Check for updates

Data paper

The International Journal of Robotics Research
2022, Vol. 41(1) 3–11
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02783649211050677
journals.sagepub.com/home/ijr
SAGE

HARMONIC: A multimodal dataset of assistive human–robot collaboration

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

Abstract

We present the Human And Robot Multimodal Observations of Natural Interactive Collaboration (HARMONIC) dataset. This is a large multimodal dataset of human interactions with a robotic arm in a shared autonomy setting designed to imitate assistive eating. The dataset provides human, robot, and environmental data views of 24 different people engaged in an assistive eating task with a 6-degree-of-freedom (6-DOF) robot arm. From each participant, we recorded video of both eyes, egocentric video from a head-mounted camera, joystick commands, electromyography from the forearm used to operate the joystick, third-person stereo video, and the joint positions of the 6-DOF robot arm. Also included are several features that come as a direct result of these recordings, such as eye gaze projected onto the egocentric video, body pose, hand pose, and facial keypoints. These data streams were collected specifically because they have been shown to be closely related to human mental states and intention. 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` *and* `YAML` *files.*

Keywords

Human–robot interaction, shared autonomy, intention, multimodal, eye gaze, assistive robotics

1. 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 perceive, understand, and predict human mental states (the human's goals, intentions, and future actions, often unknown to external observers) that determine what assistance the robot should provide.

Concretely, when people complete physical tasks, their external behaviors, such as their eye gaze, can reveal insights about their internal mental states. An assistance system that can understand how these behaviors relate to the task can predict what objects and locations of a visual scene the human deems to be task relevant. The system can also use these behaviors to determine whether or not interactions with these objects or locations will take place, and qualities that describe these interactions. This information is not known to the system prior to completing a task, and is not relayed to the system by the human via traditional means (e.g., verbal or written communication). Thus understanding these mental states in order to assist the human requires perceiving and interpreting the human's behavior during human–robot collaborations.

One example of a behavior that has been well studied in physical tasks is eye gaze. People almost exclusively fixate their eye gaze on objects or locations involved in their current task (Hayhoe and Ballard, 2005), thereby ignoring task irrelevant parts of a scene. Should these objects or locations require a direct interaction, people fixate their gaze on these objects and locations prior to moving their hands to complete the interaction (Land and Hayhoe, 2001), thus revealing the intended interaction object in advance of any physical contact. Gaze also lingers on key points in the task, such as obstacles, revealing certain landmarks of manipulation (Johansson et al.,

[*]These authors contributed equally.
[1]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[2]University of Washington, Seattle, WA, USA

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

2001). In addition, people gaze at objects before uttering verbal references, which others can use to disambiguate and predict speech (Admoni et al., 2014; Boucher et al., 2012).

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.

In addition, pupil size has been correlated with cognitive load (Beatty, 1982; Bednarik et al., 2018; Krejtz et al., 2018), and understanding current human body posture can both reveal desired tasks and help to avoid potentially dangerous collisions (Mainprice et al., 2015).

In this paper, we present the Human And Robot Multimodal Observations of Natural Interactive Collaboration (HARMONIC) dataset. The HARMONIC dataset contains 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 six-degree-of-freedom (6-DOF) Kinova Mico robot arm is controlled in three dimensions via a two-axis joystick and manual mode-switching. In some cases the robot provides additional assistance through shared autonomy (Javdani et al., 2018).

Though the data were collected during an assistive eating task, their usefulness extends beyond the specific domain of eating. The manual condition can be used to study human teleoperation in the general case, for example, with tasks using simplified grippers such as vacuum tooling. When combined with the shared autonomy conditions, these data can be used to study co-manipulation across individuals and varying levels of robot agency. Included in the data are a wide array of non-verbal behaviors situated in a real-world task defined with a clear goal and, thus, is relevant for a variety of human–robot collaborations.

Human behavioral data include egocentric RGB videos, eye gaze positions relative to these videos, infrared (IR) videos of both eyes, stereo, third-person video of the participant, and EMG recordings on the joystick-controlling arm. Robot related data include joystick control inputs from the user, the control input and belief distribution calculated by the assistance algorithm, and the robot position. Environmental data include the 3D locations of the food morsels as well as the locations of fiducial markers. Further information as well as an explanation as to how to access these data is offered in the following sections.

