Hand-Eye-Object Tracking for Human Intention Inference

Adebayo, S., McLoone, S., & Dessing, J. C. (Accepted/In press). Hand-Eye-Object Tracking for Human Intention Inference. In Intelligent Control and Automation Sciences - 6th ICONS 2022: Proceedings International Federation of Automatic Control.

Published in:
Intelligent Control and Automation Sciences - 6th ICONS 2022: Proceedings

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2022 IFAC.
This manuscript is distributed under a Creative Commons Attribution-NonCommercial-NoDerivs License (https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits distribution and reproduction for non-commercial purposes, provided the author and source are cited.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person’s rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.
Hand-Eye-Object Tracking for Human Intention Inference

Samuel Adebayo Seán McLoone Joost C. Dessing

Centre for Intelligent Autonomous Manufacturing Systems
Queen’s University Belfast, Northern Ireland, UK
(e-mail: {sadebayo01, s.mcloone, j.dessing}@qub.ac.uk)

Abstract: Optimizing human-robot interaction in the performance of collaborative tasks is a challenging problem. An important aspect of this problem relates to the robot’s understanding of human intentions. Empowering robots with accurate intention inference capabilities sets the stage for more natural, safe, and efficient interactions and greater confidence in the Human-Robot Interaction domain. Intentions can be deduced by observing human cues as they interact with the environment, but currently, there is no clear-cut method for achieving this in an effective way. Here, we present a novel method for intention inference based on the integration of three visual cues, namely, hand movement, eye fixation, and object interaction, coupled with a bidirectional LSTM neural network for the classification of human intention. Experimental studies evaluating our approach against two-visual cue alternatives confirm the utility of our approach.

Keywords: Intention inference; Hand-Eye-Object Interaction; Vision Systems; Big data, Data Engineering; Machine Learning.

1. INTRODUCTION

The last decade has seen a tremendous advance in Human Robot Interaction (HRI) research, driven by the desire to have human-centric, human-machine collaborative systems, which some are referring to as the 5th industrial revolution. Despite this, the effectiveness and efficiency of existing HRI systems for collaborative tasks is often limited when compared to human-human interaction due to the inability of robots to infer human intention. To achieve natural and effective Human-Robot co-working it is essential to endow robots with perceptive abilities to enable them to infer human intention. Consequently, developing an optimal framework to infer intention is a very active area of research (Li and Zhang, 2017).

To achieve intelligent intention inference, there is a need to consider all observable quantities and corresponding encodings involved in interaction. Doing so should give insights into the required information for active collaboration. Motion estimation is an active component in the planning and execution of a robot’s decision making algorithm (Zhang et al., 2021). Hence, accurate human pose estimation is required to tackle and improve HRI frameworks.

Human Hand-Eye (H-E) coordination is the ability to carry out tasks that require the simultaneous use of the hands and eyes - such as activities that use the hands to coordinate motion based on information perceived by the eyes. The human eyes aid the brain to determine the spatial position of objects in the environment as well as body segments. In a human task involving H-E coordination, the eye gaze usually precedes hand motion. Therefore, based on visual spatial information from the eyes, the hands carry out the set task. Fixations (attention to an anticipated target) on objects are unique and can be extracted and related to task features. Hence, Human Hand-Eye-Object (H-E-O) interactions involve the interdependence of hand spatial position (while in motion) in interaction with object(s) following observable (encoded) perception from the eyes. The application of modern machine learning techniques to achieve a vision-based approach to intention inference is a relatively new area of research. However, the advent of cheap cameras, high performance computing, and modern machine learning techniques have opened up new opportunities for vision researchers to explore the possibilities for developing improved frameworks for HRI.

In HRI the use of human gaze has been shown to improve the quality of interactions, increase persuasive ability as a storyteller, and enhance robot to human hand over tasks (Ham et al., 2015). Moreover, there are studies backing the use of gaze cues as a method of improving conversations and enabling natural interactions with humans (Willems et al., 2018). The hand is also actively involved in HRI, due to its function in collaborative assembly tasks: handover, picking, grasping, gripping, etc. Therefore, it makes sense for all task-relevant components of interaction and assembly - hands, eyes, and object - to be considered when inferring intention. Although, hand-eye coordination involving two or more robots (robot-to-robot) has been researched extensively using several methods and approaches (Levine et al., 2018), to date little work has been done on the incorporation of vision-based hand-eye-object signals for the purpose of inducing intention in HRI frameworks.

