ACTA IMEKO
ISSN: 2221-870X
December 2021, Volume 10, Number 4, 124 - 131

Human identification and tracking using ultra-wideband-vision data fusion in unstructured environments

Alessandro Luchetti[1], Andrea Carollo[1], Luca Santoro[1], Matteo Nardello[1], Davide Brunelli[1], Paolo Bosetti[1], Mariolino De Cecco[1]

[1] *Department of Industrial engineering, University of Trento, Sommarive, 9 - 38123 Trento, Italy*

ABSTRACT
Nowadays, the importance of working in changing and unstructured environments such as logistics warehouses through the cooperation between Automated Guided Vehicles (AGV) and the operator is increasingly demanded. The challenge addressed in this article aims to solve two crucial functions of autonomy: operator identification, and tracking. These tasks are necessary to enable an AGV to follow the selected operator along his path. This paper presents an innovative, accurate, robust, autonomous, and low-cost operator real-time tracking system, leveraging the inherent complementarity of the uncertainty regions (2D ellipses) between ultra-wideband (UWB) transceivers and cameras. The test campaign shows how the UWB system has higher uncertainty in the angular direction. In contrast, in the case of the vision system, the uncertainty is predominant along the radial coordinate. Due to the nature of the data, a sensor fusion demonstrates improvement in the accuracy and goodness of the final tracking.

Section: RESEARCH PAPER

Keywords: Human-robot interaction; following system; ultra-wideband technology; 3D vision system; sensor fusion

Citation: Alessandro Luchetti, Andrea Carollo, Luca Santoro, Matteo Nardello, Davide Brunelli, Paolo Bosetti, Mariolino De Cecco, Human identification and tracking using ultra-wideband-vision data fusion in unstructured environments, Acta IMEKO, vol. 10, no. 4, article 21, December 2021, identifier: IMEKO-ACTA-10 (2021)-04-21

Section Editor: Roberto Montanini, Università di Messina and Alfredo Cigada, Politecnico di Milano, Italy

Received July 26, 2021; **In final form** December 6, 2021; **Published** December 2021

Corresponding author: Alessandro Luchetti, e-mail: alessandro.luchetti@unitn.it

1. INTRODUCTION

Cooperation between mobile robots and people is playing a significant role in the modern economy while its demand is increasing worldwide, aided also by the growing use of autonomous and smart mobile robots. In particular, co-bots can increase human resources while reducing physical and mental load, increasing operational safety and productivity, in industrial environments. In this context, the human-following function, called "follow-me", is crucial. It consists of identifying and following the assigned operator even in unstructured environments.

There are different approaches to achieve such a task. The most commonly used technologies for tracking include vision [1], [2], time-of-flight (TOF)-Camera [3], LiDAR [4], light-emitting device (LED) [5] and UWB transceivers [6]-[131].

For each of these technologies, there are several advantages and disadvantages. LiDAR technology, while faster and more accurate than TOF camera, is much more expensive and not used unless strictly necessary. With the LED detection method there are few applications from the literature because of the low robustness due to the robot's inability to detect the light-emitting device frequently.

The best candidates for tracking operations with low cost are still 3D vision and UWB systems. The main literature contributions use them independently to solve the "follow-me" task in unstructured environments without solving their disadvantages. For example, the disadvantage of traditional vision systems is the limited field of view (FoV) compared to UWB systems and lighting influence, especially outdoor. In contrast, UWB can be applied both indoor and outdoor but suffers from higher uncertainties especially for measurements of less than one meter and in the presence of obstruction by people between the transceivers [9], [10]. However, UWB systems measurements can be made up to 80 m, in contrast to TOF cameras where after 10 m there is mostly noise while are more precise than the UWB below. Furthermore, the shape of the uncertainties of the two systems is different but complementary.

With this work, we overcome the disadvantages of these technologies by combining them to improve the robustness and reduce the uncertainty of the measurement result.

The only literature works that apply Sensor Fusion between UWB and vision systems work in structured environments with fixed UWB transceivers [11], [12] or with both fixed UWB and vision systems [13]. Their solutions are not dynamic to changes, are more expensive because of the number of devices to be used, and must be calibrated for each environment.

This work is organized as follows. An overview of the involved measurement systems is provided in the next section. The human identification algorithms to define which operator to follow are provided in Section 3; Section 4 describes the uncertainties models of the UWB and vision systems. The operator localization results through Sensor Fusion approach are discussed in Section 5 followed by conclusions in Section 6.

2. MEASUREMENT SYSTEMS

For our application, we used Decawave's UWB development board DWM10011 [14]. It has the advantages of low cost, low power consumption, and strong penetration. The receivers were programmed with a two-way ranging (TWR) architecture [15] that allows them to work in an unstructured environment, with two anchors on the robot, one master and one generic, and a tag for the tracked operator, Figure 2. The tag communicates first with the generic anchor to calculate the distance d_1, i.e. the distance between the tag and the generic anchor. Then with the master anchor to obtain the distance d_2, i.e. the distance between the tag and master anchor. Successively the generic anchor will send the estimated distance d_1 to the master, and finally, the master shares the two distances d_1, d_2 to the computer within a total time of 9 ms. The distance between the two anchors d_3 is fixed and constant. The position of the tag in the environment using only the UWB system is ambiguous because the two radii d_1 and d_2 of the circumferences can intersect at two different points. To solve this problem, the robot is equipped with a camera, which determines the uniqueness of the measurement (i.e., whether the operator is in front of the robot or not).

The selected camera is the Intel Realsense™ Depth Camera D4552 [16]. The device includes an RGB camera, two infrared (IR) cameras, a laser projector, and an inertia measurement system (IMU). The vision system can extract both a 3D point cloud of the scene and a traditional RGB image. The RGB is used to apply artificial intelligence (AI) algorithms that perform human skeleton detection, while the point cloud allows localizing the key points of the skeleton in 3D.

Figure 1 shows the designed system. In particular, the UWB and vision systems onboard the mobile robot for operator identification and tracking.

