HARMONIC: A Multimodal Dataset of Assistive Human-Robot Collaboration

Benjamin A. Newman[*1]**, Reuben M. Aronson**[*1]
Siddartha S. Srinivasa[2]**, Kris Kitani**[1]**, Henny Admoni**[1]

Abstract

We present HARMONIC, a large multi-modal dataset of human interactions in a shared autonomy setting. The dataset provides human, robot, and environment data streams from twenty-four people engaged in an assistive eating task with a 6 degree-of-freedom (DOF) robot arm. From each participant, we recorded video of both eyes, egocentric video from a head-mounted camera, joystick commands, electromyography from the participant's forearm used to operate the joystick, third person stereo video, and the joint positions of the 6 DOF robot arm. Also included are several data streams that come as a direct result of these recordings, namely eye gaze fixations in the egocentric camera frame and body position skeletons. This dataset could be of interest to researchers studying intention prediction, human mental state modeling, and shared autonomy. Data streams are provided in a variety of formats such as video and human-readable csv or yaml files.

Introduction

In human-robot collaborations, robots need to perceive, understand, and predict the effects of their own actions as well as the actions of their human partners. This is especially important for assistive robots, which perform actions toward a (sometimes implicit) human goal. To successfully produce these assistive actions, the robot system must understand and predict human mental states—the human's goals, intentions, and future actions—that determine what assistance the robot should provide. Understanding these mental states requires perceiving and interpreting human behavior during human-robot collaborations.

When people complete physical tasks, their external behaviors—like their eye gaze—can reveal a lot about their internal mental states. For example, people almost exclusively fixate their gaze on objects or locations involved in their current task[?]. People fixate an object with their gaze before they even begin moving their hand toward it[?]. Gaze lingers on key points in the task, such as obstacles, revealing certain landmarks of manipulation[?]. Additionally, people gaze at objects before uttering verbal references, which others can use for disambiguating and predicting speech[?,?].

Other human behaviors can also reveal current mental states. Electromyography (EMG) signals, which record the electrical stimulation of muscle fibers, can indicate what action people are attempting to complete with their hands. Pupil size has been correlated with cognitive load[?]. And understanding current human body posture can both reveal desired tasks and help to avoid potentially dangerous collisions[?].

In this paper, we present the Human And Robot Multimodal Observations of Natural Interactive Collaboration (HARMONIC) dataset. The HARMONIC dataset contains

Figure 1. The HARMONIC dataset provides multimodal human, robot, and environmental data collected during an assistive human-robot collaboration.

human, robot, and environment data collected during the human-robot collaborative task (Figure 1). In this task, people control an assistive robot arm to pick up bites of food in a simple eating scenario. The robot is controlled through

[*]Denotes equal contribution
[1]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA
[2]University of Washington, Seattle, WA

Corresponding author:
Benjamin A. Newman, Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213
Email: newmanba@cmu.edu

a 2 axis joystick, and in some cases it provides additional assistance through shared autonomy[?].

Human behavioral data in this dataset include an egocentric RGB video with eye gaze fixation position, IR videos of both eyes, an external view of the participant through stereo video, and EMG recordings on the joystick-controlling arm. Additional data includes continuous robot position, joystick control inputs from the user, and 3D locations of the food morsels. These data streams are detailed in the sections below.

We expect this dataset to be useful for robotics researchers investigating human-robot interaction and collaboration, as well as computer perception researchers studying multimodal human behavior. For example, researchers could use this dataset to learn correlations between eye gaze and joystick control, in order to improve the predictions of shared autonomy algorithms. Others might be interested in the dynamics of joystick inputs during different amounts of robot assistance. This dataset is currently being used to identify human errors by learning a normative gaze behavior model and identifying anomalies[?].

