Benchmark Movement Data Set for Trust Assessment in Human Robot Collaboration

Matthias Rehm
Technical Faculty of IT and Design
Aalborg University
Denmark
matthias@create.aau.dk

Kasper Hald
Technical Faculty of IT and Design
Aalborg University
Denmark
kh@create.aau.dk

Ioannis Pontikis
Technical Faculty of IT and Design
Aalborg University
Denmark
ipo@create.aau.dk

ABSTRACT

Trust is a factor that is becoming more prominent in human robot interaction research. Only few approaches so far tackle the challenge of data-driven trust assessment. In this paper, we present a data set consisting of motion tracking data from an industrial human robot collaboration task. The data is collected during a trust manipulation experiment that has been designed to elicit different trust levels in the participants. Additionally, participants filled out a standard trust questionnaire. The data set allows for developing and testing data-driven trust assessment algorithms.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; • **Computing methodologies → Machine learning**.

KEYWORDS

trust assessment, human-robot collaboration, motion tracking, machine learning

ACM Reference Format:
Matthias Rehm, Kasper Hald, and Ioannis Pontikis. 2024. Benchmark Movement Data Set for Trust Assessment in Human Robot Collaboration. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24), March 11–14, 2024, Boulder, CO, USA*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610977.3637472

1 STUDY OVERVIEW

While trust is an emergent factor in human robot interaction research ([6–8]), most trust measurements rely on post-hoc, subjective questionnaire data. This allows for using trust as an evaluation criterion for the system and interaction design similar like it has been used in automation for a long time (e.g. [5]). If trust could be linked to observable behavior of the user instead, then trust could become a factor that allows for adapting robot behavior in real-time in order to match or regulate the user's trust level. First attempts have been made for such a real-time trust assessment (e.g. [4, 9, 12]), but so far there are no data sets available for developing, testing, and comparing data-driven trust assessment.

HRI '24, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0322-5/24/03…$15.00
https://doi.org/10.1145/3610977.3637472

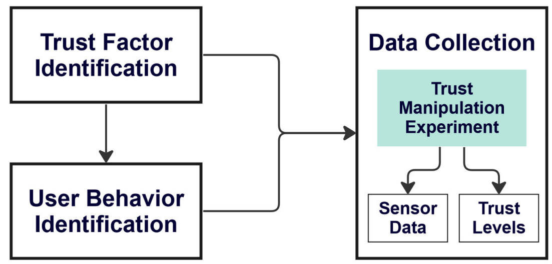

Figure 1: Collection of benchmark data base depends on identification of relevant trust factors and behaviors in a given application context.

In the Drapebot project[1], we are developing such a data-driven trust estimation system to ensure safe collaboration between a large industrial manipulator and a human worker in the context of draping of carbon-fibre pieces. To this end we have collected an extended data set of full body movement data and trust ratings in different task contexts, the transport of large textile cut pieces and a simulated draping of these cut pieces (see Section 2 for a detailed description). Because trust is a multidimensional concept that is affected by a broad range of different factors both relating to the user, the robot, and the environment [3], a number of criteria have to be fulfilled for creating such a data set, which are highlighted in Figure 1 and explained for our data set in the following:

Trust factor identification. : According to [3], we can distinguish between a large number of human-, robot-, and environment-related trust factors. For the given context it has to be clear which factors influence trust ratings, e.g. in an industrial setting with a collaborative robot proximity of the robot or it's speed would be relevant, whereas in a social setting with a small table top robot anthropomorphism or robot personality could be more relevant. From previous experiments [2], we know that speed of the robot plays a crucial role in perceived trust levels as well as in trust dynamics, e.g. for a robot with high speed starting trust levels are lower and it takes longer to reach a stable trust rating. Thus, the factor used in the data collection is the speed of the robot.