3. HUMAN IDENTIFICATION

Human detection and human pose estimation are done by applying a neural network provided by Intel in the OpenVINO™ toolkit, called *human-pose-estimation-0001*. The toolkit enables Convolution Neural Networks (CNN)-based deep learning inference and contains an optimized version of OpenCV libraries for Intel hardware [17]. This network is based on OpenPose [18] approach with tuned MobileNet v1 [19] as a feature extractor. This network results in two different outputs. The first is composed of 18 probability maps, called heat-maps, that provide all the key-points on the image: ears, eyes, nose, neck, shoulders, elbows, wrists, hips, knees, and ankles, Figure 3a. The second is 19*2 layers called part affinity fields, and it gives us information on how to match the key-points that correspond to a single person, Figure 3b. Heat-maps provide the probability for each pixel to be at the position of a key-point. Performing a threshold on the heat-maps it is possible to select the pixels with a probability higher than 50 %, from which it is possible to evaluate for each cluster of pixels the average point and the covariance matrix referring to each key-point location, Figure 3a. From these covariance matrices, the average uncertainty of each pixel was calculated as three pixels.

The image resolution was set at 848 × 480 (CxR), and the network for the *human-pose-estimation-0001* as INT8 running on the CPU extracts each frame up to 18 key-points per person. The resolution was chosen as a good trade-off between resolution of the image, and so linked to the uncertainty for the human positioning, and the speed of execution of the network. The net is available in three different resolutions, i.e. INT8, FP16, and FP32. There is no difference in performance in our test when running on the CPU, but the INT8 version is slightly faster than the others for this estimation.

Only the operator, entitled to use the robot, must be tracked. To reach this, we introduced face identification. The steps to identify the operator are two: find all the faces within the RGB frame and then detect the operator among them. For these steps, we used two CNN available on the OpenVINO™ toolkit. The face detection CNN used is *face-detection-retail-0005* and to extract the features of each face and compare them with the operator features stored in a database we use *face-reidentification-retail-0095*. The matching is made through the cosine similarity, equation (1), which is defined as the cosine of the angle θ between two features' vectors A and B on dimension R^n. Where A is the features' vector acquired in real-time and B is the one stored as ground truth. The lower the cosine, the higher the probability that the features vector are similar.

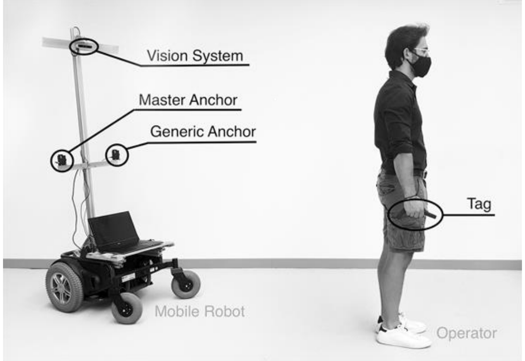


Figure 1. Real system with mobile robot and operator.

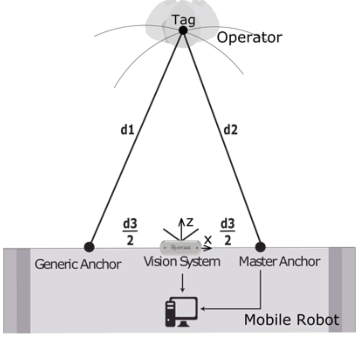
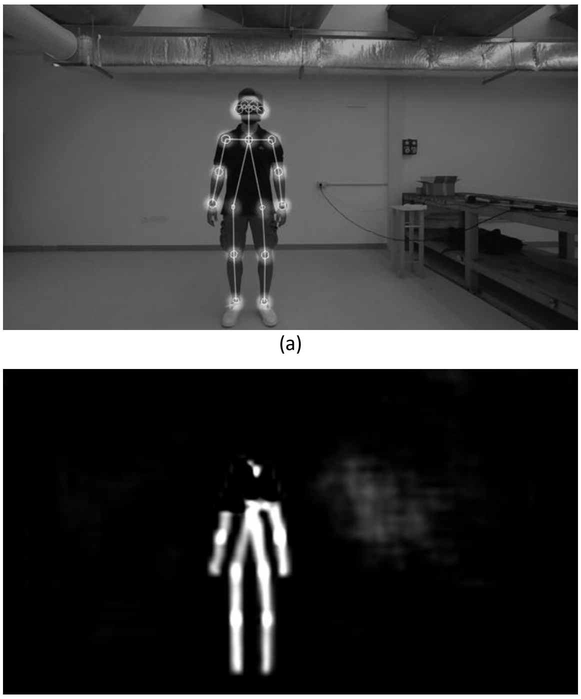


Figure 2. System configuration (Top view).



(a)

(b)

Figure 3. Heat-maps with covariances of key-points (a) and affinity map (b) of 848x480 image from Camera D455 with operator at two meters.

$$Similarity = \cos(\theta) = \frac{A \cdot B}{||A||\,||B||} = \\ = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2}\,\sqrt{\sum_{i=1}^{n} {B_i}^2}}\,, \tag{1}$$

The operator, once identified, turns on the spot, and the software saves all the features again with a neural network, called *person-reidentification-retail-0031*, for different poses. As previously stated, the correspondence is made with the cosine similarity between the feature vectors of people's bodies in the frames.

Table 1 shows the overall times for each network inference calculated on an Intel CPU i7-7700HQ with 16 *GB* of RAM.

4. HUMAN LOCALIZATION: 2D UNCERTAINTY MODELING

From UWB and vision system are extracted the same information about the operator's position with two measurements systems that are both referred to the robot base but whose uncertainty regions are complementary as explained in this section.

4.1. UWB system

Table 1. Overall times.

Network	Image resolution	Net resolution	Inference time (ms)
face-detection-retail-0005	848 × 480	INT8	9.42
face-reidentification-retail-0095	848 × 480	FP16	5.38
person-reidentification-retail-0031	96 × 48	FP16	6.50
human-pose-estimation-0001	848 × 480	INT8	115

To maintain consistency with the information received from the vision system, the reference frame of the UWB is fixed on the camera and the anchors are placed at its sides with equal distance $d_3/2$, Figure 2. In particular, the X coordinate is related to the lateral axis, the Z coordinate to the depth, and the Y coordinate to the vertical axis. For the localization of the operator in the space, we are interested only in the X and Z coordinates, since we project everything on the ground floor. The equations of the circumferences from the UWB devices become:

$$\left(x - \frac{d_3}{2}\right)^2 + z^2 = {d_1}^2, \tag{2}$$

$$\left(x + \frac{d_3}{2}\right)^2 + z^2 = {d_2}^2\,, \tag{3}$$

From equations (2) and (3), we obtain the closed-form solution for the tag position:

$$x = \frac{d_2^2 - d_1^2}{2 d_3}, \tag{4}$$

$$z = \pm \frac{1}{2} \sqrt{-\frac{(d_1^2 - d_2^2)^2}{d_3^2} - d_3^2 + 2(d_1^2 + d_2^2)}\,. \tag{5}$$

As discussed in Section 2, we only keep the positive value for the Z-coordinate because it is guaranteedby the vision system.

