Toward the Detection of the Human Intention to Interact with a Service Robot

Gabriele Abbate[1], Alessandro Giusti[1], Viktor Schmuck[2], Oya Celiktutan[2] and Antonio Paolillo[1]

***Abstract*— In this extended abstract, we propose a classifier to enable service robots to proactively detect the "intention to interact" of human users before the interaction actually begins. To this end, we use information about the user's motion, such as their planar pose and linear velocity, that can be easily detected by available state-of-the-art sensors. We report preliminary experiments of our detection module, validated using a dataset comprising 3442 sequences collected in an everyday-life scenario. The analysis carried out on this dataset opens interesting questions and challenges, that pave the road of our development towards novel scenarios to investigate.**

I. Introduction

Social robots are very useful to provide services like reception [1]; hospitality [2] or home assistance [3]; navigation guidance [4]; personal care [5]; object delivery [6]. In these contexts, the understanding of the "human intention to interact" is crucial to make services proactive and friendly, and increasing their social acceptance. In fact, many human operators normally interpret other people's body language and social cues, and thus anticipate the needs of who is manifesting the intention to interact. A service robot able to replicate such human behavior would be effective and well-accepted by users. To this end, it should be able to (i) keep track of nearby people; (ii) predict when an approaching person intends to interact with it; and (iii) react accordingly. The first skill can be solved by off-the-shelf tools. The second, instead, is more challenging and it is the problem that we decide to tackle. Then, once the intention is detected, reaction strategies to implement the third skill can be developed according to the specific robot.

This extended abstract describes preliminary results for a learning-based method that allows the robot to classify whether each tracked person intends to interact. The user's body motion, detected by off-the-shelf systems, is used to compute in real-time the probability that the person will interact. Some questions remain open and challenges need to be addressed, as discussed in the conclusions.

II. Approach

Let us consider a robot standing in an environment shared with humans. During normal operations, people routinely pass nearby the robot, entering and exiting the robot's working space; occasionally, some users engage with the robot. We also assume that the robot is equipped with sensors capable of detecting and tracking people. In particular, using the Microsoft Azure kinect [7], it is possible to measure the planar pose, head orientation, and linear velocity of the people entering the sensor's field of view. These pieces of information are indicative of how the users move nearby the robot. Interpreting this kind of data, we tackle the problem of predicting the intention of a person to interact with the robot before the interaction actually begins.

Fig. 1. In a human-robot interaction context, when properly approaching the robot, a user is classified as intending to interact (first two snapshots); otherwise it is classified as non-interacting (right).

To solve this problem, we train a binary classifier that takes as input the tracked person's motion and outputs the probability that that person will interact with the robot. The classifier is trained on a dataset composed of several sequences. A sequence represents a person tracked by the robot over time and is composed of multiple samples (one per timestep). The sequence begins when the person is first seen by the sensor; it ends when the person either begins their interaction with the robot or exits the sensor's field of view without interacting. All the samples in a sequence are marked with the same label (true or false) according to whether the user interacted or not. We collect a real-world dataset of human-machine interactions in which humans behave naturally: in particular, we place a sensor alongside an espresso coffee machine placed in a break area neighboring a corridor of an office building. During the day, many people pass through the corridor, some stop in the break area, and some of them approach the machine to take a coffee. The data about the people's motion extracted by the sensor fill a dataset of 3422 unique sequences of tracked users, accounting for more than 12 hours of recorded data.

In our scenario, sequences should be ideally labeled by

This work was supported by the European Union through the project SERMAS.

[1]Dalle Molle Institute for Artificial Intelligence (IDSIA), USI-SUPSI, Lugano, Switzerland name.surname@idsia.ch

[2]Department of Engineering, King's College London, UK name.surname@kcl.ac.uk

considering when a user operates the machine, e.g. by pressing a button on it. However, in our case, we do not have access to the machine firmware and we can not read its internal state. Therefore, we rely on the sensor used for data collection to automatically generate labels. To do so, we use the following heuristic: interaction is detected when a person stays very close (i.e. within 1 m) to the coffee machine for an uninterrupted period of 5 seconds; we assume that the interaction takes place at the end of this period; all samples in the preceding 10 seconds are labeled as positives.