We hypothesize that the introduction and integration of gaze cues, hand tracking, and object tracking in HRI using a purely vision-based system enhanced by deep learning
Motion tracking is largely stochastic and cannot be readily predicted and et al., 1995). The authors argued that a human’s behaviour led as a function of human behavior such as in (Inagaki and conclusions in Sections 4 and 5, respectively. 

The remainder of this paper are: 

(1) A novel approach to intention inference involving integration of Hand, Eye, and Object signals. Although, there are several approaches and techniques that involve integration of different signals to infer intention (see Section 2), this is the first attempt at inferring intention by integrating these three visual cues.

(2) A learning-based classification approach to intention inference.

(3) An efficient data generation pipeline for Hand-Eye-Object integrated systems.

The closest work to our approach is Trombetta et al. (2020) who present a fusion of gaze cues and hand motion to predict a human’s true intention in an analogous robot-collaborative task. As will be seen, our method presents as input a fusion of specific 2D locations of the hands (right and left), objects and eye fixation. Additionally, our approach is a simultaneous fusion of the aforementioned visual cues within a deep neural network classifier using end-to-end learning, while their approach involves preprocessing gaze points using a Kalman Filter before fusing it with hand motion data for neural network training.

Recently, HRI research has focused more on optimising the technology pipeline to enhance the interactivity of robots (Chatila et al., 2018). Human-centered robots have been mostly based on passive sensing, characterised by machine learning methods, and in particular deep learning techniques, to provide robots with capabilities such as speech recognition (Matsui et al., 1999; Okuno et al., 2001; Lauria et al., 2002; Spiliotopoulos et al., 2001), gesture recognition (Luo and Wu, 2012; Triesch and von der Malsburg, 2006; Chakraborty et al., 2018; Liu and Wang, 2018), facial and emotion recognition (Kute et al., 2019; Messer and Fausser, 2019), gaze tracking (Li and Zhang, 2017) and pose estimation (Tanaka et al., 2019). Notably, it was believed until recently that obtaining near perfect vision-based human pose estimation was a difficult task to tackle. However, the advent of modern deep neural networks for pose estimation has made obtaining, recording, and utilization of human motion easier and accessible. A recent state-of-the-art pose estimation model (Mediapipe) developed by the Perception team at Google Research achieves impressive performance and runs in real-time on relatively modest standard computing platforms (Zhang et al., 2020). The backbone of mediapipe hand pose estimation consists of 2 state-of-the-art models - Blazepalm (Bazarevsky et al., 2019), a version of the Blazeface model for detecting palms in any input image, and the hand landmark model that processes the output of the palm detector to output 21 hand landmarks. Mediapipe hands was trained on a large diverse dataset consisting of synthetic, real and in-the-wild images. The model is robust, fast, and able to detect hands in low light conditions. Lightweight and heavyweight versions of the model have

Fig. 1. Hand-Eye-Object tracking for intention inference

Fig. 2. Proposed Hand-Eye-Object Intention Deduction Framework

will improve intention inference, compared to solutions based on only one or two dependent signals. To test this hypothesis we develop an experimental framework in which the proposed fusion of vision signals are the inputs to a multi-class classifier that predicts the intended action from a set of candidate actions. The block diagram of our proposed framework is presented in Fig. 2. We present a data generation pipeline, and a machine learning model that learns to classify intention from a combination of different tracked visual components. The main contributions of the paper are:

2. RELATED WORK

Vision-based intention inference has been researched extensively over the last few decades. Some of the earliest vision-based approaches to intention inference were modelled as a function of human behavior such as in (Inagaki et al., 1995). The authors argued that a human’s behaviour is largely stochastic and cannot be readily predicted and programmed. Hence, they proposed intention inference to be induced from observable behavioural quantities, by programming the why, what, how, when, who, and when questions. These observable quantities were measured with a perception sensor in a closed coordinated structured environment. Intention inference was then achieved by computation of these processed quantities using Fuzzy Logic. Other early vision frameworks for intelligent robots were in social robotics. These robots take visual cues such as object detection, face recognition, etc. to induce social awareness (Fong et al., 2003; Feil-Seifer and Mataric, 2009).

