Hand-Eye Coordination Primitives for Assistive Robotic Co-Manipulation

Benjamin A. Newman[1], Kris M. Kitani[1], Henny Admoni[1]

Abstract— **Robots can help augment human performance in teleoperation tasks that are difficult to complete, for example, assisting a user with motor impairment to eat independently. To do so, robots must intelligently recognize the user's activity state and offer the appropriate type of assistance. Prior work has shown that a user's teleoperation input (such as a joystick control signal) can be used to predict their activity. However, basing such assessments only on direct control signals misses the opportunity to use rich human behavior signals that can further reveal user state, specifically user intent. For example, eye gaze is tightly linked to the target and timing of manipulation, and has been shown to predict a users actions and identify errors in teleoperation. We propose a semantically meaningful dictionary of hand-eye coordination primitives to characterize a user's state during co-manipulation. We perform extensive analysis of a human-robot collaboration dataset, HARMONIC, to extract frequently occurring hand-eye coordination primitives, and identify a set of five action primitives (*exploration, pursuit, correction, mode switching*, and *toggle*) that characterize the user's state. Additionally, we design data-driven models to automatically classify these primitives. Preliminary experiments with both synthetic and real data reveal the potential and limitations of state-of-the-art learning approaches.**

I. INTRODUCTION

Without any algorithmic support, it can be extremely taxing for an individual to directly control, *i.e.*, teleoperate, such a robot. This is because the number of DOFs of the robot being controlled is generally far greater than those of the input device. This challenge is compounded when considering people with motor disabilities, for whom the expressivity of the input device may decrease further in order to meet the requirements of their limited mobility. Additionally, it is not sufficient to allow the robot to complete these tasks autonomously, as previous work have shown that retaining the users explicit control is especially important in assistive domains, where users strongly prefer systems that allow them to stay in control of assistive robots, even if they are less efficient at completing the task [1], [2].

To address this, researchers have developed shared autonomy algorithms that combine user control with autonomous robot behavior, resulting in co-manipulation of the robot [1], [3], [4]. For shared autonomy algorithms to be successful, they must have the ability to accurately characterize a user's state to support complex and high-dimensional co-manipulation tasks, such as assisting a user with motor impairment to eat with a 6 degree of freedom (DOF) arm. These algorithms use direct control signals, such as the operator's joystick inputs, to predict the operator's goals so that the robot can take a cooperative action to assist the user.

[1]Benjamin A. Newman, Kris M. Kitani, and Henny Admoni are with The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, correspondence to newmanba@cmu.edu

However, basing such goal predictions only on direct control signals, such as explicit joystick behavior, misses the opportunity to use rich human behavior signals that can further reveal user state, specifically user intent. For example, eye gaze is tightly linked to the target and timing of manipulation actions in people [5], [6]. In human-robot co-manipulation tasks, here cooperative eating with a 6 DOF robot arm, eyes can be used to predict a user's actions or identify errors in teleoperation [7], [8], [9]. Eye gaze is therefore a natural mechanism to supplement the human goal prediction that takes place during shared autonomy.

To successfully incorporate this signal into the shared autonomy paradigm, it is first necessary to understand hand-eye coordination in co-manipulation. Specifically, we need to determine the basic building blocks of co-manipulation that will allow us to coordinate between a user's eye gaze and how they control the robot's end-effector (here by using their hand to manipulate a joystick). Coordination allows us to relate the varying task-relevant information contained in the different data streams to each other. For example, in assisted eating, the joystick can reveal the immediate vector in which a person wants the robot to travel, but eye gaze can reveal the ultimate bite of food the user wishes to spear; here, it is important to perform an action that does not move too far away from the immediate action, while still optimizing for the overall goal. By relating the joystick and eye gaze streams in this example, we can get a more complete vision of the user's state: not just where they want to go, but how they would like to get there.

Before coordinating the data though, it must be processed. From an algorithmic perspective, processing human behavior signals like eye gaze, head pose, or joystick inputs is non-trivial. These signals are noisy, and different data streams provide different task-relevant features. In addition to the algorithmic complexities, simply obtaining these signals is challenging, as collecting data from multiple sensors in order to train data-driven models can require burdensome engineering efforts to set up, calibrate, and synchronize. Fortunately, there have been a few large-scale dataset collection efforts for teleoperation and human-in-the-loop co-manipulation tasks [10], [11].

In this paper, we use a large-scale multi-modal data set called HARMONIC in order to identify the basic building blocks of hand-eye coordination exhibited by people during co-manipulation tasks (Fig. 1). We identify semantic, multi-modal action primitives that establish the basis for a user's state. Then, we apply modern data-driven techniques to classify multi-modal, multi-view real world data into these action primitives, in order to verify our choice in primitives as well as show that coordination between eye gaze and joystick is possible. Finally, we justify the multi-modal problem by showing that uni-modal analyses are not

sufficient to explain joint behaviors.

