Pointing Gestures for Human-Robot Interaction in Service Robotics: A Feasibility Study

Luca Pozzi[1(✉)], Marta Gandolla[2], and Loris Roveda[3]

[1] Mechanical Department, WE-COBOT Lab (Polo Territoriale di Lecco), Politecnico di Milano, Lecco, Italy
luca.pozzi@polimi.it
[2] Mechanical Department, Politecnico di Milano, Milano, Italy
[3] Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Scuola Universitaria Professionale della Svizzera Italiana (SUPSI), Università della Svizzera italiana (USI), Lugano, Switzerland

Abstract. Research in service robotics strives at having a positive impact on people's quality of life by the introduction of robotic helpers for everyday activities. From this ambition arises the need of enabling natural communication between robots and ordinary people. For this reason, Human-Robot Interaction (HRI) is an extensively investigated topic, exceeding language-based exchange of information, to include all the relevant facets of communication. Each aspect of communication (*e.g.* hearing, sight, touch) comes with its own peculiar strengths and limits, thus they are often combined to improve robustness and naturalness. In this contribution, an HRI framework is presented, based on pointing gestures as the preferred interaction strategy. Pointing gestures are selected as they are an innate behavior to direct another attention, and thus could represent a natural way to require a service to a robot. To complement the visual information, the user could be prompted to give voice commands to resolve ambiguities and prevent the execution of unintended actions. The two layers (perceptive and semantic) architecture of the proposed HRI system is described. The perceptive layer is responsible for objects mapping, action detection, and assessment of the indicated direction. Moreover, it has to listen to uses' voice commands. To avoid privacy issues and not burden the computational resources of the robot, the interaction would be triggered by a wake-word detection system. The semantic layer receives the information processed by the perceptive layer and determines which actions are available for the selected object. The decision is based on object's characteristics, contextual information and user vocal feedbacks are exploited to resolve ambiguities. A pilot implementation of the semantic layer is detailed, and qualitative results are shown. The preliminary findings on the validity of the proposed system, as well as on the limitations of a purely vision-based approach, are discussed.

Keywords: Human-Robot Interaction · Pointing · Service robotics · Action detection

K. Miesenberger et al. (Eds.): ICCHP-AAATE 2022, LNCS 13342, pp. 461–468, 2022.
https://doi.org/10.1007/978-3-031-08645-8_54

1 Background

The term *robot* was introduced in 1920 by the Czech playwright Karel Čapek in his sci-fi drama "R.U.R" [15], adapting a Slavic word for *forced labor*. Although the term remains unchanged, robots trascended their original role of tireless factory workers and nowadays promise to assist humans in many different contexts. Enhanced hardware functionalities can be combined with artificial intelligence algorithms, so to let the robots work autonomously in unstructured environments. As a consequence, Service Robotics (SR) is gaining popularity and robots are being deployed in public places such as hotels, airports or hospitals, as they can increase productivity, guaranteeing service consistency [14]. With particular reference to the healthcare sector, burdened by the ageing of the world population and professional staff shortage, SR offers an appealing solution to relief the medical personnel and to improve the quality of the cares [10].

Though non-social service robots exist (*e.g.* for cleaning, monitoring...), the use of SR might have a great impact in supporting heterogeneous and personalized tasks execution in an unstructured environment. In this sense, Human-Robot Interaction (HRI) and SR become closely linked fields. In particular, for some applications, HRI arises as a mandatory requirement, as the robot must be able to understand the user's commands. The most relevant communication means are referable to the senses of hearing, sight, and touch, either alone or combined [8]. When thinking about communication, speech is usually the first concept that comes to mind, as it is a behavior we voluntarily practise every day. However, available text-to-speech and speech recognition models are characterized by a strong computational load *vs* naturality trade-off, limiting their application [3]. A simple yet effective solution is represented by touch screens, although the resulting interaction is impoverished [3]. Beside that, motion and touch sensors can be used to feel the user's proximity and possibly infer his/her behavior (*e.g.* hugging a toy robot) [3]. However, physical interaction introduces some issues related to sanitization when the robot is shared among several users, particularly relevant in healthcare settings. One further valuable communication mean is represented by gestures, either as a source of information about the user's attitude (*i.e.* body language) or as voluntary commands sent to the robot [6]. The latter (*i.e.* action detection) still represent an open research issue and it has been selected as the preferred human machine interaction strategy in the proposed framework.

