Journal of Intelligent & Robotic Systems (2024) 110:30
https://doi.org/10.1007/s10846-023-02037-4

REGULAR PAPER

Fusion of Time-of-Flight Based Sensors with Monocular Cameras for a Robotic Person Follower

José Sarmento[1,2] · Filipe Neves dos Santos[1] · André Silva Aguiar[1] · Vítor Filipe[1,2] · António Valente[1,2]

Received: 12 April 2023 / Accepted: 13 December 2023 / Published online: 3 February 2024

Abstract
Human-robot collaboration (HRC) is becoming increasingly important in advanced production systems, such as those used in industries and agriculture. This type of collaboration can contribute to productivity increase by reducing physical strain on humans, which can lead to reduced injuries and improved morale. One crucial aspect of HRC is the ability of the robot to follow a specific human operator safely. To address this challenge, a novel methodology is proposed that employs monocular vision and ultra-wideband (UWB) transceivers to determine the relative position of a human target with respect to the robot. UWB transceivers are capable of tracking humans with UWB transceivers but exhibit a significant angular error. To reduce this error, monocular cameras with Deep Learning object detection are used to detect humans. The reduction in angular error is achieved through sensor fusion, combining the outputs of both sensors using a histogram-based filter. This filter projects and intersects the measurements from both sources onto a 2D grid. By combining UWB and monocular vision, a remarkable 66.67% reduction in angular error compared to UWB localization alone is achieved. This approach demonstrates an average processing time of 0.0183s and an average localization error of 0.14 meters when tracking a person walking at an average speed of 0.21 m/s. This novel algorithm holds promise for enabling efficient and safe human-robot collaboration, providing a valuable contribution to the field of robotics.

Keywords Person detection · UWB · Machine vision · Sensor fusion · HRC · Agricultural robotics

1 Introduction

Human-robot interaction is an essential research topic due to the increasing presence of robots in workspace environments [1, 2]. One of the most important capabilities that robots should have is path-following. This allows robots to autonomously perform certain tasks alongside humans without manually asking the robots to change their position in the workspace [3]. Implementing person-following algorithms on autonomous robots is challenging because they should consider the dynamics of the environments where the interaction occurs. Thus, approaches should consider occlusions, appearance changes, pose changes, crouching, or illumination changes [4]. In outdoor environments, such as agriculture, robot mobility is also conditioned by the irregularities of the terrain, which the robot should take into account to travel along safe paths [5–7]. When performed successfully, person-following algorithms significantly contribute to human labour conditions in these environments. For example, robots can be used in harvesting to assist humans in transporting fruit or vegetables to the storage area, thus reducing physical effort [8]. The primary objective of this paper is to advance the state-of-the-art in Human-Robot Collaboration (HRC) systems, focusing on the precise recognition and tracking of humans. The aim is to enable seamless human-robot collaboration by implementing a fully functional system capable of accurately determining the relative

✉ José Sarmento
jose.m.sarmento@inesctec.pt

Filipe Neves dos Santos
filipe.n.santos@inesctec.pt

André Silva Aguiar
andre.s.aguiar@inesctec.pt

Vítor Filipe
vfilipe@utad.pt

António Valente
avalente@utad.pt

[1] INESC Technology and Science (INESC TEC), 4200-465 Porto, Portugal

[2] Department of Engineering, University of Trás-os-Montes e Alto Douro (UTAD), 5000-801 Vila Real, Portugal

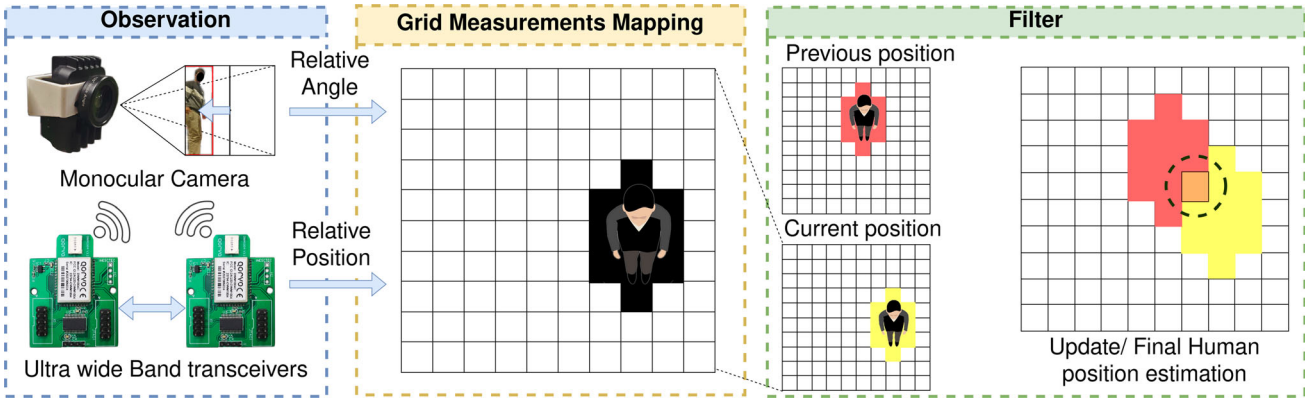


Fig. 1 Proposed system overview

position of a person with respect to a robotic platform, as depicted in Fig. 1. Building upon previous work [9], where a robot that follows humans was developed using Ultra-WideBand (UWB) transceivers and a histogram filter-based approach for localization, we present a significant improvement to enhance the UWB transceiver's uncertainty.

Our proposed system leverages sensor fusion, combining data from UWB transceivers and monocular cameras employing Deep Learning for person detection. Specifically, we investigate the method of person localization using monocular cameras and address the challenge of low field of view (fov) in such camera systems.

The key contributions of this work are as follows:

- Development of a novel approach to determine the relative position of a person with respect to the robotic platform, employing onboard UWB transceivers and a pair of monocular cameras.
- Introduction of sensor fusion through grid interception and a 2D histogram filter-based approach, significantly enhancing the accuracy of person localization.
- Creation of a publicly available dataset containing person detection data from both UWB and monocular cameras, facilitating further research and benchmarking in this domain [1,2].

By addressing these crucial aspects, we aim to pave the way for more efficient and safe human-robot collaboration systems, contributing significantly to the field of robotics.

The remainder of this paper is structured as follows. Section 2 presents the current state of the art in this field. Section 3 provides a description of the hardware used. Section 4 describes the implementation of the proposed sensor fusion using the interception between UWB transceivers and monocular cameras in a 2D grid and the use of the Range Histogram Filter to obtain relative person localisation. Section 5 evaluates the proposed methods using real work data and an in-house built robot. And finally, Section 6 provides the main conclusion extracted from the developed work.

2 Related Work

2.1 Single Sensor Localisation

Recent developments in person localization have delved into diverse sensor technologies, including Color and Depth (RGB-D) cameras [10–13], Light Detection and Ranging (LiDAR)[14–16], and Ultra Wide Band (UWB)[17–21].

RGB-D cameras are a popular choice for single-sensor person localization as they can use deep learning techniques from monocular cameras [22–24] to detect people and stereo vision to extract depth information, enabling person localization. However, RGB-D cameras have a limited field of view, making it difficult to track a person moving out of the camera's view. LiDAR is another popular choice for single-sensor person localization. LiDAR sensors can provide accurate range measurements, which can be used to estimate the position of a person. However, LiDAR sensors are frequently costly, challenging to deploy, and notably vulnerable to subject occlusion. UWB, a technology that has been in existence for several decades, has recently seen a surge in personal location. UWB sensors excel in precise distance measurements, even in obstructed environments, making them highly effective for HRC systems. In addition, their cost-effectiveness and ease of deployment have contributed to their growing adoption. A key advantage of UWB technology is its ability to penetrate obstacles, such as walls and structures, which can hinder LiDAR and visual detection methods. Unlike machine vision, which often faces limitations in its field of view, UWB offers a wider detection range.