Prior Work

Multimodal datasets have garnered much interest in many different communities, such as psychology, computer vision, human-robot interaction, and natural language processing. These datasets, though, can be difficult to collect on a large scale due to the increasing engineering demand required with each additionally desired signal. As such, many multimodal datasets have either few participants or instances, or few signals. This dataset attempts to give researchers access to a dataset that has a substantial number of individual instances and datastreams. Due to the high number of data streams, this dataset has the potential to impact research in many different fields by utilizing any subset of the offered modalities. This impact, and the work to which is relates, is described in the following section.

Robotic Control

Eyegaze, EMG, and body pose have all been useful signals for robotic control. Since eyegaze is a rich signifier of intention during manipulation, both by hand[? ? ?] and by robot[?], its use has been explored through numerous robotic collaboration settings, including anticipating which object a user will request[?], triggering assistive aid during autonomous driving[?]. Electromyography signals have been used for robot control[?] and task monitoring[?]. By making this dataset available, we intend to enable further research into these control methods.

Social Gaze

The field of social gaze studies both the understanding and synthesis of gaze patterns during social interactions. Much of this work is done in settings when multiple humans are interacting with each other or a number of robots. From these studies we can begin to understand what role gaze plays in conversation. Gaze, though, is an extremely informative signal in many contexts.

Theory of Mind

Also critical to successful human-robot interaction is the ability of the robot to understand what the human desires in a given task. Recently, the Charades-Ego datasets introduced the problem of perspective taking, or discovering links between first and third person videos. HARMONIC builds on this idea by allowing researchers to discover correlations between any number of the offered streams, potentially giving an agent a more complete view of the humans intentions.

Eye Gaze Mechanics

In addition to developing systems that improve specific subsets of HRI, we believe that this dataset could also provide an opportunity for researchers to develop, test, and improve eye gaze perception algorithms. While there exist natural, egocentric, eye gaze datasets many of these include the participants hands in the majority of frames. HARMONIC gives researchers an opportunity to develop eye gaze prediction algorithms in scenarios where the participants eyes and hands have differing foci of attention.

Data Collection Procedure

This section presents a brief overview of the user study and robot system in order to explain the conditions under which the data streams were recorded.

Participants

Twenty-four participants were recruited from the Pittsburgh area. Of these, 13 were women and 11 were men. 17 were 18–24 years old, 4 were 25–30 years old, 1 was 31–35 years old, and 2 were 41–45 years old. Participants were screened so that those who had prior experience using this robot arm in this type of eating study were not included in this dataset. Thus, the participants were novices at the task. The experiment took place in the Human And Robot Partners Lab (HARP Lab) on the Carnegie Mellon University campus, and participants were compensated $15 for their time.

Protocol

Participants performed a task which consisted of controlling a robot arm to position a fork above one of three marshmallows on a plate (see Fig. 1). They controlled a robot with a joystick using modal control: the joystick dimensions moved the end-effector of the robot in x and y, z and yaw, or pitch and roll, and a button on the joystick allowed the participant to cycle between those control configurations. When the participant had placed the fork above their desired marshmallow, or otherwise determined that the task was complete, they held down another button on the joystick. The robot would autonomously move down to the height of the plate and spear the marshmallow if it was in the right place, and then move the arm in a serving motion towards the participant's mouth. The button press concluded the trial, and the robot was reset to the starting configuration.

Participants were given a brief introduction as to the purpose of the study and then began a five-minute familiarization period, in which the participant controlled the robot in the teleoperation mode and data was not

recorded. Next, participants were fitted with eye gaze and EMG sensors (described below). They performed the task five times in sequence for each of four assistance modes (described in the next section). Assistance mode order was fully counterbalanced among participants. After each block of five trials for an assistance mode, participants were given a brief survey to record their subjective perceptions about the algorithm.

Assistance Conditions

Participants operated the robot in each of four different assistance conditions: fully teleoperated, two different levels of assistance according to the shared autonomy framework[?], and a version of assistance in which the robot is fully controlled autonomously and user input is used only for goal inference.

