# Self-Supervised Detection of Human Intention to Interact with Robots: Overview, Results and Perspectives

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

***Abstract*— Modern robots must have advanced perception skills to interact with interested people successfully. In particular, it is important to predict human behavior to provide the relevant service in an accommodating way. To this end, we have built a perception module using self-supervised techniques to detect the humans' intention to interact with a service robot. We provide an overview of our current work, presenting the achieved results and our strategy for future developments.**

## I. Introduction

The growing use of service robots in modern society necessitates safe and efficient interactions with humans, posing challenges to developing machines with social skills. Human-Robot Interaction (HRI) applications are showcasing the robots' potential to deliver relevant services, e.g. domestic [1] or hospitality assistance [2]. Nonetheless, significant research effort is still required to equip robots with advanced perception skills for predicting the behavior of nearby users.

For instance, in a scenario where a service robot is required to assist customers in a shop or guests in a hotel, robots should understand human intentions autonomously. A critical aspect regards the *prediction* of the users' intention to interact with a nearby robot. It is indeed important to detect such intentions well before the interaction begins to ensure a smooth and satisfactory interaction. In the considered scenarios, individuals may be far from the robot and act in cluttered and noisy environments. Therefore, nonverbal communication, such as body language or proxemics [3], [4], has a crucial role. However, extracting useful information about users' behavior is not straightforward in HRI, particularly when dealing with nonverbal communication [5]. A considerable body of work aims to estimate human intention, e.g., in navigation contexts [6], collaborative tasks [7], [8], or for interpreting social behaviors [9]–[11]. In this regard, the information enclosed in the user gaze has utmost importance for predicting human intentions [12], [13]. Several accurate gaze trackers are available in the literature but they often struggle to track at significant distances, presenting a challenge in the context of HRI. Although long-range methods have been proposed, they typically require cumbersome hardware unsuitable for mobile robots [14], [15]. A simpler task is the detection of mutual gaze, which is generally defined as mutual eye contact between individuals or between a person and a robot's camera sensor. This task has been extensively studied [16], also in the robotics community [17]. However, the performance of these methods decreases at far distances, which is a requirement for the considered HRI applications. Approaches tackle the users' intention detection relying solely on gaze cues [12] or combining with body motion information [18]–[22].

Fig. 1. Service robots are more and more expected to interact with people in everyday life tasks. In these contexts, robots must have advanced perception skills to predict the intentions of interested users.

Following these lines of work, we developed our self-supervised classifier to detect the users' intention to interact with a robot. Our approach is self-supervised as it can autonomously label the sequences without external supervision (i.e., understand whether the subject interacted), add them to the classifier training set, and retrain itself when inactive, progressively adapting to the new domain. The first version of our classifier only uses body motion cues [23]. To leverage users' gaze information at long distances from the robot, we have developed our long-range mutual gaze classifier [24]. Its output, paired with body motion data and facial features, is used in our latest work [25] to enhance the performance of a more sophisticated interaction intention detector solution.

In this abstract, we give an overview of our human intention perception module, present our vision of AI-enhanced HRI, and draw lines for our future activities.

## II. Self-supervised human intention detection

Our research considers the practical application of a service robot in public spaces, awaiting to assist potential users in need. These situations are difficult to model and require strong adaptive capabilities; machine learning, and AI in general, can provide powerful tools to augment the perception capabilities of robots in HRI contexts. To this end, we assume that a service robot can enhance the perceived user interaction quality by predicting nearby users' intentions

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.

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

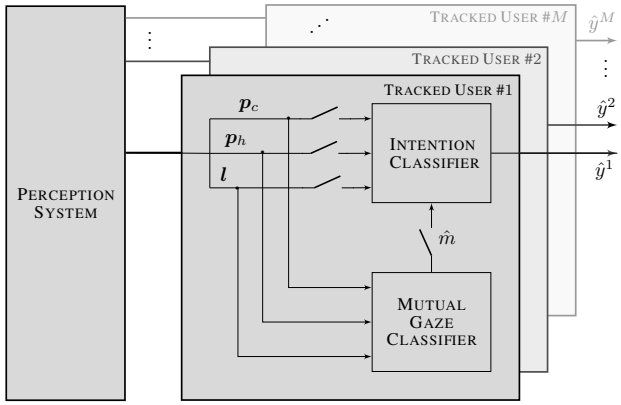


Fig. 2. Intention to interact detection architecture.