[1] https://doi.org/10.5281/zenodo.7779131

[2] https://doi.org/10.5281/zenodo.7818328

Moreover, compared to LiDAR and visual detection methods, UWB can deliver a higher update rate and provides more detailed distance data at the decimetre level.

2.2 Multi-sensor Localisation

Multi-sensor person localization techniques have emerged as effective solutions to enhance the accuracy of person tracking, surpassing the limitations of single-sensor approaches. One notable example is the work of Jinu et al. [25], who fused an Angle of Arrival (AoA) sensor with a 2D LiDAR sensor to improve the accuracy of person tracking. The AoA sensor is able to reliably detect the target, but it gives very noisy results. The 2D LiDAR sensor, on the other hand, has a lower error but is more prone to misidentifying the target. The fusion of these two sensors using a Kalman filter approach was able to reduce the overall uncertainty of the system. Liu et al. [26] proposed a system for person localization that is similar to the one developed in this paper, utilizing UWB transceivers and a monocular camera. The main difference between their system and ours is the placement of cameras and anchors. In their work, the cameras and anchors are fixed and confined to a specific space, whereas in our system, they are mounted on a robotic platform. This adaptability enables us to track people in more dynamic environments. Luchetti et al. [27] and Conejero et al. [8] present similar systems that use sensor fusion with a Bayes filter between UWB transceivers, which have higher uncertainty in the angular direction, and RGB-D cameras, in which the uncertainty is predominant along the radial coordinate, to obtain a more accurate and robust dynamic person tracking system. These systems are very similar to the one proposed in this paper in the sense that they use cameras and UWB anchors fixed to the robot, enabling dynamic person tracking. However, these systems differ from ours in the use of stereo systems instead of monocular cameras. Furthermore, in our system, we utilize a fusion technique between cameras and UWB transceivers known as the Range Histogram Filter (RHF) [9] for person tracking. Unlike the Bayes filter, the RHF allows us to utilize non-Gaussian probability models, making it more effective in handling uncertainty.

2.3 Proposed Approach

In conclusion, this paper proposes a novel approach for robust person detection by utilizing a UWB transceiver with fixed anchors on the robot and a tag on the target. UWB technology was chosen for its capabilities, such as obstacle penetration, higher update rates, and decimeter-level measurements, which outperform other person detection techniques such as LiDAR and visual detection.

(a) Front view. (b) Side view.

Fig. 2 Experimental robotic platform. The green circle surrounds the UWB tag, the red circles surround the UWB anchors, the yellow circles surround the OAK-1 cameras and the blue circle surrounds the Livox Mid-70 Lidar

However, UWB is prone to significant radial errors. Therefore, a multi-sensor approach is proposed that also incorporates visual person detection, similar to works such as [8, 27], but with the added advantage of using multiple monocular RGB cameras to increase the field of view. This is because using monocular cameras allows for a larger field of view with smaller dimensions and costs than stereo cameras.

Finally, the Range Histogram Filter (RHF) [9] will be utilized for the UWB and camera fusion, as well as for person tracking.

3 Hardware

The hardware components utilized in this study are outlined in this section. The custom-built robot, referred to as Model-E, is equipped with QORVO DWM1004 transceivers for UWB ranging. For person detection, Luxonis OAK-1 cameras with a performance of 1.4 Trillion Operations Per Second (TOPS) are used for AI applications. Additionally, for data collection in Section 5, a Livox Mid-70 LiDAR is also mounted on the robot. All mounted devices can be seen in Fig. 2.

Table 1 Sensor placement measurements relative to the centre of the robot

Sensor	Position	x (m)	y (m)	yaw (rad)
Camera	1	0.267	0.037	0.459
	2	0.267	-0.037	-0.459
UWB	1	-0.235	0.0	0.0
	2	0.235	0.280	0.0
	3	0.235	-0.280	0.0
LiDAR		0.25	0.0	0.0

The numbering of the sensors, i.e., Camera 1, corresponds to the sensor positions illustrated in Fig. 3

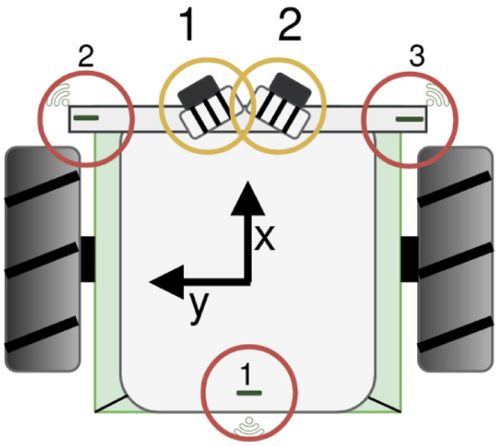


Fig. 3 Positioning of sensors with respect to the centre of the robot. Yellow and red circles surround the cameras and UWB transceivers, respectively

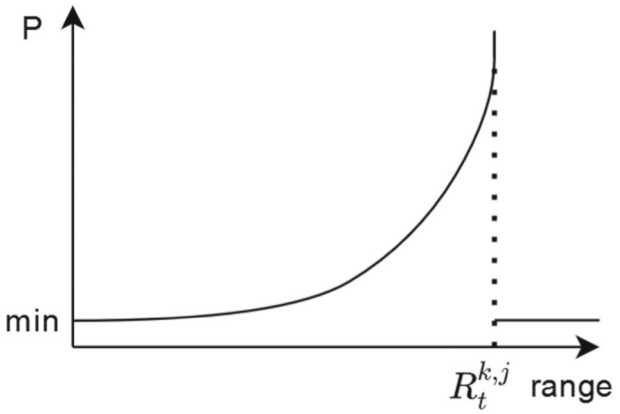


Fig. 4 Probability density function of UWB transceiver range measurements for range sensor mapping onto the grid

The relative positions of all sensors to the centre of the robot were determined through measurements. The custom support holding the cameras was intentionally set to the desired angle. These values were recorded in Table 1 and are illustrated in Fig. 3.

4 Person Localisation from the Fusion of UWB as Range Sensor and Monocular Person Detection: The Main Approach

4.1 UWB as Range Sensors

UWB transceivers can be used as range sensors through techniques such as Time of Arrival (ToA) or Angle of Arrival (AoA). Compared to other localization techniques, UWB transceivers offer significant advantages in their robustness against obstacles, making them an ideal choice for person localization in Human-Robot Collaboration (HRC) systems. Their ability to penetrate walls and other barriers allows for precise range measurements even in obstructed environments, enabling accurate tracking of individuals even behind obstacles. This unique capability enhances the effectiveness and reliability of person localization in various real-world scenarios.

