Proceedings of the 2006 IEEE
International Conference on Robotics and Biomimetics
December 17 - 20, 2006, Kunming, China

Vision and Laser Data Fusion for Tracking People with a Mobile Robot

Nicola Bellotto and Huosheng Hu
Department of Computer Science, University of Essex
Colchester CO4 3SQ, U.K.
Email: nbello@essex.ac.uk, hhu@essex.ac.uk

Abstract— **In this paper we present a multi-sensor fusion system for tracking people with a mobile robot, which integrates the information provided by a laser range sensor and a PTZ camera. We introduce the algorithms used for detecting legs from laser scans and faces from video images, then we illustrate a human motion model for the estimation of people position, orientation and height. The ego-motion of the robot is also taken into account and the information fused using an implementation of the Unscented Kalman Filter. Finally, multiple human tracks are generated and maintained thanks to an appropriate data association procedure. The results of several experiments are illustrated, proving the effectiveness of our approach, and some considerations drawn.**

I. Introduction

In recent years there has been an increasing interest towards the social aspect of mobile robots, with a growing number of applications which involves interaction between robots and general public. A relatively new subject, called Human Robot Interaction (HRI), has attracted the attention of a big part of the research community as well as industry. For these social robots, interacting does not only mean having good communication skills, but also reacting properly to the people and be aware of their presence. In this context, a system able to track the persons around the robot is helpful, if not necessary. Several techniques for multi-target tracking, used in the past mainly for aviation and military applications, can be applied to the sensor equipment of a mobile robot to track nearby people. The task is particularly challenging because the human motion is very unpredictable. Other factors related to the environment and the sensor performance may also put strict limits on the capacity to detect people and distinguish individuals.

In literature, there is a consistent number of solutions for tracking people with a mobile robot, but most of them can be categorized using the following criteria: a) tracking *one vs. many* people and b) from a *fixed vs. moving* platform. The work presented in [1] is an example of people tracking from a fixed position where they use stereo-vision to detect and track multiple persons, each of them is assigned a Kalman filter. Same filter but different sensor are used instead in [2], where the static robot can track people with a 2D laser using motion patterns previously learnt. A laser is also used in [3] to detect walking humans: the difference in this case is that the robot can move but only one person at a time can be tracked. Stereo-vision and Kalman filter are used again in [4] for the tracking of a single human while the robot moves to accomplish a following behaviour. As expected, there are less examples where the robot can move and at the same time track multiple humans. Among them there is the solutions adopted in [5], where they use particle filters to perform simultaneous localization and people tracking based on laser range readings. This approach however is based on a simple Brownian model of the human motion, which seems not particularly robust in case of clutters (e.g. two people walking closely). Laser and particle filters are also used while the robot is moving in [6], performing multiple tracking with Joint Probabilistic Data Association (JPDA) and adopting a linear motion model of the humans to reduce the clutter problem.

In all these cases, tracking is performed using one single device, laser or camera. When both are present, the laser is normally preferred for the tracking part, while the camera is only used to extract some features which help to identify the person. Our approach instead shows how the data from the two devices can be successfully fused using an implementation of Unscented Kalman Filter, obtaining additional information and improving the tracking performance.

This paper is organized as follows. Section II introduces our sensors and algorithms for people detection; Section III describes the human state estimator; Section IV explains the data association procedure; Section V illustrates several experimental results and analysis; finally Section VI presents a brief conclusion and future work.

II. People Detection

The components of a tracking system are the sensors used to detect the targets and the algorithms for elaborating the information provided by them. The robot we use is equipped with a laser range sensor and a PTZ camera. The laser, which covers the semi-circular area in front of the robot, is placed a few decimeters from the floor, so that the scanning can detect human legs in most of the cases. The camera instead is mounted on a special support, approximately 1.5m high, in a good position to spot faces.