User behavior identification: Again, relevant user behaviors are context-dependent, in the case of an industrial collaborative robot, movement data of the upper body could be relevant, in case of the

[1]https://drapebot.eu

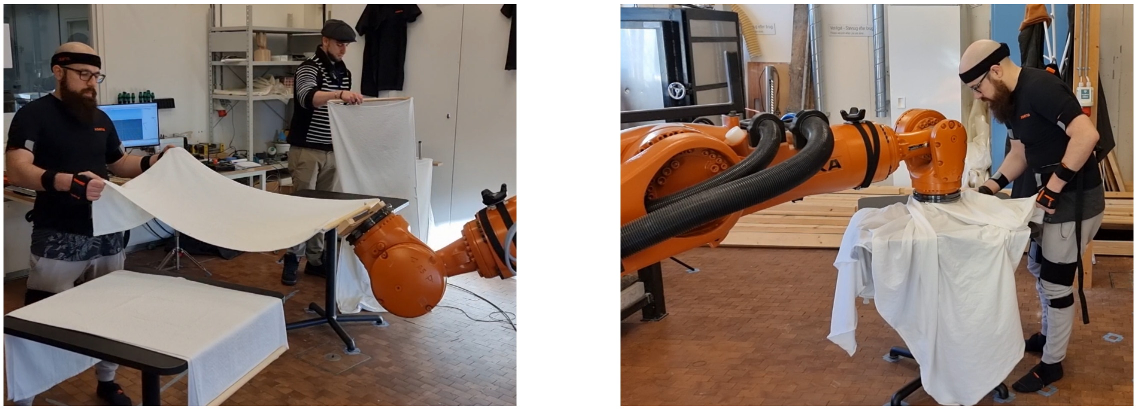

Figure 2: The collaborative transportation and draping task performed in the experiment.

social robot interaction, EEG data or facial expressions could be more relevant. For the Drapebot data set, the user has to move textile material across a room together with the robot. Thus, we collected full body movement data, hypothesizing that movement patterns will be different between low and high trust towards the robot. This hypothesis was drawn in analogy to emotion recognition, where it has been shown that movement expressivity can be linked to the user's emotional state (e.g. [10]).

Principled manipulation of trust during the data collection: When context-relevant trust factors have been identified together with their impact on trust evaluations by the user, then this knowledge allows for the elicitation of specific trust levels during the experiments, which can later be used for automatic labeling of the collected data. Based on our knowledge about the dynamics of trust development over time in relation to the robot speed [2], the experimental setup manipulated this factor across participants and tasks (see Section 2).

Extended collection of sensor data: While many questionnaire based studies rely on single interactions, we would argue that it is necessary to collect data over several task iterations to capture the dynamics of trust development.

Ground truth collection with standardized questionnaire. For each task, one of the standard trust questionnaires should be used for collecting a ground truth for the trust ratings of the individual participants. These ratings can be used for labeling purposes as well as for personalization purposes. For the Drapebot data set, we used the widely-used Trust perception scale - HRI [11].

The potential use of the data set is first and foremost for the development and evaluation of data-driven trust assessment models. With such models in place, trust monitoring and regulation become feasible, enabling safe and fluent human robot collaboration. Other uses include the in-depth analysis of movement quality and movement dynamics in extended repetitive interactions. For instance, during the data collection we observed different distinct walking styles in the transport task, from which some seemed to be more stable than others.

So far, the data has been used to derive motion descriptors based on Laban movement analysis [1] and to train a deep learning trust assessment algorithm for collaboration with large industrial manipulators. Result of the trust assessment is a trust score that is used to visualize the worker's current trust to the system so that s/he can adapt behavior accordingly, e.g. be more careful, when the system signals that the worker trusts the robot too much.

2 METHODS

The data was collected in an experiment where participants performed collaborative tasks with a Kuka KR 300 R2500 ultra robot. The tasks were designed to emulate collaborative transport and draping of carbon fibre cut pieces in the production of parts, such as outer parts for vehicles[2]. The collaborative transportation task consisted of the robot and the participant grabbing each their end of a towel laid out on a table before carrying it and laying it down on another table two meters away. This task is shown in Figure 2 (left). The collaborative draping task consisted of the robot bringing a cloth down above a table, leaving it bunched up on the table surface. Once the robot was at a complete stop the participants were instructed to approach the table and tug at the cloth to spread it as evenly across the table surface as they could before returning to their starting position, after which the robot would pull the cloth away and start the task over. This task is shown in Figure 2 (right). Both of these tasks were performed repeatedly by the participants as many times as they could within a ten-minute time limit. At no point in the duration of the experiment were the participants within collision reach of the moving robot. Safe areas were marked on the floor and the participants were monitored by the test conductor, who also had an emergency button within reach at all times.