The most typical configuration for UWB localization is to use fixed anchors in relation to the world to locate a tag. However, in our case, we want to locate people relative to a moving robot. This means that the anchors must be static relative to the robot. In previous work [9], this approach was validated. The relative distance between the UWB transceivers fixed on the robot (anchors) and the mobile UWB transceiver on the person (tag) was obtained using Time of Arrival (TOA) techniques. To improve the accuracy of the localisation, a novel algorithm called Range Histogram Filter (RHF) was developed. The RHF is a probabilistic filter that works

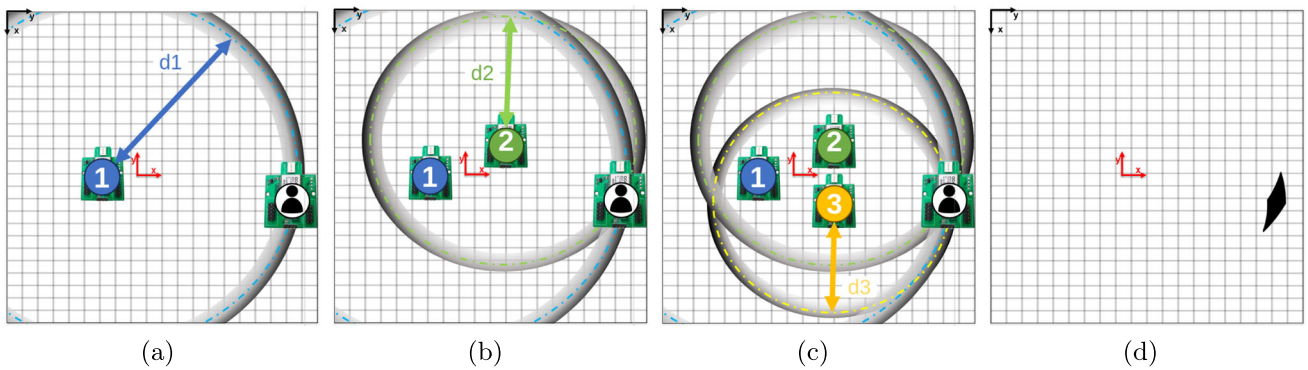


Fig. 5 Stages of UWB Person Localization observation grid creation: (**a**), (**b**), and (**c**) show all UWB anchor detections of the tag mapped in the 2D grid, and (**d**) depicts the intersection of all UWB transceivers. Please note that in all these grids, the black and red x and y axes represent the grid and the robot-center axes, respectively

on 2-Dimensional grids centred on the robot. This filter is explained in [9] and will be further explained in this paper in Section 4.3.

There are three main grids in the range histogram filter: the observation grid, the innovation grid, and the update grid. In this subsection, we will only discuss the observation grid.

The observation grid is where the UWB range detections are mapped with an exponential probability distribution (Fig. 4).

Each UWB transceiver generates a circular region with an exponential probability distribution along the radial axis, as shown in Figs. 5a, b and c. In these figures, d1, d2, and d3 represent the UWB range measurements from transceivers 1, 2, and 3, respectively, to the person tag (according to Table 1). The intersection of all UWB transceivers is shown as the dark area in Fig. 5d. This is the region where the presence of a person is most likely. Note that the width of this area is greater along the angular direction than the radial direction.

The aim of this paper is to reduce this angular error by intersecting the UWB observation grid with the observation grid obtained from person detections acquired by the cameras. Subsequently, the intersection of the two grids will undergo filtering using the RHF, enabling the determination of the position of the target. By leveraging this novel approach, it will be possible to enhance the accuracy and reliability of person localization.

4.2 Angle Estimation with Monocular Vision

To complement UWB person localisation and reduce angular uncertainty, an approach based on Deep Learning for person detection is proposed. This method uses cameras to detect people and determine their bounding boxes. The angle between the subject and the camera is then determined from the bounding box, which can then be used to determine the angle between the person to track and the robot. This fusion proves advantageous for vision as well, as UWB signals exhibit greater resistance to occlusion compared to vision. This capability addresses the inherent occlusion challenges present in vision-based methods while reducing UWB person detection angular error.

4.2.1 Camera Setup

Maximizing visual coverage is essential to complement UWB person detection. Therefore, using multiple monocular RGB cameras to increase the field of view is an effective approach. Compared to stereo cameras, monocular cameras have a wider field of view, are more compact, and are less expensive. Additionally, as UWB person detection is already accurate in the radial direction, stereo cameras would not provide a significant advantage. As a result, multiple monocular cameras are a simpler and less expensive solution for person detection. The proposed vision system comprises of two monocular cameras with additional lenses for maximum field of view coverage, as shown in Fig. 6. The cameras are strategically arranged to minimize intersection, ensuring robust detections without dead zones while maximizing the field of view.

To enable angle estimation, determining the focal length of the cameras is essential. Kalibr, an interactive framework for camera calibration, was used for this purpose [28, 29]. The resulting horizontal field of view of both cameras combined was 146 degrees, which is considered sufficient since the primary goal of the human-tracking robot is to face the person within this field of view in most cases.

Furthermore, as the proposed UWB and camera fusion algorithm (RHF) uses 2D grids, it would be easy to add more cameras, as each new camera would simply add to the camera observation grid. However, the current field of view is sufficient for the scope of this paper, which focuses on the fusion between monocular vision and UWB.

Moreover, the UWB fusion can compensate for scenarios where the robot is temporarily out of the camera's sight or occluded, as person detection remains possible. This raises the question of whether additional cameras would be effective in such cases.

4.2.2 Angle Estimation using Deep Learning techniques

The proposed method for calculating the angle between an object and a robot requires determining the relative position of the bounding box, which is obtained from Deep Learning-based object detection in the images captured by the cameras. Deep learning object detection employs neural networks

Fig. 6 OAK-1 with a 120 degrees lens

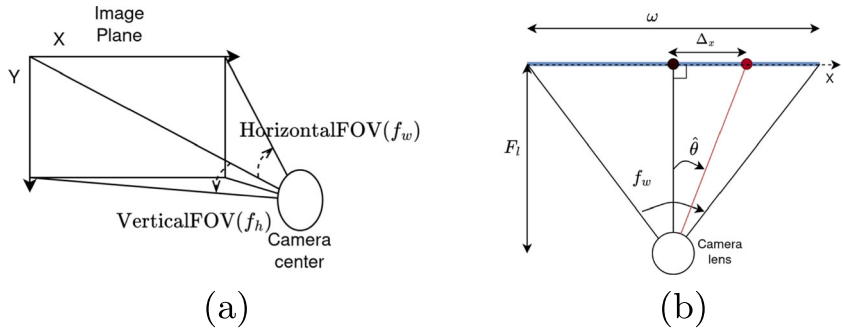


Fig. 7 Illustration of the Angle Estimation Method: (**a**) Depicts Vertical and Horizontal Perspectives. (**b**) Illustrates the Camera's Horizontal View with the Red Line Indicating the Measured Angle.

trained on labelled datasets to detect objects in images or videos. As demonstrated in Section 2, person detection using Deep Learning has been extensively researched, resulting in an abundance of resources. Notably, large datasets and pre-trained models, which simplify its implementation. To enable object detection, two OAK-1 cameras were used. These cameras can perform Deep Learning object detection offline. The company that developed the cameras, depthai, provides pre-trained Deep Learning object detection models and implementation examples [3]. Since the goal is to detect people, a pre-trained MobileNetv2-SSD model provided by depthai is used. This model is designed for object detection and is the combination of a MobileNetv2 model[30] and the SSD architecture[31]. In this context, the model has been trained on a dataset of labelled images which includes the class person.

To obtain an angle from the bounding box of the person detection, the horizontal distance in pixels between the reference and the detection must be converted to an angle. Therefore, the pinhole camera model is considered, as represented in Fig. 7. Considering only the horizontal axis, as shown in Fig. 7b, it can be concluded that the angle can be obtained from the displacement of the detection towards the centre of the image (Δ_x). Knowing that the focal length F_l is constant, the following relation is obtained in Eq. 1.

$$\hat{\theta} = \arctan\left(\frac{2\Delta_x \cdot \tan\left(\frac{f_w}{2}\right)}{\omega}\right) \tag{1}$$

where f_w represents the image's horizontal field of view and ω the image width in pixels.

The ultimate goal of the system is to intercept UWB measurements with camera angle measurements using RHF. Therefore, the camera measurements must be placed in the observation grid. For this purpose, the angles from the centre of the bounding box and its extremities will be discovered. Then the angle that derives from the centre of the bounding box is assumed to have the highest probability of being the person. The angles covered by the bounding box region then follow a normal probability distribution, as shown in Fig. 8, and are mapped into the 2D camera observation grid as shown in Fig. 9.