A. Legs Detection

Laser range sensors are often used to detect persons [3], [7]. However in most of the cases they consider only moving objects, which means a static human cannot be detected. Moreover, with such an approach it becomes problematic to observe walking people while the robot is also moving,

1-4244-0571-8/06/$20.00 ©2006 IEEE

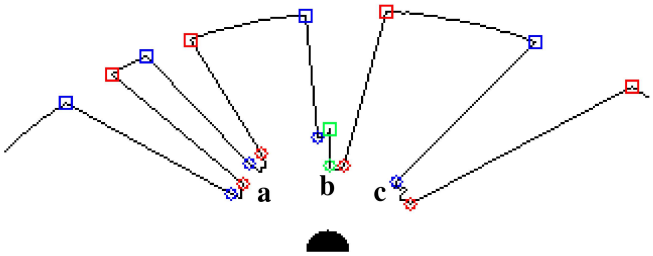

Fig. 1. Legs patterns extracted from a laser scan. Patterns (a) and (b) are very seldom confused with other objects, while (c) can be a little ambiguous for particular environments.

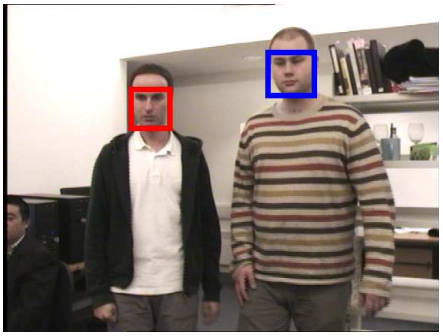

Fig. 2. Example of face detection. From the x-y position of the face on the image, bearing and elevation relative to the camera are easily computed.

because the ego-motion must be compensated. Our algorithm instead, derived from a previous work [8], is based on the recognition of three typical legs patterns which are extracted from a laser scan, as illustrated in Fig. 1. Of course these patterns are constrained by some dimensional limits. Thanks also to the high precision of the laser device, the procedure returns accurate bearing b^l and distance r^l of the people within a range of several meters.

B. Face Detection

The face detection system is an improved version of our previous work [8], which is based on the object detection algorithm of [9] and performs well in real time even under challenging conditions. Thanks to simple transformations, the position of each face in the image is converted to get the relative bearing b^f and elevation e^f with respect to the current camera orientation. An example of face detection is illustrated in Fig. 2.

III. State Estimation

Tracking a walking person is a challenging task, even more complex if performed from a mobile platform. The Kalman filter [10] is a well known Bayesan estimator which provides an elegant way to fuse the information from different sources. We introduce briefly a recent variant of this estimator which is normally used for non-linear systems: the Unscented Kalman Filter (UKF) [11]. Then we illustrate the prediction and observation models used by this filter to estimate the human state.

A. Unscented Kalman Filter

A typical non-linear system can be written as follows:

$$\begin{aligned} \mathbf{x}_k &= f(\mathbf{x}_{k-1}, \mathbf{u}_{k-1}, \mathbf{w}_{k-1}) \\ \mathbf{z}_k &= h(\mathbf{x}_k, \mathbf{v}_k) \end{aligned} \tag{1}$$

where $\mathbf{x}$ and $\mathbf{z}$ are respectively the state and the observation vectors, $\mathbf{u}$ is the control input, $\mathbf{w}$ and $\mathbf{v}$ are noises and k is the current time step. Instead of using a linear approximation like in the EKF [10], the UKF models the state uncertainty with a set of weighted sample points, called also *sigma points*, in order to capture the true mean and covariance for any non-linear system. These points $\boldsymbol{\mathcal{X}}_i$ and associated weights W_i are calculated as follows:

$$\begin{aligned} \boldsymbol{\mathcal{X}}_0 &= \bar{\mathbf{x}} & W_0 &= \beta/(n+\beta) \\ \boldsymbol{\mathcal{X}}_i &= \bar{\mathbf{x}} + \left(\sqrt{(n+\beta)\,\mathbf{P}_{xx}}\right)_i & W_i &= 1/2\,(n+\beta) \\ \boldsymbol{\mathcal{X}}_{i+n} &= \bar{\mathbf{x}} - \left(\sqrt{(n+\beta)\,\mathbf{P}_{xx}}\right)_i & W_{i+n} &= 1/2\,(n+\beta) \end{aligned} \tag{2}$$