2 Related Work and Paper Contribution

Pointing gestures are a popular choice for HRI, as they are a natural way to drive another's (*i.e.* the robot) attention. Nickel and Stiefelhagen [7] fused skin-color map and disparity images to achieve communication with a domestic robot. The findings of their work identified pointing as a viable way for object selection in an household scenario (90% of correct target identification), despite a limited geometrical accuracy (average error of 16.9). To achieve more accurate detection

performance, available methods often introduce constraints, thus limiting the problem complexity at the cost of generality. Some of them require the user to wear a sensor, such as in Gromov and colleagues work [5], in which an inertial measurement unit placed on the wrist allows to control a drone's flight. This kind of methods is hardly scalable, as the number of users grows, and requires to know the operators in advance. Another possibility is to constrain the gesture execution to ease its recognition. Azari and colleagues [1] tested their method on a mobile robot, achieving a 0.5 m error at a 5.5 m distance. However, the pointing is assumed to be done putting the index finger between the eyes and the point of interest, thus reducing the user's freedom of movement. Likewise, the problem can be simplified assuming a fixed (and optimal) relative position between the camera and the user. This has been done in the work from Showers and Si [13], who fused visual and audio information to discriminate between objects in close vicinity. The idea of multimodal communication is appealing, as it could both improve the accuracy of the detection and help in achieving a more natural interaction. Voice could be used as a trigger for the gesture detection, as in Bolt's pioneering work [2]. In his study, a user moves a cursor on a screen by pointing at it and, with the voice command "there", can select the current position as the desired one.

In this work, we propose the HRI system to leverage on color and depth images to detect pointing gestures, allowing the user to request for a service. The system is complemented with audio capabilities to concur in resolving ambiguities and to alternatively send feedbacks to the robot or the user. The goal is not limited to achieve a robust identification of the object/area selected through pointing, but also to trigger the proper robot action, based on the object's characteristic and the surrounding context. Thus, the main contribution of the present paper are:

- the development of a HRI system
  - enabling a natural interaction with the robot (*i.e.* not imposing any constraint nor on the user behavior nor on his/her position),
  - relying exclusively on sensors that can be easily mounted on a service robot (namely an RGBD camera and a microphone), hence not requiring to place any sensor on the user;
- an application of said HRI system on a service robot.

3 Materials and Methods

3.1 Two-Layers HRI System

The conceived HRI system is composed of two levels, namely the *perceptive* and the *semantic* layer, as depicted in Fig. 1.

The perceptive layer has three main functions:

i mapping the objects available for interaction, combining autonomous navigation and object detection;

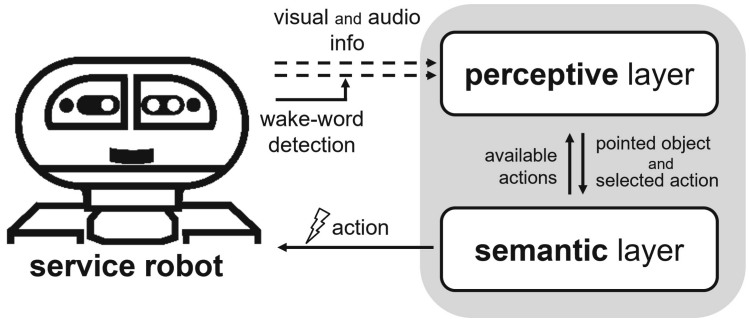


Fig. 1. Schema of the application functioning. The HRI system, with its two layers, is represented against a grey background. Arrows represent the messages exchanged between the building blocks. Dashed arrows represent conditional connections, as the stream of audio and video information is subordinate to the wake-word detection.

ii detecting the indicated direction, after having applied user tracking and action detection methods;
iii listening to the user's audio feedback, *e.g.* when the user is prompted to give a voice command to select the actions listed by the semantic layer.

The semantic layer is in charge of determining which actions, in the current context, are available for a given object. The actions may be linked to the object itself (*e.g.* a fetch-and-retrieval task), or depend on the presence of other items (*e.g.* the action *pour* when pointing to a bottle, is made available only if there is a cup in the surroundings), or require collaboration with the user (*e.g.* to unscrew the bottle cap with a one-armed robot).

To avoid having the described system running continuously, a wake-word detection system (similarly to what happens with Google Assistant or Amazon Echo) could be introduced. This would be beneficial both to relief the computational load on the robot's computer, as well as to avoid privacy issues. Indeed, triggering the a robot interaction by saying the wake-word, the user would give an implicit consensus to be tracked and listened.