Our dataset will help researchers study the complex human–robot dynamics of assistive teleoperation, which can vary across individual and across different levels of robot autonomy. For example, researchers could use this dataset to learn correlations between eye gaze and joystick control, in order to improve the goal inference predictions made by shared autonomy algorithms. Others might be interested in modeling and forecasting the dynamics of joystick inputs under differing amounts of robot assistance. Previous research using similar data has proposed identifying unexpected events (e.g., human errors or task failure) by learning a normative gaze behavior model and identifying anomalies (Aronson and Admoni, 2018); the higher-quality data provided in this dataset could continue this line of research as well as extend it to situations where the robot provides variable levels of assistance within a unified framework.

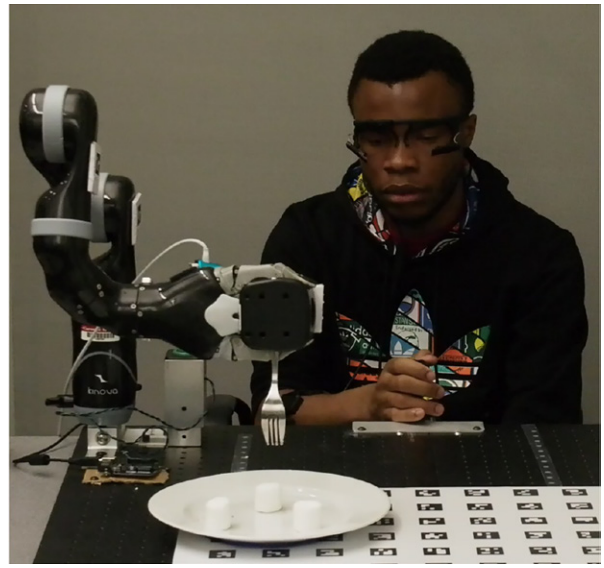

Fig. 1. The HARMONIC dataset provides multimodal human, robot, and environmental data collected during an assistive human–robot collaboration.

2. Prior work

2.1. Human interaction for robotic control

Eye gaze, EMG, and body pose have all been useful signals for robotic control. As eye gaze is a rich signifier of intention during manipulation, both by hand (Hayhoe and Ballard, 2005; Johansson et al., 2001; Land and Hayhoe, 2001) and by robot (Aronson et al., 2018), its use has been explored through numerous robotic collaboration settings, including anticipating which object a user will request (Huang and Mutlu, 2016), and triggering assistive aid during autonomous driving (Braunagel et al., 2015). Electromyography signals have been used for robot control (Artemiadis and Kyriakopoulos, 2010) and task monitoring (DelPreto et al., 2018).

In addition, there has been work in learning and leveraging human policies (using keyboard input) (Reddy et al., 2018a,b) and attention models (using keyboard input and eye gaze) (Zhang et al., 2018) for both assisted and shared robot control in Atari games in an arcade learning environment (Bellemare et al., 2012). HARMONIC provides a more realistic environment for studying such interactions. By making this dataset available, we intend to enable further research into these control methods.

2.2. Multimodal datasets in human–robot interaction

Multimodal datasets have garnered interest in many different communities, such as psychology (Xu et al., 2013), computer vision (Damen et al., 2018; Fathi et al., 2012; Pirsiavash and Ramanan, 2012; Shu et al., 2016; Sigurdsson et al., 2018a,b), human–robot interaction (Azagra et al., 2016; Ben-Youssef et al., 2017; Jayagopi et al., 2013; Sheikhi and Odobez, 2012; Stefanov and Beskow, 2016), and natural language processing (Bastianelli et al., 2014). These datasets, though, can be difficult to collect at a large scale. This can be due to the increasing engineering demand required with each additionally desired modality, physically collocating robots and humans, and the need to respect humans' privacy rights. This leads to many multimodal datasets including either few participants or few data modalities. In addition, these datasets are rarely designed to study direct, physical human–robot collaborations in which the human and robot act in similar spaces. HARMONIC gives researchers the opportunity to study direct human–robot collaboration in the form of a large-scale dataset in both the number of available modalities as well as the number of participants. Here, we compare how HARMONIC relates to other multimodal human–robot interaction in order to illustrate these distinctions and the potential use of HARMONIC.

