A multisensor based approach using supervised learning and particle filtering for people detection and tracking

Eugenio Aguirre *, Miguel García-Silvente, and Daniel Pascual

Department of Computer Science and A.I., CITIC-UGR.
E.T.S. Ingenierías en Informática y en Telecomunicaciones
University of Granada. 18071 - Granada, Spain
{eaguirre,M.Garcia-Silvente}@decsai.ugr.es
dpascual@correo.ugr.es
http://decsai.ugr.es

Abstract. People detection and tracking is an interesting skill for interactive social robots. Laser range finder (LRF) and vision based approaches are the most common although both present strengths and weaknesses. In this paper, a multisensor system to detect and track people in the proximity of a mobile robot is proposed. First, a supervised learning approach is used to recognize patterns of legs in the proximity of the robot using a LRF. After this, a tracking algorithm is developed using particle filter and the observation model of legs. Second, a Kinect sensor is used to carry out people detection and tracking. This second method uses a face detector in the color image, the color of the clothes and the depth information.The strengths and weaknesses of the second proposal are also commented. In order to put together the strengths of both sensors, a third algorithm is proposed. In this third approach both laser and Kinect data are fused to detect and track people. Finally, the multisensory approach is experimentally evaluated in a real indoor environment. The multisensor system outperforms the single sensor based approaches.

Keywords: People detection and tracking, multisensor based tracking, social robot, human-robot interaction.

1 Introduction

In order to focus its attention on humans, a social robot needs to be aware of their presence. Therefore, people detection and tracking is an interesting skill. This is a challenging task because people move freely by the environment around the robot, moreover the sensory systems could suffer from the presence of false positives, noise and vagueness in the sensorial data.

Several approaches have been proposed to carry out people detection and tracking using different kinds of sensors, being the most popular sensors for those tasks laser sensors and cameras. Regarding laser based approaches, for

* This work have been partially supported by the Spanish Government project TIN2012-38969.

instance, in [7] authors propose detection and tracking schemes for human legs by the use of a single LRF. In [15] a systematic comparative analysis of laser based tracking methods, at feet and upper-body height, is performed. In [16] a system for tracking a variable number of pedestrians in crowded scenes by exploiting laser range scanners is proposed. On these works some conclusions arise. Compared with vision approaches, the use of laser sensors is advantageous since they are robust against illumination changes in the environment and the tracking algorithms are faster and more efficient. However, laser sensors have some limitations because the robot only can obtain distance information from a 2d-plane located at a certain height. A 3D-laser could solve this limitation but other problems arise as the cost of this device or the time of the data acquisition. In regards to the vision based approaches, mono, stereo cameras and RGB-Depth, as the Kinect sensor [12], have been used to detect and track people. Stereo and RGB-Depth cameras provide color and depth information. In [13] a fuzzy algorithm for detection and tracking of people in the proximity of a robot by using stereo vision is proposed. In [19] the depth information obtained from a Kinect sensor is used to detect humans. A 2-D head contour model and a 3-D head surface model is shown. Then a tracking algorithm is proposed based on their detection results. Vision based approaches also have some limitations. Illumination conditions can affect the performance of these methods, depth information is not always reliable and false positives in the detection methods are possible. In order to improve the results of laser and vision based approaches, multisensor solutions propose to use several kinds of sensors to achieve a more robust solution. In [3] a people following behavior is developed fusing information provided by a laser sensor and a stereo camera. In [5] authors propose multisensor data fusion using a laser scanner and a monocular camera. In [17] a multiple sensor fusion approach is proposed using three kinds of devices, Kinect, laser and a thermal sensor mounted on a mobile platform. It is shown that combination of different sensory cues increases the reliability of their people following system.

In this paper we propose a multisensor system to detect and track people in the proximity of a mobile robot. First, a supervised learning approach is used to identify patterns of legs in the proximity of the robot. This method analyses certain geometric features present in the laser data in order to detect possible legs of people. A classifier is trained using Support Vector Machines (SVM) [6] to classify the data obtained from laser using instances of patterns of legs. After this, a tracking algorithm is developed using particle filter and the observation model of the legs. The tracking algorithm is experimentally evaluated and some strengths and weaknesses are commented. Second, a Kinect sensor is used to carry out the people detection and tracking. This second method uses a face detector in the color image in order to perform people detection and the color of the clothes and the depth information is used to track people. Again a particle filter is developed using only the Kinect sensor to compare the results against the first proposal. The strengths and weaknesses of the second method are commented as well. In order to put together the strengths of both sensors, a third algorithm is proposed. In this third approach both laser and Kinect are used to detect and track people in the proximity of the robot. In the same way, the multisensory approach is experimentally evaluated. In the current work the

robot is standing still and looking the motion of people but in next works this limitation will be removed taking into account the required changes.

