Body Gesture Recognition to Control a Social Mobile Robot

Javier Laplaza
javier.laplaza@upc.edu
Institut de robòtica i Informàtica Industrial de Barcelona
Barcelona, Spain

Ramón Romero
Institut de robòtica i Informàtica Industrial de Barcelona
Barcelona, Spain

Alberto Sanfeliu
Institut de robòtica i Informàtica Industrial de Barcelona
Barcelona, Spain

Anais Garrell
Institut de robòtica i Informàtica Industrial de Barcelona
Barcelona, Spain

ABSTRACT

In this work, we propose a gesture-based language to allow humans to interact with robots using their body in a natural way. We have created a new gesture detection model using neural networks and a new dataset of humans making a collection of body gestures to train this architecture. Furthermore, we compare body gesture communication with other communication channels to demonstrate the importance of adding this knowledge to robots. The presented approach is validated in diverse simulations and real-life experiments with non-trained volunteers. This attains promising results and establishes that it is a valuable framework for social robotic applications, such as human robot collaboration or human-robot interaction.

CCS CONCEPTS

• **Human-centered computing** → **Gestural input**; • **Computing methodologies** → *Machine learning*; • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

KEYWORDS

datasets, neural networks, HRI

ACM Reference Format:
Javier Laplaza, Ramón Romero, Alberto Sanfeliu, and Anais Garrell. 2023. Body Gesture Recognition to Control a Social Mobile Robot. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3568294.3580126

1 INTRODUCTION

Finding natural and new efficient communication channels is essential in Human-Robot Interaction (HRI). If we take a look at the way humans communicate with each other, we see that about 70% of the communication is non-verbal communication [12]. Moreover, when humans want to communicate with other agents with whom they do not share a common spoken language –foreigners, babies or animals– most of the communication is non-verbal [2, 5].

When it comes to communication with robots, it is possible to establish a set of gestures to communicate certain ideas in a similar way that gesture language works between humans. But, similarly to gesture language, this approach requires that both agents know which gestures compose the language and what meaning each gesture has. Furthermore, previous attempts of creating such body language with robots are generally designed by people used to working with robots. We argue that such kinds of languages cannot be expected to find success when robots face humans with no previous experience with robots.

This inspired us to seek a natural way for humans to communicate with robots using gestures. Our goal is to create a natural gesture dictionary and explore how humans rate gesture based communication versus other communication channels.

To achieve this, we first collected a dataset using human volunteers making a set of communication gestures. Then, we create a gesture detection model to allow the robot to identify human gestures. Finally, we wanted to study how convenient using gestures is for communication. To do so, we carried out several experiments using non-trained volunteers and the IVO robot [8], see Fig.1, in order to compare gesture based communication to other communication channels.

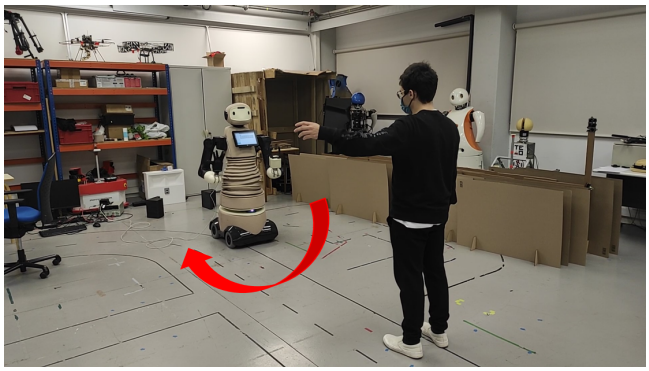

Figure 1: (©Javier Laplaza) By raising the left arm, the human is able to tell the IVO robot to turn to the left side.

Work supported under the Spanish State Research Agency through the Maria de Maeztu Seal of Excellence to IRI (MDM-2016-0656) and the EU project CANOPIES (H2020- ICT-2020-2-101016906).

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9970-8/23/03…$15.00
https://doi.org/10.1145/3568294.3580126

2 RELATED WORK

Most of the literature related to gesture detection focuses on hand gesture detection. Great examples of this are [1, 4] and [16]. Touchless interaction methods that will not use sound or speech for the communication need to somehow sense the human body's position.