III. Preliminary Results

We use the "coffee machine" dataset to train simple classifiers. To evaluate the results, we consider all frames from all the sequences to compute the Area Under the ROC Curve (AUROC): a robust binary classification metric that does not depend on a choice of the classifier's threshold, and ranges between 0.5 (for a non-informative classifier, e.g. one always reporting the majority class) and 1.0 (an ideal classifier). When computed on all testing samples pooled together, the classifiers score very high (the AUROC is larger than 0.9): the reason is that the person's distance from the device is a very strong cue of whether the person ends up interacting with it. Then, we split all the samples into seven distance bins and compute metrics for each bin, to evaluate the ability of our approach to classifying a person's intention to interact *independently* on their distance from the device. The main conclusion that we derive is that rich information about the user motion (planar pose, head orientation, and linear velocity together) yields better results (AUROC > 0.7) w.r.t. considering a subset of such information. However, when using rich sensory information, predicting performance at short distances can result in a more complicated task (AUROC ≈ 0.65) than at long distances (AUROC ≈ 0.8). We believe that this is due to our dataset: it is difficult to understand whether someone close to the machine is there to interact or to do something else. By contrast, people that are approaching from afar, exhibit clearer intention in their body language and gaze, making additional features much more valuable in that case. For a qualitative evaluation of the classifier, see Fig. 1, where we deployed our approach in an online experiment; both the external view on the experiment scene and the corresponding snapshots taken by the sensor, showing also the classifier output, can be evaluated.

IV. Discussion and Open Challenges

We have presented a classifier to predict the user's intention to interact with a robot, using a dataset of tracked users interacting with the system in a real-life scenario. The system has been validated on the testing partition of the dataset and with an online experiment. Our preliminary study opens several questions and challenges that we aim at solving in the future, paving the road toward further developments.

An important aspect of our study is the evaluation of the results, not only from the technical perspective of the used methodology but also at the level of social acceptance among users. To this end, we plan to enrich our human-robot interaction analysis using psychological tools, which have been used to investigate the interactions happening between humans. On the same line, we will better investigate proxemics notions, such as social spaces [8], or nonverbal communications modalities [9], [10] and other tools from the human-robot interaction community.

A more extensive data collection campaign will be surely beneficial to our work. Some datasets are presented [11] and available online [12] for studying social interactions. However, the specificity of the problems that we want to address requires the collection of ad hoc information. In particular, we plan to collect datasets in public environments and different social contexts.

Finally, we plan to test our framework with different robotic platforms and in different interaction contexts. With this regard, in the current study, we did not consider how the appearance of the robot would affect the interaction [13] and, consequently, the intention of the user. In the future, we will consider this aspect by proposing experiments also with human-sized robots. Furthermore, we will challenge our method with more complex scenarios, where multiple interacting users need to be detected and classified.

References

[1] 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.

[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] 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.

[4] L. Palopoli, A. Argyros, J. Birchbauer, A. Colombo, D. Fontanelli, A. Legay, A. Garulli, A. Giannitrapani, D. Macii, F. Moro *et al.*, "Navigation assistance and guidance of older adults across complex public spaces: the DALi approach," *Intelligent Service Robotics*, vol. 8, pp. 77–92, 2015.

[5] J. Mišeikis, P. Caroni, P. Duchamp, A. Gasser, R. Marko, N. Mišeikienė, F. Zwilling, C. de Castelbajac, L. Eicher, M. Früh, and H. Früh, "Lio-A Personal Robot Assistant for Human-Robot Interaction and Care Applications," *IEEE Robot. and Autom. Lett.*, vol. 5, no. 4, pp. 5339–5346, 2020.

[6] D. Lee, G. Kang, B. Kim, and D. H. Shim, "Assistive delivery robot application for real-world postal services," *IEEE Access*, vol. 9, pp. 141 981–141 998, 2021.

[7] "Microsoft Azure," https://azure.microsoft.com/en-us/products/kinect-dk, Accessed: 2023.

[8] N. Marquardt and S. Greenberg, "Informing the design of proxemic interactions," *IEEE Pervasive Computing*, vol. 11, no. 2, pp. 14–23, 2012.

[9] 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.

[10] 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.

[11] D. M. Nguyen, M. Nazeri, A. Payandeh, A. Datar, and X. Xiao, "Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset," *arXiv preprint arXiv:2303.14880*, 2023.

[12] "Multi-Modal Social Human Navigation Dataset (MuSoHu)," https://cs.gmu.edu/ xiao/Research/MuSoHu, Accessed: 2023.

[13] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," *IEEE Robot. Autom. Mag.*, vol. 19, no. 2, pp. 98–100, 2012.