The covariance matrix of the coordinates of the tag position with the UWB system is equal to:

$$C_{\text{UWB}} = J_{\text{dist}} \cdot C_{\text{dist}} \cdot J_{\text{dist}}^{T}\,, \tag{6}$$

where C_{dist} is the covariance matrix of the measured distances d_1, d_2, and d_3, (Figure 2), with their standard deviations σ_1, σ_2, and σ_3

$$C_{\text{dist}} = \begin{pmatrix} {\sigma_1}^2 & 0 & 0 \\ 0 & {\sigma_2}^2 & 0 \\ 0 & 0 & {\sigma_3}^2 \end{pmatrix}. \tag{7}$$

The UWB devices used to test the overall system were calibrated in different indoor and outdoor scenarios to find the corresponding systematic offset at different distances and the related uncertainties. Furthermore, the master and generic anchors were calibrated independently because, although the device type is the same, their internal crystals and thus their responses are not the same. In particular, we carried out measurements from 1 m to 20 m every 0.5 m with random order both indoor and outdoor. We collect data at 52 Hz for three minutes. Between different distances, we wait three minuteswith all the devices turned off. The tests were performed for each distance, one with the line of sight (LOS) between the devices free (Figure 4a) and the other two with noise elements: in one 3-4 people simultaneously walked randomly back and forth between the devices throughout the test time (Figure 4b), in the other large static metal elements were placed in random positions between the devices (Figure 4c). This was done to understand how much these scenarios affect the measurements.

Figure 5a shows an example of the box-plot result for an indoor test at 7 m in different scenarios. In particular, in all the tests done, the standard deviation of the measurements made with the walking people is significantly higher than the ones with free LOS and metal objects. The data acquired with walking people generate an overestimation, possibly because the occlusion of the LOS causes the reflected radio waves to be caught as the free LOS.

From the tests, the standard deviations d_1 and d_2 were calculated at 0.05 m outdoor with free LOS.This value was used

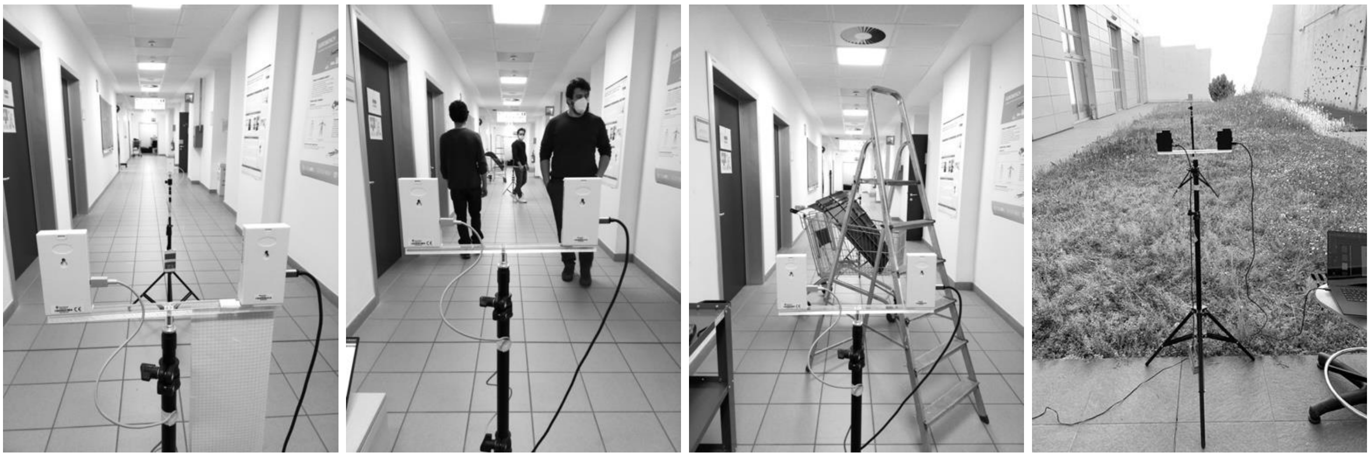
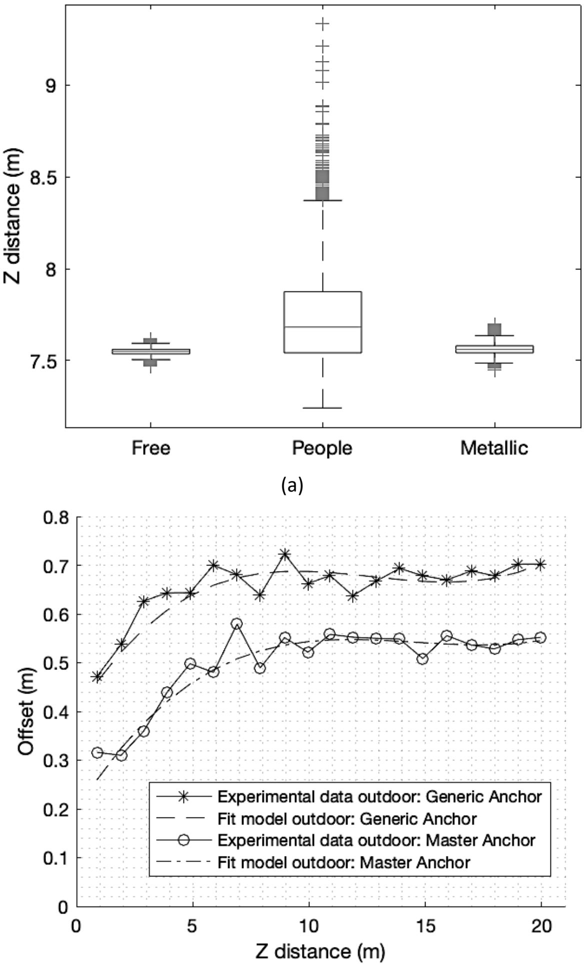

(a) (b) (c) (d)

Figure 4. Indoor scenarios: free LOS (a), people walking (b), metal objects (c); Outdoor scenarios (d).

and set in our application; instead, the standard deviation for d_3 was set at 0.01 m to take into account human error during fixing.