Hybrid techniques based on the fusion of vision systems and natural language processes to achieve specific high level tasks have also shown promising results (Moladande and Madhusanka, 2019). Speech recognition has been a key enabler in achieving intention inference in these systems. For the most part, they involve the human workers conversing with the robots using a set of speech commands to carry out specific tasks, hence intention to carry out the tasks is deduced.

Recently, HRI research has focused more on optimising the technology pipeline to enhance the interactivity of robots (Chatila et al., 2018). Human-centered robots have been mostly based on passive sensing, characterised by machine learning methods, and in particular deep learning techniques, to provide robots with capabilities such as speech recognition (Matsui et al., 1999; Okuno et al., 2001; Lauria et al., 2002; Spiliotopoulos et al., 2001), gesture recognition (Luo and Wu, 2012; Triesch and von der Malsburg, 2006; Chakraborty et al., 2018; Liu and Wang, 2018), facial and emotion recognition (Kute et al., 2019; Messer and Fausser, 2019), gaze tracking (Li and Zhang, 2017) and pose estimation (Tanaka et al., 2019). Notably, it was believed until recently that obtaining near perfect vision-based human pose estimation was a difficult task to tackle. However, the advent of modern deep neural networks for pose estimation has made obtaining, recording, and utilization of human motion easier and accessible. A recent state-of-the-art pose estimation model (Mediapipe) developed by the Perception team at Google Research achieves impressive performance and runs in real-time on relatively modest standard computing platforms (Zhang et al., 2020). The backbone of mediapipe hand pose estimation consists of 2 state-of-the-art models - Blazepalm (Bazarevsky et al., 2019), a version of the Blazeface model for detecting palms in any input image, and the hand landmark model that processes the output of the palm detector to output 21 hand landmarks. Mediapipe hands was trained on a large diverse dataset consisting of synthetic, real and in-the-wild images. The model is robust, fast, and able to detect hands in low light conditions. Lightweight and heavyweight versions of the model have
been developed, the former for deployment on edge and mobile devices and the later for deployment on high-performance machines. We used the on-device lightweight implementation in our experiments.

3. METHODS

To support our hypothesis, we developed a data generation pipeline for Hand-Eye-Object tracking and interaction to train and evaluate the proposed human action predictor. In this section we describe the materials and methods employed, including hardware requirements and setup, signal integration, and data collection. The data generation pipeline is described in Fig. 5.

3.1 Experimental Design

Our setup includes three parts - eye tracking, hand tracking, and object tracking (Fig. 5). A single Nexigo N930AF 1080P USB camera was used as a vision system for both the hand and object tracking, while a Pupil Core Eye-tracker (Kassner and Patera, 2012) was used to track the eyes. The pupil core eye-tracker is a head-mounted system with two adjustable infrared cameras positioned just below both eyes to capture the pupil position. The camera was mounted on the eye-tracker and positioned between the eyes so as to capture the world scene (hands and objects). The resolution of the world scene camera was 360 × 240 and the resolution of the cameras inferring eye fixation was 190 × 190. This configuration also facilitated calibration of the eye tracker fixation point. A Raspberry Pi was used to fuse and time synchronise the camera and Pupil Core data streams.

The software component of the setup includes the program to run pupil core, a scene fetching algorithm, a lightweight deep learning hand tracker (Mediapipe), and an object recognition model (YOLOv5). Since Pupil Core’s software is publicly available, we built our script on it, and integrated all the code to work as one build. An object dataset consisting of 250 images each of three objects (single book, a set stacked books, and a mug) was compiled and used to train a custom object detection model via transfer learning using YOLOv5 provided by Ultralytics (Ultralytics, 2021). The resulting model detects custom objects and outputs rectangular bounding boxes that correspond to the location boundary coordinates of the detected objects. It is able to perform inference in real-time and at a rate of 22 frames per second.

With Mediapipe Hands 21 2.5D keypoint locations can be plotted for the detected hands. The keypoints are in x, y, z (relative depth) coordinates. We extracted these keypoints at every data collection run. To extract the gaze location, the eye tracker was calibrated with 4-circled position corresponding to the rectangular mapping of the world scene camera. Pupil Core allows for recording and extraction of the gaze location in the world scene camera, which means one is able to deduce where in the scene users are looking.