Some work on full body gesture communication focus on identifying where the human is pointing at to identify a specific object or region, one example of this being [13]. Other works focus on identifying emotions expressed by the body posture, such as [11]. Moreover, Some gestures are very restricted to specific tasks, in [7] the approach is to detect whether a human is willing to collaborate or not during a hand-over operation according to his gestural expression.

The work proposed in [10] is similar to our approach, but they use the relative position of hands and faces to define poses. More importantly, they use UAV as the platform to interact, which has different dynamics than ground robots and also a different perspective of humans when using a camera.

Another similar work is presented in [3], focusing in optical flow techniques used in order to extract features of the human body.

Finally, in [15] a set of gestures are proposed for HRI operations. While this research is very similar to our approach, in our work we focus on allowing very general gestures instead of defining them beforehand. Also, we study how human volunteers rate the interaction with the robot using gestures.

3 MODEL ARCHITECTURE

In order to properly identify different body gestures, we created a model that consists of a neural network classifier using the body joint positions as input paramaters and returning a gesture class probability as output.

Although the model input are body joint positions, the pipeline inputs are videos. In the work presented in this paper, the temporal component has been omitted and only the last frame of each video is used as input. This means that we only take into account static gestures, since the skeleton position during these gestures don't change.

MediaPipe [9] was used to extract the full body skeleton from the image frames, obtaining a 3D skeleton from an RGB image. This model predicts the location of 33 pose landmarks, each following the $(x, y, z, visibility)$ structure that will be used as our Classifier input.

The proposed neural network uses four $Linear\ y = xA^T + b$ fully connected layers of size $256, 128, 64$ and 8 combined with $ReLU(x) = (x)^+ = \max(0, x)$ activation functions between each layer.

The outputs from our classifier consists of a one dimensional vector with length equal to the number of gestures to be predicted, making each vector element the model score for each specific gesture.

From this vector, a $Softmax$ layer is applied in order to normalize the model scores and the maximum gesture probability is then considered to be the output gesture.

For training, the user 2 was arbitrarily chosen as test dataset, whereas the rest of the volunteers were used as training data.

Figure 2: (©Javier Laplaza) Some samples of static gestures recorded in the dataset.

The loss function used for training is the Cross-Entropy Loss, and the chosen optimization algorithm for this network is the ADAM optimizer with a learning rate $\alpha = 0.01$ and $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$ hyperparameters.

4 DATASET

Since we couldn't find a suitable dataset for our goal, we decided to create a dataset in our laboratory.

The main feature of our gesture based communication dictionary is naturalness. We want everyone to be able to communicate with the robot, not only people who are already familiar with robots.

We divide the defined gestures in two groups: static and dynamic gestures (see Fig. 2):

- **Static gestures**
  - **Attention**: Catch the robot's attention to give him an order.
  - **Right**: Order the robot to turn right.
  - **Left**: Order the robot to turn left.
  - **Stop**: Order the robot to stop its trajectory.
  - **Yes**: Approve a robot's information.
  - **Shrug**: Inform the robot that you don't understand his information.
  - **Random**: Random gesture, not necessarily a communication gesture.
  - **Static**: Human is standing still.
- **Dynamic gestures**
  - **Greeting**: Greet the robot.
  - **Continue**: Order the robot to continue its path after telling him to stop.
  - **Turn-back**: Order the robot to turn 180 degrees.
  - **No**: Deny a robot's information.
  - **Slowdown**: Order the robot to reduce its speed.
  - **Come**: Order the robot to reach your position.
  - **Back**: Order the robot to move back.

As for data recording, each human volunteer was recorded using an RGB camera. When human volunteers were asked to make a gesture they were provided with a vague explanation of the gesture intention. This was done to collect data that felt most natural to

each volunteer. There was no restriction on which arms should be moved in each gesture whatsoever. Thus, different volunteers could make the same gesture in a very different way, using one arm or the other, or even both of them.