The rest of this paper is organized as follows. Section 2 describes the hardware that has been used in our system. Section 3 briefly describes the human legs detection algorithm, including the supervised learning algorithm, and the laser based tracking proposal. Section 4 shows the people detection and tracking algorithm based on Kinect. In Section 5 the multisensor approach is shown and experimental results are compared with both previous approaches. Finally some concluding remarks and future works are commented in Section 6.

2 System description

Our hardware system comprises a PeopleBot mobile robot equipped with a LRF SICK LMS200 and a Kinect sensor. Laser sensor scans 180º with a 1º resolution at 75 Hz. Its maximum range of distance in the current operation mode is 8 meters. It is mounted at a height of 30 cm above the ground. The Kinect features a RGB camera, a depth sensor and a multi-array microphone. Kinect uses an infrared projector and an infrared camera which are able to compute depth [14]. The Kinect depth sensor range is: minimum 800 mm and maximum 4000 mm. The resolution of both color and depth images is 640x480 at 30 fps. Because it uses IR, Kinect will not work under direct sunlight, e.g. outdoors. Since our system is intended to allow human robot interaction in indoor environments, these features of Kinect are suitable for our specifications. More information on Kinect is available in [20]. A laptop has been used to run the software due to the onboard computer is not powerful enough to perform video processing. The laptop has an Intel Core i5 with 4 GB DDR3 RAM and it is wired connected to the onboard computer. The laptop receives the laser data from the onboard computer while the Kinect sensor is directly connected to the laptop.

3 People detection and tracking using laser sensor (LRF based method)

The objective of this work is the design of a system capable to be used in Human Robot Interaction (HRI) applications. People interested in establishing interaction with the robot should be close to the robot; thus, an operation range of 1 to 3 m is defined. The first approach to detect and track people is based on a previous work by the authors [1] so that only a brief description is given below.

3.1 Leg detection method

The idea is to detect the legs when people are both moving or static. It is a challenging task because legs patterns are different in both situations. To do so, the laser measurements are clusterized and their geometrical properties are then analyzed. The considered properties comprise width, depth and size and all of them have been used successfully by others authors [7]. In our approach, a SVM classifier is trained by using the properties of the detected clusters and

a large data set that contains positive and negative instances of patterns of legs. Positive instances were registered with people walking and standing in the proximity of the robot. Negative instances include objects such as table legs, bins, boxes and various kinds of fire extinguishers. Note that some of this object could have geometrical properties similar to those of human legs. A balanced dataset containing 7802 instances of both, positive and negative samples, was used to train the SVM classifier.

In order to apply the SVM classifier, LibSVM was used [6]. Different kernels have been considered and the best precision is obtained with radial basis function (RBF). A wide grid-search using cross-validation has been performed in order to find the optimal value for these parameters, obtaining a precision of 89% which is suitable for this kind of application. Table 1 shows the results of a 10-fold cross validation. Results are acceptable since the rates of true positives and true negatives are high.

Table 1. Contingency table for the SVM classifier

		Observation class	
		Positive	Negative
Predicted class	Positive	88.71 %	10.62 %
	Negative	11.29 %	89.38 %

3.2 Particle filter based tracking

Particle filter is well known for its many applications in tracking. Target tracking problem is expressed by recursive Bayesian estimation. Essentially, two steps are given in each iteration: prediction and estimation. Both steps take into account the information of an observation model. Equations of particle filter are well known [4]. The vector of state, the definition of state transition and the model of noise is described in [1].

The LRF based method uses the leg detection algorithm as observation model so that each laser reading set is analyzed and the positions of possible legs are obtained. The probability for each particle is computed taking into consideration the distance between the position of the nearest detected legs to the evaluated particle and the last known position of legs of the tracked person. Details are also described in [1].