[3] https://github.com/luxonis/depthai-ros-examples

4.3 UWB Transceivers and Monocular Camera Fusion with the Use of RHF

This subsection discusses the implementation of the Range Histogram Filter (RHF) [9] for sensor fusion. RHF is based on a histogram filter, which proves highly advantageous for fusion in this context. It can effectively handle exponential (person localization by UWB) and Gaussian (person localization by the camera) systems, allowing the representation of complex distributions. The filter can be decomposed into three stages:

- **Observation** - Combination of the UWB and camera measurements;
- **Innovation** - Estimation of the next iteration;
- **Update** - Interception of the estimation and measure parts of the filter.

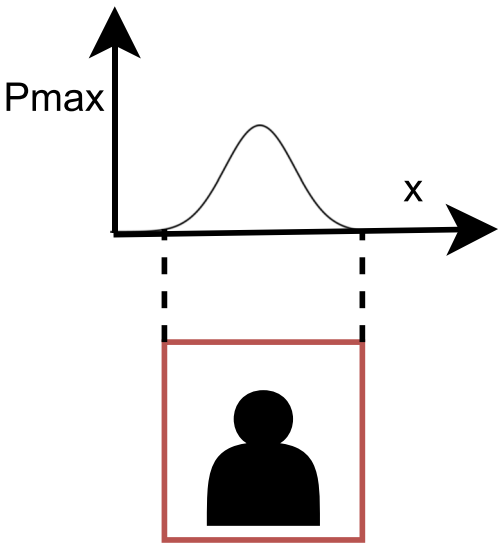


Fig. 8 Normal distribution mapping to 2D grid from bounding box. P_{Max} is the Max probability value and x is the horizontal coordinate of the image

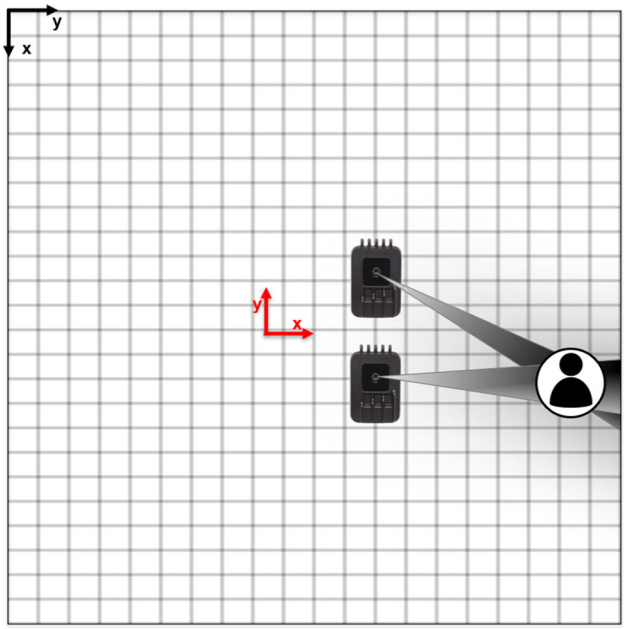


Fig. 9 Mapping of Angles Obtained from Monocular Camera Person Detections onto a 2D Grid. Please note that in all these grids, the black and red x and y axes represent the grid and the robot-center axes, respectively

The filter can be described by the Eq. 2, where the grid M is the interception of the sensor observations M^O (Eq. 4) and the innovation M^I (Eq. 7) grids. Note that M^I is from the previous time step $t-1$, whereas M^O and M are from the current time step t. This is because M^I at time $t-1$ represents the predicted position of the person at time t.

$$M_t = M_t^O \cap M_{t-1}^I \tag{2}$$

After creating a grid, the map probability undergoes normalization, making the probability values comparable for appropriate grid weighting during interception. Additionally, a minimum detection probability value is defined to give more weight to grids more likely to contain the person. These operations are described by Eq. 3, where probabilities exceeding the threshold $P_{threshold}$ are set to p/P_{Max}, with P_{Max} being the maximum probability value on the grid. If the probability value is below $P_{threshold}$, it is set to P_{min}. A smaller P_{min} value increases the significance of the grid.

$$f(x) = \begin{cases} p/P_{Max} & p > P_{threshold} \\ P_{min} & p < P_{threshold} \end{cases} \tag{3}$$

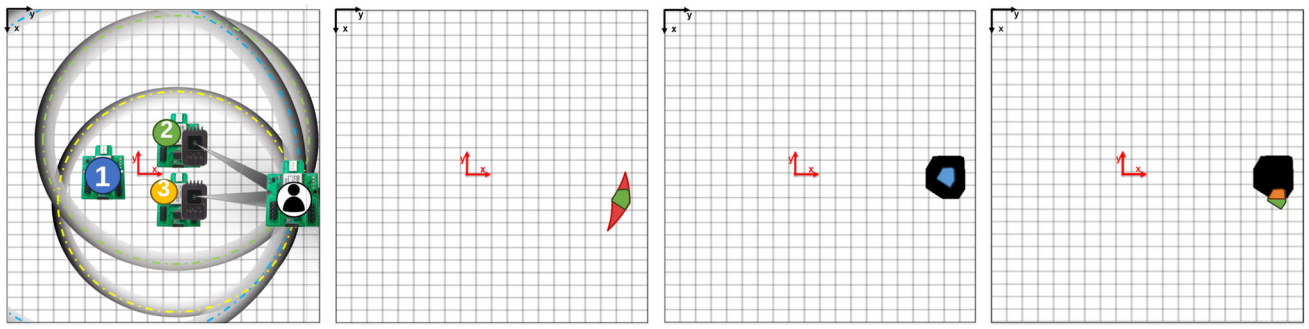
The observation grid M^O (Eq. 4) is formed by intersecting the grids generated by UWB transceivers measurements M^{uwb} (Eq. 5) and the person angle estimation from the cameras M^{cam} (Eq. 6).

Figure 10a illustrates the intersection of M^{uwb} and M^{cam}, resulting in the M^O grid, which is highlighted in green in Fig. 10b. Additionally, Fig. 10b shows a red area representing the intersection of only UWB measurements, as previously shown in Fig. 5d. Importantly, the angular dimension of the green area is smaller than that of the red, indicating that the monocular vision fusion with UWB transceivers successfully reduced the error, as anticipated.

$$M_t^O = M_t^{uwb} \cap M_t^{cam} \in [P_{min}^{obs}, 1] \tag{4}$$

$$M_t^{uwb} \in [P_{min}^{uwb}, 1] \tag{5}$$

$$M_t^{cam} \in [P_{min}^{cam}, 1] \tag{6}$$



(a) (b) (c) (d)

Fig. 10 Integration of UWB and Monocular Camera Data for Person Detection: **(a)** Sensor values from UWB transceivers and camera detections integrated into the grid. The grid's centre aligns with the robot's centre. **(c)** Depicts the intersection of the measurements from UWB transceivers and camera detections depicted in **(a)**. **(d)** Represents the previously generated innovation. **(e)** Displays the intersection of the current measurement with the previous innovation. Please note that in all these grids, the black and red x and y axes represent the grid and the robot-center axes, respectively

The primary objective is for the system to track a designated target precisely. In this context, UWB transceivers excel at isolating the target, a capability that vision-based person detection lacks. Therefore, the UWB grid is assigned greater importance in the resulting intersection than the camera grid. Consequently, the minimum probability of the camera-generated grid must also be higher: $P^{uwb}min < P^{cam}min$.

