2024 IEEE International Conference on Robotics and Automation (ICRA)
May 13-17, 2024. Yokohama, Japan

2024 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-8457-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRA57147.2024.10610289

# Predicting the Intention to Interact with a Service Robot: the Role of Gaze Cues

Simone Arreghini, Gabriele Abbate, Alessandro Giusti, and Antonio Paolillo

***Abstract*— For a service robot, it is crucial to perceive as early as possible that an approaching person intends to interact: in this case, it can proactively enact friendly behaviors that lead to an improved user experience. We solve this perception task with a sequence-to-sequence classifier of a potential user intention to interact, which can be trained in a self-supervised way. Our main contribution is a study of the benefit of features representing the person's gaze in this context. Extensive experiments on a novel dataset show that the inclusion of gaze cues significantly improves the classifier performance (AUROC increases from $84.5\,\%$ to $91.2\,\%$); the distance at which an accurate classification can be achieved improves from $2.4\,\mathrm{m}$ to $3.2\,\mathrm{m}$. We also quantify the system's ability to adapt to new environments without external supervision. Qualitative experiments show practical applications with a waiter robot.**

Fig. 1. A service robot predicts if a nearby person intends to interact, so to proactively enact a friendly behavior.

## I. Introduction

The increasing use of service robots demands safe and efficient interactions with humans, which challenges scientists to build machines with social skills. Human-Robot Interaction (HRI) applications are demonstrating the potential of robots in providing relevant services, such as home assistance [1], reception [2], and hospitality [3]. However, much research effort still needs to be carried out to endow robots with advanced perception capabilities for predicting the behavior of nearby people. Consider, for instance, the scenario where a service robot has to assist the customers of a shop, or the guests of an hotel, who ask for information. In these circumstances, it is desirable that robots can understand human intentions on their own, well before the interaction actually starts. In this way, the assisting robot can enact friendly approaching behaviors so that even the hesitant, shy, or skeptical user can be well accommodated. To make this possible, it is crucial to provide the robot with the ability to predict the intention to interact of potential nearby users. In this very initial phase of the interaction, where the person is far away from the robot, possibly in a cluttered and noisy environment, nonverbal communication, such as body language or proxemics [4], [5], plays a crucial role.

This work uses a self-supervised approach that allows a service robot to predict, for any person in its Field of View (FoV), if that person intends to interact. The approach is based on a sequence-to-sequence classifier that, given a human's pose and gaze, continuously predicts whether it is going to eventually interact. It is self-supervised because, after the subject eventually moves away, the robot can, without external supervision: reconsider the entire sequence; label it with the corresponding ground truth, i.e. whether the subject actually interacted or not; add the labeled sequence to the classifier's training set; retrain its classifier.

In a previous work [6], we explored a self-supervised approach using features of the person's body motion. The **main contribution** of this work is to assess the improvement due to gaze cues. To this end, we develop a classifier that, given sensing data, predicts mutual gaze, i.e. when someone is actively looking at the robot camera. We record in 3 distinct locations a novel dataset including 84 positive and 105 negative sequences with a service robot. Experimental results show excellent prediction performance when a Recurrent Neural Network (RNN) uses as input both the subject's pose and gaze. Most notably, gaze information enables the classifier to achieve high accuracy ($95.2\%$) at an average subject distance of 3.2 m, compared to a much shorter distance (2.4 m) at which a classifier without gaze cues achieves the same accuracy. Additional experiments quantify the approach's ability to adapt to new environments in a self-supervised way and demonstrate human-friendly behaviors that a waiter robot can enact when using the proposed model.

The remainder of the paper is organized as follows: Sec. II discusses the related literature, whereas our approach is described in Sec. III. Section IV introduces the experimental setup used for the evaluation, whose results are presented in Sec. V. Finally, concluding remarks are reported in Sec. VI.

## II. Related work

Nonverbal communication is fundamental component in the context of HRI, for both humans and robots [7], [8].

This work was supported by the European Union through the project SERMAS, by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00247, and by the Swiss National Science Foundation (grant n. 213074).

All the authors are with Dalle Molle Institute for Artificial Intelligence (IDSIA), USI-SUPSI, Lugano, Switzerland name.surname@idsia.ch

Nonetheless, the extraction of useful information about the users' behavior is not an easy task in HRI, especially when dealing with nonverbal communication [9]. A body of work aims at estimating the human intention, e.g. in the context of navigation [10], collaborative tasks [11], [12], or for social behavior interpretation [13]–[15]. An important role, in accurately predicting the human intention to interact, is played by the information enclosed in the user gaze [16], [17]. The information about a person's gaze can be summarized into two types: gaze direction estimation and mutual gaze detection. Accurate enough gaze tracker methods can be found in the literature since a long time; such methods try to solve the problem of gaze vector regression. Some Deep Learning (DL) based trackers [18]–[20] were developed in the past, however, most of the time they fall short in tracking at high distances, which is a problem the context of HRI. Long-range methods have been proposed [21], [22] but they usually require cumbersome hardware that is difficult to mount on mobile robots. Recent developments [23] showed promising performances in solving this problem, however, user implementation is not yet available, with training and deployment requiring complex setup. A much simpler task is the problem of detecting mutual gaze, which is generally defined as mutual eye contact between individuals, or between a person and a robot camera sensor. It has been studied with great results both outside [24] and inside the robotics community [25]. However, these methods again fall short in tracking at high distances. Specifically to the HRI domain, many approaches base the intention recognition only on gaze cues [16] or combined with other features like body motion cues [26]–[30]. Our approach extends, adding mutual gaze cues, what has been proposed in [6] where we devised a self-supervised algorithm for detecting the intent to interact of a human with a robot leveraging only body motion cues. It is worth mentioning that in the DL literature, the term Self-Supervised Learning (SSL) denotes the practice of using pretext tasks [31]–[33] for learning representations from big unlabeled data. Instead, we refer to the meaning used in the robotics literature since the mid-2000s: self-generating supervision by leveraging data collected in previous experiments by the robot's own sensors, a paradigm successfully applied to several robotic applications, see e.g. [34]–[36].