We provide several novel contributions toward an understanding of hand-eye coordination in co-manipulation. We first define macro action primitives, which segment user actions into meaningful sequences of individual user states, and discuss how they differ from physiological gaze primitives (Section III-A). In order to evaluate our data-driven recognition models for macro actions, we create a synthetic dataset that contains these action primitives so that we have full control over the generative process and have perfect access to ground truth annotations (Section III-C). We show how these semantic macro action primitives can be modelled using both the synthetic and real raw data (Section IV), and provide a thorough experimental analysis of our models (Section V).

II. RELATED WORK

Prior work has considered the use of direct control signals in order to provide robotic assistance. These direct signals can take the form of deliberate verbal interaction [12], [13], joystick input [3], [14], or even calculated hand and arm motions that show a robot how to complete a manipulation task [15]. While these signals vary drastically in their modalities, all of the control signals require calculated and deliberate actions from the participant.

Previous work has also modelled uni-modal and multi-modal primitives in order to aid in human robot co-manipulation. Uni-modal approaches have focused on recording human behavior, generating primitives from these recordings, implementing them as robot actions, and then testing by having a human complete the task with the robot [16], [17]. Multi-modal approaches have focused on generating primitives from recordings of various views of the same or similar direct control input (*e.g.*, EMG and human arm manipulability) in periodic tasks (*e.g.*, sawing a board with a robot) [18], [19]. Regardless of the modality, these approaches focus solely on the direct control inputs.

While these approaches have shown success, they neglect other signals that people naturally display while engaging with the world, with no added burden to the user. For example, previous work has shown that humans naturally elicit a wide range of nonverbal behaviors [20], [21] when performing collaborative tasks. In particular, eye gaze is tightly liked to hand movements, especially during manipulation tasks. When reaching for an object, gaze to that object typically precedes hand motion by about 600ms [6]. Gaze typically moves to the next object before the hand reaches its target [21], and gaze rarely rests on objects that are not involved in the current task [5].

Nonverbal behaviors have been previously used for direct control in remote robot navigation [22], drone teleoperation [23], and human-robot co-manipulation in a table carrying task [24]. However, these approaches do not consider using naturalistic gaze as an indirect, and supplementary control method, as we do in this work.

In human robot co-manipulation, prior work has begun to characterize hand-eye coordination when operating a robot under shared autonomy [7]. Though this work characterized many important interactions, it did not provide formal primitives or exhaustively analyze the relationships between joystick and eye gaze signals in a data-driven manner.

To study this problem in a naturalistic environment, we use the Human and Robot Multimodal Observations of Natural Interactive Collaboration (HARMONIC) dataset [10]. This dataset is described in detail in III-B.

III. PROBLEM DOMAIN

We build models of hand-eye coordination in human-robot co-manipulation to better understand user state during high dimensional co-manipulation tasks such as assisted eating. To describe hand-eye coordination, we define action primitives that provide a semantic understanding of user state. We draw our action primitive definitions by semantically analyzing the HARMONIC dataset, which is described briefly here. Additionally, to test these primitives in a systematic fashion, we construct a synthetic dataset, described below.

A. Defining Action Primitives

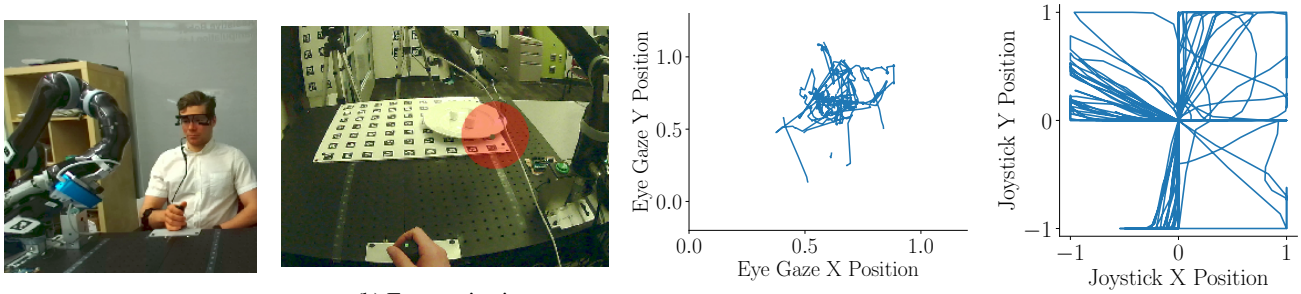
In our current work, we use the term *micro actions* to refer to three low-level gaze action primitives that can be used to understand attention, or user state. *Fixations* are eye movements that focus eye gaze on a single point in space, and are used to gather visual details. *Saccades* are fast, point-to-point movements of the eyes that bring a new area into the center of vision. Finally, *smooth pursuits* are when the eyes track a moving object to keep it in the center of vision.