Video streaming, time synchronization, and data collection scripts were developed and incorporated as a plugin within the Pupil Core software. The overall system achieves a data collection frame rate of 59 fps. The data is preprocessed to a form which allows for optimal accessibility and processing.

3.2 Data Capture

In order to reduce time complexity, and minimize algorithmic reliance on long term dependencies, we assume that the human intention to perform a collaborative task remains the same throughout the task. We consider 7 tasks in total:

1. handover
2. pickup
3. set down
4. move right (single hand)
5. move right (both hands)
6. move left (single hand)
7. move left (both hands)

with each one taking between four and six seconds to execute. The intention to perform a task is then deduced by observing the visual cues over the first one second of the task execution (59 frames). Further details on how this is achieved are presented in subsequent paragraphs.
The first one second of data was collected for 30 repetitions of each task, with each repetition performed with a slightly different orientation and choreography. This resulted in an overall dataset of 210 examples. We split the data into 90% training data and 10% test data. Due to Covid-19 restrictions, data collection was restricted to one person performing the tasks.

During each data collection run, a volunteer wears the head-mounted eye tracker, with the world scene camera mounted in such a way as to face the scene of the assembly task. Before data collection Pupil Core was calibrated to synchronize with the world scene view - following calibration the participant was asked where they were focused on, to ensure convenient matching of eye fixations in the world frame to actual fixation. The participant was then instructed on the choreographed tasks - a mixture of hand demonstrations, with specified instructions on employing normal movement as they would when carrying out such low-level tasks. For tasks involving interaction with an object, the participant was instructed on the motion around the objects in order to avoid occlusion.

For each data collection sequence, key points corresponding to locations of fixations of the eyes in the scene \((x, y)\), normalized coordinates of the object (rectangular bounding box), and normalized \(x, y, z\) location of the 21 annotated points on each human hand were recorded. One Hand-Eye-Object data point, denoted \(\{H, E, O\}\), is made up of the concatenations of normalized coordinates corresponding to the hands, eye and object in the world view. Here \(H, E\) and \(O\) are sets representing the key points of the hands in the interacting scene, the fixation points in the world view frame, and the rectangular bounding boxes respectively. Specifically,

\[
E = (e_x, e_y) \quad (1)
\]

\[
H = \{(h_{ix}^i, h_{iy}^i, h_{ix}^i, h_{iy}^i, h_{ix}^i, h_{iy}^i) \mid i = 1, ..., 21\} \quad (2)
\]

\[
O = (O_x, O_y, O_w, O_h) \quad (3)
\]

where \(e_x\) and \(e_y\) represent the normalized coordinates of the fixation in the world frame,

\[
H = \{(h_{ix}^i, h_{iy}^i, h_{ix}^i, h_{iy}^i, h_{ix}^i, h_{iy}^i) \mid i = 1, ..., 21\} \quad (2)
\]

where \(h_{ip}^i\) and \(h_{rj}^i\) denote the left and right \(p\)-th coordinate of the \(i\)-th key point, and

\[
O = (O_x, O_y, O_w, O_h) \quad (3)
\]

where \(O_x, O_y, O_w, O_h\) denote the normalized coordinates of the rectangular bounding boxes of the detected objects. In total 136 key points are recorded at each sampling instant (frame).

### 3.3 Bidirectional LSTM for Intention Prediction

Bidirectional LSTMs learn sequential features using both forward and backward passes, leading to quicker and more accurate networks (Siami-Namini et al., 2019). Each layer is made up of two LSTMs, one for each direction. As illustrated in Fig. 6 our neural network is made up of 7 layers, comprising of 3 Bidirectional LSTM layers and 4 fully connected layers with a softmax for classification. The dimension of the input to the network is 59 x 136 (59 frames x 136 key points). The resulting neural network has 1.589 million parameters.
4. RESULTS

To assess the performance of the proposed H-E-O intention prediction model we first evaluate its classification accuracy on the training and test data sets. Then we compare its performance to two-visual cue alternatives at real-time intention inference. Our model achieved 100% accuracy on the training data and 90.47% accuracy on the test data with regard to classifying the true sub-task intention. A detailed analysis of the classification performance on the test data is presented in Fig. 8 and Table 1. Fig. 8 shows the confusion matrix for the results while Table 1 reports the precision, recall and the $F_1$ score for each of the intention classes.

