2023 IEEE International Conference on
Autonomous Robot Systems and Competitions (ICARSC)
Tomar, Portugal – April 26-27, 2023

2023 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC) | 979-8-3503-0121-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICARSC58346.2023.10129582

Follow me: interleaving human tracking and interacting with a new assistant robot

Gerardo Pérez
RoboLab research group
Universidad de Extremadura
Cáceres, Spain

Noé José Zapata
RoboLab research group
Universidad de Extremadura
Cáceres, Spain

Sergio Barroso
RoboLab research group
Universidad de Extremadura
Cáceres, Spain

Alejandro Torrejón
RoboLab research group
Universidad de Extremadura
Cáceres, Spain

Pablo Bustos
RoboLab research group
Universidad de Extremadura
Cáceres, Spain

Pedro Núñez
RoboLab research group
Universidad de Extremadura
Cáceres, Spain

***Abstract*—The development of socially-aware autonomous robots for being part of our daily lives is a pressure matter. Although almost all current robots use multiple sensors, the limitations of these autonomous systems to coexist with people are still evident: they have difficulties fusing all this information in an agile and consistent way and thus understand human intentions and act accordingly. The main consequence of these limitations is that today's robots do not respond adequately to these human behaviours and, for example, result in forced and unnatural navigation. Today's robots strive to have complete knowledge of the surrounding environment to make the right decisions at the right time. However, this purpose requires a perspective that integrates sensors of different natures and, based on it, a multi-modal perception of actuation. This paper presents a multi-modal and distributed architecture for the perception of the environment that uses information given by the robot sensors. The results of the proposed architecture are validated in the Follow Me use a case in a real environment.**

***Index Terms*—CORTEX, Cognitive architecture, Human-Robot Interaction, Multi-modal perception, Distributed Robot Systems**

I. Introduction

Socially aware robotics is a term that has become a strong force in the scientific community in the last decade. The fact that future robots will coexist with humans is accepted in almost all forums and for this reason, they must behave as people would. For example, if a robot moves in a human-populated environment, should it include some specific features in its navigation algorithm to improve the possibility of being accepted? Or, if this robot is going to have a conversation with some human companion, does it have to add new features in its Human-Robot Interaction (HRI) algorithms to improve acceptance during the dialogue?

What is evident in this type of robot is that it should have the skill to detect the pose of the person in a non-invasive method. For this purpose, since their origins, robots have been provided with sensors to estimate the people in their surroundings. The task of detecting and tracking people is not trivial, as it is subject to the complexity of the environment and its dynamic content, the uncertainty of the algorithms and the sensors themselves. Moreover, if the robot moves, this displacement also reduces the accuracy of the results.

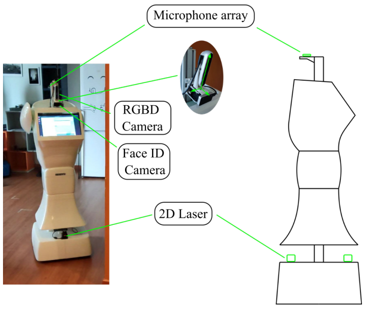

Fig. 1. The Storki robot and the set of sensors for our multi-modal social awareness people tracking algorithm.

The behavior of robots when interacting with people is a crucial factor that affects their acceptance [1]. It is hypothesized that robots can behave more like humans if they possess an attentional mechanism system capable of 1) detect and follow people; 2) respond in a synchronized and natural manner; and 3) use a combination of sensors such as video and audio to recognize a person's pose in dynamic and constantly changing environments.

This paper presents a multi-modal social awareness system that can detect and track individuals within a robot's surroundings. The system operates through the use of the CORTEX cognitive architecture, which integrates information from three distinct sensors: an RGBD camera, a face identification camera, and a microphone array that determines the position of a person in situations where they may not be visible in the image. The attentional mechanism, understood as an active perception system, consists on and RGBD camera placed on a 1 DoF servo that pans around its vertical axis at the head of the robot. In coordination with the robot's base, the orientable camera keeps the person followed centred on the stream of images. After calibrating the full system,

979-8-3503-0121-2/23/$31.00 ©2023 IEEE

the control architecture integrates the information from all sensors to estimate the person's pose with whom the robot is interacting. We validate this approach in the 'Follow me' use case, in ideal conditions (good light, spacious areas). Figure 1 shows a schema of the mobile platform and the proposed active perception system.

