A Long-Range Mutual Gaze Detector for HRI

Simone Arreghini
simone.arreghini@supsi.ch
Dalle Molle Institute for Artificial Intelligence (IDSIA)
USI-SUPSI, Lugano, Switzerland

Gabriele Abbate
gabriele.abbate@supsi.ch
Dalle Molle Institute for Artificial Intelligence (IDSIA)
USI-SUPSI, Lugano, Switzerland

Alessandro Giusti
alessandro.giusti@supsi.ch
Dalle Molle Institute for Artificial Intelligence (IDSIA)
USI-SUPSI, Lugano, Switzerland

Antonio Paolillo
antonio.paolillo@supsi.ch
Dalle Molle Institute for Artificial Intelligence (IDSIA)
USI-SUPSI, Lugano, Switzerland

ABSTRACT

The detection of mutual gaze in the context of human-robot interaction is crucial for the understanding of human partners' behavior. Indeed, the monitoring of the users' gaze from a long distance enables the prediction of their intention and allows the robot to be proactive. Nonetheless, current implementations struggle or cannot operate in scenarios where detection from long distances is required. In this work, we propose a ROS2 software pipeline that detects mutual gaze up to 5 m of distance. The code relies on robust off-the-shelf perception algorithms.

CCS CONCEPTS

• **Computer systems organization** → Robotic autonomy; • **Human-centered computing** → Code release.

KEYWORDS

Social robotics, Service robots, Human-centered perception

ACM Reference Format:
Simone Arreghini, Gabriele Abbate, Alessandro Giusti, and Antonio Paolillo. 2024. A Long-Range Mutual Gaze Detector for HRI. 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/10.1145/3610977.3637470

1 INTRODUCTION

In Human-Robot Interaction (HRI), nonverbal communication is given crucial importance [11, 20]. In particular, the user gaze is a very powerful indicator of human thinking processes and intentions, even before they are expressed [6, 7]. Nonetheless, capturing nonverbal cues for HRI purposes is not an easy task [19]. For gaze detectors, technical limitations usually restrict usage to close distances [13] [23], where eyes and face landmarks are more easily captured. In the context of social HRI, it is desirable to endow robots with perception skills that allow them to detect human behavior from far, well before the start of a potential interaction [5].

The process of extracting gaze information can be divided into two main problems: estimation of the gaze direction and detection of mutual gaze. The former is usually tackled by using precise gaze tracker algorithms. Solutions based on Deep Learning (DL) are available [13, 22, 23]; however, they lose performance at long distances. Alternative approaches implementing longer-range detectors [8, 12] usually require specific hardware (like infra-red projectors), which is not available in the standard sensory equipment of social robots. The tracking performance of such cumbersome sensors can be replicated using simpler RGB cameras [14] at the expense of keeping the maximum operating distance at around 1.8 m. Recent solutions [21] showed encouraging results in estimating the gaze direction from long distances. However, to the best of our knowledge, currently, there is no released implementation. On the other side, the detection of the mutual gaze is a much simpler problem. In general, mutual gaze defines direct eye contact between individuals, and in the HRI context of this work, it refers to whether a person is looking at a robot. This problem has been previously studied with

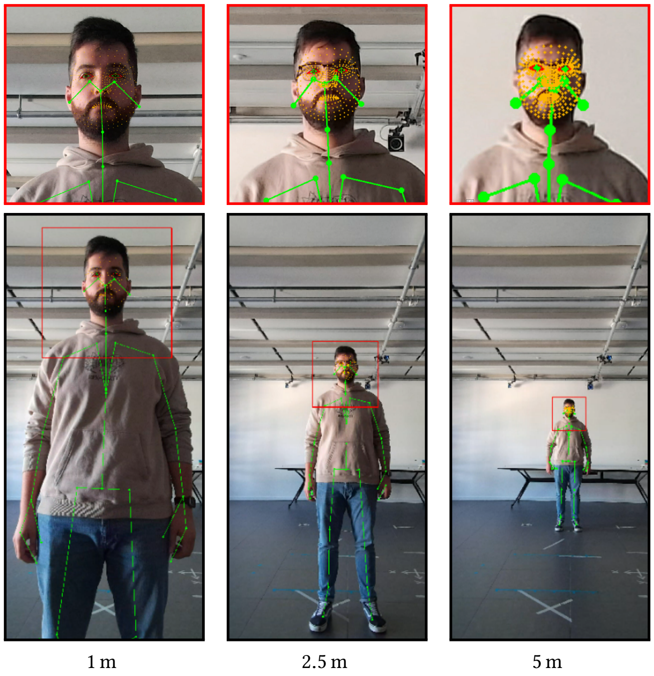

Figure 1: Our approach implements a long-range mutual gaze detector fusing the information from the users' body frames (in green) and the facial landmarks (in orange).