## III. Model

### A. Problem formulation and solution approach

Consider a service robot stationary in a public space, awaiting a possible assistance request from a user. The robot is assumed to be equipped with exteroceptive sensing capabilities, e.g. provided by an RGB-D sensor. The representative frame of the robot is chosen at its camera sensor frame and is denoted with $\mathcal{F}_s$. The subject freely moves in the environment, i.e., they can randomly enter or exit the scene, and possibly interact with the robot. The information about the person's behavior is described by properly chosen body frames and facial landmarks. The body frames of interest are one located in the middle of the person's chest (denoted with $\mathcal{F}_c$) and on the person's head ($\mathcal{F}_h$). The 3D

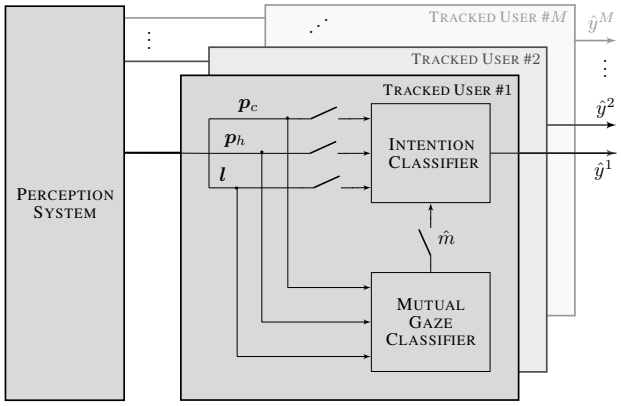


Fig. 2. System architecture.

poses of the frame $\mathcal{F}_c$ and $\mathcal{F}_h$ w.r.t $\mathcal{F}_s$ are denoted with $\boldsymbol{p}_c \in \mathbb{R}^9$ and $\boldsymbol{p}_h \in \mathbb{R}^9$, respectively[1]. We assume that such metric information, indicative of the proxemics of the subject, is measurable by the robot sensor. Furthermore, we also assume that the camera RGB images allow the detection of facial landmarks, which mainly consist of the projected locations, on the image plane, of specific points of interest on the person's face. Multiple facial landmarks are detected at once, each of which is denoted with the variable $\boldsymbol{\rho}_i \in \mathbb{R}^3$. Each landmark contains the 2D point coordinates on the image with an additional component corresponding to the depth of the landmark w.r.t. the face center of gravity. The facial landmarks, together with the body information of the subject's chest and head, are fed to a pre-trained *mutual gaze classifier* that outputs a score $\hat{m}$ representing the probability that the subject is looking at the robot. Finally, the subject's intention to interact is indicated with the boolean $y$.

The perception problem tackled in this work is the prediction of a potential user's intention to interact with the robot, by monitoring their body motion, facial landmarks, the mutual gaze, or a combination of them. The block diagram of the proposed solution is shown in Fig. 2: our system is a classifier that, given the information about a potential user, provides an estimate $\hat{y}$ of its probability of interaction with the robot. The approach relies on $(i)$ existing modules providing information on people motion, and $(ii)$ another specifically designed classifier, which computes the mutual gaze. Multiple subjects are handled, as different instances of the approach can be instantiated in parallel. In our analysis, we compare different combinations of input taken from the robot's lower-level perception pipeline, as well as different implementations of the classifier. In this section, we present the data structure and the feature sets required to train our intention to interact prediction model.

### B. Dataset

The training dataset $\mathcal{D}$ is organized into *sequences*, each one related to a subject who appears in the FoV of the sensor.

[1]Following best practices in machine learning, each Euler angle of the orientation representation is encoded as its $\sin$ and $\cos$ functions to avoid discontinuities around zero (in this case the pose size is 9).

Each whole sequence is given a boolean label, which reflects the occurrence of an interaction (true) or not (false). The sequences are composed of *frames*, i.e. sample data at each timestamp; frames belonging to the same sequence share the same label. More in detail, the dataset has this shape:

$$\mathcal{D} = \{\boldsymbol{f}_{i,j}, y_j\}_{i=1,j=1}^{N_j,S} \tag{1}$$

where $S$ is the number of sequences, and $N_j$ is the number of frames in $j$-th sequence. The features vector $\boldsymbol{f}_{i,j}$ contains the information about the person at the $i$-th frame of the $j$-th sequence; our study compares different choices of feature sets, as presented in Sec. III-C. The variable $y_j$ denotes the label for the $j$-th sequence. Note that, with the assumption that the robot is able to sense when an interaction occurs, any given past sequence can be labeled automatically using the robot's own experience, without any external supervision.

