M. R. Dogar, A. D. Dragan, R. A. Knepper, T. Niemueller, K. Strabala, and M. V. Weghe, "Herb 2.0: Lessons learned from developing a mobile manipulator for the home," *Proceedings of the IEEE*, vol. 100, no. 8, pp. 2410–2428, 2012.

[60] Q. Jenkins and X. Jiang, "Measuring trust and application of eye tracking in human robotic interaction," in *IIE Annual Conference. Proceedings.* Cancun, Mexico: Institute of Industrial and Systems Engineers (IISE), 2010, p. 1.

[61] G. Westerfield, A. Mitrovic, and M. Billinghurst, "Intelligent Augmented Reality Training for Motherboard Assembly," *International Journal of Artificial Intelligence in Education*, vol. 25, no. 1, pp. 157–172, Mar. 2015. [Online]. Available: http://link.springer.com/10.1007/s40593-014-0032-x

[62] F. E. Truitt, C. Clifton, A. Pollatsek, and K. Rayner, "The Perceptual Span and the Eye-Hand Span in Sight Reading Music," *Visual Cognition*, vol. 4, no. 2, pp. 143–161, Jun. 1997. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/713756756

[63] X. Anguera, R. Macrae, and N. Oliver, "Partial sequence matching using an unbounded dynamic time warping algorithm," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on.* Dallas, TX, USA: IEEE, 2010, pp. 3582–3585.

[64] L. Montesano and M. Lopes, "Learning grasping affordances from local visual descriptors," in *2009 IEEE 8th International Conference on Development and Learning*, Jun. 2009, pp. 1–6.

[65] X. Wang, A. Haji Fathaliyan, and V. J. Santos, "Toward Shared Autonomy Control Schemes for Human-Robot Systems: Action Primitive Recognition Using Eye Gaze Features," *Frontiers in Neurorobotics*, vol. 14, p. 567571, Oct. 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnbot.2020.567571/full

[66] S. S. Groothuis, S. Stramigioli, and R. Carloni, "Lending a helping hand: toward novel assistive robotic arms," *IEEE Robotics Automation Magazine*, vol. 20, no. 1, pp. 20–29, Mar. 2013, conference Name: IEEE Robotics Automation Magazine.

[67] B. Driessen, H. Evers, and J. v Woerden, "MANUS—a wheelchair-mounted rehabilitation robot," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 215, no. 3, pp. 285–290, Jan. 2001. [Online]. Available: http://journals.pepublishing.com/content/pr833615804485th/

[68] V. Maheu, J. Frappier, P. S. Archambault, and F. Routhier, "Evaluation of the JACO robotic arm: Clinico-economic study for powered wheelchair users with upper-extremity disabilities," in *2011 IEEE International Conference on Rehabilitation Robotics.* Zurich: IEEE, Jun. 2011, pp. 1–5. [Online]. Available: http://ieeexplore.ieee.org/document/5975397/

[69] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. LaViola, and R. Sukthankar, "Exploring the Trade-off Between Accuracy and Observational Latency in Action Recognition," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 420–436, Feb. 2013. [Online]. Available: https://doi.org/10.1007/s11263-012-0550-7

[70] L. R. Hochberg, D. Bacher, B. Jarosiewicz, N. Y. Masse, J. D. Simeral, J. Vogel, S. Haddadin, J. Liu, S. S. Cash, P. van der Smagt, and J. P. Donoghue, "Reach and grasp by people with tetraplegia using a neurally controlled robotic arm," *Nature*, vol. 485, no. 7398, pp. 372–375, May 2012, number: 7398 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nature11076

[71] O. Rogalla, M. Ehrenmann, R. Zollner, R. Becher, and R. Dillmann, "Using gesture and speech control for commanding a robot assistant," in *11th IEEE International Workshop on Robot and Human Interactive Communication Proceedings*, Berlin, Germany, Sep. 2002, pp. 454–459.

[72] L. Bi, A. >. Feleke, and C. Guan, "A review on EMG-based motor intention prediction of continuous human upper limb motion for human-robot collaboration," *Biomedical Signal Processing and Control*, vol. 51, pp. 113–127, May 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1746809419300473

[73] L. Bi, X. Fan, and Y. Liu, "EEG-Based Brain-Controlled Mobile Robots: A Survey," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 2, pp. 161–176, Mar. 2013.