II. Related work

People detection is a common problem in most social robotics applications. As we know from the theory of proxemics[1], the person's position determines where the robot should be positioned for interacting or navigating in a socially acceptable way [1], [2]. Recent studies are usually focused on methods to keep human-robot engagement, such as the work presented in [3], where authors define a multi-modal perception model for HRI and where this detection is an essential task. In accordance with this idea, this paper presents a multi-modal system that enables a socially aware robot to actively follow the person it interacts with, improving its social behaviour [4].

There is an extensive literature of scientific works related to human detection ([5]–[7]). Most of these detector algorithms suffer from a high computational load, mainly if they are based on image processing [7]. In recent years, the number of works based on neural networks has improved the accuracy of the results. The latter, coupled with the fact that more and more devices support this computational load, has enabled performance improvement [8]. The problem of human tracking is not new either, and many solutions have been presented in this field. In the papers [9], and more recently [10], we find different reviews of the best-known methods for people tracking. Usually, these systems use one or several sensors to track objects of interest ($e.g.$, people) with different strategies: from classical computer vision algorithms to neural networks. The main disadvantage, again, is their computational cost. Moreover, a single nature of data is insufficient in real situations when the robot moves in a highly populated environment: people disappear for a long time ($e.g.$, when crossing a door) or there are many people in the environment. Our proposal contributes to this research line by using different sources of information that, combined, strengthen the tracking system, always under the prism of HRI.

Socially aware robots must provide trust when interacting with humans as a critical factor in maintaining assertive engagement [3]. Based on this idea, our tracking proposal is integrated into an active perception system that aims to have the people of interest in focus. Active perception has been widely used in robotics and computer vision. However, its application to human tracking and its use in social robotics applications is not widespread. In the work of [11], a network of fixed cameras is used to track people to improve surveillance. Meanwhile, in [12], they use a multi-modal system for tracking people and mixing audio and video information. The authors start the active search with audio information and then visual information. Our objective is different, using information from both sources to keep the robot's attention on the person, prioritizing one or the other source depending on the specific situation. Our tracking approach involves direct action on the robot's base and the camera itself and using the audio information to improve the results in case of loss.

Traditionally, most active perception systems work with moving elements using only one or two degrees of freedom (pan-tilt) in the camera [13]. Our solution provides the coordinated movement between the robot base and the RGBD camera, as a person would do during a real interaction. The person is the centre of interest, and the robot moves and rotates to achieve this goal. As a contribution to this work, we include the whole formulation for this active perception system that integrates the motion of the robot base and the RGBD camera.

III. Overview of the social awareness people tracking pipeline and problem definition

Our system addresses the problem of multi-modal tracking embedded in a social robot. Both the robot base and the RGBD camera are mobile devices that track the person of interest to maintain their attention. To this end, the active perception system moves the base and camera to center the person in the acquired image. If the person disappears from the robot's view, the robot will start a dialogue to attract their attention. Once the person responds, their relative orientation is estimated using the microphone array, and the robot rotates according to the acquired information until the person is centered. Our social-awareness tracking pipeline is integrated into a human-aware navigation system and in all HRIs performed by the robot. Figure 2 provides an overview of the system and its interconnections. A more detailed description of each element is provided in subsequent sections.

The CORTEX cognitive architecture has been recently proposed as a tool to design modular software to control intelligent robots. It is based on the notion of specialized memories (i.e., working, episodic, semantic, etc.) that are interconnected by processing modules called agents. In CORTEX, robot activities correspond to flows of information across these memories that are fostered by the agents. The central element in this architecture is the Working Memory (WM)[2]. All the agents share the WM, and it is their only means of communication among them. Other memories are controlled by specific agents that create and maintain flows of information between them and the WM. The WM stores the robot's internal state and a representation of its environment relevant to the current mission.