3.3 Experimental results of LRF based method

In order to test the accuracy of the LRF based method several experiments have been carried out in a real indoor environment. A set of five paths on the floor were defined taking into account different trajectories. Two trajectories are straight, one is a circle and the last two are curves. The experiments consist of tracking people whom are following those trajectories. The trajectories have been manually mapped to serve as ground truth on people motion. Five persons participated in the experiments. It is important to acquire data from different

people since each person has a particular gait. Every person walked on each trajectory three times. Thus, 75 different samples can be analyzed to measure the accuracy of the proposal. Notice that laser and images were collected at the same time to build a dataset which is used to evaluate the three approaches shown in this work.

The performance of the LRF based tracking on a trajectory J, is measured taking into account for each time t, the euclidian distance from the hypothesis computed by the tracker h_t, to the real position $p_t \in J$. The correspondence between h_t and p_t should not be made if its distance d_t exceeds a certain threshold H. If $d_t > H$ then the tracker has missed the person in the time t. The tracking error $TE_1(J)$, given a trajectory J, is computed by $TE_1(J) = \frac{\sum_{t=1}^{m_j} d_t}{m_j}$ where m_j is the total number of matches made in the trajectory J.

The algorithm has been evaluated using different numbers of particles: 50, 100, 150, 200 and 250. Table 2 shows the results obtained for each trajectory T and each number of particles. Tracking error $TE_1(J)$ and standard deviation Std_1, in mm, are indicated. The error of tracking decreases when the number of particles increases. For a number of particles higher than 200 the error decreases at lower rate, hence, the final algorithm uses 200 particles. Using 200 particles, the average tracking error for all the trajectories is computed obtaining a value of 34,33 mm and its standard deviation is 9,93 mm. The processing framerate obtained using 200 particles has been 40,02 Hz. Therefore the LRF based approach has a good precision to track a person and it can be used in real time.

Table 2. Results in mm for $TE_1(J)$

	Number of particles									
	50		100		150		200		250	
J	TE_1	Std_1	TE_1	Std_1	TE_1	Std_1	TE_1	Std_1	TE_1	Std_1
1	28.82	47.60	28.90	51.03	27.26	50.71	27.11	47.90	25.32	39.52
2	57.04	53.53	55.60	56.40	49.43	51.75	49.38	63.16	47.83	28.83
3	28.95	30.75	31.18	31.61	27.47	28.56	24.14	31.08	23.87	39.82
4	43.11	39.80	38.65	39.12	32.10	38.88	33.32	44.75	33.07	34.12
5	49.62	40.24	46.67	41.39	42.36	42.27	37.68	40.14	36.84	44.97

Strengths of LRF based approach are the precision, performance rate and wide field of view of the sensor. However if the model of observation is not sufficiently discriminatory then it is not possible to distinguish between two people when their trajectories intersect. In such situations, the tracker can confuse the targets. This problem is illustrated by Fig. 1. In this figure, two persons are tracked by the system. Red and green points represent two different persons and blue points are the laser readings. From up left to down right, the two first scenes show the system properly tracking both people. The two scenes situated below show the situation when two persons are intersecting their trajectories and the system is confused as both trackers end up following the same person. Some proposals try to overcome this problem by including a model of human walking motion [7] or by using a more complex state and observation models and then applying data association techniques [5]. However false positives and tracking

errors are still possible. In this work, a multisensor based tracking is proposed to overcome this problem, and therefore, to achieve multiple people tracking.

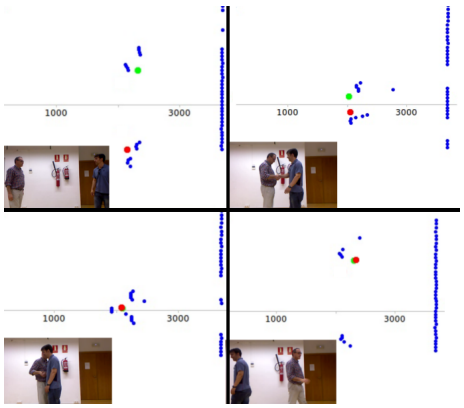

Fig. 1. LRF based approach fails to discriminate between people that get close enough.

4 People detection and tracking using Kinect sensor

Kinect sensor provides both 640x480 distance map and color image. Therefore both color and depth information can be used for people detection and tracking. The Kinect based approach used in this work has been adapted from a previous work by the authors [13]. In the previous work, a traditional stereo camera was used and the color and depth information was fused using fuzzy logic. Now, Kinect is the sensor used and color and depth information is fused using a particle filter. Below Kinect based approach is described.