About the offset between the measured distances with UWB devices and the real distances, the values were modeled from the outdoor tests in free LOS, Figure 5b. All the real distances, taken as ground truth, were measured with a laser meter type Fervi ML80 [20], able to measure distances from 0.05 m to 80.00 m with an accuracy of 0.02 m.



(a)

(b)

Figure 5. (a) Box-plot of data at 7 m indoor in different scenarios; (b) Offset model of master and generic anchors outdoor in free line of sight (LOS).

The generic anchor and the tag were powered by a Varta power bank type 57962, while a USB cable powered the master anchor from the PC. From that USB cable was possible to communicate in serial and save the estimated distances from the two anchors.

Previously we saw how the free outdoor LOS was chosen as the reference environment for the offset and standard deviation values. To take into account also other scenarios each time we check the Channel Impulse Response (CIR) [21] provided by the UWB modules. In case of some human obstruction, this parameter decreases, and we do not update the distance information waiting for a reasonable value of CIR.

The jacobian matrix J_{dist} with respect to the distances is:

$$J_{\mathrm{dist}} = \begin{pmatrix} -\frac{d_1}{d_3} & \frac{d_2}{d_3} & \frac{a_2}{2\,d_3^{\,2}} \\ \frac{4\,d_1 - \frac{4\,d_1\,a_2}{d_3^{\,2}}}{a_1} & \frac{4\,d_2 + \frac{4\,d_2\,a_2}{d_3^{\,2}}}{a_1} & -\frac{2\,d_3 - \frac{2\,a_2^{\,2}}{d_3^{\,3}}}{a_1} \end{pmatrix}, \tag{8}$$

with:

$$a_1 = 4\,\sqrt{2\,d_1^{\,2} + 2\,d_2^{\,2} - d_3^{\,2} - \frac{a_2^{\,2}}{d_3^{\,2}}}$$

$$a_2 = d_1^{\,2} - d_2^{\,2}$$

Equation (6) becomes:

$$C_{\mathrm{UWB}} = \begin{pmatrix} \frac{\sigma_3^{\,2}\,a_6^{\,2}}{4\,d_3^{\,4}} + \frac{d_1^{\,2}\,\sigma_1^{\,2}}{d_3^{\,2}} + \frac{d_2^{\,2}\,\sigma_2^{\,2}}{d_3^{\,2}} & a_1 \\ a_1 & \frac{\sigma_3^{\,2}\,a_3^{\,2}}{16\,a_2} + \frac{\sigma_1^{\,2}\,a_5^{\,2}}{16\,a_2} + \frac{\sigma_2^{\,2}\,a_4^{\,2}}{16\,a_2} \end{pmatrix}, \tag{9}$$

with:

$$a_1 = \frac{d_2\,\sigma_2^{\,2}\,a_4}{4\,d_3\,\sqrt{a_2}} - \frac{d_1\,\sigma_1^{\,2}\,a_5}{4\,d_3\,\sqrt{a_2}} - \frac{\sigma_3^{\,2}\,a_6\,a_3}{8\,d_3^{\,2}\,\sqrt{a_2}}$$

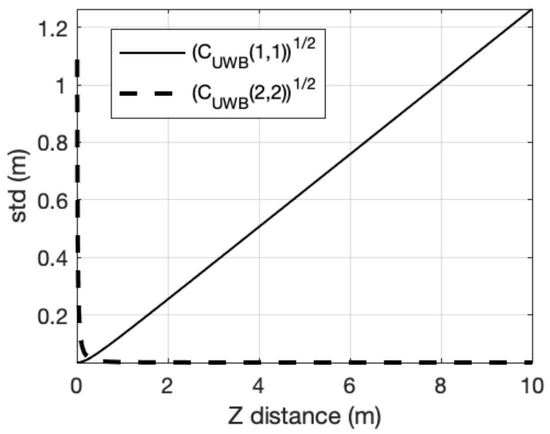


Figure 6. Standard deviation (std) of each eigenvalue in the two principal direction X and Z for UWB system from 0 m to 10 m with the off-diagonal terms of covariance matrix C_{UWB} equal to 0.

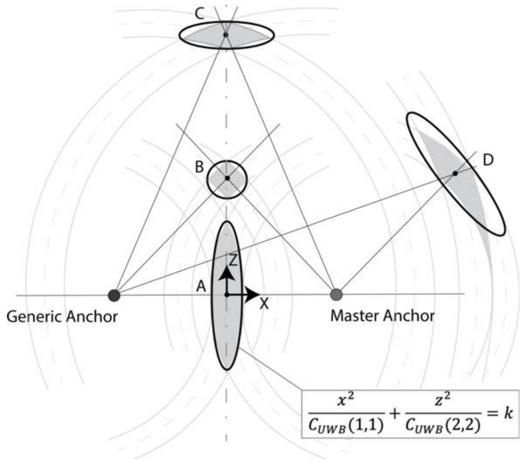


Figure 7. UWB system uncertainty ellipses for different tag positions.