The innovation grid M_t^I (Eq. 7) is a prediction of the possible next person positions, taking into account the last M_t. The prediction is made assuming that the person is moving with a maximum velocity U. This means that the updated grid M_t is dilated as a function of U. Note that M_t^I is generated at time t, but is a prediction for $t+1$. Therefore, it will only be used in the next update cycle. Figure 10c represents the innovation grid M^I at time $t-1$. The blue area is the result of the M grid at time $t-1$. The dark area is the dilation of the blue area according to a certain U.

$$M_t^I = M_{t-1} + U \in [P_{min}^{inov}, 1] \tag{7}$$

Finally, the update phase M_t (Eq. 2) is the intersection of the current observation grid M_t^O and the previous innovation grid M_{t-1}^I. However, since it is unlikely that a person moves with a speed higher than the predefined value, more importance must be given to the innovation grid: $P_{min}^{inov} < P_{min}^{obs}$. Figure 10d shows M_t (orange area), which is the intersection of M_t^O (green area) and M_{t-1}^I (dark area).

5 Results

To evaluate the person localisation achieved by the sensor fusion proposed in this article, tests were developed to assess the performance of the UWB range sensors with and without monocular vision fusion. The real experiments conducted in a laboratory scenario were performed. The computer used for testing the proposed algorithm is a laptop with an i7-7700HQ CPU running at 2.80GHz and 16GB of RAM. The proposed algorithms and the required drivers were implemented using C++ and the ROS2 framework on Ubuntu LST 20.04.

5.1 Range Evaluation

A range benchmark was performed to determine the minimum and maximum range of person following. This was done by measuring with the LiDAR the minimum and maximum distances in which measurements from UWB transceivers and a monocular camera were being acquired. Note that these results were obtained within an indoor corridor environment. This context introduces potential challenges, such as sub-optimal lighting conditions that could impact camera performance. Additionally, the presence of walls can influence UWB transceiver wave propagation, leading to reflections and potential signal interference. While walls can extend UWB range, they can also introduce outliers due to signal interactions and variations in wave path lengths, potentially leading to inaccurate measurements. To determine the position of a person with LiDAR, an algorithm was developed. The algorithm requires that the detection area be empty so that only the point cloud from the person is generated. The position is then estimated by the centre of the point cloud in 2D space, and the range measurement is determined by the Euclidean distance of the person relative to the LiDAR. Finally, the range measurements are transformed to the reference frame of each sensor. This is done by taking into account the values of sensor placement measurements in Table 1. The results are presented in Table 2.

It is crucial to highlight that the minimum distance achieved by the proposed person-tracking algorithm using LiDAR was 0.3124 meters. On the contrary, the LiDAR's maximum range wasn't reached, as it continued to measure even at the end of the corridor (15 metres). This implies that the minimum distance achievable by the UWB and camera systems could potentially be even shorter. The results reveal two significant findings. To begin, a defined limit of 12 metres must be established for the person detection algorithm. Second, the best operating range is determined when both cameras and UWB transceivers are actively gathering data, at ranges of less than 7 metres.

5.2 Standing Test

To evaluate the performance of the proposed algorithm for person localization, four specific test points were chosen strategically. Each point was positioned at a fixed distance of 1.5 meters from the centre of the robot. These test points were designed to represent scenarios where the fusion of UWB and vision measurements was both possible and not possible. A visual representation of the ground truth positions and the corresponding person detections achieved by the algorithm is provided in Fig. 11. Measurements were taken statically at these points and compared with ground truth. The average error and standard deviation were calculated to evaluate the system. To better evaluate the angular error of UWB transceivers, a common defect, and its improvement

Table 2 Range Benchmark Results

Sensor	Min. Distance (m)	Max. Distance (m)
Cameras	0.2937	7.5520
UWB	0.3924	12.7220

This table provides information on the maximum and minimum distances, recorded in meters, from the LiDAR sensor in the corresponding sensor frame where the measurement is detected

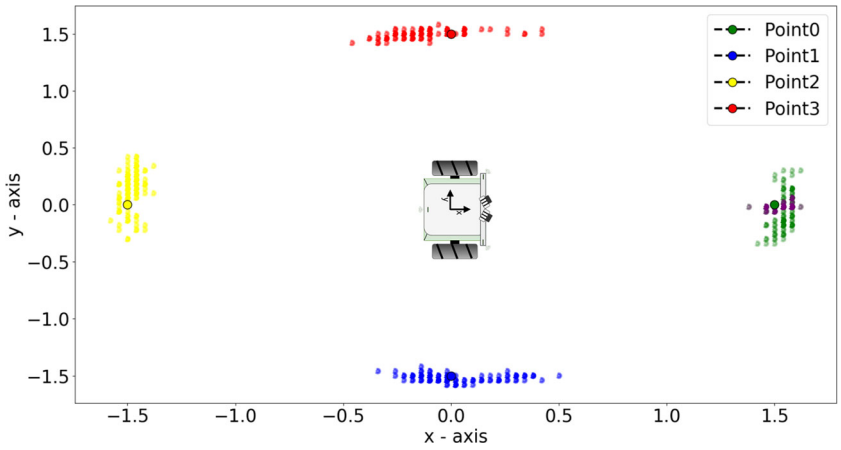


Fig. 11 Stationary person detection results. Green, blue, yellow, and red indicate visionless person localization at predefined positions. Purple represents the results of the proposed vision fusion algorithm. The robot is centered at (0,0)

with fusion, the error was calculated in both Cartesian and polar coordinates. This allowed for a better observation of the angular error. Additionally, to benchmark the improvement of the filter, the algorithm was run in a region where vision is possible (point 0), with and without fusion. This enabled the comparison of the performance of the algorithm with and without fusion. The resulting error for each position can be seen in Table 3. Considering the results at point 0, where the UWB person localisation algorithm is tested with and without the monocular vision fusion technique, the difference between the two is clear, either in the visual analysis of Fig. 11 or in the results in Table 3. The proposed fusion of UWB and monocular vision leads to a 47.75% reduction (from 0.111m to 0.058m) in the radial error and 66,67% (from 0.056rad to 0.015rad) in the angular error of person localisation compared with the results without fusion. Furthermore, it can be observed that the points where fusion is not possible (points 1, 2, and 3) since they are outside the field of view of both cameras, have significantly higher radial and angular error values, which is to be expected since the person localisation algorithm is based only on UWB.

Table 3 Error results for position estimation against a set of predefined positions

Point	Error Type	Avg. Error Without Fusion	With Fusion
0	ϵ_r (m)	0.111 ± 0.075	0.058 ± 0.023
	ϵ_θ (rad)	0.056 ± 0.056	0.015 ± 0.007
	ϵ_x (m)	0.049 ± 0.027	0.049 ± 0.027
	ϵ_y (m)	0.087 ± 0.086	0.023 ± 0.011
1	ϵ_r (m)	0.141 ± 0.099	-
	ϵ_θ (rad)	0.086 ± 0.069	-
	ϵ_x (m)	0.131 ± 0.106	-
	ϵ_y (m)	0.028 ± 0.023	-
2	ϵ_r (m)	0.187 ± 0.092	-
	ϵ_θ (rad)	0.093 ± 0.101	-
	ϵ_x (m)	0.023 ± 0.027	-
	ϵ_y (m)	0.183 ± 0.093	-
3	ϵ_r (m)	0.175 ± 0.099	-
	ϵ_θ (rad)	0.113 ± 0.067	-
	ϵ_x (m)	0.170 ± 0.100	-
	ϵ_y (m)	0.023 ± 0.023	-

The error is evaluated in both polar coordinates (r,θ) and Cartesian coordinates (x,y), where r and θ are the distance and angle between the robot and the detection. x and y are the axes of the grid

5.3 Moving Square Test

Since the ultimate goal is to obtain a robot that tracks people, it is important to evaluate the localisation of the person as a moving target. To this end, a test was conducted in which a person moved in a square perimeter covered by the cameras and the UWB transceivers. To determine the actual position of the person, LiDAR was used in the same manner as discussed in the previous test in Section 5.1. The proposed test is shown in Fig. 12.