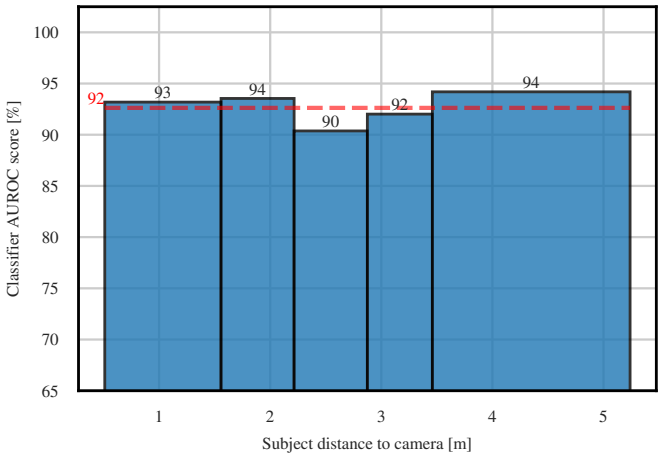


Fig. 3. Performance of the mutual gaze classifier (vertical axis) tested at 5 different distance ranges (horizontal axis).

to engage and adjusting its behavior. A block diagram representing the proposed intention to interact detection pipeline is reported in Fig. 2. The system leverages a self-supervised learning approach to forecast the likelihood of a human user engaging with a robot before any interaction unfolds. Self-supervision is achieved through the robot's autonomous sequence labeling capabilities, allowing it to gather data from its experience and progressively train itself with new information, continuously adapting to unseen environments. The user body information (i.e., the pose of the chest and head, respectively denoted with $\boldsymbol{p}_c$ and $\boldsymbol{p}_h$ in Fig. 2) is extracted using an Azure Kinect Developer Kit and more specifically the body tracking solution in its software suite. Facial landmarks (the quantity $\boldsymbol{l}$ in the block diagram of Fig. 2) are detected using RGB images via the MediaPipe python library by Google. Reliable facial feature detection even at long distances is possible thanks to the high-definition color images streamed by the Azure Kinect sensor.

Given the importance of gaze cues in the prediction process and the lack of available long-range solutions, we designed a custom mutual gaze classifier for long-range HRI applications [24]. This submodule takes as input the information about the users' body motion and facial landmarks and computes the mutual gaze $\boldsymbol{m}$, see Fig. 2. In this context, mutual gaze is intended to be the probability that someone is looking directly into the robot onboard camera. The classifier architecture is a stateless Random Forest (RF), chosen for its robustness and lightweight nature. During the design of the mutual gaze classifier, we tested different combinations of features, for the detail see [24]. Eventually, we found out that the best model is the one using both chest and head features and a subset of 21 facial landmarks. The analysis reported in Fig. 3 shows consistent Area Under Receiver Operating Curve (AUROC) (used as an index of performance of the classifiers in our work) above $90\%$ across five distance bins spacing from less than $1\,\mathrm{m}$ to more than $5\,\mathrm{m}$. The mutual gaze classifier is trained offline with an ad-hoc dataset and has to be considered a static component of the intention to interact detection pipeline. It can be easily enabled or disabled in our user's intention detection pipeline.

The first version of our work [23] aimed to test the performance of the intention classifier without the mutual gaze information. We compare different classifier architectures and explore diverse input feature sets derived only from user body motion. The effectiveness of the approach has been tested in the following scenarios: an office break area with 3442 natural sequences (both interactive and non-interactive), where a coffee machine stands in for the service robot, representing a real-world and challenging environment; the other two scenarios involve researchers interacting with service robots (with 200 and 72 sequences, respectively). Our approach can learn without external supervision and accurately predict a user's intention to interact more than 3 s before the interaction starts. Snapshots of a video showing the application of the detection pipeline are shown in Fig. 4.