$$a_2 = 2\,{d_1}^2 + 2\,{d_2}^2 - {d_3}^2 - \frac{{a_6}^2}{{d_3}^2}$$

$$a_3 = 2\,d_3 - \frac{2\,{a_6}^2}{{d_3}^3}$$

$$a_4 = 4\,d_2 + \frac{4\,d_2\,a_6}{{d_3}^2}$$

$$a_5 = 4\,d_1 - \frac{4\,d_1\,a_6}{{d_3}^2}$$

$$a_6 = {d_1}^2 - {d_2}^2$$

Noting that the elements of Equation (9) have at the denominator the measure d_3, we can say that the more distant the anchors are from each other, the more accurate the position measurement is. Another note is that the off-diagonal terms are zero if and only if the measures d_1 and d_2 are equal (considering also $\sigma_1 = \sigma_2$). In this case, Equation (9) becomes:

$$C_{\mathrm{UWB}} = \begin{pmatrix} \frac{2\,{d_1}^2\,{\sigma_1}^2}{{d_3}^2} & 0 \\ 0 & \frac{2\,{d_1}^2\,{\sigma_1}^2}{4\,{d_1}^2 - {d_3}^2} + \frac{{d_3}^2\,{\sigma_3}^2}{4\left(4\,{d_1}^2 - {d_3}^2\right)} \end{pmatrix}, \qquad (10)$$

Under these conditions, Figure 6 shows the behaviour of the standard deviations' values of each eigenvalue in the two principal directions X and Z with respect to the Z distance of the tag from the anchors, square roots of $C_{\mathrm{UWB}}(1, 1)$ and $C_{\mathrm{UWB}}(2, 2)$ respectively·

As can be seen from the behaviour of the eigenvalues in Figure 6 and shapes of the covariances in Figure 7 at the beginning (Z = 0 m) the covariance is a tight ellipse stretched in the radial direction ($C_{\mathrm{UWB}}(1, 1) < C_{\mathrm{UWB}}(2, 2)$, tag *A* in Figure 7), then when the distances d_1 and d_2 are orthogonal to each other it becomesa perfect circle ($C_{\mathrm{UWB}}(1, 1) = C_{\mathrm{UWB}}(2, 2)$, tag *B* in Figure 7) and lastly with a higher Z distance value it becomes stretched in the angular direction ($C_{\mathrm{UWB}}(1, 1) > C_{\mathrm{UWB}}(2, 2)$, tag *C* in Figure 7). If d_1 and d_2 are different, tag *D* in Figure 7, the quadratic approximation of the probability ellipse will be rotated·

4.2. Vision system

The model used to study the variance for the depth coordinate Z of the selected camera is a pinhole model. Two-IR sensors of the camera are used to evaluate the depth through a disparity matching algorithm, an RGB sensor is used instead to calculate the 2D image, all of them are co-planar and aligned. The common reference frame of the two-IR sensors is set at the same focal length of the two-IR sensors and translated by overlapping the reference frame of the right sensor above the left sensor byan amount equal to the baseline *b*, Figure 8. The charge-coupled device (CCD) resolution for each sensoris 848 pixels in columns (*c*) and 480 pixels in rows (*r*) for a total of 848 × 480 (*C* × *R*) pixels.

The expression of the depth coordinate *z* is:

$$z = f_c \cdot \frac{b}{\left|c_{Q_1} - c_{Q_2}\right|} = f_c \cdot \frac{b}{d}, \qquad (11)$$

where f_c is the focal length in X direction, *b* the baseline, and *d* the disparity, Figure 8.

From equation (11) is possible to evaluate the error in the depth estimation from the model with respect to the disparity *d*, assuming all the other parameters as known and constant values:

$$\sigma_z = \frac{\partial z}{\partial d} = \frac{b \cdot f_c}{d^2} \cdot \sigma_d\ , \qquad (12)$$

Equation (12) can be rewritten as function of the Z distance, for better understanding, by substituting the valueof the disparity *d* with a formulation obtainable from equation (11):

$$\sigma_z = \frac{z^2}{b \cdot f_c} \sigma_d\,, \qquad (13)$$

where the focal length *f* in pixel is:

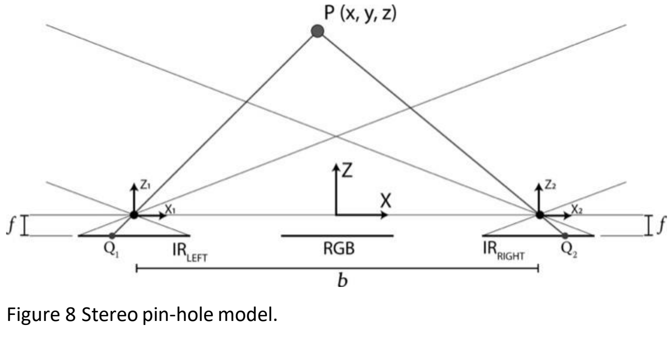


Figure 8 Stereo pin-hole model.

$$f_c = 0.5 \frac{C}{\tan(HFOV/2)}, \tag{14}$$

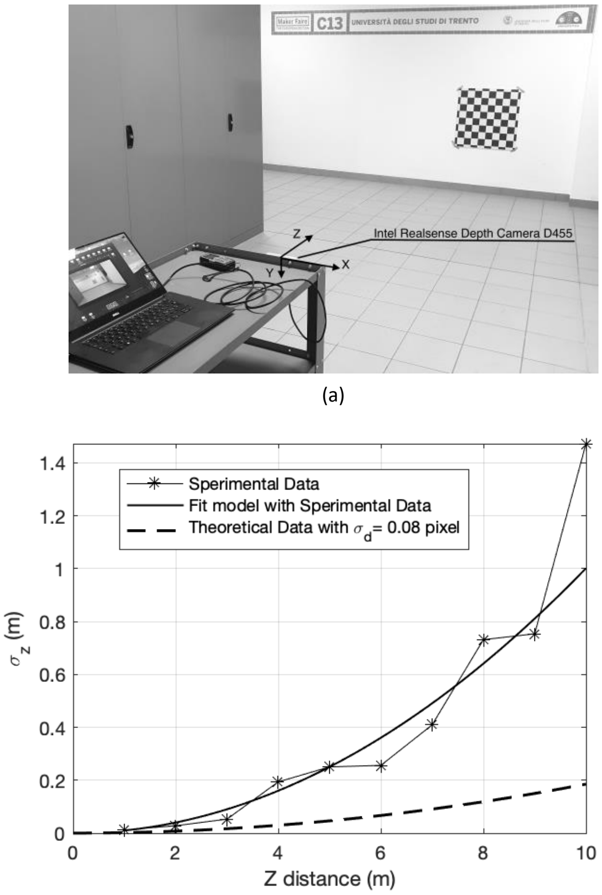
$$f_r = 0.5 \frac{R}{\tan(VFOV/2)}, \tag{15}$$

HFOV is the horizontal field of view and *VFOV* is the vertical field of view, in our case are 87° and 58° respectively. By varying the resolution, the focal length changes and so does the distance uncertainty.