### C. Feature sets

In our study, we analyze how the choice of features (introduced in Sec. III-A) impacts the prediction of the intention to interact with the robot. The first and most simple set considers only the subject's chest pose:

$$\boldsymbol{f}_{\text{C}} = \boldsymbol{p}_c \in \mathbb{R}^9 \tag{2}$$

The second set contains also the pose of the subject's head and actually serves as a baseline in our comparison since it is very similar to what has been used in [6]:

$$\boldsymbol{f}_{\text{CH}} = \left(\boldsymbol{p}_c^\top, \boldsymbol{p}_h^\top\right)^\top \in \mathbb{R}^{18} \tag{3}$$

In the third set, we only consider the estimate of the mutual gaze information, as provided by the corresponding classifier:

$$\boldsymbol{f}_{\text{M}} = \hat{m} \in [0, 1]. \tag{4}$$

The fourth one combines spatial cues with mutual gaze:

$$\boldsymbol{f}_{\text{CHM}} = \left(\boldsymbol{p}_c^\top, \boldsymbol{p}_h^\top, \hat{m}\right)^\top \in \mathbb{R}^{19}. \tag{5}$$

The last one considers also the facial landmarks:

$$\boldsymbol{f}_{\text{FULL}} = \left(\boldsymbol{p}_c^\top, \boldsymbol{p}_h^\top, \hat{m}, \boldsymbol{l}^\top\right)^\top \in \mathbb{R}^{19+3n} \tag{6}$$

where $\boldsymbol{l} = (\boldsymbol{\rho}_1^\top, \ldots, \boldsymbol{\rho}_n^\top)^\top \in \mathbb{R}^{3n}$ is the vector containing $n$ detected facial landmarks.

## IV. Experimental setup

### A. Robot, sensing and perception

In our work, we use a small wheeled omnidirectional robot (DJI RoboMaster EP[2]). Furthermore, we use the Azure Kinect RGB-D sensor[3], streaming images at a resolution of $4096 \times 3072$ pixels, with a nominal horizontal and vertical FoV equal to $90°$ and $59°$, respectively. In our setup the camera sensor is placed in the proximity of the robot, at a height from the ground of around 1 m (as in Fig. 1), which is the typical height of robots designed for HRI, see e.g. Tiago by PAL Robotics [37] or Pepper by Aldebaran [38]. The sensor's SDK enables body frame tracking of multiple subjects simultaneously, directly providing us with the measurement of $\boldsymbol{p}_c$ and $\boldsymbol{p}_h$. Facial landmarks for each subject are extracted with the MediaPipe[4] library. Its *face mesh* module returns the 2D location on the image plane of salient face points, and the corresponding predicted depth relative to the center of mass of the face. The facial landmarks originally expressed using image-normalized coordinates, are first centered around the landmarks' center of gravity and then normalized such that the mean landmark vector has unitary size. Mediapipe provides $478$ landmarks by default, from which we select $n = 39$ points representative of major structures (mouth, eye corners, irises, nose, face contours), whose normalized coordinates are concatenated in vector $\boldsymbol{l}$.

### B. Dataset collection

The dataset was collected by deploying the system and recording multiple sequences. For each sequence, a person enters the robot FoV and either interacts with it before leaving (positive sequence) or simply transits near the robot without interacting (negative sequence). An interaction consists of taking a chocolate treat that RoboMaster holds. The entire dataset contains a total of $4946$ frames and $189$ sequences ($84$ positive, $189$ negative). Data was split between $3$ different environments, using different subjects.

*Lab*: $92$ sequences ($33$ positive, $59$ negative) recorded in a mostly empty laboratory, size $9 \times 9$ m, with a single subject visible at a time. Only for this sequence, the ground truth of the subject's gaze is also collected and used to test the mutual gaze classifier.

*Office*: $42$ sequences ($15$ positive, $27$ negative) collected in a corridor of an office building, including multiple subjects simultaneously present in many frames.

*Kids*: $55$ sequences ($36$ positive, $19$ negative), collected in a break area of an office building. Subjects are $10$ to $12$ years old kids; each kid was instructed to either traverse the break area (without providing specific instructions regarding the robot) or stop and grab the chocolate treat from the robot.

All the participants or their tutors (for the kids), signed a consent form; all data is kept private; the experiments were approved by the local institution's ethical committee.

### C. Classifier architecture

We expect that the motion dynamics of body frames, facial landmarks, and the temporal evolution of the mutual gaze are important cues to predict whether a given person intends to interact with the robot. To capture these dynamics, we use a recurrent Long Short-Term Memory (LSTM) [39] neural network as a stateful sequence-to-sequence classifier. We use the implementation available in the PyTorch[5] library. All the LSTM models have 2 layers, whereas the hidden state dimension varies to accommodate the different sizes of the feature sets (from 10 for $\boldsymbol{f}_{\text{M}}$ to 65 for $\boldsymbol{f}_{\text{FULL}}$). To offer a more complete analysis, we compare the LSTM performance against a simpler, stateless model, i.e. a Random Forest (RF) classifier implemented using the scikit-learn package.