Micro actions only partially express hand-eye coordination during co-manipulation, however, because they do not capture the robot's movement. For this, we manually analyzed the HARMONIC data and identified five common *macro actions* (Fig. 2) comprised of eye gaze and joystick movements. Our five macro actions are: exploration, correction, pursuit, mode switch, and toggle.

Exploration is defined by minimal joystick activity and high eye gaze activity. Semantically, this class represents a person exploring the space with their eye gaze, preparing to make an action with the joystick. This sequence starts when the joystick moves into a period of rest and ends once the user activates the joystick. This can be seen in Fig. 3 where the joystick sits generally at the bottom of the plot, indicating no movement throughout the sequence, and the eye gaze is dispersed throughout the plot.

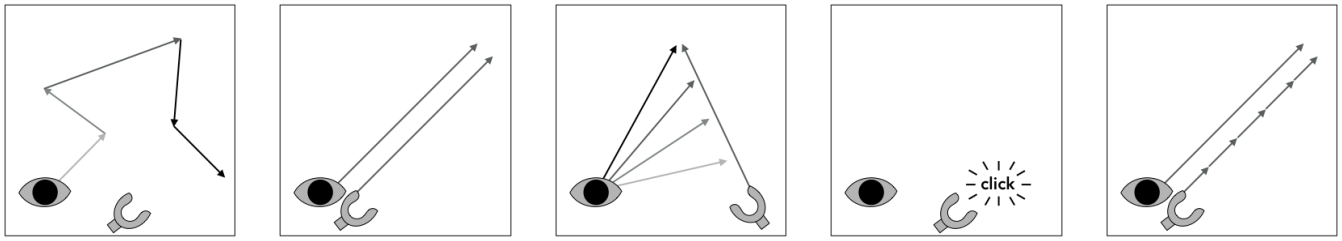
Pursuit is defined by correlated eye gaze and joystick action. In this class (which is not to be confused with the micro action primitive *smooth pursuit*), the participant moves the joystick and follows the resulting robot action with their gaze. This may result in large eye gaze movements when the robot is moving across the visual scene or little eye gaze movement when the robot's end effector is rotating. This action begins when the eye gaze begins to follow the robot's action (as resulting from the joystick activation) and ends once the eye gaze moves away from the previously fixated position. This relationship can be seen in Fig. 3 where the gaze and joystick signals are tightly coupled.

Correction can be categorized by high joystick activity and consistent eye gaze glances between a "home" point and another task relevant scene point. Prior work has called these "monitoring" glances [7]. This action can reveal an operator's goal or the target of their current control input. This sequence begins as eye gaze moves away from a previously fixated position during joystick activation. It ends once the eye gaze has travelled back to the original position (after one or more fixations elsewhere in the scene), the joystick comes to a period of long rest, or the participant



(a) Third person view (b) Egocentric view (c) Eye gaze data in one trial (d) Joystick data in one trial

Fig. 1: The HARMONIC data set contains (a) third person video, (b) egocentric video, (c) eye gaze fixations, and (d) joystick data from a human-robot co-manipulation task.



(a) Exploration (b) Pursuit (c) Correction (d) Mode switch (e) Toggle

Fig. 2: Five macro action labels capture combined eye gaze and end effector dynamics. a) Exploration denotes periods of high eye gaze movement and low joystick (robot) movement. b) Pursuit denotes periods of highly correlated eye gaze and joystick movements, where the eye gaze follows the path of the robot. c) Correction denotes successive glances between different parts of the scene while the robot is moving. d) Mode switch denotes when the user is using modal control to cycle through sets of degrees of freedom. e) Toggle denotes periods in which the joystick is being moved in rapid, short, consistent activations.

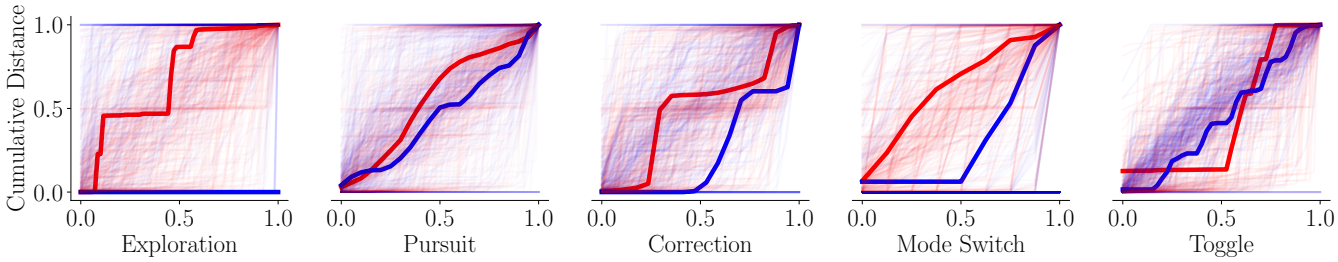


Fig. 3: A graphical description of the differences between macro action categories from HARMONIC. Eye gaze sequences are red, while joystick sequences are blue. The y-axis shows the normalized cumulative distance for each sequence. The x-axis shows normalized sequence lengths. Every trial is plotted, with a representative sequence highlighted in bold.