Formally, the WM is a directed graph defined as a pair $G = (V, E, \omega)$ comprising V a set of vertices, E a set of edges and an incidence function mapping every edge to an ordered pair of vertices, $\omega : E \mapsto \{(x, y) | (x, y) \in V \wedge x \neq y\}$. The nodes V contain instances of the concepts known to the system. Concepts can refer to physical or internal entities.

[1] Proxemics studies the spaces between people during an interaction [14]

[2] The WM is also called Deep State Representation (DSR) since it can hold heterogeneous data at different levels of abstraction

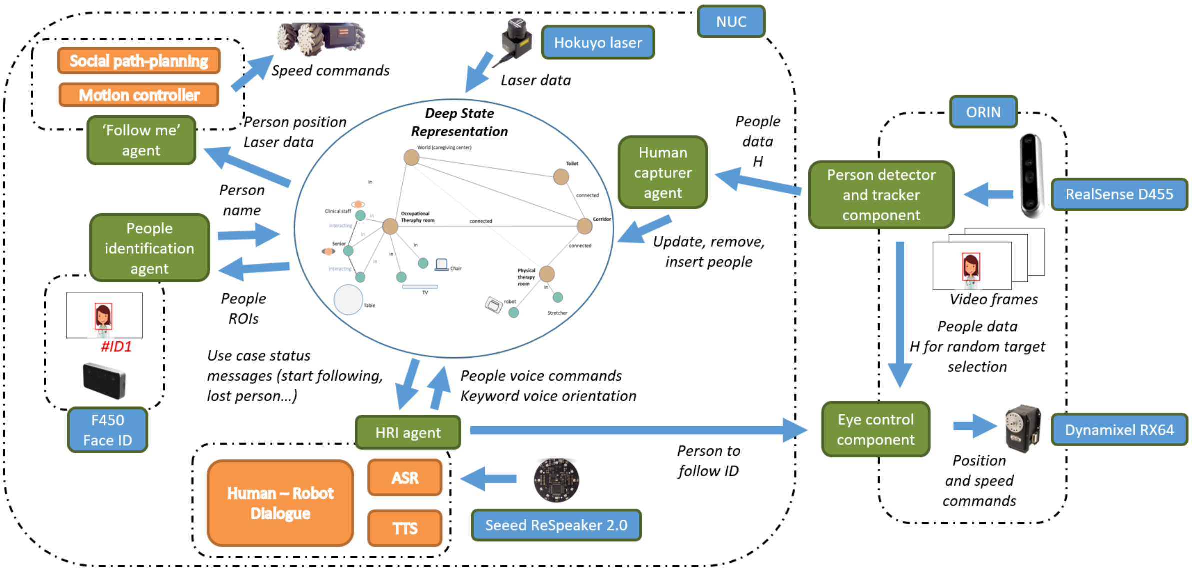

Fig. 2. Overview of the social awareness Follow Me pipeline.

When a physical concept is instantiated into the WM by some agent, it must be kept anchored by this agent to its world entity during its life span. Internal entities are generated by the agents as missions, intentions or plans and are represented in the graph so all other agents may become aware of them. The graph can be seen as an active representation of the context that is relevant to the current mission.

Our system comprises of an instance of the CORTEX architecture that runs in the NUC computer at 20Hz and a set of connected components that run in the Orin computer at up to 60Hz. These faster components implement the human detection and tracking pipeline and the orientable camera controller. A diagram illustrating these two elements and their connection can be seen in Fig. 2.