3.2 Robotic Platform

The framework is implemented and tested on a mobile service robot (*TIAGo* robot, *PAL Robotics*) [11]. The mobile base, equipped with a *SICK* laser, and the *Orbecc Astra S* RGBD camera enable the environment mapping. The same camera is exploited to capture the user's motion. For voice interaction, the *Super-Beam Stereo Array Microphone* (*Andrea Electronics*) and a 5W speaker are used [11]. The robot and the location of the mentioned devices are depicted in Fig. 2

Fig. 2. TIAGo, the service robot used to test the proposed framework. The locations of the onboard sensors used for HRI are highlighted.

4 Pilot Implementation

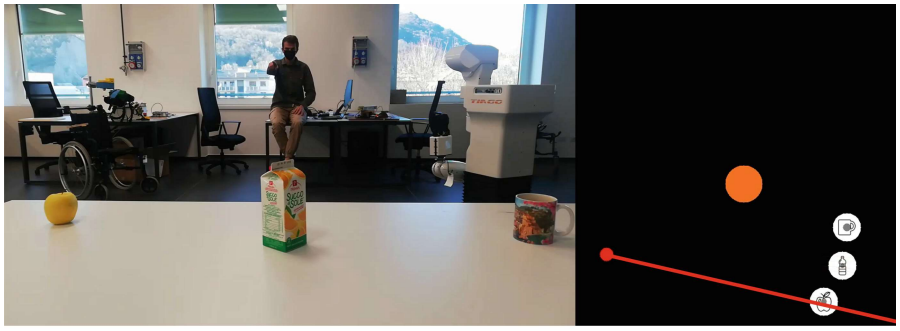
A pilot implementation of the perceptive layer, exploiting a state-of-the-art 2D keypoints extraction model and simple heuristics for action detection, led to the qualitative result shown in Fig. 3.

The Lightweight OpenPose [9] pre-trained model is exploited for keypoint detection on the RGB images. The model has been selected as it is an implementation of the popular OpenPose method [4], optimize to run in real-time on a CPU. The joints positions are then brought to the 3D space relying on the information from a synchronous pixel-aligned depth image. The RGB and depth image are considered as synchronized if the time difference between their acquisition stamps fall within a $3 \cdot 10^{-2} s$ interval. The arms movement are analyzed to estimate their kinetic energy as described by Shan and collaborators in [12]. Accordingly to the cited method, the kinetic energy estimate of a set of points at time i is defined as the sum of the points' estimates, *i.e.*

$$E(P_i) = \sum_j E(P_i^j) \tag{1}$$

where $E(P_i^j)$ is the estimated kinetic energy of a single joint, computed as

$$E(P_i^j) = \frac{1}{2}(v_i^j)^2 = \frac{1}{2}\left(\frac{P_i^j - P_{i-1}^j}{\Delta T}\right)^2 \tag{2}$$



(a) (b)

Fig. 3. Application functioning demonstration. Figure (a) The user is in the robot view and has raised his arm to point toward an object. The robot has previously mapped the object on the table, and thanks to a keypoint detection network, assesses the indicated direction and thus the selected object. Figure (b) Scene reconstruction made by the robot: white circles represent objects; the red dot is the user's wrist and the red line the pointed direction; the orange circle accounts for the robot position. (Color figure online)

where ΔT is the time elapsed between the receipt of samples $i-1$ and i. The kinetic energy estimate for each arm is therefore obtained substituting Eq. 2 into Eq. 1, *i.e.*

$$E(P_i) = \sum_j \frac{1}{2} \left(\frac{P_i^j - P_{i-1}^j}{\Delta T} \right)^2 \tag{3}$$

where $j = shoulder, elbow, wrist$. A threshold value is experimentally tuned to discriminate between the motion and rest condition of each arm.

A pointing gesture is assumed to be composed of three phases: arm lifting, stationary stance (*i.e.* pointing) and arm return. The action starts with the two arms in rest condition. If one arm starts moving upward (*i.e.* E overcomes the threshold and the wrist is moved away from the ground plane), the arm lifting phase is entered. Then, if the arm stops, the line passing through the elbow and the wrist joint is assessed as the indicated direction. The subsequent movement is recognized as the arm return phase. During the lifting and the stationary phases, the contralateral arm must be kept still, otherwise the movement is classified as a non-relevant gesture. To reduce the sensitivity to the noise in the arm poses, the transition from the rest to the motion condition is triggered only if three consecutive above-threshold values of E are received (and vice versa when switching from motion to rest). This sketched action detection system is false positives-prone, as several arm movements share the same features. However, it must be noticed that, being the interaction intentionally triggered by the user, the impact of this issue is limited.

Overall, the perceptive layer can run on a *AMD Ryzen7* 8-cores CPU at a rate of $\approx 4Hz$.