The following is a brief description of how assistance is calculated; see the journal paper[?] for a full description. The combined human-robot system is modeled as a partially observable Markov decision process (POMDP), where the participant's goal is represented as one unknown member of a small set of possible goal objects. Participant inputs via joystick are treated as observations, and the algorithm assumes that the user is noisily optimizing a cost function parameterized by their unknown goal. Therefore, the MaxEntIOC framework can be used to evaluate a belief distribution over the known goal set. From this belief state, the overall POMDP is solved by applying the QMDP approximation, which has proved reliable for similar shared control scenarios. The resulting robot action consists of a computed assistive action based on the inferred user goal distribution combined with the original applied user action.

To provide different assistance levels, the shared autonomy transition function was modified slightly. In Javdani et al.[?], the given transition function applies both user and robot control as determined by:

$$a_{applied} = u + a.$$

In order to adapt the amount of user control, the applied action was parameterized by a value γ,

$$a_{applied} = (1 - \gamma)u + \gamma a,$$

which trades off between the relative strengths of the user command and the robot assistance. Note that the original shared autonomy procedure would correspond to the case $\gamma = 0.5$.

The four conditions corresponded to four different levels of γ:

Direct teleoperation, $\gamma = 0$. The assistance signal a was computed but completely discarded, so the user had full manual control over the robot.

Low assistance, $\gamma = 0.33$. The assistance signal was combined with the direct user control, with the user signal weighted double.

High assistance, $\gamma = 0.67$. The assistance signal was combined with direct user control, but the assistance signal was more highly weighted.

Robot control, $\gamma = 1$. The user control signal was not passed through to the robot control. It was used for goal inference only, and the robot was autonomously controlled based on its goal inference results.

Sensors

Participant behavior was captured using a variety of sensors.

Eye gaze Participant eyegaze direction was captured by a Pupil Labs Pupil[? ?] sensor. This sensor consists of a glasses-like frame with an infrared camera with infrared illumination mounted below each eye for dark pupil tracking, plus a third RGB camera oriented outward to capture egocentric video. The eye cameras capture video at 120 Hz, and pupil labs software detects the pupil pixel center. Before data is captured, the pupil locations and world camera videos are calibrated by placing a marker in the field of view of the participant at several points and asking the participant to look at the center of the marker ("manual marker calibration"). For most of the participants, this calibration routine was recorded; it is available in the `calib` folder for each participant. In addition, the calibration is checked by asking the participants to look at particular places in the scene periodically (between each condition); recordings of these procedures are found in the `check` folders.

EMG Participant muscle activation while controlling the joystick was controlled using a Myo sensor[?]. Due to initialization failures, this data is only available for about 20% of the runs (see Table 1 for full details). It consists of the following signals:

- EMG message, denoting the activation of eight individual EMG sensors
- ORI message, denoting the orientation of the arm in roll/pitch/yaw
- IMU message, denoting the readings of the IMU attached to the armband

External video Participant behavior was captured using a Stereolabs[?] ZED camera. Left and right videos are stored as separate AVI files. The ZED camera was placed on a tripod at approximately the same (marked) location for each trial in order to capture a full-on view of the participant and occasional views of the scene. ZED videos are available for the 10 participants who consented to their images being released; in other cases, data is not provided, but offline skeleton tracking information run on that data is available.

Descriptive Statistics

This dataset consists of a total of 480 trials, comprising 20 trials for 24 participants. Altogether, the data represents about five hours of continuous instrumented robot control. A summary of the data available divided by type appears in Table 1.

Data Streams

The data is organized first by participant, with folders `p100`-`p123` corresponding to the twenty-four participants. Within each participant folder, there are folders for the three types of recordings: the `calib` folder contains recordings of

	Left Eye	Right Eye	Egocentric Video	ZED Camera
Total duration (h:m:s)	5:19:26	5:10:45	5:33:44	2:21:4
Total frames	2299877	2237380	600728	253921
Nominal frequency (Hz)	120	120	30	30
Frames dropped	133301	195860	7459	339155
Coverage (%)	94.52	91.95	98.77	42.81
Present (%)	100.00	100.00	100.00	40.04
Coverage if present (%)	94.52	91.95	98.77	100.00