HRI '24, March 11–14, 2024, Boulder, CO, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0322-5/24/03.
https://doi.org/10.1145/3610977.3637470

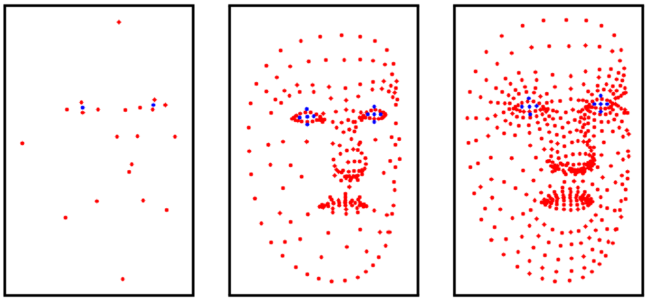

Figure 2: Different numbers of face landmarks (21 on the left, 249 in the center, 478 on the right) detected on the same face. The ones related to eye pupils and irises are in blue.

good results [9], even in the robotics domain, making use of body landmarks in an approach similar to our proposed solution [15]. However, these methods are insufficient for many social HRI applications as the maximum operating distance is below 2 m. Instead, a desirable operating zone should at least include the robot's social space, defined as a circle of 4 m around the robot [16, 19]. To this end, we use a face mesh tracking solution with a finer resolution (allowing more control over the number and location of the landmarks) and integrate it with a 3D body tracker. Fusing carefully chosen face landmarks with body information enables our mutual gaze detector to work at distances up to around 5 m.

Our approach aims at providing the HRI community with a long-range mutual gaze detector, see Fig. 1. We design a detector that only relies on a standard sensor and can be easily used in different HRI tasks and scenarios. Examples of applications are: (*i*) prediction of the intention to interact, as detecting mutual gaze from far distances is indicative of users' interaction will; (*ii*) engagement monitoring, to adapt the robot's behavior to maintain or increase engagement during tasks or conversations; (*iii*) collaborative tasks, e.g. for industrial manufacturing, as mutual gaze tracking can improve coordination. The approach leverages off-the-shelf algorithms providing information on the users' body motion and facial landmarks. It is implemented in the Robot Operating System 2 (ROS2) framework and can be easily installed and deployed.

2 APPROACH

2.1 Problem formulation

We consider a robot that performs social interaction with humans. The robot is equipped with an RGB-D sensor, and its reference frame is defined at the center of its RGB camera sensor. This frame is oriented so that its vertical axis consistently aligns with the gravity vector, while the heading angle remains unconstrained and free to follow the robot's orientation. This design choice is crucial because it allows to be independent of the camera movement and orientation, thus allowing the camera to be mounted on moving parts, e.g. a robotic head. The information about the users' body motion is expressed by monitoring two particular frames of interest: one located in the person's chest, and the other on their head. We assume that the poses of these frames w.r.t. the robot's frame are measurable by the robot perception system. Furthermore, we also assume that the camera RGB images allow the detection of facial landmarks, which consist of the projected locations, on the image plane, of specific points of interest detected on a person's face. Each landmark is defined by a 3D vector consisting of its image coordinates (see Fig. 2), and the predicted depth of the corresponding point on the user's face. In the proposed approach, facial landmarks and body information are fed to a classifier that predicts whether the subject is looking at the robot, i.e. gives an estimate of the *mutual gaze*. We aim to provide an implementation of such a mutual gaze classifier that can work for distances greater than 2 m.