for $i = 1, \ldots, n$, where $\bar{\mathbf{x}}$, $\mathbf{P}_{xx}$ and n are respectively mean, covariance and dimension of the state $\mathbf{x}$, β is a parameter for tuning the higher order moments of the approximation (normally set so that $n + \beta = 3$ for Gaussian distributions) and $\left(\sqrt{(n+\beta)\,\mathbf{P}_{xx}}\right)_i$ is the i^{th} column or row of the matrix square root of $\mathbf{P}_{xx}$. Mean and covariance of the non-linear transformation $\mathbf{y} = f(\mathbf{x})$ are calculated as follows:

$$\boldsymbol{\mathcal{Y}}_i = f\left(\boldsymbol{\mathcal{X}}_i\right) \tag{3}$$

$$\bar{\mathbf{y}} = \sum_{i=0}^{2n} W_i \boldsymbol{\mathcal{Y}}_i \tag{4}$$

$$\mathbf{P}_{yy} = \sum_{i=0}^{2n} W_i \left[\boldsymbol{\mathcal{Y}}_i - \bar{\mathbf{y}}\right]\left[\boldsymbol{\mathcal{Y}}_i - \bar{\mathbf{y}}\right]^T \tag{5}$$

This procedure, called unscented transformation (UT), is at the base of the current filter[1].

Like for the EKF, the estimation procedure of the UKF consists of two steps, prediction and correction. First of all, the state is augmented to include the process noise[2], so to have $\mathbf{x}^a = [\mathbf{x}\ \mathbf{w}]^T$, and the relative $2n^a + 1$ sigma points $\boldsymbol{\mathcal{X}}_i$ are generated from the last estimation using (2). Then the *a priori* mean and covariance of the state are predicted with the UT:

$$\boldsymbol{\mathcal{X}}_{i_k^-} = f\left(\boldsymbol{\mathcal{X}}_{i_{k-1}}, \mathbf{u}_{k-1}\right) \quad \text{for } i = 0, \ldots, 2n^a \tag{6}$$

$$\hat{\mathbf{x}}_k^{a-} = \sum_{i=0}^{2n^a} W_i \boldsymbol{\mathcal{X}}_{i_k^-} \tag{7}$$

$$\mathbf{P}_k^- = \sum_{i=0}^{2n^a} W_i \left[\boldsymbol{\mathcal{X}}_{i_k^-} - \hat{\mathbf{x}}_k^{a-}\right]\left[\boldsymbol{\mathcal{X}}_{i_k^-} - \hat{\mathbf{x}}_k^{a-}\right]^T \tag{8}$$

[1]To avoid non-positive, semidefinite covariances when $\beta < 0$, it is possible to use a modified form [12] given by $\mathbf{P}_{yy}^{MOD} = \mathbf{P}_{yy} + [\boldsymbol{\mathcal{Y}}_0 - \bar{\mathbf{y}}]\,[\boldsymbol{\mathcal{Y}}_0 - \bar{\mathbf{y}}]^T$.

[2]The observation noise $\mathbf{v}$ could be also included in $\mathbf{x}^a$.

For the correction stage, the UT is used again to predict the observation as follows:

$$\boldsymbol{\mathcal{Z}}_{ik} = h\left(\boldsymbol{\mathcal{X}}_{ik}^{-}\right) \quad \text{for } i = 0, \ldots, 2n^a \tag{9}$$

$$\hat{\mathbf{z}}_k = \sum_{i=0}^{2n^a} W_i \boldsymbol{\mathcal{Z}}_{ik} \tag{10}$$