2.2.1. Robots in conversational settings. The majority of publicly released human–robot interaction datasets study the inclusion of robots as conversational partners. To successfully incorporate robots as part of a social conversation, it is necessary to perceive human behavior, understand how this relates to the conversation, and be able to synthesize similar behavior in order to keep the conversation flowing smoothly. Much of this work surrounds determining the human's visual focus of attention (VFOA) (Jayagopi et al., 2013; Sheikhi and Odobez, 2012). In these works, VFOA is a discrete representation of eye gaze estimated from the user's head position. Other datasets are designed to capture unscripted conversations with a robot (Ben-Youssef et al., 2017) by capturing conversations through a robot's third person video recorder. In all of these works, no signal specific sensors (e.g., an eye gaze camera) were used in order to capture specific human behaviors (e.g., eye gaze).

Other conversational datasets have a linguistic focus (Bastianelli et al., 2014). This work designs an interaction in which a human commands a robot to perform a specific task, and contains many different views of the language spoken. Owing to the focus on verbal communication, this dataset does not give researchers the ability to understand how non-verbal behaviors may be utilized in order to understand the intent behind the human's command.

Finally, perhaps the most similar dataset to HARMONIC (in terms of data streams collected) again focuses on predicting the VFOA during a conversation (Stefanov and Beskow, 2016). Unlike other works focusing on VFOA, this dataset does capture eye gaze explicitly through the use of a Tobii eye tracker (Tobii AB, 2017). Thus, this work studies how gaze changes between structured and unstructured conversation with and without the presence of a robot. This dataset is not designed, however, to study these non-verbal behaviors in physical collaborations.

In all of these situations, the behaviors collected for analysis are centered around non-collaborative tasks. HARMONIC provides the opportunity to study how these behaviors may be interpreted in order to provide better assistance during collaborative tasks.

2.2.2. Robot as student. The Multimodal Human–Robot Interaction Dataset (Azagra et al., 2016) is designed for interactive object learning through human guidance. This dataset presents a situation in which a human uses a small number of task specific behaviors in order to teach the robot about object models. This dataset is intended to instruct the robot by leveraging a human's innate teaching ability, as opposed to studying physical human–robot collaboration.

Humans teaching robots has also been studied in psychology (Xu et al., 2013). Here, researchers studied how humans' eye gaze patterns changed as a robot displayed gaze patterns that were designed to emulate the gaze patterns displayed by people who employ different styles of learning. Again, this task is different from a direct collaboration such as our shared autonomy task. In addition, this dataset does not seem to be publicly available.

2.3. Machine learning and computer vision

Surprisingly, our dataset is similar to those from the machine learning and computer vision communities. The tasks studied in these datasets often include non-scripted, egocentric videos of daily activities (Damen et al., 2018), gaze prediction for egocentric videos (Fathi et al., 2012), action recognition in third person video (Pirsiavash and Ramanan, 2012; Sigurdsson et al., 2018a), relating first and third person videos as a proxy for theory of mind (Sigurdsson et al., 2018b), and learning about human social affordance from a third person view (Shu et al., 2016). These datasets include large amounts of potentially relevant data for human–robot collaboration, but most importantly, they do not contain interactions with a robot. Although these datasets may be useful for an initial understanding of human behavior, they do not provide insights into how these behaviors manifest in human–robot collaborations.

3. 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.

3.1. Participants

Twenty-four participants (13 female) were recruited from the Pittsburgh area. Seventeen were between the ages of 18 and 24, four between 25 and 30, one between 31 and 35, and two between 41 and 45. The participant pool was screened for prior experience using this robot arm in similar studies and, thus, were novices at the task. The experiment took place in the Human And Robot Partners (HARP) Lab on the Carnegie Mellon University campus. Participants were compensated US$15 for one and a half hours of their time.

3.2. Protocol

Participants controlled a robot arm, attempting to position a fork above one of three marshmallows placed on a plate (see Figure 1). They controlled a robot with a two-axis joystick using modal control: the joystick's two axes moved the end-effector of the robot in x and y, z and yaw, or pitch and roll. A joystick button allowed participants to cycle between control configurations when pressed for less than 500 ms. When the task was completed (that is, once a participant was satisfied with the fork's position or had given up on the task), the participant held down the same joystick button for longer than 500 ms. This action triggered an autonomously executed plan in which the robot moved down to the height of the plate and speared the marshmallow (conditional on the proper positioning of the fork). Finally, the robot arm moved into a serving position near the participant's mouth. This concluded the trial, and the robot automatically reset to the starting configuration.