	Joystick	Robot position	Myo EMG	Myo IMU	Myo ORI
Total duration (h:m:s)	4:56:00	5:48:05	1:10:49	1:10:53	1:10:53
Total frames	2131160	1670798	212465	212664	212659
Nominal frequency (Hz)	120	80	50	50	50
Frames dropped	114250	1680	802368	802204	802206
Coverage (%)	94.91	99.90	20.94	20.95	20.95
Present (%)	100.00	100.00	21.48	21.48	21.48
Coverage if present (%)	94.91	99.90	99.75	99.83	99.83

Table 1. Descriptive statistics of each data stream in the data set. *Total duration* and *Total frames* refer to the collective amount of data of that signal over all trials and participants. *Total duration* is extracted by dividing the total frames by the *nominal frequency*. *Frames dropped* are based on interpolating from the nominal frame rate and detecting missing data. *Coverage* is computed by dividing the number of data frames by the expected number of data frames from the nominal frequency over the whole dataset, *Present* indicates the fraction of trials that have at least one data point of that type, and *Coverage if present* is the total number of data frames divided by the expected number evaluated only if at least one data point is present in the trial.

calibration passes, the `check` folder contains intermediate gaze accuracy checks, and the `run` folder contains standard data collection runs. Each of these recording types contain numbered subfolders indicating the run sequence.

A single trial capture (a numbered folder) has the following subfolders:

- `raw_data` contains binary capture information as recorded, in bagfiles or pickle files.
- `text_data` contains exported CSV files containing the raw data. The particular data streams available there are detailed below.
- `videos` contains video files exported as AVI, in addition to the timestamps of each frame as either numpy (`*.npy`) or raw text files.
- `stats` contains a number of YAML files detailing statistical information about the trial and overall data stream, including the number of records, approximate time distances between individual records, and estimates of the times when data points may have been dropped based on the nominal data collection frame rate.
- `gaze` contains the collected gaze data in a format suitable for import, analysis, and replay through the Pupil Labs software.
- `processed` contains a number of new formats of data extrapolated from the underlying data, including a video of the egocentric recording with a dot overlaid at the gaze point, a bag file for visualization through an external package, skeleton tracking of ZED video information, etc. More processed results may become available as we continue to work with the data.

Timing and synchronization

All data points were timestamped on collection and are stored as either 32-bit or 64-bit floating point values in number of nanoseconds from the Unix epoch. The CSV files available in `text_data` provides this data in several columns for convenience. Certain data streams provide an `orig_timestamp` field; this field is a relic of the data processing step to adjust all time to a common epoch and may be ignored.

To ease the process of rectifying all of these data streams, two common indices are provided for all data streams. The `world_index` field gives the corresponding frame number of the egocentric video for each data point (i.e., the index of the frame whose timestamp is the first value that occurs after the timestamp of the data point). A second common index, `world_index_corrected`, provides a second index into the egocentric video, except with a correction for frames that were dropped in that video. Thus, the `world_index_corrected` value represents roughly a common 30Hz clock throughout the trial. For more sophisticated data alignment, please use the provided timestamps or see the data loading tool provided separately from the dataset.

Eye Gaze

Eye gaze videos are recorded at 120 Hz and located in the `videos` folder as `eye0.mp4` and `eye1.mp4`. The timestamp of the data collection of each frame is available in the corresponding NumPy binary file, `eye0_timestamps.npy` and `eye1_timestamps.npy`. The automated pupil detection results for each eye are in the `text_data` folder, under `pupil_eye0.csv` and `pupil_eye1.csv`. Field names

correspond to the output of the 3D pupil detection process in Pupil Labs; see their documentation for an explanation of the fields.

The egocentric video is available in the `videos` folder as `world.mp4`, with timestamps corresponding to individual frames in `world_timestamps.npy`. The calculated gaze position within the corresponding video frame is given in `text_data/gaze_positions.csv`. See the pupil labs documentation for a full description of the fields. We note here that the fields `norm_pos_x` and `norm_pos_y` in that file correspond to the (x, y) pixel in coordinates normalized to the egocentric video frame size, with the bottom left as $(0, 0)$ and the top right $(1, 1)$.