The disparity standard deviation σ_d for the Camera D455 is declared to be 0.08 pixel by the manufacturer. Since many interfering effects increase this value, we calibrated the camera to check if the declared model equation (13) fits the real data and if so, to obtain the real standard deviation σ_d. We made 400 consecutive depth acquisitions of a chessboard from 1 m to 10 m at random steps of one meter, Figure 9a., Figure 9b shows the theoretical and the obtained σ_Z as a function of Z distance for the image resolution *C* of 848 pixels. As can be seen in Figure 9b the behaviour of the experimental found model is higher than the theoretical one. The new experimental value for σ_d found by applying the least absolute residual robust (LAR) method is 0.4 pixel.

To evaluate the coordinates in X and Y dimension we use the Brown-Conrady distortion calibration model [22] with the intrinsic constant coefficients (k_1, k_2, k_3, p_1, p_2) of the specific selected camera:

$$u_x = \hat{x} \cdot f + 2 \cdot p_1 \cdot \hat{x}\hat{y} + p_2(r + 2\hat{x}\hat{y}) \,, \tag{16}$$


(a)

(b)

Figure 9. (a) Camera calibration test with chessboard; (b) standard deviation results on Z distance with stereocamera D455 for resolution C of 848 pixels.

$$u_y = \hat{y} \cdot f + 2 \cdot p_2 \cdot \hat{x}\hat{y} + p_1(r + 2\hat{x}\hat{y}) \,, \tag{17}$$

with:

$$\hat{x} = \frac{c - C/2}{f_c} = \frac{x_p}{f_c}$$

$$\hat{y} = \frac{r - R/2}{f_r} = \frac{y_p}{f_c}$$

$$r = \hat{x}^2 + \hat{y}^2$$

$$f = 1 + k_1 \cdot r + k_2 \cdot r^2 + k_3 \cdot r^3$$

The new expression of the X and Y position becomes:

$$x = u_x \cdot z \,, \tag{18}$$

$$y = u_y \cdot z \,, \tag{19}$$

As for the UWB system, we are interested only in the X and Z coordinates for the localization of theoperator in the space.

The covariance matrix of the coordinates of the operator with camera system is equal to:

$$C_\mathrm{cam} = J_\mathrm{cam} \cdot C_\sigma \cdot J_\mathrm{cam}^T \,, \tag{20}$$

where:

$$C_\sigma = \begin{pmatrix} {\sigma_c}^2 & 0 & 0 \\ 0 & {\sigma_r}^2 & 0 \\ 0 & 0 & {\sigma_z}^2 \end{pmatrix}, \tag{21}$$

$$J_{cam}^T = \begin{pmatrix} z\left(\frac{a_1}{f_c} + \frac{6\,p_2\,x_p}{{f_c}^2} + \frac{x_p\left(\frac{2\,k_1\,x_p}{{f_c}^2} + \frac{4\,k_2\,x_p\,a_2}{{f_c}^2} + \frac{6\,k_3\,x_p\,{a_2}^2}{{f_c}^2}\right)}{f_c} + \frac{2\,p_1\,y_p}{f_c\,f_r}\right) & 0 \\ z\left(\frac{2\,p_2\,y_p}{{f_r}^2} + \frac{x_p\left(\frac{2\,k_1\,y_p}{{f_r}^2} + \frac{4\,k_2\,y_p\,a_2}{{f_r}^2} + \frac{6\,k_3\,y_p\,{a_2}^2}{{f_r}^2}\right)}{f_c} + \frac{2\,p_1\,x_p}{f_c\,f_r}\right) & 0 \\ p_2\left(\frac{3\,{x_p}^2}{{f_c}^2} + a_3\right) + \frac{x_p\,a_1}{f_c} + \frac{2\,p_1\,x_p\,y_p}{f_c\,f_r} & 1 \end{pmatrix}, \tag{22}$$

with:

$$a_1 = k_1\,a_2 + k_2\,{a_2}^2 + k_3\,{a_2}^3 + 1$$

$$a_2 = \frac{{x_p}^2}{{f_c}^2} + a_3$$

$$a_3 = \frac{{y_p}^2}{{f_r}^2} \,.$$

C_cam depends on the focal length *f*, on the baseline *b*, and on the standard deviations of disparity σ_d and pixels σ_c, σ_r.

Once the depth frame and the RGB one are aligned, it is possible to estimate the distance to any pixel and thus project the operator's skeleton in space. Sometimes, key-points can have an incorrect distance value, i.e., the Z-coordinate, especially when the pose is very close to the viewing system. It is due to distortion caused by camera lenses and imperfect image realignment. To overcome this problem, we extract an average position of key-points with the median. In this case, if a key-point is projected too far from the other or too near, it will have no impact on the final operator position.

The off-diagonal terms of equation (20) are zero if and only if $x_\mathrm{p} = 0$, $y_\mathrm{p} = 0$. In this case equation (20) becomes:

$$C_{\mathrm{cam}} = \begin{pmatrix} \frac{z^2\sigma_c^2}{f_c^2} & 0 \\ 0 & {\sigma_z}^2 \end{pmatrix}, \tag{23}$$

Figure 10 shows the standard deviations' values of each eigenvalue in the two principal directions X and Z of equation (23) with the covariances' shapes in Figure 11 as previously done for UWB system. As can be seen in Figure 10, after 30 cm, the radial coordinate in the Z direction is always greater than the angular one in the X direction ($C_{\mathrm{cam}}(1, 1) < C_{\mathrm{cam}}(2, 2)$ i.e. the ellipse is stretched in the radial direction.

5. SENSOR FUSION

Sensor fusion was done between the position estimated by the UWB system and the position estimated by the vision system to reduce the uncertainty and improve the estimation of the operator's position inspace. Figure 12 shows two examples in two different tag positions where the uncertainties of the position estimation through the UWB system are more significant along the angular direction (X-coordinate), centring the reference system on the origin of the camera. Notice that in the case of the vision system,the uncertainties are predominant along the radial direction (Z-coordinate). By fusing the information with Bayes' theorem [23], it is possible to reduce their uncertainties in both directions.