The human detection pipeline starts with the camera RGBD stream acquisition, which is then processed by a YOLOv7 DNN optimized for Nvidia[3] to extract person-labelled regions. Then, the trt-pose DNN[4] extracts the skeleton of each region and adds it to the list of detected people. Finally, the list is processed by the ByteTrack algorithm to create a list of known people and stabilize their identifiers by performing Kalman prediction on the unmatched ROIs and an optimal alignment with the Hungarian algorithm and the Intersection over Union (IoU) metric. Additionally, the Orin computer runs another component controlling the servo to engage and track a selected person from the list of known people. When the Orin starts, these components take control of the camera, implementing an inhibition of return (IOR) dynamics. The IOR mechanism selects a human target randomly and tracks them for a period of time before switching to another individual who has not been recently visited. This process continues until the CORTEX instance is initiated, at which point the list of recognized individuals is obtained from the Orin computer and the agents prepare to begin a mission. The Follow Me mission commences when an individual approaches the robot, captures its attention, is identified by the Realsense Face ID camera, and issues the command ”Follow Me”. The robot then designates that person as the new target and proceeds to follow them.

The CORTEX instance used here includes a WM shared by a set of agents:

- Human capturer agent: receives information about detected individuals and updates the graph nodes accordingly. Existing data is updated, disappeared individuals are deleted, and new individuals are inserted into the graph.
- Human identification agent: extracts and stores the names of individuals appearing in an image as an attribute in the corresponding node representing that person.
- HRI agent: engages in conversations with the individual in case of loss or when the individual initiates dialogue or a new mission. It also detects the keyword for robot call and the direction of the voice signal.
- 'Follow me' agent: oversees and controls the ongoing mission, issuing commands to the robot base. It modifies edges in the WM to reflect the mission's progress, particularly during interactions initiated by humans. Additionally, the agent responds to changes in the WM, such as the loss of the tracked person, by modifying edges to inform other agents of the situation.

IV. Multi-modal People Tracking Algorithm

This section contains the core technical work. The different modalities include the person tracking system, person identification and human localization based on audio signals. We then describe the data fusion from different sources and conclude

[3] https://github.com/Linaom1214/TensorRT-For-YOLO-Series
[4] https://github.com/NVIDIA-AI-IOT/trt_pose

with the displacement of the camera base and servo that make up the active perception system proposed in the paper. Let H be the set of all the persons h_i detected by the robot. Each $h_i \in H$, is defined in the CORTEX architecture as a person node with the following information $h_i = \{p_i, s_i, id_i, f_i, d_i\}$, where p_i is the 6D pose of the person, $(x, y, \theta)_i$, s_i the set of 3D joints that make up its skeleton, id_i is the person's identifier, f_i is the identifier obtained from the facial recognizer, and d_i is the distance from the human to the robot.

A. Human detection and tracking based on video modality

The video modality consists of a skeleton point-tracking algorithm. Our social robot uses an orientable RGBD camera. The extrinsic parameters of the camera are calibrated after positioning them on the robot. Human pose detection and tracking: to track a human in front of the robot robustly, we use a method that takes the output of the YOLOv7 DNN [17] and feeds it to the ByteTracker [18] tracking algorithm. The system accepts, as input at time τ, a color image T of size w x h and generates, as output, the 3D location and orientation of each person in the image p_i and it's track identifier id_i. The image is processed by YOLOv7, which returns the people's ROIs $B_i = \{a_i, b_i\}$, being a_i and b_i the top-left and bottom-right points. When a person is detected, a track is started using the ByteTracker algorithm, obtaining the id_i. The 3D pose from the robot's perspective is obtained from the central pixel $O_i = \left\{ a_x i + \frac{b_x i - a_x i}{2}, a_y i + \frac{b_y i - a_y i}{2} \right\}$ value of the ROI of the stereo image.