Notice that in this work people detection and tracking comprise separate processes. When new people are detected then independent trackers are created for every one. People detection is made by using the frontal face detector of OpenCV [10] based on the Viola and Jones' method [18]. Once a person is detected, a model of that person is built by our method. This model is an elliptical region placed at the height of the person chest. Standard anthropomorphic measurements have been taken into account to build this model. The model is resized depending on the distance of people to the Kinect sensor. The center of this elliptical model will be the target to track on the color image. This is done by analyzing its color and depth information. To do so, a color histogram $\hat{q}$ of the elliptical region is calculated using the HSV color space [8]. HSV color space is relatively invariable to illumination changes. A color histogram $\hat{q}$ comprises $n_h n_s$ bins for the hue and saturation. However, chromatic information cannot be considered reliable when the value component is too small or too big. Therefore, pixels on this situation are not used to describe the chromaticity. Because of these pixels might have important information, the histogram includes also

n_v bins to capture its luminance information. Thus, the resulting histogram is composed by $m = n_h n_s + n_v$ bins.

As stated above, we consider an elliptical region of the image to create the color model whose horizontal and vertical axes are h_x and h_y respectively. Let p_c be the ellipse center and $\{p_j\}_{j=1,...,n}$ the locations of the interior pixels. Let's also define a function $b : \Re^2 \rightarrow 1, ..., m$ which associates to the pixel at location p_j the index $b(p_j)$ of the histogram bin corresponding to the color u of that pixel. It is now possible to compute the color density distribution for each bin $\hat{q}(u)$ of the elliptical region with:

$$\hat{q}(u) = \frac{1}{n} \sum_{j=1}^{n} k[b(p_j) - u], \tag{1}$$

where the parameter k is the Kronecker delta function. Please notice that the resulting histogram is normalized, i.e., $\sum_{u=1}^{m} \hat{q}(u) = 1$. After calculating the color model, the Bhattacharyya coefficient as described in [2] can be computed. In the case of a discrete distribution it can be expressed as indicated in Eq. 2. The result expresses the similarity between two color models in the range of $[0, 1]$ where 1 means that they are identical and 0 means that they are completely different.

$$\rho(\hat{q}, \hat{q}') = \sum_{u=1}^{m} \sqrt{\hat{q}(u)\hat{q}'(u)}. \tag{2}$$

Once the Bhattacharyya coefficient is computed, two models of color $\hat{q}$, $\hat{q}'$ can be compared through the Bhattacharyya distance [2]:

$$BD(\hat{q}, \hat{q}') = \sqrt{1 - \rho(\hat{q}, \hat{q}')} \tag{3}$$

It provides values near 0 when two color models are similar and tends to 1 as they differ. An important feature of ρ is that both color models, $\hat{q}$ and $\hat{q}'$, can be compared even if they have been created using regions of different sizes.

4.1 Particle filter based tracking

A particle filter is again used to achieve a robust tracking of detected people. In this case the state at time t is defined as a pair of coordinates (x, y) on the image plus the information of depth Z of that pixel. These coordinates correspond to the pixel centered on the elliptical region that is used to model the detected person. That is, the people position S_t is represented by the state model $S_t = [x_t, y_t, Z_t]$. The prediction is carried out by the model of the state transition. The state transition is defined as $S_t = S_{t-1} + R_{t-1}$ where S_{t-1} is the previous state vector and R_{t-1} is the process noise. A model of people velocity is not explicitly considered in order to manage the unpredictability of human behaviors. The noise is modeled using a Gaussian with average μ_R and standard deviation σ_R. Experimental data have been taken into account to establish the values of μ_R and σ_R in order to model the conditions of the real world.

Condensation algorithm [11] is used to generate a weighted set of particles $(s_i(t), \Pi_i(t))$ where $s_i(t)$ represents an hypothesis of the position of the person being tracked, and $\Pi_i(t)$ is a factor called *importance weight* which provides an

estimation of the observation. At the beginning the algorithm is provided by an initial sample $(s_i(0), \Pi_i(0))$ of N equally weighted particles. At each iteration, the algorithm uses the sample set $(s_i(t-1), \Pi_i(t-1))$ to create a new one. A resample mechanism is used to solve the divergence problem by eliminating particles having low importance weights. Afterwards, the model of state transition is used to predict the motion of the person obtaining the prediction of the state S'_t. The weight $\Pi_i(t)$ of each particle is computed based on the new observation $O(t)$. Then the weights are normalized so that $\sum_{i=1}^{N} \Pi_i(t) = 1$.