The experiment was performed with speed as a between-subjects condition. Participants performed either the transportation or the draping task first. Second, during the transportation task, the robot would move at either a slow speed setting or a fast speed setting. Before the first tasks, as well as after both the first and second task, participants filled out Schaefer's[11] 14-item human-robot trust questionnaires.

At the beginning of the experiment the participants were dressed in the Xsens MVN Awinda tracking suit for full-body tracking.

[2]A demo video of the data collection tasks is available here: https://youtu.be/5pAXF7t1Us4

High quality tracking requires measurements of the participants body. Before putting on the tracking suit we measured height, foot length, shoulder height, shoulder width, elbow span, wrist span, arm span, hip height, hip width, knee height and ankle height. The suit consists of a tight-fitting shirt, straps, a headband and 17 inertial measurement units (IMU). The IMUs were applied by the test conductor before an N-pose calibration was performed. Figure 3 provides an overview of the placement of the IMU units on the body of the participant.

Figure 3: Placement of IMU units.

In addition to the IMUs we utilized a SteamVR tracking space with HTC Vive Trackers for position-aiding in order to counter position drift over time. In addition, one HTC Vive Tracker was placed on the robot itself for tracking its movements, allowing us to separate the steps of the collaboration in the recorded tracking data afterwards (see Section 3.2). All tracking sessions were refined after the experiments with the build-in HD reprocessing in the Xsens tracking sotware.

3 DATASET

Data has been collected in the transport and draping tasks (counterbalanced) from 20 participants, 7 female and 13 male, average age 25 (SD = 4.0). Average height was 1.74 meters (SD = 0.1). One session consists of 24 trials on average for the transport task with an average duration of 19.4 seconds per repetition and of 23.6 trials for the draping task with an average duration of 18.6 seconds. This resulted in 479 repetitions for the transport task (9276 seconds) and 472 for the draping task (8777 seconds). For all sessions, body tracking was performed using the Xsens MVN Awinda tracking suit. It consists of a tight-fitting shirt, gloves, headband, and a series of straps used to attach 17 IMUs to the participant. After calibration the system uses inverse kinematics to track and log the movements of the participant at a rate of 60 Hz. The measurements include linear and angular speed, velocity, and acceleration of every skeleton tracking point[3].

For each task (transport and draping) the data set consists of a folder with 20 individual files, one for each participant. The naming convention for these files follows a pattern wherein "P01SD" denotes a participant in the draping task (with "D" representing draping), and "P01ST" represents a participant in the transport task (with "T" indicating transport). Each of these files is in xlsx-format. In addition to these task-specific files, three xlsx-files are available. Two of these files are dedicated to annotating the collected data, one for draping and one for transport. The third xlsx-file contains the trust scores associated with each task for each participant respectively. This organization ensures a systematic approach to data management and analysis within the research repository. In the following, we describe each file type in detail.

3.1 Tracking Data

For each file generated by the Xsens tracking system, the data are categorized into sessions, each with its unique characteristics:

(1) Segment Orientation - Quaternion, (2) Segment Orientation - Euler, (3) Segment Position, (4) Segment Velocity, (5) Segment Acceleration, (6) Segment Angular Velocity, (7) Segment Angular Acceleration, (8) Joint Angles ZXY, (9) Joint Angles XZY, (10) Ergonomic Joint Angles ZXY, (11) Ergonomic Joint Angles XZY, (12) Center of Mass, (13) Sensor Free Acceleration, (14) Sensor Magnetic Field, (15) Sensor Orientation - Quaternion, (16) Sensor Orientation - Euler.

For each of the 17 sensors, this data is provided for the X, Y, and Z axes, as well as combined data. For more detailed information regarding the specific data types and sensors, please refer to the XSENS manual.

3.2 Data Annotation

The two task consist of different phases, which are exemplified in Figure 4. In the transport task, we can distinguish between picking up the textile, transporting it to the other table, dropping it on the table and returning to the starting point. Draping task consists of approaching the table, draping the textile and returning to the starting point. For each task, a separate file named "sorted_draping.xlsx" and "sorted_transport.xlsx" exists. These files follow the follwing format:

1. column: frame number in the corresponding tracking data file, e.g. P01SD.xlsx
2. column: annotations for participant 01 on a frame by frame basis
3. column: annotations for participant 02 on a frame by frame basis

   ...