enters into one of the other semantic categories. This can be seen in Fig. 3 where the eye gaze initially takes a stair step approach indicating fast movement initially, a pause and then fast movement again. This pattern is then followed in the joystick channel.

Mode switch represents when the participant switches control modes. This class is programatically generated by taking the five frames before and after the button press that causes a robot control mode switch.

Toggle is defined by quick, successive joystick taps with the eye gaze path closely following the end effector. This begins when the participant makes short bursts with the joystick, and ends either when the joystick comes to a period of inactivity or consistent activity. This can be seen in Fig. 3, where the cumulative distance of the joystick takes a stair step pattern, while the gaze initially lags behind and then catches up at the end of the sequence.

B. HARMONIC Dataset

As the source of real-world human-robot co-manipulation data, we used the previously released HARMONIC dataset [10]. HARMONIC includes data from 24 people operating a Kinova MICO robot in an assistive eating task. Among many data streams, the dataset provides binocular eye gaze at 120Hz and joystick signal at 120Hz/100Hz (moving/hold) with time stamps allowing for the ability to flexibly re-sample.

The full HARMONIC dataset was recorded to examine people's use of robot assistance in co-manipulation. Participants controlled the end effector of a 6-DOF robot using a 2-axis

joystick to complete a food spearing task (Fig. 1). Because the joystick input has fewer dimensions than the robot end effector, people operated the robot two DOFs at a time, and pressed a button on the joystick to switch modes. Participants each completed five trials of the food spearing action under four different assistance levels, for a total of twenty trials per person.

As participants completed the eating co-manipulation task, their eye gaze, joystick inputs, robot position, and other signals were recorded. Eye gaze was recorded using a Pupil sensor [25], [26], which captured pixel position of gaze fixation on an egocentric view of the world (Fig. 1). Joystick signal was recorded as a two dimensional value, where both the x and y coordinates range from -1 to 1.

Though the full dataset includes approximately five hours of data, for the current analysis, we are investigating teleoperation only. Therefore, we only included the five trials per participant where people were fully teleoperating the robot (*i.e.*, the robot assistance signal was set to zero).

C. Synthetic Dataset

Real-world data is noisy, so we developed a synthetic dataset that allows us to test our models with hypothetically perfect inputs. Additionally, it provides an opportunity to control and experiment under a variety of noise parameters, prototype experiments at scale, have full control over the data generator, and have access to actual ground truth labels.

This synthetic dataset was designed to mimic the task in the HARMONIC dataset. The robot end effector navigates to a virtual goal, while a virtual eye gaze stream is simultaneously overlaid on the scene. As seen in Fig. 4, the robot is represented as a triangle, the eye gaze as a square, and the goal as a circle. The robot aims to navigate to within a threshold of the goal. Trials with fewer than 200 or more than 1000 frames were discarded. For further information on how these data were generated, please refer to the supplemental materials.

Table I shows the distribution of micro and macro actions in both HARMONIC and synthetic datasets. We can see that the number of sequences is relatively balanced in the HARMONIC dataset, with there being fewer toggle sequences overall.

TABLE I: Distribution of class labels in HARMONIC and synthetic datasets for both the micro and macro classification tasks.

		HARMONIC	Synthetic
Micro	Saccade	0.1796	0.4321
	Smooth Pursuit	0.5541	0.1743
	Fixation	0.2663	0.3936
Macro	Exploration	0.2319	0.3377
	Correction	0.2424	0.0993
	Pursuit	0.1765	0.1140
	Mode Switch	0.2181	0.3377
	Toggle	0.1311	0.1113

IV. Method

A. Micro and Macro Action Labeling

Micro action labels were automatically classified using Bayesian Decision Theory Identification (I-BDT) [27] which classifies gaze-actions in online settings. For an explanation of this algorithm, we refer readers to the original paper, as well as our supplementary material.

Macro actions were hand labeled for ten participants in the HARMONIC dataset. Sequences with significant amounts of low confidence eye gaze calculations (given by the eye tracker). Two of the ten labelled participants (p102 and p103) were discarded completely due to significant missing data. For the remaining labelled participants, we dropped 62 of 2931 total sequences (2.1% of sequences) or 2931 of 48135 frames (3.4% of frames) because of missing gaze data. Supplementary information contains more details about our exclusion criteria, as well as our subsampling method.