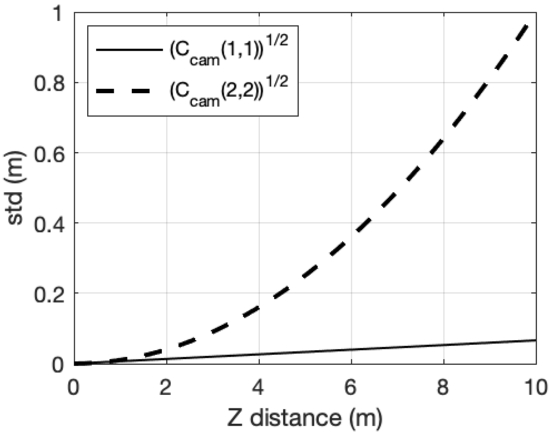


Figure 10: Standard deviation (std) of each eigenvalue in the two principal direction X and Z for camera system from 0 m to 10 m with the off-diagonal terms of covariance matrix C_{cam} equal to 0.

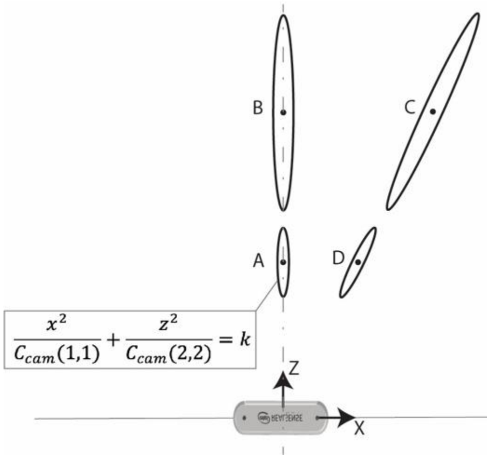
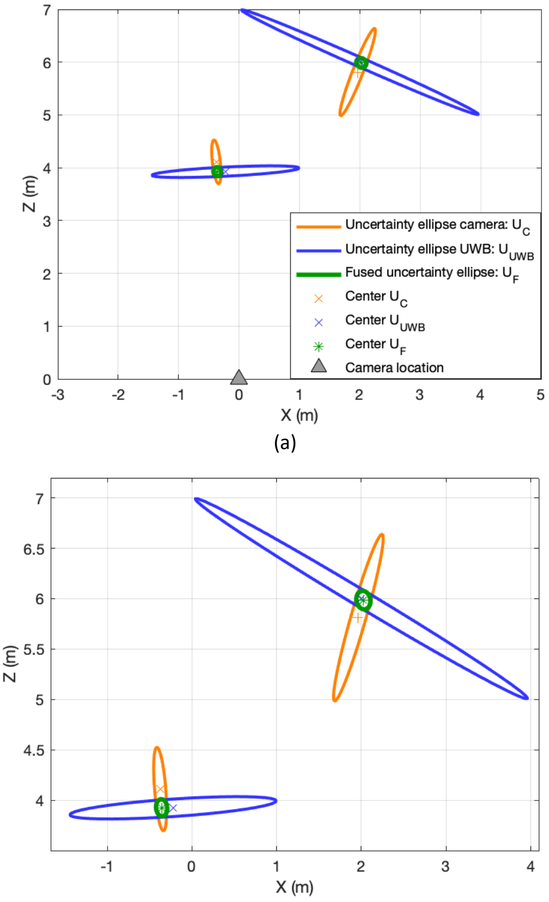


Figure 11. Camera system uncertainty ellipses for different tag positions.



(a)

(b)

Figure 12. (a) Two examples of uncertainty ellipses (95 % with *k*= 2.4478 [24]) in two different tag positions (x=0 m, z=4 m; x=2 m, z=6 m) from camera and UWB system; (b) zoom of the results.

We can summarize the fused information shown in Figure 12 through the following expression:

$$C_{\mathrm{fused}} = [[C_{\mathrm{UWB}}]^{-1} + [C_{\mathrm{cam}}]^{-1}]^{-1}, \tag{24}$$

$$P_{\mathrm{fused}} = C_{\mathrm{fused}}[[C_{\mathrm{UWB}}]^{-1}P_{\mathrm{UWB}} + [C_{\mathrm{cam}}]^{-1}P_{\mathrm{cam}}], \tag{25}$$

where:

- P_{UWB} and P_{cam} are the estimated mean position of the UWB system and camera system respectively;
- C_{UWB} and C_{cam} are the covariance matrices of the estimated position of the UWB system and camera system respectively.

6. CONCLUSIONS

In this work, an innovative and robust method to identify and track an operator from a mobile robot in real-time was developed. In particular, the operator can be identified through a convolution neural network tool and tracked through a designed localization method obtained by fusing the information from low-cost sensors such as UWB transceivers and a depth camera. It was implemented to locate the operator's position with less uncertainty. The test campaign shows the UWB system has a higher uncertainty in the angular direction, contrary to the camera, where the uncertainty is higher in the radial direction. Specifically, C_{UWB} is affected by the distances measured between the tag and the two anchors (d_1, d_2) and between the anchors themselves (d_3) with their corresponding uncertainties. C_{cam} instead depends on the focal length f, on the baseline b, and on the standard deviations of disparity σ_{d} and pixels σ_{c}, σ_{r}.

Our solution makes the final system robust and more precise due to the complementarity of information between the covariance matrices of the UWB system and the vision system.

To obtain a better result for the real application, the UWB transceivers were calibrated following a test campaign that provided the correct behaviour of the systematic offset in the measurements and their standard deviations. Offsets change for a specific UWB device and increase with Z distance. The standard deviation for the distances between the devices was calculated at 0.05 *m* in the scenario with a free line of sight and outdoor.

Another test campaign, this time on the vision system, was carried out to evaluate the theoretical model of the standard deviation σ_z as a function of Z distance and the value of camera disparity standard deviation σ_{d}. The σ_{d} calculated for our Camera D455 is 0.4 pixels. Moreover, the average uncertainty of each pixel σ_{c}, σ_{r} was calculated as three pixels through heat-maps analysis of the convolution neural network used.