When obtaining individual ROIs, the person 2D pose estimation of each one is obtained through NVidia $TRTpose$ two-branch CNN, returning the candidate body joints $J_h^\tau(u, v)$. The set $J_h^\tau(u, v)$ has a 3D correspondence, which is referenced to centre of the robot, $J_h^\tau(x, y, z) = \{J_1, J_2, ..., J_N\}$. These values are also calculated directly from stereo images. From the set $J\tau_h(x, y, z)$, we select the subset $s_\tau \subset J_h(x, y, z)$, composed of those values that have been observed by the camera in the previous k_t frames, as long as their value do not exceed the threshold Θ_{J_i}:

$$\frac{(J_i^\tau - \dot{J})}{J_i^\tau} \leq \Theta_{J_i} \tag{1}$$

where $\dot{J} = \frac{1}{k_T} \sum_{\tau=0}^{\tau=k_t} J_{\tau-k_t}^i$ is the mean values of this key point in the k_t previous frames. From the set s_τ, we estimate the orientation θ_i, we first define the vectors α_i, β_i, where α_i is that of the mean values of the left skeleton joints with the centroid of the person, and β_i is the equivalent on the right side. Finally, θ_i is obtained through the vector product defined by $\theta_i = \alpha_i \times \beta_i$.

In our algorithm, each new human in the CORTEX architecture (node in the WM) is assigned a unique identifier, id_i. Figure 3 illustrates the bounding box of the persons detected by the algorithm.

B. Human identification

The goal of this stage is to achieve personalized HRI that enhance social acceptance. Person recognition is performed using a Face ID camera, and facial authentication is obtained using a dedicated system-on-chip. The result of this algorithm is an identifier (f_i) for each person previously recognized by the camera, which updates the corresponding node (h_i) in the DSR.

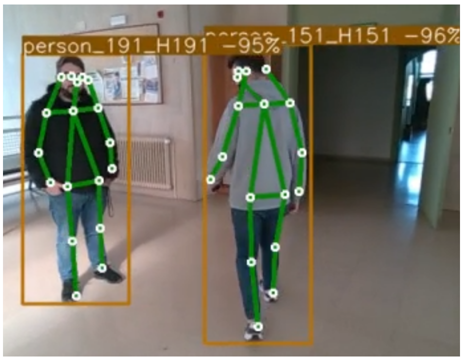

Fig. 3. People's ROIs with the associated $J\tau_h(x, y, z)$ joints.

C. Human localization based on audio modality

The audio modality aims to estimate the position of the person interacting with the robot. The proposed algorithm uses a Voice Activity Detector (VAD), with the goal of discriminating people talking from the rest of the ambient sound. Once this voice signal is detected, the algorithm estimates its localization. While most existing solutions work with a single microphone, the provided proposal uses a far-field microphone array device capable of detecting voices up to 5m away.

To localize the sound source, we use the well-known SRP-PHAT method (Steered Response Power-Phase Transform) [15], [16]. This algorithm uses the PHAT (Phase Transform) weighting to compute the General Cross Correlation (GCC) technique for each pair of microphones within the array.

D. Fusing multi-modal inputs for human tracking

The next stage merges the information from both modalities so that the resulting information has less uncertainty than would be possible when these sources are used individually. The direct fusion of the data is not performed on the total data of the person's position since the information provided by the audio modality only affects the person's orientation concerning the robot. In this sense, the person's localization $p_i(x, y)$ is estimated only from the video modality. The orientation information $p_i(\theta)$ provided by the audio modality will only be added if there is no information about the person in the video mode.

E. Attention control

The detection and subsequent tracking method of people proposed in this work are integrated within an attention mechanism that aims to provide the robot with social skills during its navigation and interaction with people. The perception of the environment and the search for the person are done actively. Thus, once the person's location is estimated, the proposed

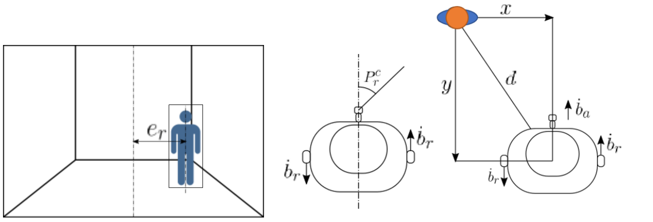

Fig. 4. a) Starting situation: the position of the person is known to the robot; b) The attentional mechanism modifies the speeds and final positions of the robot base and the camera servo.

system addresses the problem of focusing its attention on the person. The mathematical formula that provides the robot with this social ability is described below. Our system focuses the interest in the person through the combined movement of the robot base and this camera. We formulate this combined displacement for both the base and the camera: the result is a smooth and socially aware motion, just as a person would do during a conversation.