[2]https://www.dji.com/ch/robomaster-ep-core

[3]https://learn.microsoft.com/en-us/azure/kinect-dk/system-requirement

[4]https://developers.google.com/mediapipe/solutions

[5]https://pytorch.org/

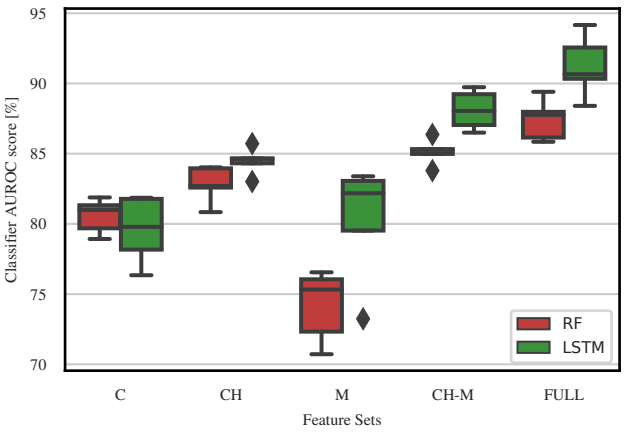


Fig. 3. AUROC of RF and LSTM classifiers with the different feature sets.

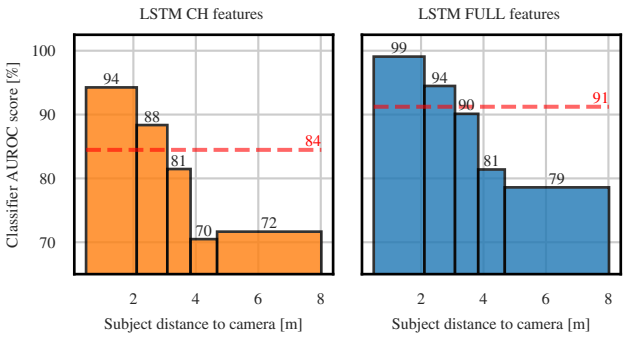


Fig. 4. AUROC for the LSTM using $\boldsymbol{f}_{\text{CH}}$ (left) and $\boldsymbol{f}_{\text{FULL}}$ (right) for different human-robot distance quantiles.

All the models are evaluated using a 5-fold stratified cross-validation strategy; folds are computed by splitting the set of sequences in training and testing sets, such that all frames for a given sequence stay in the same set.

### D. Mutual gaze classifier

In our setup, some feature sets include mutual gaze information through the variable $\hat{m}$ (see Sec. III-C). Off-the-shelf algorithms for mutual gaze estimation [25] are not designed for distances larger than 1 m; therefore, we trained an ad-hoc mutual gaze classifier tailored to our sensing setup. To this end, we collected a dataset $\mathcal{D}_{\text{gaze}}$ in which subjects stood in front of the robot in predefined positions, arranged in such a way to cover the sensor's FoV. For each position, the subject moves their head and torso while alternating periods of looking at the robot and periods of looking elsewhere; ground truth is collected by the subject himself, by keeping a button pressed when and only when the gaze is on the robot. More in detail, with reference to Fig. 2, our mutual gaze module takes as input $\boldsymbol{p}_c$, $\boldsymbol{p}_h$ and $\boldsymbol{l}$. Data thus collected produced a training set of 12849 samples. Since we expect that face landmark dynamics are not useful to predict mutual gaze, we rely on a simple stateless RF classifier, which can provide good robustness while being very lightweight. More details about our mutual gaze classifier can be found in [40].

## V. Experimental Results

### A. Mutual gaze estimation performance

When evaluated using 5-fold cross-validation on $\mathcal{D}_{\text{gaze}}$, the mutual gaze classifier yields an Area Under Receiver Operating Curve (AUROC) value of 91.9%, an accuracy of 83.3%, and an average precision of 86.3%. We further verify the performance of the classifier in our relevant environments by evaluating it on the frames of the interaction dataset $\mathcal{D}$ (Sec. IV-B) that were collected in the *Lab* environment, for which we have ground truth for subject gaze. On this testing set, the mutual gaze classifier yields: 90.4% AUROC, 82.4% accuracy, 82.2% average precision.

### B. Intention to interact: frame-level performance

We first evaluate the algorithm performances frame-by-frame using the AUROC metric, comparing the two different model architectures (RF and LSTM) introduced in Sec. IV-C. The plot in Fig. 3 shows that the stateful LSTM classifiers consistently outperform the simpler stateless RF counterparts for every input feature set. In the following, we focus our comparison on two models: $(i)$ the LSTM model using $\boldsymbol{f}_{\text{CH}}$, which we refer to as the *baseline*, as it is most representative of the model introduced in [6]; and $(ii)$ the LSTM model using $\boldsymbol{f}_{\text{FULL}}$ that we refer to as *our model*. The latter differs from the former since it includes gaze and face landmark cues. The plot in Fig. 3 shows that adding gaze information significantly improves classifier performance, increasing AUROC from 84.5% for the baseline to 91.2% for our model.