Participants were given a brief introduction as to the purpose of the study and then began a 5-minute familiarization period, in which they controlled the robot in teleoperation mode and data were not recorded. Next, participants were fitted with eye gaze and EMG sensors (described in the following). 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, participants were given a brief survey to record their subjective perceptions about the algorithm. Once the final survey was completed, participants were presented with a survey that compared all conditions through ranked preference as well as free response.

3.3. 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 (Javdani et al., 2018), and a fully autonomous robot.

The following is a brief description of how assistance is calculated: a full description is available in a prior publication (Javdani et al., 2018). The combined human–robot system is modeled as a partially observable Markov decision process (POMDP) (Kaelbling et al., 1998; Sondik, 1978), where the participant's goal is represented as one unknown member of a small set of possible goals. Participant inputs via joystick are treated as observations. The algorithm assumes that the user is noisily optimizing a cost function parameterized by their unknown goal. Therefore, the Maximum Entropy Inverse Optimal Control (MaxEntIOC) (Ziebart et al., 2008) 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 (Littman et al., 1995) approximation, which has proved reliable for similar shared control scenarios. Our implementation changed the original formulation slightly in order to remove the inherent living reward, which can cause the robot to converge on a goal even in the absence of any positive joystick actuation. 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 from prior work. In the work of Javdani et al. (2018), the given transition function applies both user and robot control as determined by $a_{\text{applied}} = u + a$.

In order to adapt the amount of user control, the applied action was parameterized by a value γ: $a_{\text{applied}} = (1-\gamma) u + \gamma a$, 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 $\gamma = 0.5$ and normalizing the vector a_{applied}.

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.

Autonomous 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.

3.4. Sensors

3.4.1. Eye gaze. Participant eye gaze direction was captured by a Pupil Labs Pupil (Kassner et al., 2014; Pupil

Table 1. Descriptive statistics of each data stream in the dataset. *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 datum 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 datum is present in the trial.

	Left eye	Right eye	Egocentric video	ZED camera	
Total duration (h:min:s)	5:19:26	5:10:45	5:33:44	4:44:45	
Total frames	2,299,877	2,237,380	600,728	512,569	
Nominal frequency (Hz)	120	120	30	30	
Frames dropped	133,301	195,860	7,459	94,431	
Coverage (%)	94.52	91.95	98.77	84.44	
Present (%)	100.00	100.00	100.00	87.25	
Coverage if present (%)	94.52	91.95	98.77	94.83	
	Joystick	Robot position	Myo EMG	Myo IMU	Myo ORI
Total duration (h:min:s)	4:56:00	5:48:05	1:10:49	1:10:53	1:10:53
Total frames	2,131,160	1,670,798	212,465	212,664	212,659
Nominal frequency (Hz)	120	80	50	50	50
Frames dropped	114,250	1,680	802,368	802,204	802,206
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

Labs, Inc., 2017) sensor. This sensor consists of a glasses-like frame with two infrared cameras 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 were captured, the pupil locations and world camera videos were calibrated by asking the participant to look at the center of the marker held in front of them by the researcher (‘‘manual marker calibration’’). This calibration routine was recorded for most participants and is made available in the `calib` folder. The calibration is verified between each condition by asking participants to look at particular places in the scene. These checks are recorded and made available in the `check` folders.

3.4.2. EMG. Participant muscle activation while controlling the joystick was captured using a Myo sensor (Thalmic Labs, Inc., 2018). Owing to initialization failures, these data are only available for about 20% of the runs (see Table 1 for full details). It consists of the EMG message, denoting the activation of eight individual EMG sensors, the ORI message, denoting the orientation of the arm in roll/pitch/yaw, and the inertial measurement unit (IMU) message, denoting the readings of the IMU attached to the armband.

3.4.3. External video. Participant behavior was captured using a Stereolabs (Stereolabs, Inc., 2018) ZED camera. Left and right videos are stored as separate `MP4` 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 all cases, offline skeleton and face tracking information is available.

4. Descriptive statistics

This dataset consists of 480 trials, comprising 20 trials for 24 participants. The data represent about five hours of continuous instrumented robot control. A summary of the data available appears in Table 1.