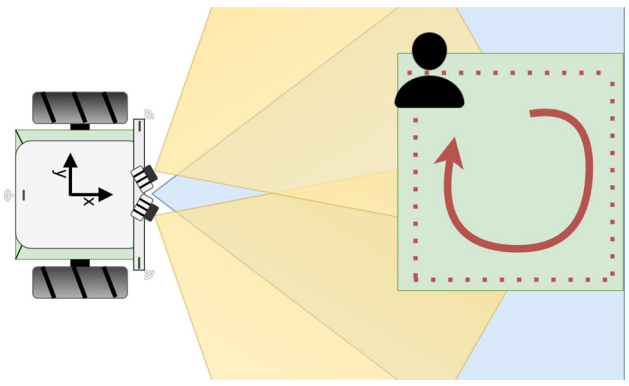


Fig. 12 Test setup diagram: Yellow, blue, and green areas denote the camera's fov, LiDAR's operational range, and the person's movement zone, respectively

Table 4 Position estimation error results when the person is at a slow pace varying expected person speed

Expected Person Speed (m/s)	Min. Error (m)	Max. Error (m)	Avg. Error (m)
1.94	0.0041	0.5869	0.1404 ± 0.0883
1.39	0.0041	0.5853	0.1388 ± 0.0877
0.83	0.0041	0.5453	0.1378 ± 0.0869

In this test, the average speed travelled by the subject was 0.2071 ± 0.0785 m/s, with the highest value being 0.3966 m/s and the lowest being 0.0685 m/s. The error values are presented in meters

Table 5 Position estimation error results when the person is at a fast pace varying expected person speed

Expected Person Speed (m/s)	Min. Error (m)	Max. Error (m)	Avg. Error (m)
1.94	0.0039	0.7601	0.2247 ± 0.1442
1.39	0.0039	0.7670	0.2205 ± 0.1389
0.83	0.0039	1.1264	0.2884 ± 0.2089

In this test, the average speed travelled by the subject was 0.7147 ± 0.1201 m/s, with the highest value being 0.9210 m/s and the lowest being 0.3994 m/s. The error values are presented in meters. The robot is located on the left side of the graph at (0,0)

In this figure, it is possible to observe that the person follows a predefined square route in a bounded area. Since the range of the UWB transceiver is much larger than that of the camera and the LiDAR, this area is limited not only by the field of view of the camera but also by the LiDAR, which has a limited field of view of 70 degrees. Although the LiDAR is very accurate, the UWB tag was not placed exactly in the centre of the target as required by LiDAR target acquisition. This introduces errors in the test but is still useful to evaluate the detection of the moving target. To evaluate the performance of the actual speed of the target compared to the expected speed input parameter of the innovation part of the filter, a test was first performed at slow speeds, achieving an average speed of 0.2071 m/s. The results can be found in Table 4 and Fig. 13a. Subsequently, a test with increased speed was carried out, achieving an average speed of 0.7147 m/s. The results can be found in Table 5 and Fig. 13b. In addition, Table 6 shows the execution times of the proposed algorithm. The datasets recorded in these experiments, both for slower[1] and for higher[2] speeds, have been made publicly available.

In the first test, where the targeted person moves at lower speeds, the expected speed value of the innovation part of the filter has no significant impact on the selected group. However, in the second test, where the subject speed is increased to values higher than the selected expected speed parameter values, this causes the system to have a higher error than in the first test. Moreover, the error in this second test decreases significantly as the value of the expected velocity increases. From this, selecting a higher expected velocity is preferable. However, the value should not be set excessively high, as this would result in a performance loss of the proposed filter. Analysis of Figs.13a and b also suggests that human detection has an optimal following range, as the error tends to increase the farther the human is from the robot. Based on the recorded times from Table 6, it can also be concluded that this algorithm can be run offline.

5.4 Comparison with State-of-the-Art

In this subsection, the results presented in the previous subsections will be compared to the approaches discussed in the Related Works section (Section 2).

LiDAR exhibits the highest accuracy for person localization. However, it relies on a single technology, which renders it less robust, particularly in scenarios involving occlusion. Additionally, LiDAR is associated with a higher cost.

RGB-D cameras demonstrate accuracy in the angular direction but experience reduced performance in the radial

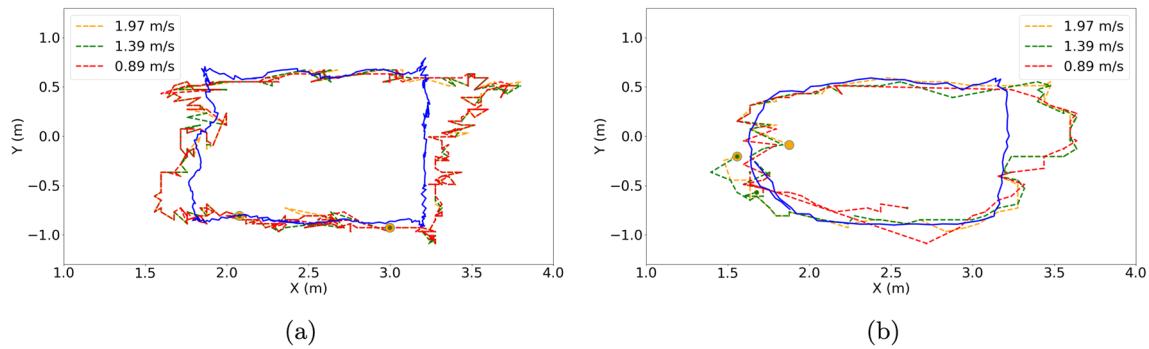


Fig. 13 Results of moving person detection. The orange, green, and red colours represent the moving person detection by the proposed algorithm using different input values for the expected speed of the person for the proposed filter. The blue line is the LiDAR-generated ground truth of a person moving at an average speed of 0.2 m/s **(a)** and 0.7 m/s **(b)**

Table 6 Execution Times of each Section of the Proposed Algorithm in Seconds

Test speed	Algorithm sections	Min. Time) (s)	Max. Time (s)	Avg. Time (s)
Slow	Filter	0.0096	0.0483	0.0183 ± 0.0069
	UWB	0.0097	0.0508	0.0192 ± 0.0069
	Vision	0.0034	0.2500	0.0076 ± 0.0083
Fast	Fiter	0.0095	0.0426	0.0179 ± 0.0062
	UWB	0.0097	0.0508	0.0192 ± 0.0067
	Vision	0.0034	0.0210	0.0074 ± 0.0025

In the algorithm sections, Vision refers to the creation and updating of the camera grid, UWB refers to the creation and updating of the UWB transceivers grid, and finally the filter section refers to the histogram filter, which includes the fusion of both grids, the innovation, and update parts

dimension. Their single-sensor nature leads to limitations in robustness and occlusion handling. Nonetheless, RGB-D cameras are more cost-effective than LiDAR systems.

UWB technology provides accuracy in the radial direction but shows performance degradation in angular measurements. Compared to LiDAR and RGB-D, UWB stands out for its superior target recognition capabilities and resilience to occlusion. Moreover, UWB solutions are cost-effective compared to RGB-D and LiDAR.

The fusion of RGB-D and UWB constitutes a compelling solution, capitalizing on the angular precision of RGB-D cameras and the radial accuracy of UWB. This multi-sensor approach significantly enhances robustness, especially in scenarios prone to occlusion. It effectively counteracts the camera's susceptibility by relying on the UWB's reliability for localization, even though there may be trade-offs in angular precision. Importantly, this combined approach provides a cost-effective alternative to LiDAR.