Fig. 4 reports an experiment in which we split the testing data into 5 distance quintiles, and evaluate the classifier separately on each. All reported AUROC values are significantly greater than 0.5, which indicates that, even among subjects that are at approximately the same distance from the robot, the classifier is effective at differentiating those who are likely to interact and those who are not; i.e., even though subject distance from the robot is a powerful feature, it is not the only aspect that is considered by the models. The figure further shows that the improvement of the performance introduced by the contribution of the gaze is uniform across the whole range of subject-robot distances. In the bin of the closest distance (the one from 0.48 to 2.10 m), the gain in performance due to the additional gaze information and facial landmarks is about 5%. This improvement increases to 9% for the intermediate bin (containing distances from 3.08 to 3.83 m), and to 7% for the farthest distances (above 4.68 m).

Fig. 5 shows the statistics from positive and negative sequences for the whole dataset. In the left and center plots, all sequences are temporally aligned in such a way that $t = 0$ denotes the time of interaction, in the case of positive sequences, and the time in which the subject is at the closest distance from the robot, in case of negative sequences. We observe from the left plot that negative sequences reach, on average, a distance from the robot of 1.6 m before moving further. The center plot shows that at

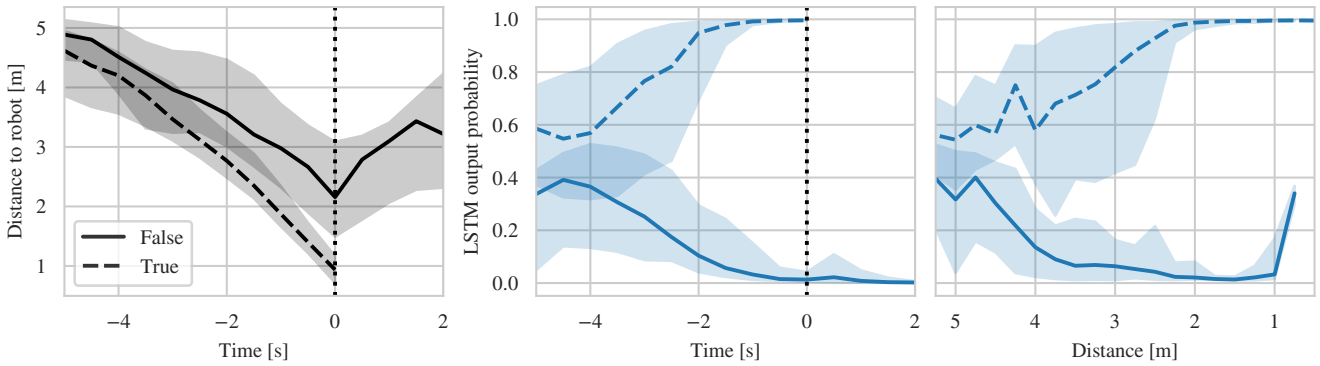


Fig. 5. Median distance to the robot (left) and median predicted probability of interaction (center) as a function of time. Time $t = 0$ is defined for each sequence as the moment when the subject either interacts, for positive sequences (dashed line), or the moment in which the subject is closest to the robot, for negative sequences (continuous line). The rightmost plot reports the predicted probability of interaction as a function of distance to the camera, ignoring negative samples with $t > 0$. Shaded areas represent the interquartile range.