[74] A. F. Salazar-Gomez, J. DelPreto, S. Gil, F. H. Guenther, and D. Rus, "Correcting robot mistakes in real time using EEG signals," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, May 2017, pp. 6570–6577.

[75] Z. C. Chao, Y. Nagasaka, and N. Fujii, "Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkey," *Frontiers in Neuroengineering*, vol. 3, 2010. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fneng.2010.00003/full

[76] W. Wang, J. L. Collinger, A. D. Degenhart, E. C. Tyler-Kabara, A. B. Schwartz, D. W. Moran, D. J. Weber, B. Wodlinger, R. K. Vinjamuri, R. C. Ashmore, J. W. Kelly, and M. L. Boninger, "An Electrocorticographic Brain Interface in an Individual with Tetraplegia," *PLOS ONE*, vol. 8, no. 2, p. e55344, Feb. 2013. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0055344

[77] S. E. Ghobadi, O. E. Loepprich, F. Ahmadov, K. Hartmann, O. Loffeld, and J. Bernshausen, "Real Time Hand Based Robot Control Using Multimodal Images," *IAENG International Journal of Computer Science*, vol. 35, no. 4, pp. 110–121, 2008. [Online]. Available: http://www.iaeng.org/IJCS/issues_v35/issue_4/IJCS_35_4_08.pdf

[78] J. L. Raheja, R. Shyam, U. Kumar, and P. B. Prasad, "Real-Time Robotic Hand Control Using Hand Gestures," in *2010 Second International Conference on Machine Learning and Computing*, Bangalore, India, Feb. 2010, pp. 12–16.

[79] M. M. Hayhoe, A. Shrivastava, R. Mruczek, and J. B. Pelz, “Visual memory and motor planning in a natural task,” *Journal of Vision*, vol. 3, no. 1, pp. 6–6, Jan. 2003. [Online]. Available: https://jov.arvojournals.org/article.aspx?articleid=2158157

[80] C.-S. Lin, C.-W. Ho, W.-C. Chen, C.-C. Chiu, and M.-S. Yeh, “Powered wheelchair controlled by eye-tracking system.” *Optica Applicata*, vol. 36, pp. 401–412, 2006. [Online]. Available: http://opticaapplicata.pwr.edu.pl/article.php?id=2006230401

[81] P. S. Gajwani and S. A. Chhabria, “Eye motion tracking for wheelchair control,” *International Journal of Information Technology*, vol. 2, no. 2, pp. 185–187, 2010. [Online]. Available: http://csjournals.com/IJITKM/PDF%203-1/2.pdf

[82] S. Li, X. Zhang, and J. D. Webb, “3-D-gaze-based robotic grasping through mimicking human visuomotor function for people with motion impairments,” *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 12, pp. 2824–2835, 2017.

[83] S. Dziemian, W. W. Abbott, and A. A. Faisal, “Gaze-based teleprosthetic enables intuitive continuous control of complex robot arm use: Writing & drawing,” in *2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*. IEEE, 2016, pp. 1277–1282.

[84] S. Li and X. Zhang, “Implicit Intention Communication in Human–Robot Interaction Through Visual Behavior Studies,” *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 4, pp. 437–448, Aug. 2017, conference Name: IEEE Transactions on Human-Machine Systems.

[85] M.-Y. Wang, A. A. Kogkas, A. Darzi, and G. P. Mylonas, “Free-View, 3D Gaze-Guided, Assistive Robotic System for Activities of Daily Living,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Madrid, Spain: IEEE, 2018, pp. 2355–2361.

[86] H. Zeng, Y. Shen, X. Hu, A. Song, B. Xu, H. Li, Y. Wang, and P. Wen, “Semi-Autonomous Robotic Arm Reaching With Hybrid Gaze–Brain Machine Interface,” *Frontiers in Neurorobotics*, vol. 13, 2020, publisher: Frontiers. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnbot.2019.00111/full?utm_source=S-TWT&utm_medium=SNET&utm_campaign=ECO_FNINS_XXXXXXXX_auto-dlvrit

[87] A. Shafti, P. Orlov, and A. A. Faisal, “Gaze-based, Context-aware Robotic System for Assisted Reaching and Grasping,” in *2019 International Conference on Robotics and Automation (ICRA)*, Montreal, QC, Canada, May 2019, pp. 863–869, iSSN: 2577-087X, 1050-4729.