21. column: annotations for participant 20 on a frame by frame basis

For the transport task, there are four annotations: "pick," "transport," "drop," and "return". In the case of the draping task, there are three distinct annotations: "approach," "draping," and "return".

[3] see the Xsens manual for a detailed description of available measurements: https://base.movella.com/s/article/Output-Parameters-in-MVN-1611927767477

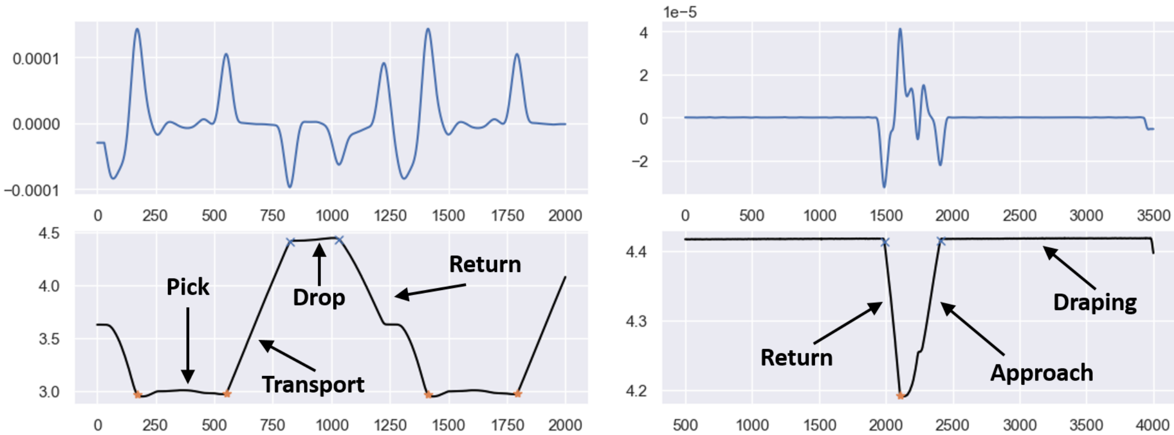

Figure 4: Position tracking of the robot allows to annotate the different phases of the transport (left) and draping task (right).

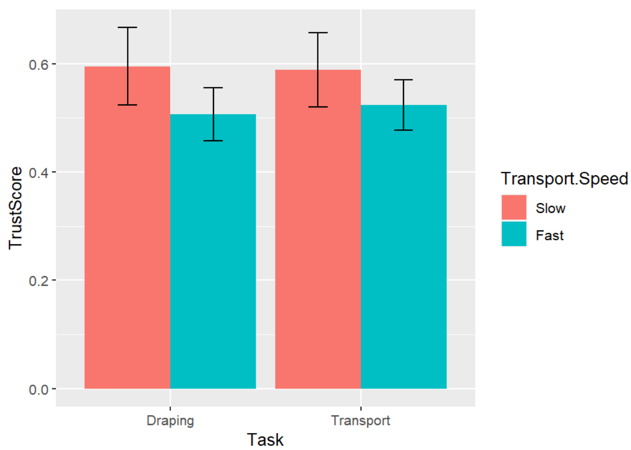

Figure 5: Distribution of trust scores in the data set.

3.3 Trust Scores

The trust scores associated with the data are based on the participants' answers to the questions of the Trust perception scale HRI. Figure 5 gives an overview of the trust score distribution across the different tasks and speeds. The table contains two entries for each participant, one for the transport and one for the draping task. The file consists of two parts. First, some demographic information is provided for each participants:

- Subject: Denoting the participant number
- Transport Speed: Indicating whether the session was conducted at a fast or slow pace
- Age: Reflecting the participant's age
- Gender: Identifying the participant's gender
- Dominant Hand: Specifying the participant's dominant hand
- Height: Providing information about the participant's height

Then the answers to each question of the 14 questions from the instrument are given:

- Which % of time does the robot (1) Function successfully, (2) Act consistently, (3) Communicate with people, (4) Provide feedback, (5) Malfunction, (6) Follow directions, (7) Meet the needs of the task, (8) Perform exactly as instructed, (9) Have errors, (10) Provide appropriate information.
- Which % of the time is the robot (1) Unresponsive, (2) Dependable, (3) Reliable, (4) Predictable.