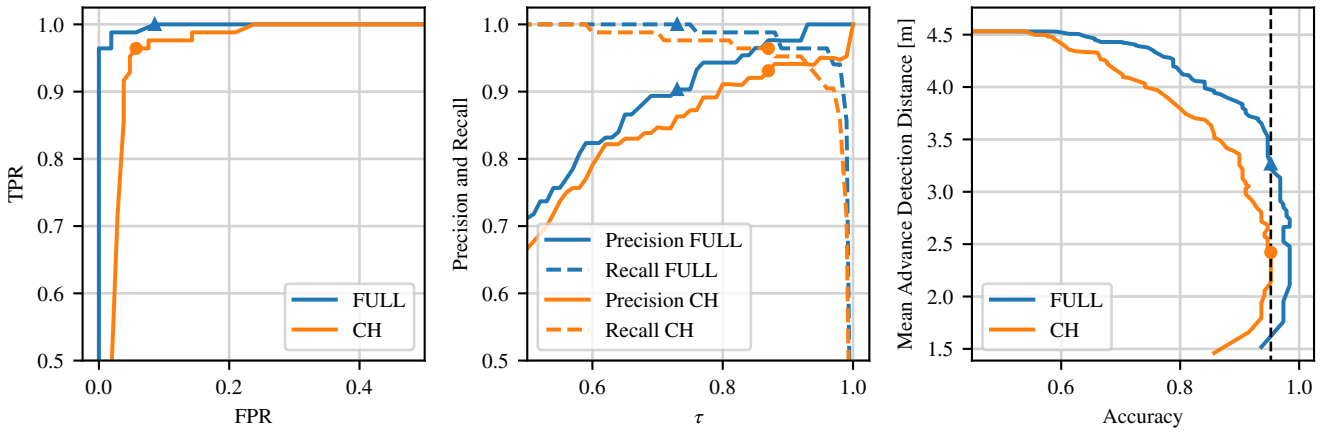


Fig. 6. Sequence-level performance metrics for the LSTM approach with (blue) and without (orange) gaze and face landmark features. Left: ROC curve. Center: Precision and Recall as a function of threshold $\tau$. Right: Mean Advance Detection Distance (vertical axis) vs. achieved accuracy (horizontal axis) for different values of $\tau$. The orange circles denote a threshold value of $\tau = 0.87$ for the baseline model set to achieve maximum accuracy. Conversely, the blue triangles denote a threshold value of $\tau = 0.73$ needed by our model to display the same level of accuracy.

$t = 0$ (vertical dotted line), the model yields very sharp predictions. The distribution of the output probabilities for positive and negative sequences start diverging at $t = -4$ s, and clearly separate at $t = -3$ s. The rightmost plot reports the same data but with distance to the robot on the horizontal axis. Negative sequences that reach distances below 1 m are few, so the rightmost data is noisy.

### C. Intention to interact: sequence-level performance

We now report the performance of our models when evaluating them at the level of entire sequences. For each sequence, we simulate that the model is applied to each frame, and when exceeding a threshold $\tau$, the robot takes an irreversible decision to enact a given behavior (e.g. facing, approaching or greeting the user). Negative sequences in which the output probability never exceeds $\tau$ are true negatives (i.e., the robot correctly ignored a non-interacting subject); positive sequences in which the output probability exceeds $\tau$ for at least a single frame are true positives; only for true positives, from the earliest frame whose classifier output exceeds $\tau$, we compute the *advance detection time* (i.e. the amount of anticipation) and *advance detection distance* (i.e. the distance of the user when the robot reacted). False negatives denote sequences for which the robot did not react to an interacting user; false positives denote sequences in which the robot incorrectly reacted to a non-interacting subject. Given these definitions, we compute sequence-level metrics: False Positive Rate (FPR), True Positive Rate (TPR), precision, recall and accuracy. Figure 6 reports these metrics for both the baseline model (which uses $\boldsymbol{f}_{\text{CH}}$), in orange, and our model (that relies on the more complete information contained in $\boldsymbol{f}_{\text{FULL}}$), in blue. We observe that the latter outperforms the former in all metrics, regardless of the threshold value $\tau$. In particular, the center plot highlights that high threshold values are key to obtaining very high precision and recall performance, with our model consistently outperforming the baseline in both metrics. Nevertheless, the plot does not show the complete picture: high threshold values yield high performance because, in

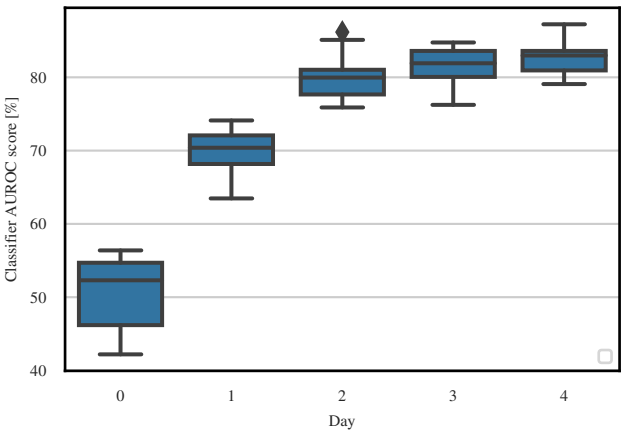


Fig. 7. AUROC during self-supervised adaptation to a new environment.

this case, the model does not commit to a decision until very late in the sequence, when most positive sequences yield very high probabilities; this behavior is not useful in practice. The right plot in Fig. 6 studies the trade-off between sequence classification accuracy, on the horizontal axis, and mean advance detection distance, on the vertical axis, controlled by $\tau$. Low values of $\tau$ yield early, distant but inaccurate detections (top left). Increasing $\tau$ decreases the mean Advance Detection Distance but improves accuracy up to a maximum value; further increases of $\tau$ lead to a marked increase in false negatives, and negatively impact both Advance Detection Distance and Accuracy, as can be also seen from the drops in the recall value. The orange dots denote a threshold ($\tau = 0.87$) yielding maximum accuracy ($95.2\%$) for the baseline classifier, and the blue triangles denote the threshold ($\tau = 0.73$) needed to get to the same accuracy for our model. At this threshold, our model yields a significantly better advance detection distance (3.27 m) w.r.t. the baseline (2.42 m): an improvement of $0.85$ m.

### D. Self-supervised adaptation to new environments

To test the self-supervised ability of the proposed approach, we consider the situation of a robot with a model that is pre-trained in two environments (*Lab* and *Office*) and then deployed in a new one (*Kids*), which has a different layout and where subjects behave differently. We use $1/5$ of the sequences in the deployment environment as our fixed testing set. The remaining sequences of the deployment environment are split into four groups (*Day 1* to *Day 4*) which are assumed to be collected by the robot in a self-supervised way during its first few days of deployment. Figure 7 reports the frame-level AUROC computed on the fixed testing set: day 0 represents the model trained only on the training data, which has not yet adapted to the new environment; the subsequent entries represent the classifier trained on data collected from the training environments plus the deployment environment up to day $n \in \{1, 2, 3, 4\}$. Results are reported for $10$ replicas, obtained by different random splits of the data in the deployment environment.