B. Models and Input Representations

We built models for micro and macro actions using a two-layer Gated Recurrent Unit Recurrent Neural Network (GRU-RNN) [28] to encode a given input sequence using the PyTorch neural network library [29]. We tested four input families: eye gaze only, joystick only, an early fusion of eye gaze and joystick, and a late fusion of eye gaze and joystick (Table II. We further considered different hidden sizes for these models, which are shown in the *hsize* column of Table II.

To decode these sequences into classification vectors, we collect the context vectors for each step in the sequence, and then feed this into a three layer Multi-Layer Perceptron. In the late fusion models, the eye gaze and joystick signals are each encoded by two separate encoders, and then the context vectors are concatenated prior to being decoded. This is in contrast with the early fusion models, in which the eye gaze and joystick sequences are concatenated along the feature axis and then jointly encoded.

The first layer of the decoder is the product of the maximum sequence length and the hidden layer size. The second layer is half that, and the final layer is the number of classes. This decoder model is fully connected, and ReLU [30] is used for non-linearity after the first and second layers. All models were trained using the Adam optimizer [31] using a learning rate of 1e-3 and cross-entropy loss weighted by the class distribution (given in Table I).

The *x,y* embedding indicates that the inputs are being given to the model as is, with no modifications. For synthetic data, this is the x,y position of both eye gaze and joystick. For HARMONIC data, eye gaze includes the confidence score from the eye tracker, while joystick additionally includes a one hot vector indicating the current control mode. The *dx,dy* embedding indicates that the difference (or discrete derivative) of the signal is taken before passing the input to the model. Finally, the *binary* representation divides the input space into a 10x10 grid and generates a one hot vector indicating the pixel closest to the real valued number. This vector is then passed to the model as input.

C. Problem Setup

For both tasks, we consider a supervised classification problem. Our goal here is to show a correlation between the segmented raw data and our provided macro labels. Outperforming chance and the zero-rule (guessing the majority class) shows that the chosen macro labels are good segmentations of the raw data. In future work, the representations learned by these models could

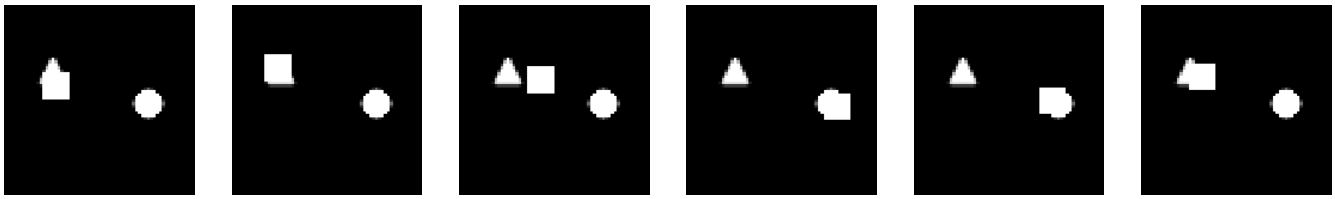


Fig. 4: Our synthetic dataset was modeled as a simplified version of the eye gaze and joystick signals from the HARMONIC task. Here, an example of the exploration action, with simulated eye gaze (square), robot position (triangle), and goal (circle).

TABLE II: Results for the synthetic and HARMONIC datasets on the micro and macro classification tasks. We report accuracy (acc) and mean average precision (mAP). We test a variety of input streams: eye gaze only (eye only), joystick only (joy only), an early fusion of eye gaze and joystick (eye+joy (e)), and a late fusion of eye gaze and joystick (eye+joy (l)). Additionally, we test different embeddings (embed) of the inputs: binary, real (x, y), and difference (dx, dy). Finally, we test two different hidden unit sizes (hsize): 16 and 256.