The data used to calibrate between pupil data and gaze point is stored for each run in the text files `pupil_cal_eye0.csv`, `pupil_cal_eye1.csv`, and `world_cal_positions.csv`. This data is the same between runs of the same participant and is provided as a convenience to recalculate a calibration if desired. The details of how the current calibration is derived from this data can be found in the Pupil Labs software documentation.

Third Person Video

ZED videos were recorded using the Stereolabs ZED software, version 1.1.0. Data was stored as an internal Stereolabs SVO file, which includes separate left and right videos, as well as a common timestamp. The videos were extracted to the `videos` directory as `zed_left.avi` and `zed_right.avi`, and the timestamps were rescaled to the Unix epoch and stored as an integer number of nanoseconds from the epoch in `zed_ts.txt`, as well as in the same floating-point NumPy format as the other videos in `zed_timestamps.npy`. The `zed_corrs.csv` stores the correlations to a common index, as explained above.

Additional sensor data

The following data streams are also available, all in the `text_data` directory. They have been extracted or calculated from the binary storage in the `raw_data` directory.

- `control_mode.txt` contains a single character referring to the assistance condition of that trial, where 0 represents direct teleoperation and 3 represents robot control.
- `morsel.yaml` is a YAML file with the transforms for each detected morsel positions in the robot base frame.
- `ada_joy.csv` stores the raw joystick input provided by the user. Note that the joystick input is only provided when it is unchanged from the previous message, so the raw data has inconsistent timing. For ease of use, the joystick data has been resampled to a common 120 Hz frequency, with missing data filled in by holding the previous value. Duplicate data can be noted by seeing that the header fields are unchanged when the data is held.
- `input_info.csv` contains information about the user input to the robot. The `robot_mode` field denotes which control mode the robot is in (x/y, z/yaw, or pitch/roll), and the rest of the fields denote the applied twist corresponding to the user's joystick input.
- `assistance_info.csv` contains the outcome of the shared autonomy algorithm. It stores the current probability inferred for each goal and the resultant twist applied to the robot at that timestep.
- `joint_states.csv` contains the information for each joint of the robot.
- `robot_position.csv` contains the cartesian position of each of the robot links, as calculated from the forward kinematics using the data from `joint_states.csv`.
- `myo_emg.csv` contains the raw EMG output of the Myo sensor.
- `myo_imu.csv` contains the data from the IMU on the Myo sensor.
- `myo_ori.csv` contains the orientation data received from the Myo sensor.

Known Issues

Missing Data

Due to computational load, certain data streams may have periodic dropouts. The `stats` directory contains some info on when and how often these occur, and overall statistics are given in Table 1. The missing data is particularly exacerbated for the Myo signal. Due to an initialization failure, the Myo data is unavailable for certain participants. In these cases, the text files are present for ease of access but contain no data. Finally, due to permissions restrictions, the ZED video capture is only available for certain participants. Within those participants, some initialization failure means that videos of certain trials are occasionally missing.

Accessing the Data

The data will be hosted on the HARP Lab website: http://harp.ri.cmu.edu/harmonic. Several files are provided for download:

- `harmonic_all.tar.gz`, a compilation of all of the data, roughly 430 Gb.
- `harmonic_data.tar.gz`, consisting of the `text_data`, `videos`, and `stats` directories, approximately 230 Gb.
- `harmonic_sample.tar.gz`, consisting of all of the data for a single participant, roughly 30 Gb.

The data sets will be versioned using semantic versioning, and that page will maintain a log of all changes that may be made to the dataset after release.

Conclusion

In this paper, we present a dataset of humans performing a food acquisition task by controlling a robot manipulator. During this task, a variety of types of participant data were collected, including eyegaze information, electrymography of the controlling arm, stereo video, and robot controller information. This dataset can enable research into human-robot collaboration and multimodal human behavior analysis.

Acknowledgements