5 Preliminary Findings

The initial work on the perceptive layer allowed the authors to better understand the limits and the potential of a fully vision-based approach. Though the described system is able to assess a good estimate of the indicated direction, the perception layer must be refined to improve the accuracy and, most importantly, the robustness. Indeed, the current implementation is sensitive to variations in light condition and user-robot relative position. As far as performance is concerned, the proposed implementation can work in real-time for the task at hand on a standard CPU. The speed rate, indeed, is sufficient to recognize pointing gestures, characterized by a considerable static phase. Nevertheless, the perceptive layer would benefit of a faster pose estimation algorithm as it would enable to use the same approach on more dynamic gestures. Moreover, it would be useful to associate a confidence score to the geometrical information, *e.g.* the variance of the detected direction in the last n samples (as in [13]). In addition, the implementation of a vocal feedback system would allow to make the robot aware of any mistake in perception. This would both prevent the execution of undesired actions and pave the way for a learning-based improvement of the performance.

6 Conclusions

The authors acknowledge the limitations of a preparatory study lacking of a robust experimental evaluation. Nevertheless, the proposed architecture for a HRI system is sound and the provided pilot implementation of the perceptive layer represents an intriguing starting point for the project development. The proposed HRI framework would only relying on the robot's onboard instrumentation (*i.e.* RGBD camera, microphone and speaker), enabling a natural connection with a mobile service robot. Indeed, the method introduces a minimal overhead to trigger the interaction (as easy as saying a couple of words) and does not impose any constraint nor on the user behavior nor on the ambient. Ultimately, the spontaneity of the interaction represents the greatest strength of the presented HRI framework, as it is a fundamental requirement to achieve users' acceptance.

References

1. Azari, B., Lim, A., Vaughan, R.: Commodifying pointing in HRI: simple and fast pointing gesture detection from RGB-D images. In: 2019 16th Conference on Computer and Robot Vision (CRV), pp. 174–180 (2019). https://doi.org/10.1109/CRV.2019.00031
2. Bolt, R.A.: "Put-That-There": voice and gesture at the graphics interface. In: Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques, pp. 262–270, SIGGRAPH 1980. Association for Computing Machinery, New York, NY, USA (1980). https://doi.org/10.1145/800250.807503

3. Bonarini, A.: Communication in human-robot interaction. Curr. Robot. Rep. **1**(4), 279–285 (2020). https://doi.org/10.1007/s43154-020-00026-1
4. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: CVPR (2017)
5. Gromov, B., Abbate, G., Gambardella, L.M., Giusti, A.: Proximity human-robot interaction using pointing gestures and a wrist-mounted IMU. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 8084–8091 (2019). https://doi.org/10.1109/ICRA.2019.8794399
6. Ji, Y., Yang, Y., Shen, F., Shen, H.T., Li, X.: A survey of human action analysis in HRI applications. IEEE Trans. Circuits Syst. Video Technol. **30**(7), 2114–2128 (2020). https://doi.org/10.1109/TCSVT.2019.2912988
7. Nickel, K., Stiefelhagen, R.: Visual recognition of pointing gestures for human-robot interaction. Image Vis. Comput. **25**(12), 1875–1884 (2007)
8. Onnasch, L., Roesler, E.: A taxonomy to structure and analyze human–robot interaction. Int. J. Soc. Robot. **13**(4), 833–849 (2020). https://doi.org/10.1007/s12369-020-00666-5
9. Osokin, D.: Real-time 2D multi-person pose estimation on CPU: lightweight OpenPose. arXiv preprint arXiv:1811.12004 (2018)
10. Ozturkcan, S., Merdin-Uygur, E.: Humanoid service robots: the future of healthcare? J. Inf. Technol. Teach. Cases 20438869211003905 (2021). Prepublished 23 June 2021
11. Pagès, J., Marchionni, L., Ferro, F.: TIAGo: the modular robot that adapts to different research needs (2016)
12. Shan, J., Akella, S.: 3D human action segmentation and recognition using pose kinetic energy. In: 2014 IEEE International Workshop on Advanced Robotics and its Social Impacts, pp. 69–75 (2014). https://doi.org/10.1109/ARSO.2014.7020983
13. Showers, A., Si, M.: Pointing estimation for human-robot interaction using hand pose, verbal cues, and confidence heuristics. In: Meiselwitz, G. (ed.) SCSM 2018. LNCS, vol. 10914, pp. 403–412. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91485-5_31
14. Wirtz, J., et al.: Brave new world: service robots in the frontline. J. Serv. Manage. **29**, 907–931 (2018). https://doi.org/10.1108/JOSM-04-2018-0119
15. Čapek, K., R.U.R.: Rossum's Universal Robots. Aventinum (1920)