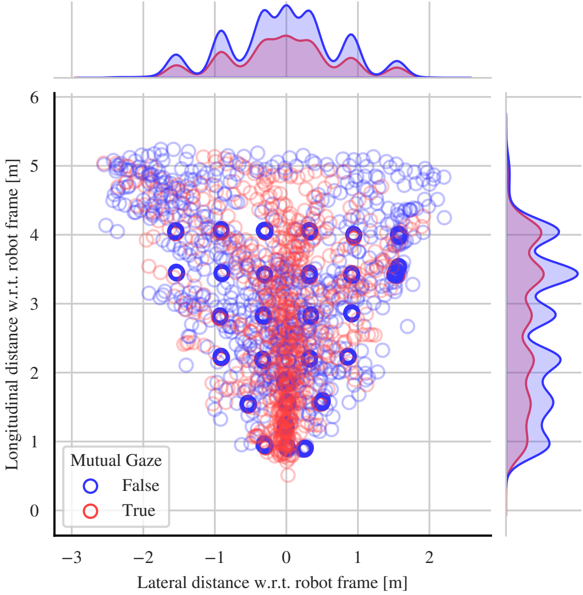

Figure 3: Users' position w.r.t. the robot frame during the data collection and the normalized per-axis distributions.

2.2 Data structure

Previously available algorithms for mutual gaze estimation [15] are not designed for distances larger than 2 m. Publicly available gaze datasets do not contain data acquired at further distances and are not fit for our needs (see, e.g., [10] or the dataset related to [23]). Therefore, we train our classifier on a dataset collected ad hoc. Such a dataset is composed of two subsets differing in the users' motion patterns during the acquisition campaign. The first subset is called *Standing set* and gathers data of users standing in front of the robot in 28 predefined positions. Such positions are arranged on a grid pattern to cover the entire sensor's Field of View (FoV) ranging from 0.8 m to 4.2 m. In each position, the user does not walk but moves their head and torso while alternating periods of looking at the robot and elsewhere. For each instant, we record the users' body frame poses, the facial landmarks, and the mutual gaze ground truth, switching position after about 90 s. Data are labeled by the subjects themselves, by pressing a wireless remote button when they directly look into the sensor. Data thus collected produced 12373 samples (34.8% true and 65.2% false). The Standing set provides a lot of data, useful to build the core capabilities of the classifier, but it does not offer much variability. Therefore, the

Passing by set is also recorded. In this dataset, the subjects pass in front of the robot with different random moving patterns. This set contains 1449 samples (56.5% are true and 43.5% false). Figure 3 shows the pattern of the users' position for each sample contained in the whole training dataset, with the darker positions showing the grid pattern used for the Standing set. The complete dataset is composed of 13822 samples; 522 of them (all belonging to the Passing by set) are used as a test set; the remaining 13300 are used for training. The local institutional ethical committee has cleared the data-gathering campaign and the experiments.

2.3 Implementation details

The RGB-D sensor is placed at a height of around 1.10 m that is typical of many commercially available social robots, e.g. Tiago by PAL Robotics [17] or Pepper by Aldebaran [18]. The sensor used in this work is Azure Kinect RGB-D sensor [1], streaming images at the maximum resolution of 4096 × 3072 pixels at 5 Hz; with horizontal and vertical FoV equal to 90° and 59°, respectively.

The sensor's Software Development Kit (SDK) provides the tracking of 32 salient body frames composing a skeleton of the user, as shown in green in Fig. 1. From such a skeleton, we select the head and chest frame information. As for the facial landmarks, we use a modified version of MediaPipe library [2]. This package offers detection of up to 478 facial landmarks from RGB images, providing a very detailed mesh of the users' faces. However, the vanilla implementation lacks complete support for multiple users and long distances simultaneously. To overcome this issue, we use the head frame information from the Azure Kinect to manually track the users' faces in the input image. The users' face regions are then cropped from the image and rescaled to 200 × 200 pixels, i.e. the internal input resolution of MediaPipe's face landmark detector. This operation does not impact the precision of the extracted landmarks thanks to the high-resolution image of the sensor. Indeed, the dimension, in the image space, of a normal-sized human face at around 5 m is only marginally smaller than the resizing window. This process also comes with the additional benefits of increased speed and stability of the tracking process, even at close range. The cropped region around the users' heads can be seen as red squares in Fig. 1. The classifier is tested with different subsets of facial landmarks, spanning from the full mesh of 478 landmarks, to around half size with 249 landmarks, and finally the smallest subset of 21 landmarks, i.e. about the 4% of the original set. The different facial landmark sets are shown in Fig. 2. Regarding the classifier architecture, we rely on a simple stateless Random Forest (RF), which can provide good robustness while being very lightweight. We use the scikit-learn [4] implementation of such algorithm leaving the default settings for training as well as the number of 100 trees.