The observation model is required to carry out the update. As model of observation, position (x, y) on the image, depth Z and color information are used. On one hand, let $f_{x,y}$ be the euclidian distance in pixels between the position of the particle $s_i(t)$ on the image and the last known state S_t and f_Z the difference of depth between both positions. On the other hand, let $BD(\hat{q}, \hat{q}')$ be the Bhattacharyya distance (Eq. 3) between the corresponding elliptical regions centered on the particle $s_i(t)$ and the last known position S_t. Then, the importance weight of each particle is computed by:

$$\Pi_i(t) = e^{-\frac{1}{2}\left(\frac{f_{x,y}}{\sigma_1}\right)^2} \cdot e^{-\frac{1}{2}\left(\frac{f_Z}{\sigma_2}\right)^2} \cdot (1 - BD(\hat{q}, \hat{q}')). \tag{4}$$

Parameters σ_1, σ_2 correspond to the standard deviations of two zero centered normal distributions, respectively. σ_1, σ_2 have been experimentally tuned. The final person position corresponds to the mean of the state $\mathcal{E}[S(t)]$, calculated as $\mathcal{E}[S(t)] = \sum_{i=1}^{N} \Pi_i(t) s_i(t)$. Please, note that face detection is only used in the detection phase but it is not used in the tracking phase. Therefore, once a person is detected, this person can be tracked using its people model although his or her face is not again detected.

4.2 Experimental results of Kinect based method

The goal is to compare the results of the Kinect based approach with the LRF based approach. Therefore the same dataset collected to test the LRF based approach is used. However both systems use different coordinates systems. As Kinect sensor provides the information of depth, the estimated position on the image x, y can be projected to the LRF coordinates system. Also it is required to have into account that the ground truth was built by measuring the positions of the middle point between the legs and now the target is located at the height of the chest. All these details have been taken into consideration in order to achieve comparable results. The algorithm has been evaluated using different numbers of particles: 50, 100, 150, 200 and 250. Table 3 shows the results obtained for each trajectory T and each number of particles. Tracking error $TE_2(J)$ and standard deviation Std_2 in mm are computed in a similar way to that of the first approach. Once again, the tracking error decreases as the number of particles increases. Using 200 particles, the average tracking error for all the trajectories is computed obtaining a value of 66,17 mm and its standard deviation is 29,29 mm. The processing framerate obtained using 200 particles has been 2,46 Hz. These results point out that the precision of this approach is lower than LRF based approach although it is enough to develop Human-Robot Interaction applications. The main problem is the processing framerate since which is low due to the computing

time required to process the color image and the usage of face detector on each frame in order to detect a new people in the frame.

Table 3. Results in mm for $TE_2(J)$

	Number of particles									
	50		100		150		200		250	
J	TE_2	Std_2	TE_2	Std_2	TE_2	Std_2	TE_2	Std_2	TE_2	Std_2
1	73.63	37.90	58.02	67.69	58.61	72.12	52.01	40.46	50.33	49.72
2	127.27	135.03	111.41	130.71	100.18	126.65	97.80	123.60	97.18	124.62
3	114.44	99.18	98.24	93.33	92.87	81.71	96.45	100.35	86.96	89.32
4	44.04	67.36	39.35	32.86	38.50	32.62	32.92	25.47	33.15	20.15
5	89.38	90.98	67.90	68.50	51.03	45.87	51.68	56.88	51.04	34.21

However the Kinect based approach has a main advantage over the LRF based approach. When two or more people are tracked and the color of their vests are different, the Kinect based approach can still track people without confusing them and it can cope as well with certain level of occlusion. This situation is shown by Fig. 2.

Fig. 2. Tracking two people using the Kinect based approach.

5 People detection and tracking using a multisensor approach

On one hand, LRF based approach is precise and fast but it can confuse the targets when two or more people are being tracked. On the other hand, the Kinect based approach has a lower precision and is slower but it can distinguish people by using color and depth information. Both approaches can suffer from false positives detection. That is, the SVM classifier can recognize laser data as legs in a false way and the OpenCV face detector can recognize faces in

the color image erroneously as well. The multisensor approach fuses information from both sensors in order to achieve a more robust people detection and tracking system. In the detection phase, first the leg detector is used to recognize possible pairs of legs in the proximity of the robot. Second, the possible detected faces are matched to the possible legs and both observations have to be coherent to consider that a new person has been detected. Notice that the fields of view of both devices are different. That is, there can be legs detected but if the person is out of the field of view of Kinect then it is not possible to find the corresponding face. Only when both, legs and face, are detected the system creates a new tracker if the person was not already being tracked.