We first describe the final position of the camera and the servo velocity during the movement. Let be the case described in Fig. 4, where the person is displaced e_r concerning the centre of the image. This e_r is defined as the rotational distance between the person's centre and the image, measured in radians. Then, we formulate the *rotational servo speed* (rad/s) $\dot{P}_r$ as:

$$\dot{P}_r = (k_{e_r} e_r + k_{\dot{e}_r} \dot{e}_r) \frac{k_d}{d_i} \tag{2}$$

where $\dot{e}_r$ is the difference between the last and actual rotational distance (rad), and k_{e_r}, $k_{\dot{e}_r}$ are adaptable coefficients for e_r, and $\dot{e}_r$, respectively. The value of both k_{e_r}, $k_{\dot{e}_r}$ must be tuned according to the features of the servo. The coefficient $\frac{k_d}{d_i}$ is intended to regulate the speed, increasing it when the distance of the person from the robot d_i is less than k_d. Its purpose is to deal with the fact that the shorter the distance, the faster the movement of the servo is required to keep the person in the image.

We estimate the final camera position (*i.e.*, final servo position, in radians) from the equation:

$$P_r^c = o_r - k_{e_r} e_r \tag{3}$$

being o_r the current servo position. The Eq. 3 shows how the motion of the camera servo is directly coupled with the displacement of the person. Next equations describe both $\dot{b}_r$ (rad/s), $\dot{b}_a$ (m/s) and $\dot{b}_s$ (m/s), the rotational, advance and side velocities of the robot, respectively:

$$\dot{b}_r = arctan(\frac{p_y}{p_x}) \tag{4}$$

where p_x and p_y are the x and y positions of the human relative to the robot.

$$\dot{b}_a = \dot{b}_{max} \cdot G(\dot{b}_r) \cdot H(\dot{d}) \tag{5}$$

Algorithm 1 Calculation of servo velocity and position

```
$e_r \leftarrow (p_v - v)/2$
$\dot{e}_r \leftarrow e_r^{\tau} - e_r^{\tau-1}$
$P_r^c \leftarrow o_r - k_{e_r} e_r$
$P_r^e \leftarrow k_{e_r} e_r + k_{\dot{e}_r} \dot{e}_r$
$\dot{P}_r = \begin{cases} max\_value, & \text{if } abs(e_r) \geq \Theta_{e_r} \\ \dot{P^e}_r, & \text{otherwise} \end{cases}$
```

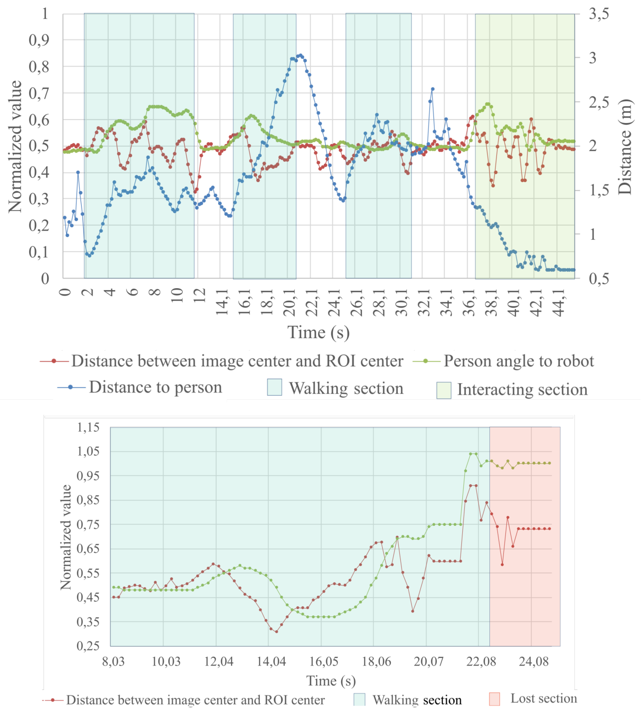

Fig. 5. Metrics obtained from the experiment.