[88] T. B. Moeslund, A. Hilton, and V. Krüger, “A survey of advances in vision-based human motion capture and analysis,” *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90–126, Nov. 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314206001263

[89] F. Lv and R. Nevatia, "Recognition and Segmentation of 3-D Human Action Using HMM and Multi-class AdaBoost," in *Computer Vision – ECCV*, ser. Lecture Notes in Computer Science, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer, 2006, pp. 359–372.

[90] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, Jun. 2012, pp. 1290–1297.

[91] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA: IEEE, Jun. 2014, pp. 588–595. [Online]. Available: http://ieeexplore.ieee.org/document/6909476/

[92] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, Jun. 2015, pp. 1110–1118. [Online]. Available: http://ieeexplore.ieee.org/document/7298714/

[93] Y. Li, Z. Ye, and J. M. Rehg, "Delving into egocentric actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 287–295.

[94] M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA: IEEE, 2016, pp. 1894–1903.

[95] Y. Zhang, "Edinburgh Handedness Inventory (revised)," 2012. [Online]. Available: http://zhanglab.wikidot.com/handedness

[96] J. Heikkila and O. Silven, "A four-step camera calibration procedure with implicit image correction," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. San Juan, Puerto Rico, USA: IEEE, Jun. 1997, pp. 1106–1112, iSSN: 1063-6919.

[97] M. A. A. Haseeb and R. Parasuraman, "Wisture: RNN-based Learning of Wireless Signals for Gesture Recognition in Unmodified Smartphones," *arXiv:1707.08569 [cs]*, Jul. 2017, arXiv: 1707.08569. [Online]. Available: http://arxiv.org/abs/1707.08569

[98] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *International Conference for Learning Representations*, San Diego, CA, USA, 2015, pp. 1–13. [Online]. Available: https://dblp.org/db/conf/iclr/iclr2015.html

[99] N. Japkowicz, "The Class Imbalance Problem: Significance and Strategies," in *Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI)*, Las Vegas, Nevada, USA, 2000, pp. 111–117.

[100] Y. S. Aurelio, G. M. de Almeida, C. L. de Castro, and A. P. Braga, "Learning from imbalanced data sets with weighted cross-entropy function," *Neural Processing Letters*, pp. 1–13, 2019.

[101] D. C. Wells, "The Mode Filter: A Nonlinear Image Processing Operator," in *Instrumentation in Astronomy III*, vol. 0172. Tucson, AZ, USA: International Society for Optics and Photonics, May 1979, pp. 418–421. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/0172/0000/The-Mode-Filter-A-Nonlinear-Image-Processing-Operator/10.1117/12.957111.short

[102] M. F. Land, "Eye movements and the control of actions in everyday life," *Progress in Retinal and Eye Research*, vol. 25, no. 3, pp. 296–324, May 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1350946206000036

[103] G. Hoffman, "Evaluating Fluency in Human–Robot Collaboration," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 3, pp. 209–218, Jun. 2019, conference Name: IEEE Transactions on Human-Machine Systems.

[104] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *2011 International Conference on Computer Vision*, Barcelona, Spain, Nov. 2011, pp. 1036–1043.

[105] Y. Li, M. Liu, and J. M. Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 619–635.

[106] B. Soran, A. Farhadi, and L. Shapiro, "Action Recognition in the Presence of One Egocentric and Multiple Static Cameras," in *Computer Vision – ACCV 2014*, D. Cremers, I. Reid, H. Saito, and M.-H. Yang, Eds. Cham: Springer International Publishing, 2015, vol. 9007, pp. 178–193, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-319-16814-2_12

[107] A. Furnari and G. Farinella, "What Would You Expect? Anticipating Egocentric Actions With Rolling-Unrolling LSTMs and Modality Attention," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, Oct. 2019, pp. 6251–6260. [Online]. Available: https://ieeexplore.ieee.org/document/9008264/

[108] S. Sudhakaran, S. Escalera, and O. Lanz, "LSTA: Long Short-Term Attention for Egocentric Action Recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 9946–9955. [Online]. Available: https://ieeexplore.ieee.org/document/8954401/