ACKNOWLEDGEMENT

This work was supported by the Italian Ministry for University and Research (MUR) under the program “Dipartimenti di Eccellenza (2018-2022)”.

REFERENCES

[1] M. Gupta, S. Kumar, L. Behera, V. K. Subramanian, A novel vision-based tracking algorithm for a human-following mobile robot, IEEE Transactions on Systems, Man, and Cybernetics: Systems 7 (2016), pp. 1415–1427.
DOI: 10.1109/TSMC.2016.2616343

[2] S.-O. Lee, M. Hwang-Bo, B.-J. You, S.-R. Oh, Y.-J. Cho, Vision based mobile robot control for target tracking, IFAC Proceedings Volumes 34(4) (2001), pp. 75–80.
DOI: 10.1016/S1474-6670(17)34276-3

[3] G. Xing, S. Tian, H. Sun, W. Liu, H. Liu, People-following system design for mobile robots using kinectsensor, 25th Chinese Control and Decision Conference (CCDC), IEEE, Guiyang, China, 25-27 May 2013, pp. 3190–3194.
DOI: 10.1109/CCDC.2013.6561495

[4] S. A. Ahmed, A. V. Topalov, N. G. Shakev, and V. L. Popov, Model-free detection and following of moving objects by an omnidirectional mobile robot using 2d range data, IFAC-PapersOnLine 51(22) (2018), pp. 226–231.
DOI: 10.1016/j.ifacol.2018.11.546

[5] Y. Nagumo A. Ohya, Human following behavior of an autonomous mobile robot using light-emitting device, in Proceedings 10th IEEE International Workshop on Robot and Human Interactive Communication. ROMAN 2001 (Cat. No. 01TH8591), IEEE, Bordeaux, Paris, France, 18-21 Sept. 2001, pp. 225–230.
DOI: 10.1109/ROMAN.2001.981906

[6] T. Feng, Y. Yu, L. Wu, Y. Bai, Z. Xiao, Z. Lu, A human-tracking robot using ultra wideband technology, IEEE Access 6 (2018), pp. 42541–42550.
DOI: 10.1109/ACCESS.2018.2859754

[7] T. G. Kim, D.-J. Seo, K.-S. Joo, A following system for a specific object using a UWB system, in 2018 18th International Conference on Control, Automation and Systems (ICCAS), IEEE, PyeongChang, Korea (South), 17-20 Oct. 2018, pp. 958–960.

[8] D.-J. Seo, T. G. Kim, S. W. Noh, H. H. Seo, Object following method for a differential type mobile robot based on ultra wide band distance sensor system, 17th International Conference on Control, Automation and Systems (ICCAS), IEE, Jeju, Korea (South), 18-21 Oct. 2017, pp. 736–738.
DOI: 10.23919/ICCAS.2017.8204325

[9] L. Santoro, D. Brunelli, D. Fontanelli, On-line Optimal Ranging Sensor Deployment for Robotic Exploration, IEEE SENSORS JOURNAL, v. 2021, (2021).
DOI: 10.1109/JSEN.2021.3120889

[10] L. Santoro, M. Nardello, D. Brunelli and D. Fontanelli, Scale up to infinity: the UWB Indoor Global Positioning System, 2021 IEEE International Symposium on Robotic and Sensors Environments (ROSE), 2021, pp. 1-8,
DOI: 10.1109/ROSE52750.2021.9611770

[11] G. Ding, H. Lu, J. Bai, X. Qin, Development of a high precision UWB/vision-based AGV and control system, 5th International Conference on Control and Robotics Engineering (ICCRE), Osaka, Japan, 24-26 April 2020, pp. 99–103.
DOI: 10.1109/ICCRE49379.2020.9096456

[12] H. Xu, L. Wang, Y. Zhang, K. Qiu, S. Shen, Decentralized visual-inertial-UWB fusion for relative state estimation of aerial swarm, IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May-31 Aug. 2020, pp. 8776–8782.
DOI: 10.1109/ICRA40945.2020.9196944

[13] F. Liu, J. Zhang, J. Wang, H. Han, D. Yang, An UWB/vision fusion scheme for determining pedestrians’ indoor location, Sensors 20(4) (2020), p. 1139.
DOI: 10.3390/s20041139

[14] Decawave dwm1001. Online [Accessed 06 December 2021]
https://www.decawave.com/product/dwm1001-development-board/

[15] M. Kwak, J. Chong, A new double two-way ranging algorithm for ranging system, 2nd IEEE International Conference on Network Infrastructure and Digital Content, IEEE, Beijing, China, 24-26 Sept. 2010, pp. 470–473.
DOI: 10.1109/ICNIDC.2010.5657814

[16] Intel RealSense Camera D455. Online [Accessed: December 2021]
https://www.intelrealsense.com/depth-camera-d455/

[17] OpenCV. Online [Accessed: December 2021]
https://opencv.org/

[18] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, Y. A. Sheikh, Openpose: Realtime multi-person 2d pose estimation using part affinity fields, IEEE Transactions on Pattern Analysis and Machine Intelligence 43(1) (2019), pp. 172- 186.
DOI: 10.1109/TPAMI.2019.2929257

[19] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: efficient convolutional neural networks for mobile vision applications. Online [Accessed 06 December 2021]
https://arxiv.org/abs/1704.04861

[20] Fervi Misuratore di distanza laser [Accessed 06 December 2021]
https://www.fervi.com/ita/strumenti-di-misura/misuratori-analogici-e-digitali/misuratore-di-distanze/misuratore-di-distanza-laser-pr-8240.htm

[21] C. Jiang, S. Chen, Y. Chen, D. Liu, Y. Bo, An UWB channel impulse response de-noising method for NLOS/LOS classification boosting, IEEE Communications Letters 24(11) (2020), pp. 2513–2517.
DOI: 10.1109/LCOMM.2020.3009659

[22] C. B. Duane, Close-range camera calibration, Photogramm. Eng, 37(8) (1971), pp. 855–866.

[23] W. Elmenreich, An introduction to sensor fusion, Vienna University of Technology, Austria 502 (2002), pp. 1–28.

[24] R. C. Smith, P. Cheeseman, On the representation and estimation of spatial uncertainty, The international journal of Robotics Research 5(4) (1986), pp. 56–68.
DOI: 10.1177/027836498600500404