5.1 Particle filter based tracking

The state definition is similar to the Kinect based approach, but it now includes the information on the position $h_t = \{hx_t, hy_t\}$ of the people legs. Thus, people position S_t is represented by $S_t = [x_t, y_t, Z_t, h_t]$. The state transition and noise models are similar to those explained in Sect. 3 and Sect. 4. The observation model includes both previous kinds of information $f_{x,y}$ and f_Z, so that the importance weight of the particle is computed by:

$$\Pi_i(t) = e^{-\frac{1}{2}\left(\frac{f_{x,y}}{\sigma_1}\right)^2} \cdot e^{-\frac{1}{2}\left(\frac{f_Z}{\sigma_2}\right)^2} \cdot (1 - BD(\hat{q}, \hat{q}')) \cdot e^{-\frac{1}{2}\left(\frac{f_h}{\sigma_3}\right)^2} \tag{5}$$

being f_h the euclidian distance between the position of the nearest detected legs to $s_i(t)$ and the last known position of legs of the tracked person. Parameters σ_1, σ_2 are the same as of those in Eq. 4 and σ_3 correspond to the standard deviation of a zero centered normal distribution. σ_3 has been experimentally tuned. The final person position corresponds to the mean of the state $\mathcal{E}[S(t)]$, calculated as $\mathcal{E}[S(t)] = \sum_{i=1}^{N} \Pi_i(t) s_i(t)$.

5.2 Experimental results of multisensor based method

The idea is to compare the results of the multisensor based approach with the two previous approaches. Therefore the same dataset is used. The transformation of the coordinates is also done in this case. The algorithm has been evaluated using different numbers of particles: 50, 100, 150, 200 and 250. Table 4 shows the results obtained for each trajectory T and each number of particles. Tracking error $TE_3(J)$ and standard deviation Std_3, in mm, are indicated. Again the error of tracking decreases as the number of particles increases. Using 200 particles, the average tracking error for all the trajectories is computed, obtaining an average value of 40,88 mm and its standard deviation is 10,18 mm. The processing framerate obtained using 200 particles has been 3,24 Hz. The results point that the precision of this approach is lower than LRF based approach but higher than the Kinect based approach. Nevertheless, we think that it is enough to develop Human-Robot Interaction applications. Also, the multisensor approach can distinguish several people depending on the color of their vest and avoid some false positives of both face and legs detectors. It is the most robust approach and certain level of occlusion can been managed by the system. The processing framerate has improved regarding to the Kinect based approach due

to that the face detector is not used for each frame but only when an additional legs are detected. However it is still slow due to the computing time required to process the color image. Although the frame rate is low, some applications on mobile robots have been developed using similar processing framerates taking into account certain limits. For instance, in [5] a multisensor human detection and tracking system at 4 Hz is used to follow people. Also, in [9] a multisensor system running a face detector at a rate of 3 Hz provides good results to follow people moving in a standard office domain.

Table 4. Results in mm for $TE_3(J)$

	Number of particles									
	50		100		150		200		250	
J	TE_3	Std_3	TE_3	Std_3	TE_3	Std_3	TE_3	Std_3	TE_3	Std_3
1	44.16	40.61	42.08	44.77	38.96	43.23	32.54	42.39	30.87	42.95
2	78.69	52.55	66.15	55.45	61.55	54.95	58.63	48.67	58.11	63.11
3	50.12	52.48	49.86	58.52	41.74	55.37	37.64	51.54	34.63	55.44
4	56.76	50.45	53.24	58.44	39.50	60.11	37.42	57.35	35.97	55.27
5	45.17	28.55	42.15	27.49	38.97	39.20	38.15	28.64	37.86	41.64