			Micro				Macro			
input	embed	hsize	acc (synth)	acc (real)	mAP (synth)	mAP (real)	acc (synth)	acc (real)	mAP (synth)	mAP (real)
eye only	binary	256	0.9186	0.6964	0.9670	0.6700	0.8310	0.4884	0.7266	0.4431
eye only	x,y	256	0.9620	0.6648	0.9949	0.6666	0.7763	0.4884	0.5981	0.4522
eye only	x,y	16	0.9316	0.6774	0.9738	0.6716	0.8539	0.4653	0.7713	0.4459
eye only	dx,dy	256	0.8656	**0.7076**	0.8703	0.7027	0.9185	0.5347	0.9237	0.4805
eye only	dx,dy	16	**0.9695**	0.7006	**0.9963**	**0.7028**	0.9195	**0.5644**	0.9050	**0.5054**
joy only	binary	256	0.8671	0.5182	0.9068	0.5037	0.7783	0.6436	0.6274	0.5886
joy only	x,y	256	0.9141	0.5561	0.9464	**0.5248**	0.8062	0.6931	0.6710	0.6797
joy only	x,y	16	0.9016	0.5610	0.9328	0.5148	0.7962	0.6502	0.6599	0.6642
joy only	dx,dy	256	0.9456	**0.5638**	0.9895	0.5112	0.8917	0.5479	0.8230	0.5241
joy only	dx,dy	16	**0.9486**	0.5372	**0.9902**	0.5079	0.9006	0.5809	0.8203	0.5666
eye+joy (e)	binary	256	0.9416	**0.6830**	0.9804	0.6668	0.7952	0.6205	0.6595	0.6244
eye+joy (e)	x,y	256	0.8971	0.6669	0.9242	0.6822	0.8082	0.6997	0.6847	0.7053
eye+joy (e)	x,y	16	0.9226	0.6767	0.9555	0.6895	0.8956	0.6832	0.8456	0.6682
eye+joy (e)	dx,dy	256	0.9575	0.6522	0.9955	0.6608	0.9652	0.5578	0.9664	0.5452
eye+joy (e)	dx,dy	16	**0.9650**	0.6669	**0.9956**	**0.6915**	0.9543	0.5248	0.9706	0.4748
eye+joy (l)	binary	256	0.9271	0.6767	0.9776	0.6864	0.8678	0.6271	0.7979	0.6357
eye+joy (l)	x,y	256	0.9386	0.6669	0.9861	0.6782	0.8111	0.6238	0.6702	0.6613
eye+joy (l)	x,y	16	0.9051	0.6697	0.9434	0.6869	0.7873	0.6964	0.6680	0.6997
eye+joy (l)	dx,dy	256	0.9640	**0.6957**	0.9954	**0.6930**	0.9612	0.6139	0.9707	0.5889
eye+joy (l)	dx,dy	16	**0.9710**	0.6613	**0.9977**	0.6709	0.9662	0.5314	0.9792	0.5257

TABLE III: Results for the HARMONIC dataset on the macro task using the best representations from Table II.

Input	gaze embed	gaze hsize	joy embed	joy hsize	acc	mAP
eye+joy(ef)	dx,dy	16	x,y	16	0.6799	0.6805
eye+joy(ef)	dx,dy	256	x,y	256	0.6403	0.6197
eye+joy(lf)	dx,dy	16	x,y	256	**0.6997**	**0.7147**

be used as context vectors that can be incorporated into the shared autonomy paradigm. We give the results on the classification problem in Section V.

V. EXPERIMENTAL RESULTS

Experiment results for both real and synthetic data sets on both micro and macro tasks are shown in Table II. Following the initial experiments, an analysis of individual differences on the real was performed for both the micro and macro task. For this, we used k-fold cross validation where each fold was a single participant, as seen in Table IV. Accuracy and mean average precision (mAP) are reported for all experiments. For both micro and macro actions, the chance values were calculated by taking the inverse of the number of classes (three for micro action classification and five for macro action classification). The majority class values are given in Table I for both data sets and tasks. These are calculated by dividing the total number of sequences of a particular class and by the total number of sequences in the entire dataset.

A. Micro Action Results

The best performance for the synthetic data on both metrics came from the joint late fusion model with a hidden size of 16 and inputs represented as the difference of the raw signal. The accuracy score of 0.9710 outperforms both guessing at chance (0.3333) and consistently guessing the majority class (0.4321). Given that these results are an idealized version of the real world data, these numbers represent a theoretical upper bound on performance.

The real data outperformed chance and guessing the majority

class (0.3333 and 0.5541), as well. The best results for these data were with the eye only model (XXX), with the best representation being the difference of the input signal. Accuracy was best under the 256 hidden size model, while mAP performed the best under the 16 hidden size model, but both models performed similarly on both metrics.

B. Macro Action Results

Synthetic data performed well on all categories for macro action classification (Table II). In all cases it out-performed chance (0.2) and guessing the majority class (0.3377). The best performance (XXX) was realized by the late fusion model with hidden size of 16 and input streams represented as the difference of the raw signal. Both fusion models significantly out-performed the single stream models.

The real data also outperformed chance (0.2) and guessing the majority class (0.2424) in all experiments, with the best performance (XXX) resulting from the early fusion model with hidden size of 256 and the original input stream as the input to the model. While this model performed the best, performance on the late fusion model and the joystick only model were similar.