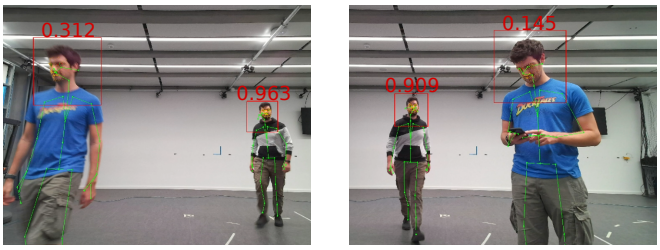


Fig. 8. Two snapshots of the robotic experiments used to evaluate our framework in real-life scenarios.

We observe that performance at day 0 is very poor (AUROC $\approx 0.5$), because the behavior of kids and the room layout is very different from the training data; performance increases as the robot accumulates additional experience.

### E. Evaluation of the intention to interact in a robotic task

Our approach has been evaluated in real experiments. Two indicative snapshots of this experiment are shown in Fig. 8. In the right image, a person is standing very close to the robot (the one with a predicted probability of $0.145$) but is not paying attention as he is looking at his phone. Our predictor recognizes this situation, by providing a low predicted probability of interaction; the other subject in the same snapshot, instead, is walking and looking toward the robot and his predicted probability is high ($0.909$). Similar conclusions can be drawn by looking at the left snapshot. Here, instead of standing, the person not interested in interacting (the man in blue) is walking past the robot and is given a low probability ($0.312$), whereas the other person shows interest in interacting (the man in the background), with a predicted interaction probability of $0.963$.

This experiment and other results can be seen in the video accompanying this paper.

## VI. Conclusions

We have shown that gaze cues improve the prediction of a potential user's intention to interact with a service robot; in line with the literature, the presented system outperforms our previous work that only uses body motion cues. The user's intention to interact prediction is an important perception task to enact proactive behaviors that yield effective and satisfactory interaction experiences for users. The implementation consists of a classifier that takes as input the features of a person's body and gaze and outputs its probability of interaction. We have compared two different architectures of classifiers (random forest and long-short term memory) with different feature sets. Additional material is available at `https://github.com/idsia-robotics/intention-to-interact-detector-gaze`.

Future work will be devoted to testing our perception module in more challenging real-world social scenarios, such as the hall of a hotel or the entrance of a shopping mall, and experimenting with robot reaction behaviors. A user study will be carried out to evaluate the users' satisfaction level with service robots in real social contexts.

## REFERENCES


[1] G. A. Zachiotis, G. Andrikopoulos, R. Gornez, K. Nakamura, and G. Nikolakopoulos, "A survey on the application trends of home service robotics," in *IEEE Int. Conf. on Robotics and Biomimetics*, 2018, pp. 1999–2006.

[2] M. K. Lee, S. Kiesler, and J. Forlizzi, "Receptionist or information kiosk: how do people talk with a robot?" in *ACM Conference on Computer Supported Cooperative work*, 2010, pp. 31–40.

[3] A. Tuomi, I. P. Tussyadiah, and J. Stienmetz, "Applications and implications of service robots in hospitality," *Cornell Hospitality Quarterly*, vol. 62, no. 2, pp. 232–247, 2021.

[4] J. Urakami and K. Seaborn, "Nonverbal cues in human–robot interaction: A communication studies perspective," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 2, pp. 1–21, 2023.

[5] L. Takayama and C. Pantofaru, "Influences on proxemic behaviors in human-robot interaction," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2009, pp. 5495–5502.

[6] G. Abbate, A. Giusti, V. Schmuck, O. Celiktutan, and A. Paolillo, "Self-supervised prediction of the intention to interact with a service robot," *Robotics and Autonomous Systems*, vol. 171, p. 104568, 2024.

[7] N. Gasteiger, M. Hellou, and H. S. Ahn, "Factors for personalization and localization to optimize human–robot interaction: A literature review," *International Journal of Social Robotics*, pp. 1–13, 2021.

[8] S. Saunderson and G. Nejat, "How robots influence humans: A survey of nonverbal communication in social human–robot interaction," *International Journal of Social Robotics*, vol. 11, pp. 575–608, 2019.

[9] J. Rios-Martinez, A. Spalanzani, and C. Laugier, "From proxemics theory to socially-aware navigation: A survey," *International Journal of Social Robotics*, vol. 7, pp. 137–153, 2015.

[10] P. Agand, M. Taherahmadi, A. Lim, and M. Chen, "Human Navigational Intent Inference with Probabilistic and Optimal Approaches," in *IEEE Int. Conf. on Robotics and Automation*, 2022, pp. 8562–8568.

[11] A. Belardinelli, A. R. Kondapally, D. Ruiken, D. Tanneberg, and T. Watabe, "Intention estimation from gaze and motion features for human-robot shared-control object manipulation," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2022, pp. 9806–9813.