Our approach integrates UWB technology with multiple monocular cameras, a deliberate choice, given the limited field of view inherent to RGB-D cameras. This combination extends the system's coverage while preserving cost-effectiveness. It consistently performs at a level comparable to RGB-D and UWB systems, which is evident when comparing the results in Section 5.2 with the stationary findings presented by [8]. This alignment with expectations is unsurprising since RGB-D cameras exhibit suboptimal radial performance. Consequently, when combined with UWB, monocular cameras fulfill a similar role, mainly focused on mitigating the angular errors resulting from UWB localization.

A summarized conclusion of these observations can be found in Table 7.

6 Conclusion

This paper presents a method for human tracking in settings involving human-robot collaboration (HRC). In particular, relative person localisation is accomplished by combining monocular vision with ultra-wideband (UWB) transceivers as range sensors.

This method is able to track an individual equipped with a UWB transceiver, utilizing these transceivers as range sensors. UWB transceivers have a large angular error, which can be problematic in human-robot collaboration (HRC) settings where accuracy is essential.

To reduce the angular error, the range information from the UWB transceivers was fused with the angle estimations obtained from a monocular vision system. The vision system uses Deep Learning object detection techniques to detect people and estimate their relative angle to the robot. This method has a very high angular accuracy, and when fused with UWB, it can significantly reduce the overall angular error of the system.

Compared to UWB localisation alone, the experiments showed that combining UWB and monocular vision using the RHF significantly reduced the angular error. The angular

Table 7 Person detection techniques

Approach	Angular	Radial	Robustness	Occlusion	Cost	Ref.
LiDAR	++	++	-	-	-	[14–16]
RGB-D	+	-	-	-	+	[10–13]
UWB	-	+	-	+	++	[9, 18–21]
RGB-D & UWB	+	+	+	+	+	[8, 27]
RGB & UWB	+	+	+	+	+	Proposed approach

error decreased by 66,67% and the radial error by 47.75%. Furthermore, the position of the target was estimated with an average inaccuracy of 0.14 m and 0.22 m at moderate velocities of 0.21 m/s and higher velocities of 0.71 m/s, respectively. From the results presented, where a maximum velocity of 0.92 m/s was achieved, we conclude that this approach can be used up to a velocity of 0.92 m/s to locate the target successfully.

The main advantage of our approach is that the target can be located relative to the robot without the need for fixed sensors in the workspace. This is particularly advantageous in unstructured environments such as agriculture. Furthermore, the approach is computationally efficient with an average processing time of 0.0183 seconds.

The robustness and safety of HRC systems will be explored in future work. This includes investigating the following:

- Robustness to disturbances: An in-depth analysis of the robustness of the approach is planned, particularly in scenarios with multiple individuals present. This is a common scenario in situations such as harvesting. Furthermore, the real-world performance of the system in diverse environments, including challenging settings such as agricultural landscapes, will also be assessed. Open fields and row crops could result in different results due to occlusion of the vision or irregular wave propagation of UWB signals. This research will involve extensive experimentation and data collection to gauge the adaptability and reliability of the technology.
- Safe Human-robot collaboration (HRC) navigation: Investigations will extend to the realm of HRC navigation. Collision avoidance techniques such as obstacle detection and avoidance and path planning will be explored to enhance the safety of human-robot interactions. These studies will involve both theoretical modelling and practical experimentation.

Author Contributions Not applicable

Funding Open access funding provided by FCT|FCCN (b-on). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No 101004085 and 101000554.
Disclaimer: The sole responsibility for the content on this [website/newsletter/publication] lies with the authors. It does not necessarily reflect the opinion of the European GNSS Agency (GSA) or the European Commission (EC). The GSA or the EC are not responsible for any use that may be made of the information contained therein.

Availability of data and materials Not applicable

Code availability Not applicable

Declarations

Ethics approval Not applicable

Consent to participate Not applicable

Consent for publication Not applicable

References

1. Murphy, R.R., Nomura, T., Billard, A., Burke, J.L.: Human-robot interaction. IEEE Robot. Autom. Mag. **17**(2), 85–89 (2010). https://doi.org/10.1109/MRA.2010.936953
2. Kosuge, K., Hirata, Y.: Human-robot interaction. In: 2004 IEEE International Conference on Robotics and Biomimetics, pp. 8–11 (2004). https://doi.org/10.1109/ROBIO.2004.1521743
3. Islam, M.J., Hong, J., Sattar, J.: Person-following by autonomous robots: A categorical overview. Int J Robot Res **38**(14), 1581–1618 (2019). https://doi.org/10.1177/0278364919881683
4. Chen, B.X., Sahdev, R., Tsotsos, J.K.: Integrating stereo vision with a cnn tracker for a person-following robot. In: Liu, M., Chen, H., Vincze, M. (eds.) Computer Vision Systems, pp. 300–313. Springer, Cham (2017)
5. Sarmento, J., Aguiar, A.S., Santos, F.N.d., Sousa, A.J.: Robot navigation in vineyards based on the visual vanish point concept. In: 2021 International Symposium of Asian Control Association on Intelligent Robotics and Industrial Automation (IRIA), pp. 406–413 (2021). https://doi.org/10.1109/IRIA53009.2021.9588722
6. Aguiar, A.S., dos Santos, F.N., Cunha, J.B., Sobreira, H., Sousa, A.J.: Localization and mapping for robots in agriculture and forestry: A survey. Robotics **9**(4) (2020)
7. Deremetz, M., Lenain, R., Laneurit, J., Debain, C., Peynot, T.: Autonomous human tracking using uwb sensors for mobile robots: An observer-based approach to follow the human path. In: 2020 IEEE Conference on Control Technology and Applications (CCTA), pp. 372–379 (2020). https://doi.org/10.1109/CCTA41146.2020.9206153
8. Conejero, M.N., Montes, H., Andujar, D., Bengochea-Guevara, J.M., Ribeiro, A.: Collaborative Harvest Robot. Lecture Notes in Networks and Systems 590 LNNS, 415–426 (2023). https://doi.org/10.1007/978-3-031-21062-4_34
9. Sarmento, J., Santos, F.N.D., Aguiar, A.S., Sobreira, H., Regueiro, C.V., Valente, A.: FollowMe - A Pedestrian Following Algorithm for Agricultural Logistic Robots, pp. 179–185 (2022). https://doi.org/10.1109/ICARSC55462.2022.9784791. Cited by: 0. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85132993872&doi=10.1109%2fICARSC55462.2022.9784791&partnerID=40&md5=7bab52d61c081d0e5b4368e5454cbe42
10. Montesdeoca, J., Toibero, J.M., Jordan, J., Zell, A., Carelli, R.: Person-following controller with socially acceptable robot motion.