C. Participant Level Cross Validation

Table IV shows accuracy and mAP scores when each participant is considered as their own test set for the micro and macro tasks. The micro task should be compared to the eye only, dx,dy, 256 model, and the macro task should be compared to the eye+joy (e), x,y, 256 model. Evidence to suggest individual differences was found in the macro action classification, but not the micro action classification.

TABLE IV: Micro action cross validation by participant ID (eye only, dx,dy, 256 model). † Data were excluded due to significant noise. ‡ Data were not labeled for macro actions.

	Micro		Macro	
ID	Accuracy	mAP	Accuracy	mAP
p101	0.7021	0.7571	0.5587	0.6551
p102	0.6651	0.6396	†	†
p105	0.7317	0.6690	†	†
p106	0.7161	0.7089	0.6818	0.7742
p107	0.6140	0.6658	0.6023	0.5704
p108	0.7901	0.7346	0.7005	0.7010
p109	0.6984	0.6458	0.6447	0.6140
p110	0.6889	0.7492	0.5209	0.5921
p111	0.5305	0.5527	0.6204	0.6111
p112	0.7105	0.7313	0.4638	0.5165
p113	0.6549	0.6851	‡	‡
p114	0.7111	0.7250	‡	‡
p115	0.7333	0.7019	‡	‡
p117	0.8023	0.6709	‡	‡
p118	0.8175	0.6650	‡	‡
p119	0.5987	0.5872	‡	‡
p121	0.7128	0.7097	‡	‡
p122	0.6942	0.6745	‡	‡
p123	0.6915	0.7144	‡	‡
Avg	0.6923	0.6429	0.5951	0.6233

VI. DISCUSSION

Improved performance in the synthetic data for both classification tasks was realized by jointly modelling eye gaze and joystick data. For real data the best results for micro action classification came by modelling eye gaze alone, with the joystick model only slightly outperforming consistently guessing the majority class. This indicates that in real data, the patterns of joystick behavior do not significantly differ across micro action sequences, further justifying the need to create and analyze the macro actions to encompass user control input signals.

In macro action classification, the joystick and both fusion models all performed similarly, with slight improvements coming from the fusion models. Thus, outperforming the joystick signal is possible on this task, but the fusion between the eye gaze and joystick signals is nontrivial and should be explored further. All models significantly outperformed guessing the majority class, indicating that eye gaze and joystick do display distinct patterns within the proposed primitives.

Another finding was that the model's preferred input representations were consistent across task and dataset. The best performance from both synthetic and real data on the micro and macro classification tasks came from representing the eye gaze as the original, real valued input. The joystick stream was best represented as the difference in the synthetic data and as a real value in the real data, but, as the joystick in the synthetic data actually represents the end effector of the robot, taking the diff of this signal actually results in a signal similar to the joystick data in the real dataset.

Additionally, we found no consistent improvement when considering wider hidden representations, indicating that smaller models can be used to achieve this task. This finding is important, as assistance algorithms must be processed online, and using smaller, lightweight models makes this approach more viable. Furthermore, there was no appreciable difference between the early and late fusion models.

Finally, Table IV shows that micro actions are consistent across participants, as the weighted average of accuracy over the 19-fold cross validation compared similarly to the accuracy of the best eye only model (though mAP underperformed). This is in contrast to the macro action categories, Table IV in which the weighted average of both accuracy and mAP significantly under performed the best early fusion model. This indicates that there are individual differences, and building effective models to account for these differences should be studied further.

VII. CONCLUSION

In this work we are motivated by the need to understand hand-eye coordination for human-robot co-manipulation. This problem is especially important for assistive robotics tasks, in which operators could benefit greatly from the introduction of indirect control signals, such as eye gaze, to assistance algorithms. We introduced a novel concept of macro actions, which are semantic action primitives that represent high-level task activities. These macro actions are complementary to micro actions, which represent low-level behavior. We defined five macro actions that combine eye gaze and joystick on an assistive eating task drawn from the HARMONIC dataset. We then developed multi-modal models of micro and macro actions, and extensively analyzed the models' performance under different parameters. Our analysis further justifies the need for semantic macro primitives, and highlights the benefit of jointly modeling eye gaze and joystick signals within a

single task. Finally, we discussed how participants show individual differences. This work will enable future research into combining indirect and direct control signals to comprehensively perceive human goals in co-manipulation settings.

REFERENCES

[1] 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, 2016.

[2] 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, 2012.

[3] S. Javdani, S. Srinivasa, and A. Bagnell, "Shared autonomy via hindsight optimization," in *Proceedings of Robotics: Science and Systems*, Rome, Italy, July 2015.

[4] A. Dragan and S. Srinivasa, "A policy blending formalism for shared control," *The International Journal of Robotics Research*, May 2013.

[5] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *Trends in Cognitive Sciences*, vol. 9, no. 4, pp. 188–194, 2005.