In the most recent version of our architecture, the body joint tracking data, facial features, and gaze cues are finally fused and fed to the intention to interact classifier which integrates the different data sources [25]. We examine the performance with different input features: a combination of users' body joint poses (chest, denoted with 'C', and/or head, 'H'), mutual gaze information ('M'), and facial features. The input feature set containing all the sources is referred to as 'FULL'. Furthermore, two different classifiers have been used: a stateful sequence-to-sequence classifier, implemented as a recurrent Long Short-Term Memory (LSTM) neural network, and a stateless model, represented by a RF. Evaluation of all model variations is conducted using a 5-fold stratified cross-validation strategy. This strategy ensures that folds are computed by splitting the set of sequences into training and testing sets while ensuring that all frames for a given sequence remain in the same set. Experimental results, reported in Fig. 5, demonstrate the consistent superiority of LSTM classifiers over RF counterparts across all input feature sets. This comparison shows that adding gaze information significantly enhances classifier performance, increasing AUROC from $84.5\%$ for the LSTM CH (baseline) to $91.2\%$ for

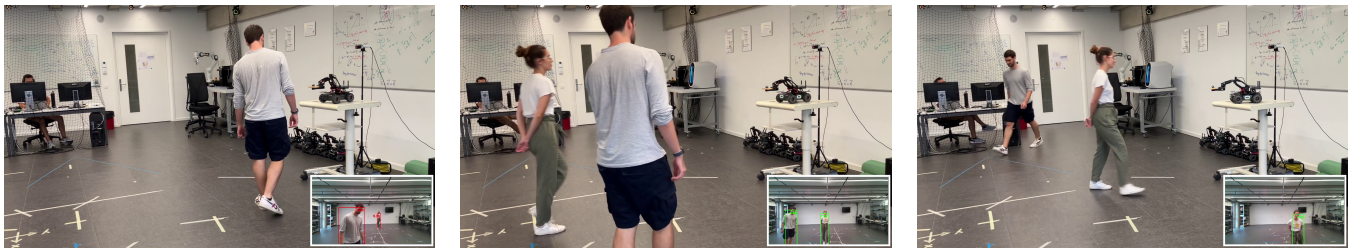

Fig. 4. First version of our detection pipeline: using body motion cues, our system can predict if a user is going to interact (shown with green markers on the sensor view) or not (in red) with the robot. Multiple users are handled with the same approach.

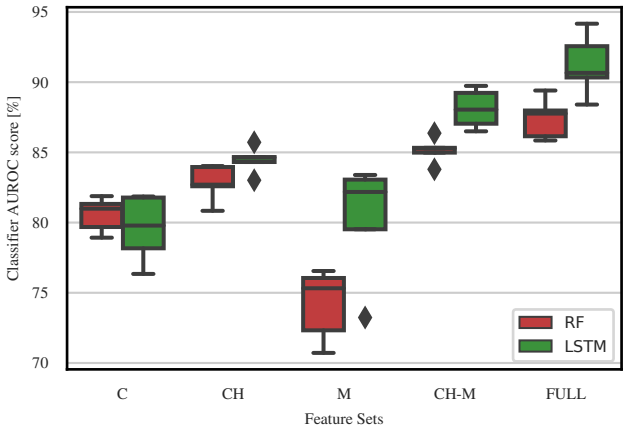


Fig. 5. AUROC of the intention classifiers with the different feature sets.