Robotics and Autonomous Systems **153**, 104075 (2022). https://doi.org/10.1016/j.robot.2022.104075
11. Zhang, R., Zhang, Y., Zhang, X.: Tracking in-cabin astronauts using deep learning and head motion clues. IEEE Access **9**, 2680–2693 (2021). https://doi.org/10.1109/ACCESS.2020.3046730
12. Algabri, R., Choi, M.T.: Target recovery for robust deep learning-based person following in mobile robots: Online trajectory prediction. Applied Sciences (Switzerland) **11** (2021). https://doi.org/10.3390/app11094165
13. Guffanti, D., Brunete, A., Hernando, M., Rueda, J., Navarro, E.: Robogait: A mobile robotic platform for human gait analysis in clinical environments. Sensors **21** (2021). https://doi.org/10.3390/s21206786
14. Zhang, Z., Yan, J., Kong, X., Zhai, G., Liu, Y.: Efficient motion planning based on kinodynamic model for quadruped robots following persons in confined spaces. IEEE/ASME Transactions on Mechatronics **26**, 1997–2006 (2021). https://doi.org/10.1109/TMECH.2021.3083594
15. Cha, D., Chung, W.: Human-leg detection in 3d feature space for a person-following mobile robot using 2d lidars. International Journal of Precision Engineering and Manufacturing **21**, 1299–1307 (2020). https://doi.org/10.1007/s12541-020-00343-7
16. Chen, X., Liu, J., Wu, J., Wang, C., Song, R.: LoPF : An Online LiDAR-Only Person-Following Framework **71** (2022)
17. Bharadwaj, R., Alomainy, A., Koul, S.K.: Experimental Investigation of Body-Centric Indoor Localization Using Compact Wearable Antennas and Machine Learning Algorithms. IEEE Transactions on Antennas and Propagation **70**(2), 1344–1354 (2022). https://doi.org/10.1109/TAP.2021.3111308
18. Otim, T., Bahillo, A., Diez, L.E., Lopez-Iturri, P., Falcone, F.: Impact of Body Wearable Sensor Positions on UWB Ranging. IEEE Sensors Journal **19**(23), 11449–11457 (2019). https://doi.org/10.1109/JSEN.2019.2935634
19. Su, Z., Pahlavan, K., Agu, E., Wei, H.: Proximity Detection During Epidemics: Direct UWB TOA Versus Machine Learning Based RSSI. International Journal of Wireless Information Networks **29**(4), 480–490 (2022). https://doi.org/10.1007/s10776-022-00577-4
20. Guler, S., Jiang, J., Alghamdi, A.A., Masoud, R.I., Shamma, J.S.: Real Time Onboard Ultrawideband Localization Scheme for an Autonomous Two-robot System. 2018 IEEE Conference on Control Technology and Applications, CCTA 2018, 1151–1158 (2018). https://doi.org/10.1109/CCTA.2018.8511568
21. Hepp, B., Tobias, N.: Omni-directional person tracking on a flying robot using occlusion-robust ultra-wideband signals (2016)
22. Qiu, R., Xu, M., Yan, Y., Smith, J.S.: A methodology review on multi-view pedestrian detection. Studies in Big Data **106**, 317–339 (2022). Cited by: 2. https://doi.org/10.1007/978-3-030-95239-6_12
23. Shen, L., Tao, H., Ni, Y., Wang, Y., Stojanovic, V.: Improved yolov3 model with feature map cropping for multi-scale road object detection. Measurement Science and Technology **34**(4), 045406 (2023). https://doi.org/10.1088/1361-6501/acb075
24. Liu, J., Chen, X., Wang, C., Zhang, G., Song, R.: A person-following method based on monocular camera for quadruped robots. Biomimetic Intelligence and Robotics **2**(3), 100058 (2022). https://doi.org/10.1016/j.birob.2022.100058
25. Jin, D., Fang, Z., Zeng, J.: A robust autonomous following method for mobile robots in dynamic environments. IEEE Access **8**, 150311–150325 (2020). https://doi.org/10.1109/ACCESS.2020.3016472
26. Liu, F., Zhang, J., Wang, J., Han, H., Yang, D.: An uwb/vision fusion scheme for determining pedestrians indoor location. Sensors (Switzerland) **20** (2020). https://doi.org/10.3390/s20041139
27. Luchetti, A., Carollo, A., Santoro, L., Nardello, M., Brunelli, D., Bosetti, P., Cecco, M.D., Montanini, R.: Acta imeko human identification and tracking using ultra-wideband-vision data fusion in unstructured environments **10** (2021)
28. Furgale, P., Rehder, J., Siegwart, R.: Unified temporal and spatial calibration for multi-sensor systems. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1280–1286 (2013). https://doi.org/10.1109/IROS.2013.6696514
29. Rehder, J., Nikolic, J., Schneider, T., Hinzmann, T., Siegwart, R.: Extending kalibr: Calibrating the extrinsics of multiple imus and of individual axes. In: 2016 IEEE International Conference on Robotics and Automation (ICRA), pp. 4304–4311 (2016). https://doi.org/10.1109/ICRA.2016.7487628
30. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: MobileNetV2: Inverted Residuals and Linear Bottlenecks (2019)
31. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: Single shot MultiBox detector. In: Computer Vision – ECCV 2016, pp. 21–37. Springer, ??? (2016). https://doi.org/10.1007/978-3-319-46448-0_2

José Sarmento is currently a researcher at the Centre for Robotics in Industry and Intelligent Systems (CRIIS) of the INESC TEC, his current research interest is in agricultural robots for safe interaction and coexistence with humans and animals. He obtained the B.Sc. degree in Electrical and Computer Engineering from the Instituto Superior de Engenharia do Porto (ISEP), the M.Sc. degree in Electrical and Computer Engineering from the Faculdade de Engenharia da Universidade do Porto (FEUP), and is currently pursuing a Ph.D. degree in Electrical and Computer Engineering from the Universidade de Trás-os-Montes e Alto Douro (UTAD).

Filipe Neves dos Santos holds a Degree (5-year degree) in Electrical and Computer Engineering in 2003 from Instituto Superior de Engenharia do Porto (ISEP), a M.Sc. in Electrical and Computer Engineering from the Instituto Superior Técnico (IST) from Universidade Técnica de Lisboa, in 2007, and received the PhD degree in Electrical and Computer Engineering at the Faculdade de Engenharia (FEUP), Universidade do Porto, Portugal, in 2014. He is a formal robotics researcher since 2003, with a large experience on navigation systems for unnamed aerial and ground vehicles. He is senior researcher and the head of the Laboratory of Robotics and IoT for Smart Precision Agriculture and Forestry at CRIIS (Centre for Robotics in Industry and Intelligent Systems) of the INESC TEC. He has more than 40 peer review papers in international conferences and international journals. Actually his priority research fields are: robotics for agriculture/forestry; human-robot interaction; Perception and, Semantic SLAM.

André Silva Aguiar received the Electrical Engineering MSc in 2019 from the Faculty of Engineering of the University of Porto, and the PhD in Electrical Engineering in 2023 from the UTAD University. He is a researcher at the Centre for Robotics in Industry and Intelligent Systems of the INESC TEC in Porto. His current research interests are focused on robotics navigation, Simultaneous Localization and Mapping and mission planning.

Vítor Filipe received the M.S. degree in informatics from the University of Minho, Portugal, in 1997, and the Ph.D. degree in electrical engineering from the University of Trás-os-Montes e Alto Douro (UTAD), Portugal, in 2003. He is currently an Associate Professor (Habilitation) in electrical engineering with the School of Science and Technology, UTAD. Since 2015, he has been a Senior Researcher with the Institute for System and Computer Engineering, Technology and Science (INESC TEC). He is the coauthor of about 100 papers published in scientific journals, book chapters, and proceedings of peer-reviewed international scientific conferences. His current research interests include computer vision and the application of artificial intelligence in images.

António Valente graduated in Electrical Engineering from University of Trás-os-Montes and Alto Douro (UTAD), Portugal in 1994, and in 1999 a MsC degree in Industrial Electronics from University of Minho, Portugal. He obtained in 2004 a PhD degree at UTAD, working in the field of micro-systems for agriculture. Presently, he is an Associate Professor with Habilitation in the Department of Engineering, UTAD, and director of the same department. He is a senior researcher at the Institute for Systems and Computer Engineering - Technology and Science (INESC TEC) and a senior member of IEEE. He was chairman of ICARSC 2015 and local organizer of Robótica 2015, Vila Real, Portugal. He is the organizer of Portuguese Micromouse Contest (robotics competition organized annually). His professional interests are in sensors, MEMS sensors, microcontrollers, and embedded systems, with application focus to agriculture.