[6] 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, 2001.

[7] 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: ACM, 2018, pp. 4–13. [Online]. Available: http://doi.acm.org/10.1145/3171221.3171287

[8] R. M. Aronson and H. Admoni, "Gaze for error detection during human-robot shared manipulation," in *RSS Workshop: Towards a Framework for Joint Action*, 2018.

[9] H. Admoni and S. S. Srinivasa, "Predicting user intent through eye gaze for shared autonomy," in *Proceedings of the AAAI Fall Symposium: Shared Autonomy in Research and Practice*. AAAI Press, 2016, pp. 298–303.

[10] B. A. Newman, R. M. Aronson, S. S. Srinivasa, K. Kitani, and H. Admoni, "Harmonic: A multimodal dataset of assistive human-robot collaboration," 2018.

[11] Z. Wang and A. Majewicz Fey, "Human-centric predictive model of task difficulty for human-in-the-loop control tasks," *PLOS ONE*, vol. 13, no. 4, pp. 1–21, 04 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0195053

[12] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," in *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, San Francisco, CA, August 2011, pp. 1507–1514.

[13] A. Broad, J. Arkin, N. Ratliff, T. Howard, and B. Argall, "Real-time natural language corrections for assistive robotic manipulators," *Int. J. Rob. Res.*, vol. 36, no. 5-7, pp. 684–698, June 2017. [Online]. Available: https://doi.org/10.1177/0278364917706418

[14] S. Javdani, H. Admoni, S. Pellegrinelli, S. S. Srinivasa, and J. A. Bagnell, "Shared autonomy via hindsight optimization for teleoperation and teaming," *The International Journal of Robotics Research*, vol. 37, no. 7, pp. 717–742, 2018. [Online]. Available: https://doi.org/10.1177/0278364918776060

[15] J. Kofman, X. Wu, T. J. Luu, and S. Verma, "Teleoperation of a robot manipulator using a vision-based human-robot interface," *IEEE Transactions on Industrial Electronics*, vol. 52, no. 5, pp. 1206–1219, 2005.

[16] H. Ben Amor, G. Neumann, S. Kamthe, O. Kroemer, and J. Peters, "Interaction primitives for human-robot cooperation tasks," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 2831–2837.

[17] G. Maeda, M. Ewerton, R. Lioutikov, H. Ben Amor, J. Peters, and G. Neumann, "Learning interaction for collaborative tasks with probabilistic movement primitives," in *2014 IEEE-RAS International Conference on Humanoid Robots*, Nov 2014, pp. 527–534.

[18] L. Peternel, N. Tsagarakis, and A. Ajoudani, "Towards multi-modal intention interfaces for human-robot co-manipulation," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 2663–2669.

[19] ——, "A humanrobot co-manipulation approach based on human sensorimotor information," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 7, pp. 811–822, July 2017.

[20] J.-D. Boucher, U. Pattacini, A. Lelong, G. Bailly, F. Elisei, S. Fagel, P. Dominey, and J. Ventre-Dominey, "I reach faster when i see you look: Gaze effects in humanhuman and humanrobot face-to-face cooperation," *Frontiers in Neurorobotics*, vol. 6, p. 3, 2012. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnbot.2012.00003

[21] 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, 2001. [Online]. Available: http://www.jneurosci.org/content/21/17/6917

[22] H. O. Latif, N. Sherkat, and A. Lotfi, "Teleoperation through eye gaze (telegaze): A multimodal approach," in *2009 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Dec 2009, pp. 711–716.

[23] 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, Sep. 2014. [Online]. Available: https://bop.unibe.ch/JEMR/article/view/2387

[24] D. J. Agravante, A. Cherubini, A. Bussy, P. Gergondet, and A. Kheddar, "Collaborative human-humanoid carrying using vision and haptic sensing," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 607–612.

[25] Pupil Labs, Inc., "Pupil labs - pupil," 2017, https://pupil-labs.com/pupil/.

[26] M. Kassner, W. Patera, and A. Bulling, "Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '14 Adjunct. New York, NY, USA: ACM, 2014, pp. 1151–1160. [Online]. Available: http://doi.acm.org/10.1145/2638728.2641695

[27] T. Santini, W. Fuhl, T. Kübler, and E. Kasneci, "Bayesian identification of fixations, saccades, and smooth pursuits," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, ser. ETRA '16. New York, NY, USA: ACM, 2016, pp. 163–170. [Online]. Available: http://doi.acm.org/10.1145/2857491.2857512

[28] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. [Online]. Available: https://www.aclweb.org/anthology/D14-1179

[29] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.

[30] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10. USA: Omnipress, 2010, pp. 807–814. [Online]. Available: http://dl.acm.org/citation.cfm?id=3104322.3104425

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.