The results of the real-time comparison with two-visual cue based models are reported in Table 2. The two-visual cue intention prediction models considered were Hand-Eye (H-E) and Hand-Object (H-O). We trained these models using the same dataset and network architecture used to train the H-E-O model. The classification performance for the H-E and H-O models on the test data set are 85.23% and 89.21%, respectively. A real-time data comparison of the classification performance of these models and the H-E-O intention prediction models was also carried out. This was achieved by running the models on a live data stream of the volunteer performing the tasks. We categorized the acquired data as being either for coordinated or uncoordinated task executions. A coordinated task execution is where the volunteer follows the same choreography sequence and time steps as employed for the original training and test dataset, while an uncoordinated task execution is one where the volunteer does not follow a prescribed choreography. We tested intention classifier performance for 30 instances of each of the 7 task classes under these conditions. The reported comparisons are based on the accuracy of prediction over the 30 instances. All tests were carried out on an Ubuntu Machine running Nvidia GeForce RTX 3050.

4.1 Discussion

From the results obtained, it can be concluded that there is a significant improvement in intention inference when using the proposed integrated hand-eye-object signals methodology when compared to H-O, which is in-turn superior to H-E. H-E-O is 13.67% and 10.79% better than H-O for coordinated tasks and uncoordinated, respectively.

| Table 1. Accuracy of Hand-Eye-Object Model |
|--------------------------------------------|
| Intention class | Precision | Recall | $F_1$ score |
|-----------------|-----------|--------|-------------|
| handover1       | 1.00      | 1.00   | 1.00        |
| pickup1         | 1.00      | 1.00   | 1.00        |
| setdown         | 1.00      | 1.00   | 1.00        |
| moveright1      | 0.67      | 1.00   | 0.80        |
| moveright2      | 1.00      | 1.00   | 1.00        |
| moveleft1       | 0.67      | 1.00   | 0.80        |
| moveleft2       | 1.00      | 0.67   | 0.80        |
| accuracy        | 0.90      |        |             |

| Table 2. Comparison of H-E and H-O and H-E-O models for classification of coordinated and uncoordinated tasks |
|-------------------------------------------------------------------------------------------------------------|
| H-E | H-O | H-E-O |
|-----|-----|-------|
| % Coord Task (30) | 0.7216 | 0.7755 | 0.9122 |
| % UnCoord Task (30) | 0.6332 | 0.7111 | 0.8190 |

5. CONCLUSION AND FUTURE DIRECTION

In this paper we present a novel approach to learning-based intention inference and a methodology for collecting hand-eye-object datasets. Experimental results confirm that our deep learning model is able to learn the statistical relationship between the hands, eyes, and interacting object for intention prediction in reaching type tasks. We believe that our proposed end-to-end learning and visual cue fusion methodology will reduce the complexity of integrating signals in vision-based HRI and provide the basis for advancing the state-of-the-art in intention aware robots. It is noted that the results presented are based on a preliminary dataset. Ongoing work is targeted at collecting a much large intention classification dataset involving multiple participants and a broader range of tasks to enable a much more general intention inference model to be developed.

ACKNOWLEDGEMENTS

This research is funded by the Engineering and Physical Sciences Research Council (EPSRC), UK.
REFERENCES

Bazarevsky, V., Kartyunik, Y., Vakunov, A., Raveendran, K., and Grundmann, M. (2019). BlazeFace: Sub-millisecond neural face detection on mobile GPUs. *CoRR*, abs/1907.05047. URL http://arxiv.org/abs/1907.05047.

Chakraborty, B.K., Sarma, D., Bhuyan, M.K., and MacDorman, K.F. (2018). Review of constraints on vision-based gesture recognition for human–computer interaction. *IET Computer Vision*, 12(1), 3–15.

Chatila, R., Renaudo, E., Andries, M., Chavez-Garcia, R.O., Luce-Vayrac, P., Gottstein, R., Alamii, R., Clodic, A., Devin, S., Girard, B., et al. (2018). Toward self-aware robots. *Frontiers in Robotics and AI*, 5, 88.

Feil-Seifer, D. and Matarić, M.J. (2009). Toward socially assistive robotics for augmenting interventions for children with autism spectrum disorders. In O. Khatib, V. Kumar, and G.J. Pappas (eds.), *Experimental Robotics*, 201–210. Springer Berlin Heidelberg, Berlin, Heidelberg.