6 Conclusions and future work

In this paper a new multisensor system to detect and track people in the proximity of a mobile robot has been proposed. The multisensor approach tries to put together the strengths of both LRF and Kinect sensors. To explain the develop of the multisensor system and its advantages, first the LRF based approach is shown and experimentally evaluated. This method analyses certain geometric features present in the laser data in order to detect possible legs of people. A classifier is trained using SVM to classify the data obtained from laser using instances of patterns of legs. The LRF based approach is briefly described because is based on a previous work by the authors. Second, a new Kinect based approach has been developed for this work. The second approach has been also experimentally evaluated and results show less precision and more computation time than the first one but it can distinguish people using the color of their vests. The best results have been obtained by the multisensor system. The multisensor based approach is able to detect and track people in a real indoor environment obtaining average tracking error of approximately 4 cm. The main contributions of this work are the development of the multisensor based approach and the method to fuse color and depth information of the Kinect sensor with distance information of a LRF sensor using a particle filter. As future work the goal will be to improve the processing framerate. To do so, for instance, one possibility is to reduce the resolution of the images so that less data have to be processed. Another idea is to execute the face detector only on certain parts of the images instead of on the whole frame. Also, parallel computing can be used to improve the speed. Finally, depending on the required precision, the number of particles used can be lowered in order to reduce the processing time.

References

1. E. Aguirre, M. Garcia-Silvente, and J. Plata. Leg detection and tracking for a mobile robot and based on a laser device, supervised learning and particle filtering. In *ROBOT2013: First Iberian Robotics Conference*, volume 252 of *Advances in Intelligent Systems and Computing*, pages 433–440. Springer Int. Publis., 2014.
2. F. Aherne, N. Thacker, and P. Rockett. The bhattacharyya metric as an absolute similarity measure for frequency coded data. *Kybernetica*, 32:1–7, 1997.
3. A. Ansuategui, A. Ibarguren, J.M. Martínez-Otzeta, C. Tubío, and E. Lazkano. Particle filtering for people following behavior using laser scans and stereo vision. *International Journal on Artificial Intelligence Tools*, 20(02):313–326, 2011.
4. M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.
5. N. Bellotto and H. Hu. Multisensor-based human detection and tracking for mobile service robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(1):167–181, 2009.
6. Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
7. W. Chung, H. Kim, Y. Yoo, C.-B. Moon, and J. Park. The detection and following of human legs through inductive approaches for a mobile robot with a single laser range finder. *IEEE Transactions on Industrial Electronics*, 59(8):3156–3166, 2012.
8. J.D. Foley and A. van Dam. *Fundamentals of Interactive Computer Graphics*. Addison Wesley, 1982.
9. J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Plötz, G. A. Fink, and G. Sagerer. Multi-modal anchoring for human-robot interaction. *Robotics and Autonomous Systems*, 43(2-3):133–147, 2003.
10. Intel-Corporation. Opencv, 2015.
11. M. Isard and A. Blake. Condensation-conditional density propagation for visual trackings. *International Journal of Computer Vision*, 29:5–28, 1998.
12. Microsoft. Kinect official webpage, 2010.
13. R. Paúl, E. Aguirre, M. García-Silvente, and R. Muñoz-Salinas. A new fuzzy based algorithm for solving stereo vagueness in detecting and tracking people. *International Journal of Approximate Reasoning*, 53:693–708, 2012.
14. Primesense. Primesense official webpage, 2005.
15. K. Schenk, M. Eisenbach, A. Kolarow, and H. Gross. Comparison of laser-based person tracking at feet and upper-body height. In J. Bach and S. Edelkamp, editors, *KI 2011: Advances in Artificial Intelligence*, volume 7006 of *Lecture Notes in Computer Science*, pages 277–288. Springer, 2011.
16. Xiaowei Shao, K. Katabira, R. Shibasaki, Huijing Zhao, and Y. Nakagawa. Tracking a variable number of pedestrians in crowded scenes by using laser range scanners. In *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*, pages 1545–1551, Oct 2008.
17. L. Susperregi, J. M. Martínez-Otzeta, A. Ansuategui, A. Ibarguren, and B. Sierra. RGB-D, laser and thermal sensor fusion for people following in a mobile robot. *International Journal of Advanced Robotic Systems*, 10:271, 2013.
18. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 511–518, 2001.
19. Lu Xia, Chia-Chih Chen, and J.K. Aggarwal. Human detection using depth information by kinect. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 15–22, June 2011.
20. Z. Zhang. Microsoft kinect sensor and its effect. *MultiMedia, IEEE*, 19(2):4–10, 2012.