being $\dot{b}_{a_{max}}$ the maximum advance speed of robotic base (m/s), $\dot{d}$ is the difference between the reference distance and the actual person to robot distance, $G(\dot{b}_r)$ is the Gaussian equation for rotational speed dependent speed control: $e^{-k_G \dot{b}_r}$, and $H(\dot{d})$ is defined as the Hyperbolic tangent equation for distance to human dependent direction and speed control:

$$H = \frac{e^{k_H d} - e^{-k_H d}}{e^{k_H d} + e^{-k_H d}} \tag{6}$$

where k_G and k_H are free parameters to modify the Gaussian and the hyperbolic tangent slope.

$$\dot{b}_s = \dot{b}_{s_{max}} k_{p_x} \tag{7}$$

Where $\dot{b}_{s_{max}}$ is the maximum side speed of robotic base (m/s), and k_{p_x} a coefficient that reduces side speed when p_x tends to 0. The algorithm 1 describes the calculation of servo position and velocity from the position of the person. In our case, we use the position in pixels of the person p_v at τ in an image of size $(u \times v)$.

V. Experimental results and discussion

A. Robotic platform

Storki is a custom build robot made of an omnidirectional base, an orientable head and a large-size tactile display, see Fig. 1. The head is equipped with an Intel RealSense D455 stereo camera, and an Intel RealSense Face ID F450 used for face identification. It also includes a commercial microphone circular array from SeeedStudio[5]. This array has 4 MEMS microphones forming a circle with a diameter of 6.48 cm. The head has been designed with a pan DoF using a Dynamixel R64 servo. The RGBD camera and the servo are connected to the Orin's USB ports. The robot is also equipped with an embedded NUC i7 computer, where the CORTEX instance is run.

B. Follow Me use case

Initially, the robot remains in a waiting state while it has no visual contact with the human. The person might call the robot at some point using a keyword, and at that moment the robot initiates the rotation process based on an estimation of the direction of the speaker. Once in alignment with the person, a two-part process of detection and recognition of the person takes place. The interaction between the robot and the person takes place in the form of a dialogue after a positive identification. The action begins when the person uses the tracking keyword (”follow me”) [6]. Fig. 5 shows, along a tracking process, both the distance between the person and the robot and the normalised values of the orientation of the base in the range $[-\pi/2, \pi/2]$, and the deviation of the centre of the ROI with respect to the centre of the image in the case of the camera tracking the person. On the other hand, the sections in which the person is moving are defined.

The independence between the visual tracking system and the motion control of the base becomes apparent as the user moves around. The visual system has greater agility and the ability to track the person against the base moving. Obstacles and a slower rotation speed increase the convergence time of the alignment to the person. Fig. 6b shows the moments before the person's loss. The visual system is able to correct the displacement of the person at a higher speed, while the base follows a similar adjustment process at a lower speed, arriving at the loss of the person as soon as he or she leaves the robot's field of vision. The tracking process resumes once the person has been re-identified. If the individual stops, as can be seen in Fig. 6a, the robot will correctly orient itself towards the individual and approach until it reaches a socially acceptable distance.

VI. Conclusions

This paper describes a multi-modal and distributed system for person tracking for social robots. The main objective is to develop an attentional mechanism that actively focuses its attention on the person with whom the robot interacts. The system uses video (RGBD sensors) and audio modalities to estimate the location of people of interest and adds a unique identifier to have personalized dialogues. The data fusion is performed in real-time and is accessible for use in high-level robot behaviours. In particular, we have validated the proposed solution on Follow Me use case.