5. Data streams

The data are organized first by participant (`p100`–`p123` reflecting the 24 participants). Each participant folder contains folders for three types of recordings: `calib` contains calibration passes, `check` contains intermediate gaze accuracy checks, and `run` contains data collection runs. These folders contain numbered subfolders indicating the run sequence. A visual representation of selected data streams can be seen in Figure 2.

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

- `text_data` contains exported `CSV` files containing the raw data. The particular raw data streams available are detailed in the following subsections. Additional to the raw data, this directory contains the body skeleton, facial, and hand keypoints generated by running OpenPose (Cao et al., 2017; Simon et al., 2017; Wei

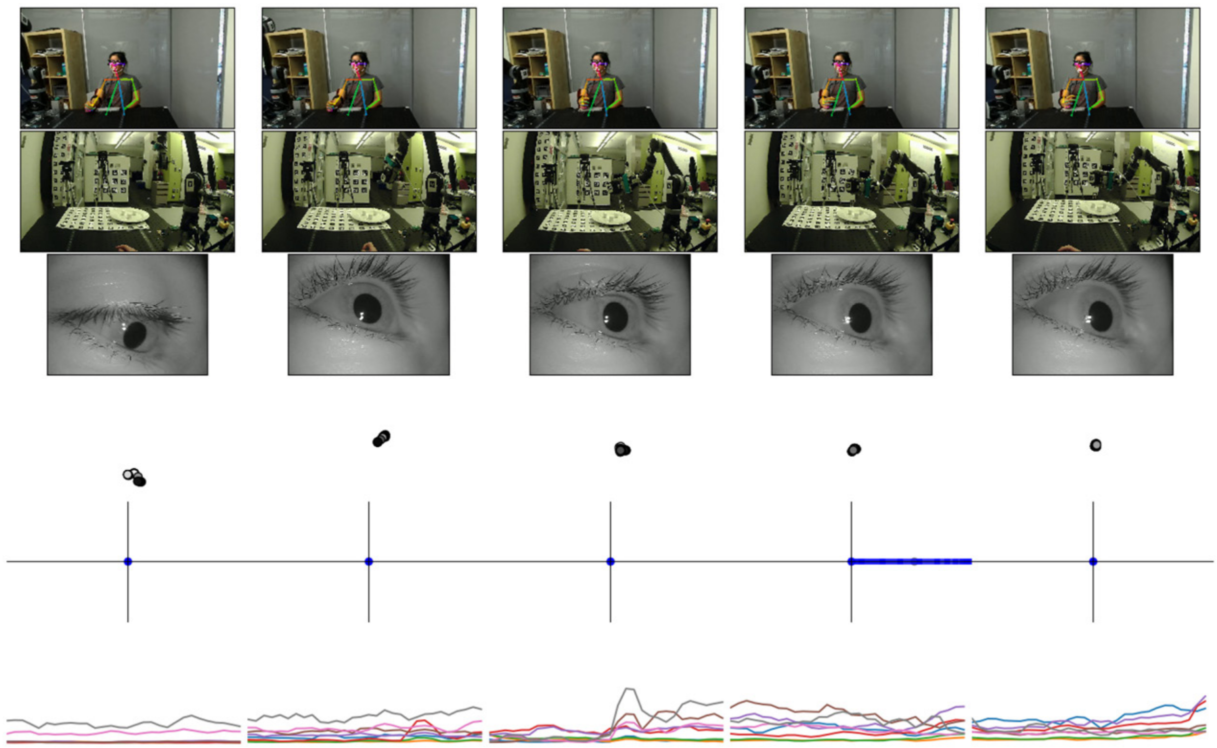

Fig. 2. A visualization of several streams from the HARMONIC dataset. The top row displays the ZED video with OpenPose skeletons overlaid, then the egocentric video captured from the Pupil camera, left eye video, 1 second of the calculated gaze dot, the trajectory of the joystick, and finally the Myo activations. For the gaze dot and the joystick, lighter colors represent more recent points in time. Each of these plots represents 1 second of data, sampled at 30 FPS.

et al., 2016) on the left and right streams of the third-person ZED videos. The outputs from OpenPose are compiled into face, right and left hand, and pose files for each stream of the depth camera. For full descriptions, please refer to the OpenPose documentation.