[109] M. Liu, S. Tang, Y. Li, and J. Rehg, "Forecasting Human-Object Interaction: Joint Prediction of Motor Attention and Actions in First Person Video,"

arXiv:1911.10967 [cs], Jul. 2020, arXiv: 1911.10967. [Online]. Available: http://arxiv.org/abs/1911.10967

[110] S. Schaal, "Dynamic Movement Primitives -A Framework for Motor Control in Humans and Humanoid Robotics," in *Adaptive Motion of Animals and Machines*, H. Kimura, K. Tsuchiya, A. Ishiguro, and H. Witte, Eds. Tokyo, Japan: Springer, 2006, pp. 261–280. [Online]. Available: https://doi.org/10.1007/4-431-31381-8_23

[111] B. Velichkovsky, A. Sprenger, and P. Unema, "Towards gaze-mediated interaction: Collecting solutions of the "Midas touch problem"," in *Human-Computer Interaction INTERACT '97: IFIP TC13 International Conference on Human-Computer Interaction, 14th–18th July 1997, Sydney, Australia*, ser. IFIP — The International Federation for Information Processing, S. Howard, J. Hammond, and G. Lindgaard, Eds. Boston, MA: Springer US, 1997, pp. 509–516. [Online]. Available: https://doi.org/10.1007/978-0-387-35175-9_77

[112] C.-M. Huang and A. L. Thomaz, "Joint Attention in Human-Robot Interaction," in *2010 AAAI Fall Symposium Series*, Nov. 2010. [Online]. Available: https://www.aaai.org/ocs/index.php/FSS/FSS10/paper/view/2173

[113] R. Eidenberger and J. Scharinger, "Active perception and scene modeling by planning with probabilistic 6D object poses," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2010, pp. 1036–1043, iSSN: 2153-0866.

[114] J. H. Kim, K. Abdel-Malek, Z. Mi, and K. Nebel, "Layout Design using an Optimization-Based Human Energy Consumption Formulation," SAE International, Warrendale, PA, SAE Technical Paper 2004-01-2175, Jun. 2004, iSSN: 0148-7191, 2688-3627. [Online]. Available: https://www.sae.org/publications/technical-papers/content/2004-01-2175/

[115] D. Ognibene and G. Baldassare, "Ecological Active Vision: Four Bioinspired Principles to Integrate Bottom–Up and Adaptive Top–Down Attention Tested With a Simple Camera-Arm Robot," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 1, pp. 3–25, Mar. 2015, conference Name: IEEE Transactions on Autonomous Mental Development.

[116] D. Ognibene and Y. Demiris, "Towards Active Event Recognition," in *Twenty-Third International Joint Conference on Artificial Intelligence*, Jun. 2013. [Online]. Available: https://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/view/6705

[117] R. M. Aronson, T. Santini, T. C. Kübler, E. Kasneci, S. Srinivasa, and H. Admoni, "Eye-Hand Behavior in Human-Robot Shared Manipulation," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '18. New York, NY, USA: Association for Computing Machinery, Feb. 2018, pp. 4–13. [Online]. Available: https://doi.org/10.1145/3171221.3171287

[118] A. D. Dragan and S. S. Srinivasa, “A policy-blending formalism for shared control,” *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 790–805, Jun. 2013. [Online]. Available: http://ijr.sagepub.com/cgi/doi/10.1177/0278364913490324

[119] D.-J. Kim, R. Hazlett-Knudsen, H. Culver-Godfrey, G. Rucks, T. Cunningham, D. Portee, J. Bricout, Z. Wang, and A. 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*, vol. 42, no. 1, pp. 2–14, Jan. 2012, conference Name: IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans.

[120] D. Gopinath, S. Jain, and B. D. Argall, “Human-in-the-Loop Optimization of Shared Autonomy in Assistive Robotics,” *IEEE Robotics and Automation Letters*, vol. 2, no. 1, pp. 247–254, Jan. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7518989/

[121] K. A. Hoff and M. Bashir, “Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 57, no. 3, pp. 407–434, May 2015. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0018720814547570

[122] J. M. McGuirl and N. B. Sarter, “Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information,” *Human factors*, vol. 48, no. 4, pp. 656–665, 2006.

[123] S. Mattsson, Å. Fasth, C. Berlin, and J. Stahre, “Describing human-automation interaction in production,” in *Swedish Production Symposium, SPS12*, 2012.