VII. Acknowledgements

This work has been partially supported by FEDER funds, the Spanish Ministry of Science and Innovation TED2021-131739B-C22, PDC2022-133597-C41 and the FEDER Project 0770_EuroAGE2_4_E (Interreg V-A Portugal-Spain (POCTEP) Program).

References

[1] Vega, A., Manso, L., Macharet, D., Bustos, P., and Núñez, P. ”Socially Aware Robot Navigation System in Human-populated and Interactive Environments based on an Adaptive Spatial Density Function and Space Affordances”, in Pattern Recognition Letters, vol. 118, pp. 72–84, 2019.

[2] Kirby, R. ”Social robot navigation”. Carnegie Mellon University, 2010.

[3] Alves, C., and Paro-Costa, P. ”Multimodal social scenario perception model for initial human-robot interaction”, in XXXII Conference on Graphics, Patterns and Images, 2019.

[4] Fong, T., Nourbakhsh, I., and Dautenhahn, K. ”A survey of socially interactive robots”, in Robotics and Autonomous Systems, vol. 42, pp. 143–166, 2003.

[5] Nguyen, D., Li, W., and Ogunbona, P. ”Human detection from images and videos: a survey”, in Pattern Recognition, vol. 51, pp. 148–175, 2016.

[6] Chen, Y., Tian, Y., and He, M. ”Monocular human pose estimation: a survey of deep learning-based methods”, in Computer Vision and Image Understand vol. 192,pp. 102–897, 2020.

[7] Ansari, M., and Kumar, D. ”Human detection techniques for real time surveillance: a comprehensive survey”, in Multimedia Tools and Applications volume 80, pp. 8759–8808, 2021.

[8] Dargan, S., Kumar, M., Ayagari, M, and Kumar, G. ”A survey of deep learning and its applications: a new paradigm to machine learning”, in Archives of Computational Methods in Engineering, pp. 1-22, 2019.

[9] Watada, J., Musa, Z., Lakhmi C., and Fulcher, J. ”Human Tracking: A State-of-Art Survey”, in International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, 2010.

[10] Iguernaissi, R., Djamal M., Aziz, K., and Drap, P. ”People tracking in multi-camera systems: a review”, in Multimedia Tools and Applications, vol. 78, pp. 10773–10793, 2019.

[11] Satsangi, Y. ”Active perception for person tracking”, in PhD Thesis, University of Amsterdam, 2019.

[12] Barıs¸, B., and Gokhan, I. ”Audio-visual Multi-person Tracking for Active Robot Perception”, in IEEE/SICE International Symposium on System Integration. 2015.

[13] Chen, S., Li, Y., and Kwok, N. “Active vision in robotic systems: A survey of recent developments”, in the International Journal of Robotics Research, vol. 30, no. 11, pp. 1343–1377, 2011.

[14] Hall, E.T. ”The Hidden Dimension: Man's Use of Space in Public and Private”. The Bodley Head Ltd, 1969.

[15] Silverman, H.F., and Brandstein, M.S. ”Microphone arrays: signal processing techniques and applications”, in Brandstein, M.S., Ward, D. (Eds.) Springer, 2001.

[16] Marti, A. ”Multichannel audio processing for speaker localization, separation and enhancement”, in Ph.D. Thesis. Universitat Politècnica de València, 2013.

[17] Wang, Chien-Yao and Bochkovskiy, Alexey and Liao, Hong-Yuan Mark ”YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors”, arXiv, 2022.

[18] Yifu Zhang and Peize Sun and Yi Jiang and Dongdong Yu and Zehuan Yuan and Ping Luo and Wenyu Liu and Xinggang Wang ”ByteTrack: Multi-Object Tracking by Associating Every Detection Box”, CoRR, abs/2110.06864, 2021

[5] https://seedstudio.com

[6] The video of the full experiment can be viewed at https://drive.google.com/file/d/1tOIdBa37hdRMPJ1ycUy_BKYVc4TbvS1x/view?usp=share_link.