- `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. `videos` contains the pupil video files (`eye0.mp4`, `eye1.mp4` and `world.mp4`) exported as MP4 files using the H.264 video codec (Richardson, 2010). Additionally included are the timestamps of each frame as either numpy (`*.npy`) files, raw text (`*.txt`), or `CSV` (`*.csv`).
- `processed` contains a number of new formats of data extrapolated from the underlying data (e.g., a video of the egocentric recording with a dot overlaid at the gaze point).

5.1. Timing and synchronization

All data points were timestamped on collection and stored as either 32- or 64-bit floating point values in number of nanoseconds from the Unix epoch. The `CSV` files in `text_data` provide these data in several columns.

For ease of use, two common indices are provided for all data streams. The `world_index` field gives the egocentric video frame number corresponding to each data point. A second common index, `world_index_corrected`, provides a second index into the egocentric video, with a correction for dropped video frames. The `world_index_corrected` value approximates a common 30 Hz clock running throughout the trial. For more sophisticated data alignment, please use the provided timestamps.

5.2. Eye gaze

Eye gaze videos were recorded at 120 Hz and located in the `videos` folder as `eye0.mp4` and `eye1.mp4`, encoded using the H.264 video codec (Richardson, 2010). Frame level timestamps are available in corresponding NumPy binary files, `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, as described in their documentation.

Egocentric video is available in the `videos` folder as `world.mp4` (encoded using the H.264 codec (Richardson, 2010)), with frame-level timestamps located in `world_timestamps.npy`. 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 fields. The fields `norm_pos_x` and `norm_pos_y` correspond to the (x, y) pixel in coordinates normalized to the egocentric video frame size, with the origin point in the top left.

Data used to calibrate between pupil data and gaze point are stored in the text files `pupil_cal_eye0.csv`, `pupil_cal_eye1.csv`, and `world_cal_positions.csv`. These data are the same between runs of the same participant and is provided as a convenience to recalculate a calibration if desired. Details of the current calibration method can be found in the Pupil Labs software documentation.

5.3. Third-person video

ZED videos were recorded using the Stereolabs ZED software, version 1.1.0. Data were initially stored as a Stereolabs SVO file, including separate left and right videos and a common timestamp. Videos were extracted to the `videos` directory as `zed_left.mp4` and `zed_right.mp4` encoded using H.264 (Richardson, 2010). 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 floating-point NumPy format in `zed_timestamps.npy`. The `zed_corrs.csv` stores the correlations to a common index, as explained previously.

5.4. Additional sensor data

The following data streams are available in the `text_data` directory, having been extracted or calculated from the original binary.

`control_mode.txt` contains one character referring to that trial's assistance condition: 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 raw joystick input provided by the user. Joystick input is only provided when changed from the previous message leading to inconsistent timing in the raw data. To rectify this, joystick data have been resampled to a common 120 Hz frequency and missing data filled by the previous value. Duplicate data are noted by unchanged headers.
- `input_info.csv` contains 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 EMG output of the Myo.
- `myo_imu.csv` contains IMU output of the Myo. `myo_ori.csv` contains orientation data received from the Myo sensor.

6. Known issues

6.1. Missing data

Owing 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 are particularly exacerbated for the Myo signal due to the data recording software failing to start. Finally, owing to permissions restrictions, unedited ZED video capture is available for 10 participants, de-identified video (video with faces blurred) is available for 13 participants, and video for 1 participant is unavailable for release. Within the released participants, some initialization failure means that videos of certain trials are occasionally missing.

7. 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_data.tar.gz`, a compilation of all of the data (~68 GB), `harmonic_minimal.tar.gz`, consisting of the `text_data`, `videos`, and `stats` directories (~15 GB), `harmonic_text.tar.gz`, consisting of the `text_data` directory (~4 GB), and finally `harmonic_sample.tar.gz`, consisting of all of the data for a single participant (~303 MB). The datasets 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. Furthermore, our GitHub contains a repository for basic processing tools located here: https://github.com/HARPLab/harmonic_cpp. Finally, for the original robot control code, follow this link to a fork of the publicly available implementation of the shared autonomy code we used: https://github.com/HARPLab/ada_meal_scenario (Javdani et al., 2015). Our robot control code is on the branch: "adjustable."