Fong, T., Nourbaksh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4), 143–166.

Ham, J., Cuijpers, R., and Cabibihan, J. (2015). Combining robotic persuasive strategies : the persuasive power of a storytelling robot that uses gazing and gestures. *International Journal of Social Robotics*, 7(4), 479–487.

Inagaki, Y., Sugie, H., Aisu, H., Ono, S., and Unemi, T. (1995). Behavior-based intention inference for intelligent robots cooperating with human. In *Proceedings of 1995 IEEE International Conference on Fuzzy Systems*, volume 3, 1695–1700.

Kassner, M.P. and Patera, W.R. (2012). *PUPIL: Constructing the Space of Visual Attention*. Master’s thesis, Massachusetts Institute of Technology.

Kute, R.S., Vyas, V., and Anuse, A. (2019). Component-based face recognition under transfer learning for forensic applications. *Information Sciences*, 476, 176–191.

Lauria, S., Bugmann, G., Kyriacou, T., and Klein, E. (2002). Mobile robot programming using natural language. *Robotics and Autonomous Systems*, 38(3-4), 171–181.

Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., and Quillen, D. (2018). Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5), 421–436.

Li, S. and Zhang, X. (2017). Implicit intention communication in human–robot interaction through visual behavior studies. *IEEE Transactions on Human-Machine Systems*, 47(4), 437–448.

Liu, H. and Wang, L. (2018). Gesture recognition for human-robot collaboration: A review. *International Journal of Industrial Ergonomics*, 68, 355–367.

Luo, R.C. and Wu, Y.C. (2012). Hand gesture recognition for human-robot interaction for service robot. In *2012 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, 318–323.

Matsui, T., Asoh, H., Fry, J., Motomura, Y., Asano, F., Kurita, T., Hara, I., and Otsu, N. (1999). Integrated natural spoken dialogue system of jijo-2 mobile robot for office services. In *AAAI/IAAI*.

Messer, U. and Fausser, S. (2019). Predicting social perception from faces: A deep learning approach. *CoRR*, abs/1907.00217.

Moladande, M. and Madhusanka, B. (2019). Implicit intention and activity recognition of a human using neural networks for a service robot eye. In *2019 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, 38–43.

Okuno, H., Nakadai, K., Hidai, K., Mizoguchi, H., and Kitano, H. (2001). Human-robot interaction through real-time auditory and visual multiple-talker tracking. In *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No.01CH37180)*, volume 3, 1402–1409 vol.3.

Siami-Namini, S., Tavakoli, N., and Namin, A.S. (2019). The performance of lstm and bilstm in forecasting time series. In *2019 IEEE International Conference on Big Data*, 3285–3292.

Spiliotopoulos, D., Androussopoulos, I., and Spyropoulos, C.D. (2001). Human–robot interaction based on spoken natural language dialogue. In *in Proc. European Workshop on Service and Humanoid Robots, 2001*, 25–27.

Tanka, R., Oshima, C., and Nakayama, K. (2019). Intention inference from 2d poses of preliminary action using openpose. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion, GECCO ’19*, 1697–1700. Association for Computing Machinery, New York, NY, USA.

Triesch, J. and von der Malsburg, C. (2006). *Robotic gesture recognition*, 233–244. Springer, Berlin, Heidelberg.

Trombetta, D., Rotithor, G.S., Salehi, I., and Dani, A.P. (2020). Human intention estimation using fusion of pupil and hand motion. *IFAC-PapersOnLine*, 53(2), 9535–9540. 21st IFAC World Congress.

Ultralytics (2021). *Yolov5: The friendliest ai architecture you’ll ever use*. "https://ultralytics.com/yolov5".

Willemse, C., Marchesi, S., and Wykowska, A. (2018). Robot faces that follow gaze facilitate attentional engagement and increase their likeability. *Frontiers in Psychology*, 9.

Zhang, D., Vien, N.A., Van, M., and McLoone, S. (2021). Non-local graph convolutional network for joint activity recognition and motion prediction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2970–2977.

Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C., and Grundmann, M. (2020). Mediapipe hands: On-device real-time hand tracking. *CoRR*, abs/2006.10214. URL https://arxiv.org/abs/2006.10214.