Supplemental Material for Hand-Eye Coordination Primitives for Assistive Robotic Co-Manipulation

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

I. Subsampling and Exclusion Criteria

Since the joystick and eye tracker recorded data at different frequencies, these two streams needed to be time aligned prior to analysis. The HARMONIC dataset provides a world time index that is aligned to the egocentric camera attached to the eye tracker, which runs at 30 Hz. For the joystick stream, the values closest to this world time index were averaged.

Prior to sub sampling the gaze data, trials with average gaze confidence less than 0.9 and velocity greater than 0.5 pixels per frame were excluded from analysis. After this, the eye gaze data points that were closest to the world time index were first filtered by the eye tracker's confidence for those data points (any individual points below 0.9 were excluded) and then the points were averaged. For both micro and macro actions, sequences with length less than two were discarded from analysis.

II. I-BDT Explanation

A movement ratio, defined as the proportion of non-zero velocities within a window to the total window size, is determined at each time step and used to calculate the initial pursuit prior. A pursuit prior can be determined as the mean of all the previous pursuit priors calculated. The priors for saccade and fixation are each assigned half of the remaining probability distribution.

The current movement ratio is used directly as the likelihood of smooth pursuit. The likelihoods of the fixation and saccade are determined by first fitting a Gaussian Mixture Model over a short sample of the eye gaze velocities to determine a means for fixation and saccade speeds. Then a noisy threshold is applied to the current eye gaze velocity in order to determine the fixation and saccade likelihoods.

Given these definitions for priors and pursuits, Bayes' Rule can be used in order to calculate the posterior for each label given the current eye gaze velocity. The velocities used for classification, as well as the classification outputs, are shown in Figure 1. To adapt this algorithm to our setting, we simply considered the entire trial's eye gaze in order to learn the saccade and fixation means, effectively turning adapting this algorithm to an offline setting.

III. Synthetic Dataset Replication Details

We built the synthetic dataset as an idealized version of the HARMONIC dataset. To construct it, we developed an autonomous virtual agent that completed a simplified task of orienting a robot end effector at a goal location using joystick signals. At the same time, eye gaze signals from the virtual agent were overlayed on the scene.

In order to progress through the task, we defined state transitions between the five possible macro actions described in the paper. The virtual agent always starts in an exploration state, immediately followed by a mode switch. Then, an action out of toggle, pursuit, or correction was randomly sampled. Once this action was complete, the virtual agent entered back into the exploration state and continued until it reached the goal condition.

As previous work has noted [1], people tend to remain axis-aligned when activating the joystick during robot control. Following this observation, we constrained the joystick to move in only the x or y direction when moving toward the goal. The first direction of movement was chosen randomly, and then was then alternated until the joystick met the goal condition.

To mimic real world data, several noise parameters were considered. The goal and agent start positions were randomly placed in the environment with a minimum distance requirement between the two. The mean and standard deviation of the robot speed were both sampled then re-sampled using these initial values upon starting any joystick action.

Similarly, several eye gaze values were sampled. As in the robot signal, the mean and standard deviation of the eye gaze moving speed were sampled at the beginning of the trial, then re-sampled using these parameters on every eye gaze action. A positional jitter was created by sampling a small value to add to the x and y position of the eye gaze at each frame. The standard deviation of these positional jitter values were sampled at the beginning of the trial. Finally, the mean and standard deviation of fixation length were sampled at the beginning of the trial, with re-sampling upon the start of a fixation.

References

[1] 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

[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

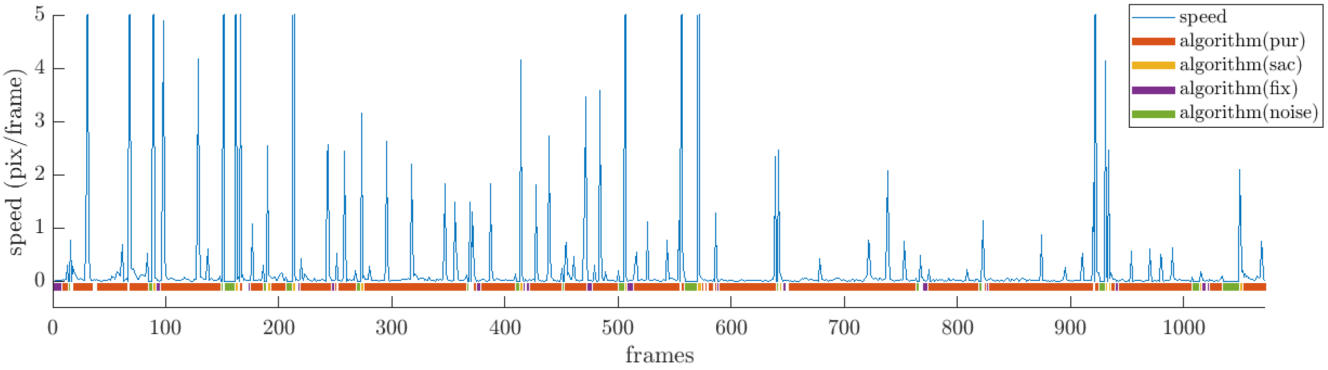

Fig. 1: The I-BDT algorithm classifies fixations (purple), saccades (yellow), and smooth pursuits (red) from gaze movement speed. Automatic detection and classification of physiological gaze labels is open research.