8. Conclusion

We have presented a dataset of humans who performed a food acquisition task by controlling a robot manipulator. During this task, a variety of types of participant data were collected, including eye gaze information, EMG of the controlling arm, stereo video, and robot controller information. This dataset enables research into human–robot collaboration and multimodal human behavior analysis.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Science Foundation (grant number IIS-1755823) and the Paralyzed Veterans of America. The first author is supported by a National Science Foundation Graduate Research Fellowship (DGE 1745016).

ORCID iDs

Benjamin Alexander N https://orcid.org/0000-0002-9651-3470
Reuben Aronson https://orcid.org/0000-0002-8615-3715

References

Admoni H, Datsikas C and Scassellati B (2014) Speech and gaze conflicts in collaborative human–robot interactions. In: *Annual Conference of the Cognitive Science Society (CogSci)*, pp. 104–109.

Aronson RM and Admoni H (2018) Gaze for error detection during human–robot shared manipulation. In: *RSS Workshop: Towards a Framework for Joint Action*.

Aronson RM, Santini T, Kubler TC, Kasneci E, Srinivasa SS and Admoni H (2018) Eye–hand behavior in human–robot shared manipulation. In: *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

Artemiadis PK and Kyriakopoulos KJ (2010) EMG-based control of a robot arm using low-dimensional embeddings. *IEEE Transactions on Robotics* 26(2): 393–398.

Azagra P, Mollard Y, Golemo F, Murillo AC, Lopes M and Civera J (2016) A multimodal human–robot interaction dataset. In: *NIPS 2016, workshop Future of Interactive Learning Machines*.

Bastianelli E, Castellucci G, Croce D, Iocchi L, Basili R and Nardi D (2014) Huric: A human robot interaction corpus. In: Choukri K, Declerck T, Loftsson H and et al. (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

Beatty J (1982) Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin* 91(2): 276.

Bednarik R, Bartczak P, Vrzakova H, et al. (2018) Pupil size as an indicator of visual-motor workload and expertise in microsurgical training tasks. In: *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA '18)*. New York: ACM Press.

Bellemare MG, Naddaf Y, Veness J and Bowling M (2012) The arcade learning environment: An evaluation platform for general agents. *CoRR* abs/1207.4708.

Ben-Youssef A, Clavel C, Essid S, Bilac M, Chamoux M and Lim A (2017) UE-HRI: A new dataset for the study of user engagement in spontaneous human–robot interactions. In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI 2017)*. New York: ACM Press, pp. 464–472.

Boucher JD, Pattacini U, Lelong A, et al. (2012) I reach faster when I see you look: Gaze effects in human–human and human–robot face-to-face cooperation. *Frontiers in Neurorobotics* 6: 00003.

Braunagel C, Kasneci E, Stolzmann W and Rosenstiel W (2015) Driver-activity recognition in the context of conditionally autonomous driving. In: *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pp. 1652–1657.

Cao Z, Simon T, Wei SE and Sheikh Y (2017) Realtime multi-person 2D pose estimation using part affinity fields. In: *Proceedings of CVPR*.

Damen D, Doughty H, Farinella GM, et al. (2018) Scaling egocentric vision: The EPIC-KITCHENS dataset. *CoRR* abs/1804.02748.

DelPreto J, Salazar-Gomez AF, Gil S, Hasani RM, Guenther FH and Rus D (2018) Plug-and-play supervisory control using muscle and brain signals for real-time gesture and error detection. In: *Robotics: Science and Systems*.

Fathi A, Li Y and Rehg JM (2012) Learning to recognize daily actions using gaze. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y and Schmid C (eds.) *Computer Vision – ECCV 2012*. Berlin: Springer, pp. 314–327.

Hayhoe M and Ballard D (2005) Eye movements in natural behavior. *Trends in Cognitive Sciences* 9(4): 188–194.

Huang CM and Mutlu B (2016) Anticipatory robot control for efficient human–robot collaboration. In: *ACM/IEEE International Conference on Human–Robot Interaction (HRI)*, pp. 83–90.

Javdani S, Admoni H, Pellegrinelli S, Srinivasa SS and Bagnell JA (2018) Shared autonomy via hindsight optimization for teleoperation and teaming. *The International Journal of Robotics Research* 37(7): 717–742.

Javdani S, Srinivasa SS and Bagnell JA (2015) Shared autonomy via hindsight optimization. In: *Robotics: Science and Systems (RSS)*.