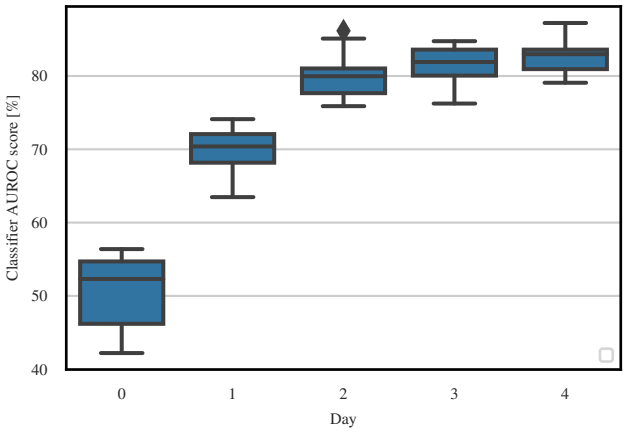


Fig. 6. AUROC during self-supervised adaptation to a new environment.

the LSTM FULL model. Finally, to test the self-supervised ability of the whole system, we consider the situation of a robot trained in previous environments and then deployed in a new one in which human 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. The evolution of the AUROC metric over the days, computed on the testing set, can be seen in Fig. 6. On Day 0 the model is trained only on the training data and, thus, 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 $n \in \{1, 2, 3, 4\}$. The system can swiftly adapt to unseen scenarios reaching a plateau in performance on Day 4. Figure 7 shows three snapshots of an experiment where users' intention to interact with a robot is detected by leveraging also gaze information.

## III. Future Directions

We have built a self-supervised pipeline to detect the human intention to interact with robots. Our analysis highlights the effectiveness of incorporating gaze and face landmark cues in predicting user interaction with the robot and the importance of considering multiple information cues in achieving accurate and early predictions of users' intentions.

In the future, we will study the impact of our algorithm on people's feelings during real-life interactions. Such experiments would investigate the influence on people's perceptions of the robot's actions and when they should be triggered. However, placing autonomous robots in the wild, freely interacting with people, poses many challenges. The most complicated regards privacy issues. The ability to adapt to a novel environment is based on the assumption of gathering data, labeling them, and subsequently retraining the classifier. Furthermore, offline processing of user data is crucial for meaningful in-depth analysis of users' attitudes and behaviors around robotic agents. However, saving personal data without explicit user consent (difficult to gather in large, uncontrolled user studies) is a delicate matter and still represents an open challenge. A quick solution would be to use less detailed, anonymized user data. However, this solution is a compromise paid at the cost of reduced algorithm performances. Another way could be directly asking for explicit consent through the robot during the interaction. However, in our experience, asking for consent during or even before the interaction usually results in a less natural experience, and could create a bias in the collected data. We believe that research efforts should tackle these open challenges, whose solution could greatly speed up the large deployment of social robots in everyday scenarios.

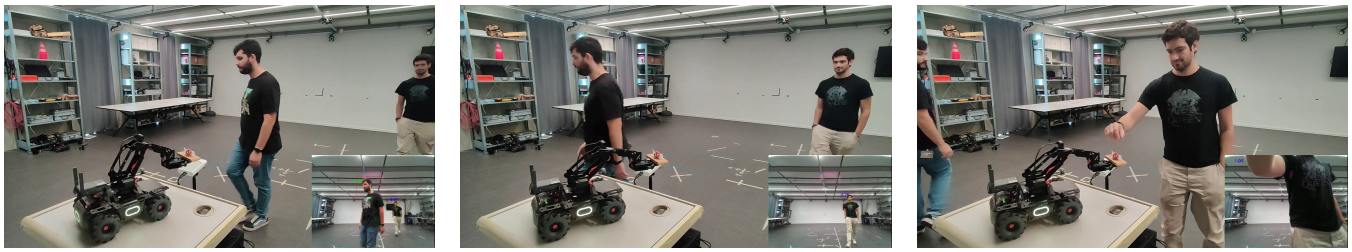

Fig. 7. The version of our detection pipeline that uses also mutual gaze information: our system can discriminate between users interested in interacting with the robot (and looking at it) from others who are just passing by very closely.

## 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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. –.

[25] ——, "Predicting the intention to interact with a service robot: the role of gaze cues," in *IEEE Int. Conf. on Robotics and Automation*, 2024, pp. –.