Each gesture was repeated three times, first 1 meter away from the camera, then 4 meters away and finally 6 meters away.Each video contains information of only one gesture, and all the videos were recorded indoors, with no body self occlusions.

Finally, we use MediaPipe [9] to extract the 3D joints of the human in the video.

We also consider a wide range of users regarding age, gender, education level and culture. Taking this into account, we created our dataset thanks to 10 human volunteers, 7 men and 3 women.

5 EXPERIMENTS

In this section, we show the obtained results with our architecture and the developed user study to demonstrate the acceptability of the presented framework.

5.1 Model results

We train the model until it starts overfitting on the test dataset. Then, we stop the training and check the results on the test dataset. The results can be seen in Table 4.

The results show that the model accuracy for certain gestures is remarkable. Specifically, *attention, right, left, shrug and static* gestures rate an F1-score above 0.8. This result was expected, since all those gestures are very different from other gestures in the dataset.

On the other hand, there are two gestures that the model particularly struggles classify: *stop and yes*. Again, this result was expected, since both gestures present an overall body pose very similar, only the hand position differs.

Finally, the *random* gesture has a F1-score of 0.64, which is understandable for a class that include a rich distribution of body poses.

We created a confusion matrix (see Fig. 6) to study inter-class miss-classifications. From this matrix, we confirm that our model isn't able to differentiate gestures *yes and stop*. Possible ways to tackle this can be increasing model complexity or using a more complete Mediapipe model that incorporates finger and face information.

5.2 User Study

The results presented in the previous section demonstrate that the robot is able to detect and recognize human natural gesture. A user study was also conducted to determine whether the body gesture recognition to control our robot enhances the usability and the comfort of the robot from the point of view of the human. We compared our method with the use of a remote controller.

The hypothesis we endeavored to test was as follows: "Participants will feel more comfortable and will perceive difference between the use of body gesture recognition and the use of a remote controller."

We asked humans to communicate different orders to the robot, specifically:

- Order the robot to move closer.
- Order the robot to move away.
- Order the robot to turn to the right.
- Order the robot to turn to the left.
- Order the robot to look up.
- Order the robot to look down.

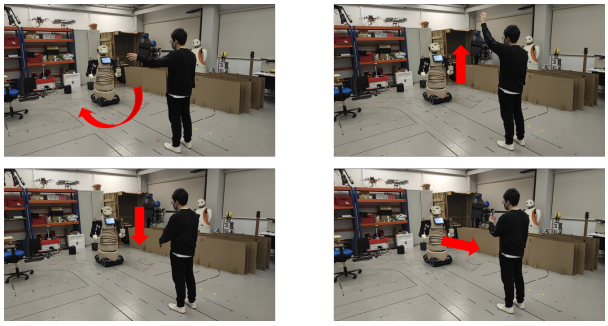

Figure 3: (©Javier Laplaza) Samples of gestures used by the user: *turn left* (top left), *look upwards* (top right), *look downwards* (bottom left) and *approach* (bottom right).

Gesture	Precision	Recall	F1-score
attention	1.00	0.77	0.87
right	1.00	0.9	0.95
left	0.89	0.83	0.86
stop	0.52	0.4	0.45
yes	0.38	0.57	0.46
shrug	0.76	0.87	0.81
random	0.66	0.63	0.64
static	0.87	0.9	0.88

Figure 4: (©Javier Laplaza) Precision, recall and F1-score for every model gesture class.

Note that even though volunteers were told what order they had to give, they weren't told how they should give the order, the same way that the dataset was collected.

In the first experiment, the human had to use the natural gestures to express to the robot what action must perform. We conducted these experiments in a Wizard-of-Oz way, since using the gesture detector may lead to missing some of the gestures, and it can cause a negative impact to the user perception of gestures as a channel of communication. The delay between the human making the gesture and the robot following the order was around 1 second.

Then, we repeated the same experiment but this time we gave a remote controller to the human, thus that he/she could tele-operate the robot after some instruction. The robot started the motion as soon as the human operated the controller.