Jayagopi DB, Sheiki S, Klotz D, et al. (2013) The Vernissage corpus: A conversational human–robot-interaction dataset. In: *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 149–150.

Johansson RS, Westling GR, Bä Ckströ A and Flanagan JR (2001) Eye–hand coordination in object manipulation. *The Journal of Neuroscience* 21(17): 6917–6932.

Kaelbling LP, Littman ML and Cassandra AR (1998) Planning and acting in partially observable stochastic domains. *Artificial Intelligence* 101(1): 99–134.

Kassner M, Patera W and Bulling A (2014) 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 (UbiComp '14) Adjunct*. New York: ACM Press, pp. 1151–1160.

Krejtz K, Duchowski AT, Niedzielska A, Biele C and Krejtz I (2018) Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLoS ONE* 13(9): 0203629.

Land MF and Hayhoe M (2001) In what ways do eye movements contribute to everyday activities? *Vision Research* 41(25): 3559–3565.

Littman ML, Cassandra AR and Kaelbling LP (1995) Learning policies for partially observable environments: Scaling up. In: Prieditis A and Russell S (eds.) *Machine Learning Proceedings 1995*. San Francisco, CA: Morgan Kaufmann, pp. 362–370.

Mainprice J, Hayne R and Berenson D (2015) Predicting human reaching motion in collaborative tasks using inverse optimal control and iterative re-planning. In: *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 885–892.

Pirsiavash H and Ramanan D (2012) Detecting activities of daily living in first-person camera views. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2847–2854.

Pupil Labs, Inc. (2017) Pupil labs - Pupil. https://pupil-labs.com/pupil/.

Reddy S, Dragan AD and Levine S (2018a) Where do you think you're going? Inferring beliefs about dynamics from behavior. *CoRR* abs/1805.08010.

Reddy S, Levine S and Dragan AD (2018b) Shared autonomy via deep reinforcement learning. *CoRR* abs/1802.01744.

Richardson IE (2010) *The H.264 Advanced Video Compression Standard*. 2nd edn. New York: John Wiley & Sons, Inc.

Sheikhi S and Odobez JM (2012) Recognizing the visual focus of attention for human robot interaction. In: *Proceedings of the Third International Conference on Human Behavior Understanding (HBU'12)*. Berlin: Springer-Verlag, pp. 99–112.

Shu T, Ryoo MS and Zhu SC (2016) Learning social affordance for human–robot interaction. In: *International Joint Conference on Artificial Intelligence (IJCAI)*.

Sigurdsson GA, Gupta A, Schmid C, Farhadi A and Alahari K (2018a) Charades-ego: A large-scale dataset of paired third and first person videos. *CoRR* abs/1804.09626.

Sigurdsson GA, Gupta A, Schmid C, Farhadi A and Alahari K (2018b) Charades-ego: A large-scale dataset of paired third and first person videos. *CoRR* abs/1804.09626.

Simon T, Joo H, Matthews I and Sheikh Y (2017) Hand keypoint detection in single images using multiview bootstrapping. In: *Proceedings of CVPR*.

Sondik EJ (1978) The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations Research* 26(2): 282–304.

Stefanov K and Beskow J (2016) A multi-party multi-modal dataset for focus of visual attention in human–human and human–robot interaction. In: *Proceedings of LREC*.

Stereolabs, Inc. (2018) Stereolabs. https://www.stereolabs.com/.

Thalmic Labs, Inc. (2018) Myo Gesture Control Armband. https://www.myo.com/.

Tobii AB (2017) Tobii pro. https://www.tobiipro.com/

Wei SE, Ramakrishna V, Kanade T and Sheikh Y (2016) Convolutional pose machines. In: *Proceedings of CVPR*.

Xu TL, Zhang H and Yu C (2013) Cooperative gazing behaviors in human multi-robot interaction. *Interaction Studies* 14: 390–418.

Zhang R, Liu Z, Zhang L, et al. (2018) AGIL: Learning attention from human for visuomotor tasks. In: *Proceedings of the 15th European Conference on Computer Vision (ECCV 2018)*, Munich, Germany, 8–14 September 2018, Part XI, pp. 692–707.

Ziebart BD, Maas A, Bagnell JA and Dey AK (2008) Maximum entropy inverse reinforcement learning. In: *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pp. 1433–1438.