The overall trust score is then calculated based on responses to these questions and provided as an additional column. The last column denotes the task as either "Transport" or "Draping". On our Github[4], we provide an example Jupyter notebook for importing the data into Python for further processing.

4 USAGE NOTES

The data set is available on Zenodo[5] as open access with the Creative Commons Attribution 4.0 International License following FAIR principles. When using the data set, we would be grateful for a reference to this paper. While the data set contains human subject data, this data is anonymous (both motion tracking and questionnaire data) and does not pose any ethical problems. The data collection and sharing has been approved by the research board at Aalborg University.

Trust is of increasing interest in the HRI community as can be seen by the number of publications in recent years. It is our conviction that a data-driven approach to trust will boost the usefulness of the concept for safe, fluent and adaptive human robot collaboration. To this end, data sets like this one are necessary for testing, evaluating and benchmarking approaches to data-driven trust assessment.

ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101006732 (Drapebot).

[4]hhttps://github.com/HRI-AAU/DrapebotExample
[5]https://zenodo.org/record/8224067

REFERENCES

[1] Kasper Hald and Matthias Rehm. 2023. Determining Movement Measures for Trust Assessment in Human-Robot Collaboration Using IMU-Based Motion Tracking. In *32nd IEEE International Conference on Robot and Human Interactive Communication*. IEEE, United States.

[2] Kasper Hald, Matthias Rehmn, and Thomas B. Moeslund. 2020. Human-Robot Trust Assessment Using Motion Tracking and Galvanic Skin Response. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 6282–6287. https://doi.org/10.1109/IROS45743.2020.9341267

[3] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[4] Wan-Lin Hu, Kumar Akash, Neera Jain, and Tahira Reid. 2016. Real-Time Sensing of Trust in Human-Machine Interactions. *IFAC-PapersOnLine* 49, 32 (2016), 48 – 53.

[5] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[6] Kin Man Lee, Arjun Krishna, Zulfiqar Zaidi, Rohan Paleja, Letian Chen, Erin Hedlund-Botti, Mariah Schrum, and Matthew Gombolay. 2023. The Effect of Robot Skill Level and Communication in Rapid, Proximate Human-Robot Collaboration. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, New York, NY, USA, 261–270. https://doi.org/10.1145/3568162.3577002

[7] Jennifer E. Martinez, Dawn VanLeeuwen, Betsy Bender Stringam, and Marlena R. Fraune. 2023. Hey?! What Did You Think about That Robot? Groups Polarize Users' Acceptance and Trust of Food Delivery Robots. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, New York, NY, USA, 417–427. https://doi.org/10.1145/3568162.3576984

[8] Nina Moorman, Erin Hedlund-Botti, Mariah Schrum, Manisha Natarajan, and Matthew C. Gombolay. 2023. Impacts of Robot Learning on User Attitude and Behavior. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, New York, NY, USA, 534–543. https://doi.org/10.1145/3568162.3576996

[9] SM Mizanoor Rahman, Yue Wang, Ian D Walker, Laine Mears, Richard Pak, and Sekou Remy. 2016. Trust-based compliant robot-human handovers of payloads in collaborative assembly in flexible manufacturing. In *2016 IEEE International Conference on Automation Science and Engineering (CASE)*. IEEE, 355–360.

[10] Matthias Rehm. 2010. *Nonsymbolic Gestural Interaction for Ambient Intelligence*. Academic Press, United States, 327–345.

[11] Kristin E. Schaefer. 2016. *Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI"*. Springer US, Boston, MA, 191–218. https://doi.org/10.1007/978-1-4899-7668-0_10

[12] Shayan Shayesteh, Amit Ojha, and Houtan Jebelli. 2022. Workers' Trust in Collaborative Construction Robots: EEG-Based Trust Recognition in an Immersive Environment. *Automation and robotics in the architecture, engineering, and construction industry* (2022), 201–215.