In each case, we asked the volunteers to make all gestures/commands in a random order. We also chose randomly between the gesture communication and the controller as the first experiment for each

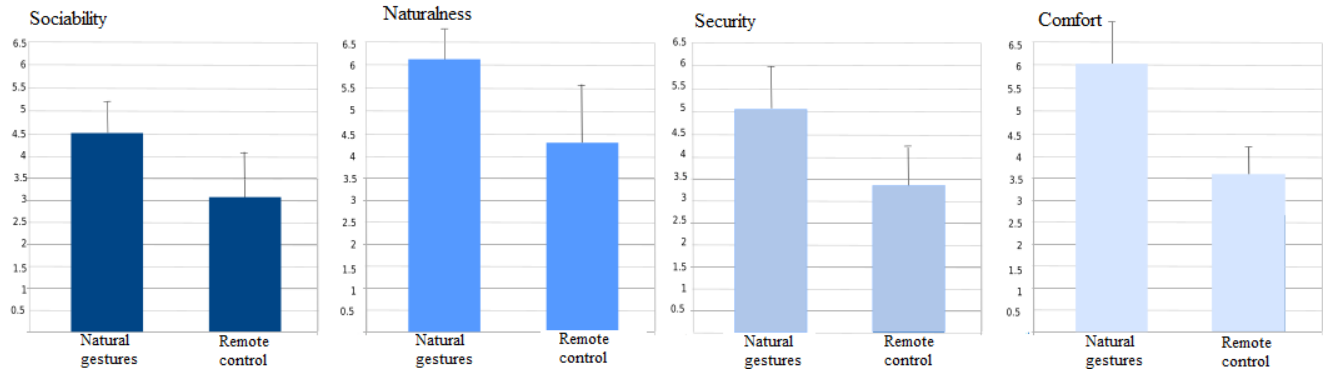

Figure 5: (©Javier Laplaza) Evaluation from 1 (low) to 7 (high) of the main aspects related to the robot behavior in body gesture recognition.

Figure 6: (©Javier Laplaza) Model confussion matrix.

volunteer in order to avoid possible biases. Refer to Fig.3 to see some samples of the gesture communication.

For the experiments, we selected 15 people (8 men, 7 women) on the University Campus. Participants ranged in age from 19 to 50 years (M=29.5, SD=9.2).

Participants were asked to complete a variety of surveys based on [14]. Our independent variables considered whether participants make use of our gesture recognition or the remote control. The main dependent variables involved participants' perceptions of the **sociability**, **naturalness**, **security** and **comfort** characteristics. Each of these fields, was evaluated by every participant using a questionnaire to fill out after the experiment based on [6].

Participants were asked to answer a questionnaire, following their encounter with the robot in each mode of behavior. To analyze their responses, we grouped the survey questions into four scales: the first measured robot's sociability, while the second naturalness, and third and fourth evaluated the security and comfort, respectively. Both scales surpassed the commonly used 0.7 level of reliability (Cronbach's alpha).

Each scale response was computed by averaging the results of the survey questions comprising the scale. ANOVAs were run on each scale to highlight differences between the three robot behaviors.

Below, we provide the results of comparing the two different methods. To analyze the source of the difference, four scores were examined: "sociability", "naturalness", "security" and "comfort", plotted in Fig. 5. For all four aspects, the evaluation score plotted in Fig. 5, pairwise comparison with Bonferroni demonstrate there were difference between the two kind of behavior approaches, $p < 0.05$.

Therefore, after analyzing these four components, we may conclude that if the robot is capable of understand people's body gesture the acceptability of the robots increases, and participants perceived the robot as a social entity.

6 CONCLUSIONS

We created a dataset of natural gestures to communicate with a robot. We also developed a new neural network architecture able to classify the static gestures in the dataset. We reach high classification accuracy (> 80%) in most of the gestures, except for gestures that look very similar. Our future work will try to add also the dynamic gestures to the model and decouple gestures that can be easily confused.

The experiments we conducted yielded conclusive results. We found that people felt their interaction with the robot was more natural when the robot communicated through gestures. Detailed analysis showed that these capacities improved the human's perception of the robot's security and sociability. Finally, volunteers perceived our robot more sociable and closer when it recognized and understood their gestures, and the interactions were longer.