2.4 Performances

We compare different versions of the classifier: using only facial landmarks (the corresponding model is denoted with F*n*, where *n* is the number of the landmarks) or in combination with chest and head pose information (CH-F*n*). The performances are evaluated with the AUROC metric calculated using a 10 fold cross-validation approach. The algorithm is first evaluated frame-by-frame; the results are shown in Fig. 4. From the results of the first 3 models (F21, F249, and F478), which only use facial landmarks, one can appreciate that richer facial features do not translate into better performances, but into a decrease of the mean AUROC from 89.1% to 87.9%. This trend is even more pronounced when combining chest and head information (model CH-F21, CH-F249, and CH-F478). Indeed, the model using the smallest facial landmarks subset (CH-F21) exhibits the best result with an average AUROC of 92.6%. The best model (CH-F21) is evaluated w.r.t. the distance from the camera sensor. The training samples are grouped into 5 bins according to the absolute distance from the sensor frame. The results in Fig. 5 show that the performance is always above 90%. Finally, to validate the generalization of our classifier w.r.t. to unseen samples, CH-F21 is tested on the smaller testing set, which contains completely separated sequences from the training set on which the classifier has an AUROC of 93.3%. This generalization capability comes from the fact that the proposed classifier works with pre-processed data and not raw inputs, thus relying on the robustness of the off-the-shelf

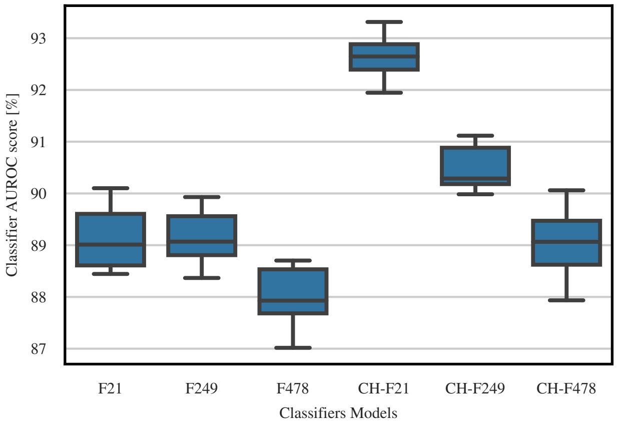

Figure 4: Area Under Receiver Operating Curve (AUROC) metric of different versions of the mutual gaze classifiers using different sets of input feature.

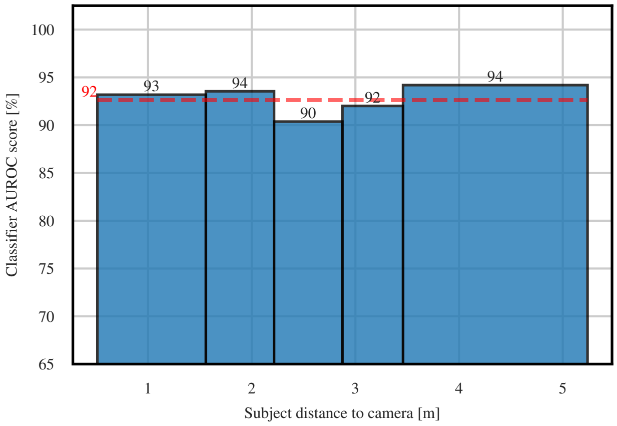

Figure 5: AUROC of the CH-F21 classifier (vertical axis) tested at different distance ranges (horizontal axis). The distances are grouped in quintiles of the test set.

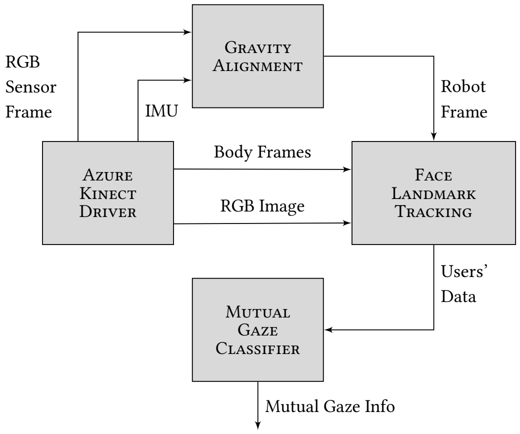

Figure 6: ROS2 architecture of our system.

components of the pipeline. The robustness has been empirically confirmed in further experiments with multiple unseen subjects. The system displays a maximum operating range of about 5 meter. This is due to limitations of the sensor in both the maximum distance for reliable body tracking and the resolution of the image, which after 5 meter causes the face region of interest to be blurred.