[12] S. Vinanzi, C. Goerick, and A. Cangelosi, "Mindreading for Robots: Predicting Intentions via Dynamical Clustering of Human Postures," in *Int. Conf. on Development and Learning and Epigenetic Robotics*, 2019, pp. 272–277.

[13] A. Zaraki, M. Giuliani, M. B. Dehkordi, D. Mazzei, A. D'ursi, and D. De Rossi, "An RGB-D based social behavior interpretation system for a humanoid social robot," in *RSI/ISM International Conference on Robotics and Mechatronics*, 2014, pp. 185–190.

[14] A. Gaschler, S. Jentzsch, M. Giuliani, K. Huth, J. de Ruiter, and A. Knoll, "Social behavior recognition using body posture and head pose for human-robot interaction," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2012, pp. 2128–2133.

[15] F. Del Duchetto, P. Baxter, and M. Hanheide, "Are you still with me? continuous engagement assessment from a robot's point of view," *Frontiers in Rob. and AI*, vol. 7, p. 116, 2020.

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

[17] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: a review," *Journal of Human-Robot Interaction*, vol. 6, no. 1, May 2017.

[18] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4511–4520.

[19] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2176–2184.

[20] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *IEEE Conference on Computer Vision and Pattern Recognition workshops*, 2017, pp. 51–60.

[21] C. Hennessey and J. Fiset, "Long range eye tracking: bringing eye tracking into the living room," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, 2012, pp. 249–252.

[22] D.-C. Cho and W.-Y. Kim, "Long-range gaze tracking system for large movements," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 12, pp. 3432–3440, 2013.

[23] M. Zhang, Y. Liu, and F. Lu, "Gazeonce: Real-time multi-person gaze estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4197–4206.

[24] E. Chong, E. Clark-Whitney, A. Southerland, E. Stubbs, C. Miller, E. L. Ajodan, M. R. Silverman, C. Lord, A. Rozga, R. M. Jones *et al.*, "Detection of eye contact with deep neural networks is as accurate as human experts," *Nature communications*, vol. 11, no. 1, p. 6386, 2020.

[25] M. Lombardi, E. Maiettini, D. De Tommaso, A. Wykowska, and L. Natale, "Toward an attentive robotic architecture: Learning-based mutual gaze estimation in human–robot interaction," *Frontiers in Robotics and AI*, vol. 9, p. 770165, 2022.

[26] M. Brenner, H. Brock, A. Stiegler, and R. Gomez, "Developing an engagement-aware system for the detection of unfocused interaction," in *Int. Symp. on Robot and Human Interactive Communication*, 2021, pp. 798–805.

[27] D. Vaufreydaz, W. Johal, and C. Combe, "Starting engagement detection towards a companion robot using multimodal features," *Robot. Auton. Syst.*, vol. 75, pp. 4–16, 2016.

[28] Y. Kato, T. Kanda, and H. Ishiguro, "May I help you? - Design of human-like polite approaching behavior-," in *ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2015, pp. 35–42.

[29] J. Bi, F.-c. Hu, Y.-j. Wang, M.-n. Luo, and M. He, "A method based on interpretable machine learning for recognizing the intensity of human engagement intention," *Scientific Reports*, vol. 13, no. 1, p. 2537, 2023.

[30] A. Gaschler, S. Jentzsch, M. Giuliani, K. Huth, J. de Ruiter, and A. Knoll, "Social behavior recognition using body posture and head pose for human-robot interaction," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2012, pp. 2128–2133.

[31] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[32] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *IEEE International Conference on Computer Vision*, 2017, pp. 2051–2060.

[33] M. Nava, A. Paolillo, J. Guzzi, L. M. Gambardella, and A. Giusti, "Learning visual localization of a quadrotor using its noise as self-supervision," *IEEE Robot. and Autom. Lett.*, vol. 7, no. 2, pp. 2218–2225, 2022.

[34] ——, "Uncertainty-aware self-supervised learning of spatial perception tasks," *IEEE Robot. and Autom. Lett.*, vol. 6, no. 4, pp. 6693–6700, 2021.

[35] A. Lookingbill, J. Rogers, D. Lieb, J. Curry, and S. Thrun, "Reverse optical flow for self-supervised adaptive autonomous robot navigation," *International Journal of Computer Vision*, vol. 74, pp. 287–302, 2006.

[36] R. Hadsell, P. Sermanet, J. Ben, A. Erkan, M. Scoffier, K. Kavukcuoglu, U. Muller, and Y. LeCun, "Learning long-range vision for autonomous off-road driving," *J. Field Robot.*, vol. 26, no. 2, pp. 120–144, 2009.

[37] J. Pages, L. Marchionni, and F. Ferro, "Tiago: the modular robot that adapts to different research needs," 2016.

[38] A. K. Pandey and R. Gelin, "A mass-produced sociable humanoid robot: Pepper: The first machine of its kind," *IEEE Robot. Autom. Mag.*, pp. 1–1, 07 2018.

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[40] S. Arreghini, G. Abbate, A. Giusti, and A. Paolillo, "A long-range mutual gaze detector for HRI," in *ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2024, pp. –.