The findings presented in the previous section reinforce the notion that the robot's ability to understand human body gestures is an important skill to master in order to achieve natural interaction with people. Overall, people interacting and communicating with the robot using natural gestures enhances the engagement between the robot and the human.

REFERENCES

[1] Samer Alashhab, Antonio-Javier Gallego, and Miguel Ángel Lozano. 2019. Hand Gesture Detection with Convolutional Neural Networks. In *Distributed Computing and Artificial Intelligence, 15th International Conference*, Fernando De La Prieta, Sigeru Omatu, and Antonio Fernández-Caballero (Eds.). Springer International Publishing, Cham, 45–52.

[2] Michael Argyle. 1972. Non-verbal communication in human social interaction. (1972).

[3] Jen-Yen Chang, Antonio Tejero-de-Pablos, and Tatsuya Harada. 2019. Improved Optical Flow for Gesture-based Human-robot Interaction. *CoRR* abs/1905.08685 (2019). arXiv:1905.08685 http://arxiv.org/abs/1905.08685

[4] Amanda Cardoso Duarte, Samuel Albanie, Xavier Giró-i-Nieto, and Gül Varol. 2022. Sign Language Video Retrieval with Free-Form Textual Queries. *CoRR* abs/2201.02495 (2022). arXiv:2201.02495 https://arxiv.org/abs/2201.02495

[5] Robert A Hinde and Robert Aubrey Hinde. 1972. *Non-verbal communication*. Cambridge University Press.

[6] Rachel Kirby. 2010. *Social robot navigation*. Carnegie Mellon University.

[7] Jun Kwan, Chinkye Tan, and Akansel Cosgun. 2020. Gesture Recognition for Initiating Human-to-Robot Handovers. arXiv:2007.09945 [cs.CV]

[8] Javier Laplaza, Nicolás Rodríguez, J. E. Domínguez-Vidal, Fernando Herrero, Sergi Hernández, Alejandro López, Alberto Sanfeliu, and Anaís Garrell. 2022. IVO Robot: A New Social Robot for Human-Robot Collaboration. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Sapporo, Hokkaido, Japan) *(HRI '22)*. IEEE Press, 860–864.

[9] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines.

[10] Sepehr MohaimenianPour and Richard Vaughan. 2018. Hands and Faces, Fast: Mono-Camera User Detection Robust Enough to Directly Control a UAV in Flight. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 5224–5231. https://doi.org/10.1109/IROS.2018.8593709

[11] Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari. 2021. Survey on Emotional Body Gesture Recognition. *IEEE Transactions on Affective Computing* 12, 2 (2021), 505–523. https://doi.org/10.1109/TAFFC.2018.2874986

[12] Tim O'Sullivan, John Hartley, Danny Saunders, Martin Montgomery, and John Fiske. 1994. *Key concepts in communication and cultural studies*. Routledge.

[13] Jagdish Lal Raheja, Mona Chandra, and Ankit Chaudhary. 2018. 3D gesture based real-time object selection and recognition. *Pattern Recognition Letters* 115 (2018), 14–19. https://doi.org/10.1016/j.patrec.2017.09.034 Multimodal Fusion for Pattern Recognition.

[14] Ely Repiso, Anaís Garrell, and Alberto Sanfeliu. 2018. Robot approaching and engaging people in a human-robot companion framework. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 8200–8205.

[15] Jia Chuan A. Tan, Wesley P. Chan, Nicole L. Robinson, Elizabeth A. Croft, and Dana Kulic. 2021. A Proposed Set of Communicative Gestures for Human Robot Interaction and an RGB Image-based Gesture Recognizer Implemented in ROS. *CoRR* abs/2109.09908 (2021). arXiv:2109.09908 https://arxiv.org/abs/2109.09908

[16] Nurettin Çağrı Kılıboz and Uğur Güdükbay. 2015. A hand gesture recognition technique for human–computer interaction. *Journal of Visual Communication and Image Representation* 28 (2015), 97–104. https://doi.org/10.1016/j.jvcir.2015.01.015