3 SOFTWARE

3.1 System Architecture

The whole system, implemented using the ROS2 infrastructure [3], is composed of 4 different nodes (see Fig. 6) described below.

3.1.1 Azure Kinect Driver. This node runs the driver of the RGB-D sensor and the body tracking offered by its SDK. It publishes the following information: (*i*) the raw signals of the Kinect's onboard Inertial measurement unit (IMU); (*ii*) the RGB image streams; (*iii*) the body frames from the sensor SDK; and (*iv*) the RGB camera frame. The frame information is standardized in ROS2 as *tf* messages.

3.1.2 Gravity Alignment. This node calculates the gravity-aligned robot frame as defined in Sec. 2.1. It takes the IMU data to calculate the difference of the RGB sensor frame orientation w.r.t. the inertial vertical direction and broadcasts the aligned robot frame. Ultimately, the robot frame is defined as the one centered in the origin of the RGB sensor frame constrained vertically to be aligned with gravity and horizontally with the camera heading.

3.1.3 Face Landmarks Tracking. This node is the core of the perception pipeline and implements the face landmarks extraction. It takes as input the user's body and robot frames information and the current RGB image. Firstly, it performs projection of the users' 3D head positions onto related 2D points on the image plane. This information is used to crop the regions of interest of the RGB image corresponding to the faces of the detected users. Such cropped images are fed to multiple instances of the MediaPipe Face Mesher, which detects the face landmarks for each user. This crucial step, introduced in Sec. 2.3, allows us to overcome the distance range limits of the MediaPipe implementation. Exploiting the robust detection of the head frames provided by the Azure Kinect SDK, we can allow the face landmarks detection by MediaPipe at distances further than 2 m. To be independent of the camera orientation, the body frame poses, originally expressed in the RGB sensor frame, are transformed into the gravity-aligned robot frame. The detected face landmarks and the transformed body frame poses of the detected users are finally time synchronized and published as a custom ROS2 topic. Such message is called *Users' Data* in the scheme of Fig. 6.

3.1.4 Mutual Gaze Classifier. This node is a ROS2 wrapper for the actual scikit-learn implementation of our classifier. It takes as input the user data custom topic provided by the Face Landmark Tracking node. As output, it publishes a simple custom topic that is denoted *Mutual Gaze Info* in Fig. 6. This message contains the IDs of the detected users and the corresponding probability of mutual gaze as computed by the classifier. This node also runs a GUI showing the real-time evolution of the predicted probability related to the user who has been tracked for the longest time.

3.2 Code release

Our software is open-sourced under the MIT license and hosted on the GitHub page of our institution, available at:

https://github.com/idsia-robotics/mutual_gaze_detector

where the README file explains in detail the procedure needed to create the right setup and test the code. Specific paragraphs address code maintenance, licensing, and deployment with an emphasis on responsible use of the software. One can choose to either create an environment on their machine or use ready-to-use Docker images, which are provided to ease the code setup by executing the code from within a container. This choice comes at the cost of a reduced execution speed of the pipeline. We provide two different images; the first can be used to run the entire system online using an Azure Kinect; the second offers a quick demo displaying the code capabilities using pre-recorded data. The data are provided as ROS2 *rosbag*, which contains frame-wise pre-processed (anonymous) user's information. This is useful if an Azure Kinect is not readily available to the user. In the README file, we also provide a video showing the launch of the code and the GUI running on such offline data.

4 CONCLUSIONS

In this work, we presented a long-range mutual gaze detector implemented as a classifier and designed for HRI applications. We trained the classifier with diverse data and achieved good performances that did not degrade as the users' distance increased. Therefore, it can be considered useful to swiftly detect mutual gaze at long distances, outperforming the options available in the literature; also it can find application beyond the HRI context. The whole framework is available online and is built on robust and popular off-the-shelf perception algorithms. The implementation is realized within the ROS2 framework and, thus, can be easily deployed for HRI tasks.

ACKNOWLEDGMENTS

This work was supported by the European Union through the project SERMAS, and by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00247.