[124] B. M. Muir, “Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems,” *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, 1994.

[125] J. D. Lee and K. A. See, “Trust in automation: Designing for appropriate reliance,” *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[126] S. T. Fiske, A. J. Cuddy, and P. Glick, “Universal dimensions of social cognition: Warmth and competence,” *Trends in cognitive sciences*, vol. 11, no. 2, pp. 77–83, 2007.

[127] E. A. Bustamante, “A reexamination of the mediating effect of trust among alarm systems’ characteristics and human compliance and reliance,” in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 53, no. 4. SAGE Publications Sage CA: Los Angeles, CA, 2009, pp. 249–253.

[128] N. Bagheri, G. A. Jamieson *et al.*, "Considering subjective trust and monitoring behavior in assessing automation-induced "complacency."," *Human performance, situation awareness, and automation: Current research and trends*, vol. 1, pp. 54–59, 2004.

[129] V. L. Pop, A. Shrewsbury, and F. T. Durso, "Individual differences in the calibration of trust in automation," *Human factors*, vol. 57, no. 4, pp. 545–556, 2015.

[130] X. J. Yang, V. V. Unhelkar, K. Li, and J. A. Shah, "Evaluating effects of user experience and system transparency on trust in automation," in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI.* IEEE, 2017, pp. 408–416.

[131] D. Holliday, S. Wilson, and S. Stumpf, "User trust in intelligent systems: A journey over time," in *Proceedings of the 21st international conference on intelligent user interfaces*, 2016, pp. 164–168.

[132] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 2013, pp. 251–258.

[133] A. S. Clare, M. L. Cummings, and N. P. Repenning, "Influencing trust for human–automation collaborative scheduling of multiple unmanned vehicles," *Human factors*, vol. 57, no. 7, pp. 1208–1218, 2015.

[134] C. L. Lisetti and F. Nasoz, "Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 11, p. 929414, Sep. 2004. [Online]. Available: https://doi.org/10.1155/S1110865704406192

[135] S. Nikolaidis, E. Kasneci, and S. Srinivasa, "Leveraging eye tracking and physiological signals for fluent human robot collaboration," in *Proc. IROS Workshop Human Robot Interaction Collaborative Manuf. Environ.(HRI-CME)*, 2017.

[136] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, vol. 35, no. 10, pp. 1243–1270, Oct. 1992. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00140139208967392

[137] R. Riedl and A. Javor, "The biology of trust: Integrating evidence from genetics, endocrinology, and functional brain imaging." *Journal of Neuroscience, Psychology, and Economics*, vol. 5, no. 2, pp. 63–91, 2012. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0026318

[138] W.-L. Hu, K. Akash, N. Jain, and T. Reid, "Real-Time Sensing of Trust in Human-Machine Interactions**This material is based upon work supported by the National Science Foundation under Award No. 1548616. Any opinions, findings,

and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation." *IFAC-PapersOnLine*, vol. 49, no. 32, pp. 48–53, 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2405896316328609

[139] A. Xu and G. Dudek, "Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2015, pp. 221–228.

[140] S. M. M. Rahman and Y. Wang, "Mutual trust-based subtask allocation for human–robot collaboration in flexible lightweight assembly in manufacturing," *Mechatronics*, vol. 54, pp. 94–109, Oct. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957415818301211

[141] E. Tulving and D. L. Schacter, "Priming and human memory systems," *Science*, vol. 247, no. 4940, pp. 301–306, 1990.

[142] A. H. Dar, A. S. Wagner, and M. Hanke, "REMoDNaV: robust eye-movement classification for dynamic stimulation," *Behavior Research Methods*, vol. 53, no. 1, pp. 399–414, Feb. 2021. [Online]. Available: https://doi.org/10.3758/s13428-020-01428-x

[143] J. J. Braithwaite, D. P. Z. Watson, R. O. Jones, and M. A. Rowe, "Guide for analysing electrodermal activity & skin conductance responses for psychological experiments," *CTIT technical reports series*, 2013.

[144] D. Ungureanu, F. Bogo, S. Galliani, P. Sama, X. Duan, C. Meekhof, J. Stühmer, T. J. Cashman, B. Tekin, J. L. Schönberger *et al.*, "Hololens 2 research mode as a tool for computer vision research," *arXiv preprint arXiv:2008.11239*, 2020.