REFERENCES

[1] Accessed: 2023. Azure Kinect sensor SDK system requirements. https://learn.microsoft.com/en-us/azure/kinect-dk/system-requirements.

[2] Accessed: 2023. MediaPipe. https://developers.google.com/mediapipe/solutions.

[3] Accessed: 2023. ROS 2 Documentation. https://docs.ros.org/en/humble/.

[4] Accessed: 2023. scikit-learn Documentation - Random Forest Classifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[5] Gabriele Abbate, Alessandro Giusti, Viktor Schmuck, Oya Celiktutan, and Antonio Paolillo. 2024. Self-Supervised Prediction of the Intention to Interact with a Service Robot. *Robotics and Autonomous Systems* 171 (2024), 104568.

[6] Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction* 6, 1 (2017).

[7] Anna Belardinelli. 2023. Gaze-based intention estimation: principles, methodologies, and applications in HRI. arXiv:2302.04530 [cs].

[8] Dong-Chan Cho and Whoi-Yul Kim. 2013. Long-range gaze tracking system for large movements. *IEEE Transactions on Biomedical Engineering* 60, 12 (2013), 3432–3440.

[9] Eunji Chong, Elysha Clark-Whitney, Audrey Southerland, Elizabeth Stubbs, Chanel Miller, Eliana L Ajodan, Melanie R Silverman, Catherine Lord, Agata Rozga, Rebecca M Jones, et al. 2020. Detection of eye contact with deep neural networks is as accurate as human experts. *Nature communications* 11, 1 (2020), 6386.

[10] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. 2014. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the symposium on eye tracking research and applications*. 255–258.

[11] Norina Gasteiger, Mehdi Hellou, and Ho Seok Ahn. 2021. Factors for personalization and localization to optimize human–robot interaction: A literature review. *International Journal of Social Robotics* (2021), 1–13.

[12] Craig Hennessey and Jacob Fiset. 2012. Long Range Eye Tracking: Bringing Eye Tracking into the Living Room. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (Santa Barbara, California) *(ETRA '12)*. Association for Computing Machinery, New York, NY, USA, 249–252. https://doi.org/10.1145/2168556.2168608

[13] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. 2016. Eye Tracking for Everyone. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 2176–2184. https://doi.org/10.1109/CVPR.2016.239

[14] Wenyu Li, Qinglin Dong, Hao Jia, Shijie Zhao, Yongchen Wang, Li Xie, Qiang Pan, Feng Duan, and Tianming Liu. 2019. Training a camera to perform long-distance eye tracking by another eye-tracker. *IEEE Access* 7 (2019), 155313–155324.

[15] Maria Lombardi, Elisa Maiettini, Davide De Tommaso, Agnieszka Wykowska, and Lorenzo Natale. 2022. Toward an attentive robotic architecture: Learning-based mutual gaze estimation in Human–Robot Interaction. *Frontiers in Robotics and AI* 9 (2022), 770165.

[16] Nicolai Marquardt and Saul Greenberg. 2012. Informing the design of proxemic interactions. *IEEE Pervasive Computing* 11, 2 (2012), 14–23.

[17] Jordi Pages, Luca Marchionni, and Francesco Ferro. 2016. Tiago: the modular robot that adapts to different research needs. In *International Workshop on Robot Modularity at IEEE/RSJ Int. Conf. on Intelligent Robots and Systems 2016*.

[18] Amit Kumar Pandey and Rodolphe Gelin. 2018. A Mass-Produced Sociable Humanoid Robot: Pepper: The First Machine of Its Kind. *IEEE Robot. Autom. Mag.* (07 2018), 1–1.

[19] Jorge Rios-Martinez, Anne Spalanzani, and Christian Laugier. 2015. From proxemics theory to socially-aware navigation: A survey. *International Journal of Social Robotics* 7 (2015), 137–153.

[20] Shane Saunderson and Goldie Nejat. 2019. How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics* 11 (2019), 575–608.

[21] M. Zhang, Y. Liu, and F. LU. 2022. GazeOnce: Real-Time Multi-Person Gaze Estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 4187–4196. https://doi.org/10.1109/CVPR52688.2022.00416

[22] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4511–4520.

[23] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017. It's written all over your face: Full-face appearance-based gaze estimation. In *IEEE Conference on Computer Vision and Pattern Recognition workshops*. 51–60.