The innovation covariance and the cross-correlation matrices are then computed using the following equations:

$$\mathbf{P}_{\nu\nu k} = \mathbf{R}_k + \sum_{i=0}^{2n^a} W_i \left[\boldsymbol{\mathcal{Z}}_{ik} - \hat{\mathbf{z}}_k\right]\left[\boldsymbol{\mathcal{Z}}_{ik} - \hat{\mathbf{z}}_k\right]^T \tag{11}$$

$$\mathbf{P}_{xzk} = \sum_{i=0}^{2n^a} W_i \left[\boldsymbol{\mathcal{X}}_{ik}^{-} - \hat{\mathbf{x}}_k^{a-}\right]\left[\boldsymbol{\mathcal{Z}}_{ik} - \hat{\mathbf{z}}_k\right]^T \tag{12}$$

Finally, the gain $\mathbf{K}_k$ is calculated and used to correct the estimation and its covariance as follows:

$$\mathbf{K}_k = \mathbf{P}_{xzk}\mathbf{P}_{\nu\nu k}^{-1} \tag{13}$$

$$\hat{\mathbf{x}}_k^a = \hat{\mathbf{x}}_k^{a-} + \mathbf{K}_k\left[\mathbf{y}_k - \hat{\mathbf{z}}_k\right] \tag{14}$$

$$\mathbf{P}_k = \mathbf{P}_k^{-} - \mathbf{K}_k\mathbf{P}_{\nu\nu k}\mathbf{K}_k^T \tag{15}$$

where $\mathbf{y}_k$ is the current measure given by the sensor.

Despite the increased complexity due to the sigma points, the advantage of the UKF with respect to the EKF is that the absence of linearization improves the estimation performance and avoid the calculus of Jacobian matrices.

B. Human Models

Modelling human motion is a difficult task because most of the time people's behaviours are unpredictable. Many applications for human tracking are based on a simple Brownian model [5], but to handle occlusions a constant velocity model is a better choice [4], [6]. The following model extends the latter case including two novelties:

- the estimation of the human height z^h;
- the assumption that the velocity v^h is always positive.

The estimation of z^h is possible thanks to our face detection system, knowing also the tilt of the camera and its height from the ground. The human height adds a third dimension to the tracking space and this is particularly useful for distinguishing and assigning the correct track to different persons (data association), rather than improving the performances of the tracking itself. Instead, the assumption that v^h is always positive takes into account the fact that a person, when is walking, is normally heading forward. This is a plausible constraint that permits the correct estimation of the human orientation, which otherwise could be subjected to an error of 180° (i.e. the person is walking backward). Of course, the orientation can be estimated only when the human target is moving (v^h not null). In addition to z^h and v^h, the model below includes then the 2D position $\left(x^h, y^h\right)$ of the human and the orientation ϕ^h:

$$\begin{cases} x_k^h = x_{k-1}^h + v_{k-1}^h \cdot \Delta t_k \cdot \cos\phi_{k-1}^h \\ y_k^h = y_{k-1}^h + v_{k-1}^h \cdot \Delta t_k \cdot \sin\phi_{k-1}^h \\ z_k^h = z_{k-1}^h \\ \phi_k^h = \phi_{k-1}^h \\ v_k^h = \left|v_{k-1}^h\right| \end{cases} \tag{16}$$

where $\Delta t_k = t_k - t_{k-1}$ is the time interval. Errors are modeled as additive zero-mean Gaussian noises and are omitted here for simplicity.

Differently from the previous ones, the human observation model is quite complex. The available measurements, coming from the face and legs detection, are bearing b_m^f and elevation e_m^f of the human face, plus bearing b^l and range r^l of the legs. Of course these observations depend on the position and orientation of the robot and its camera, therefore the model includes the following values: 2D position (x^r, y^r) and orientation ϕ^r of the robot given by the odometry, plus pan ψ^r and tilt θ^r of the camera. To complicate things is the fact that camera and laser are not aligned with the robot central axis, therefore their displacement must also be taken into account with the following transformations:

$$\begin{aligned} c_{x0} &= x^r + c_x \cdot \cos\phi^r \quad & c_{y0} &= y^r + c_x \cdot \sin\phi^r \\ l_{x0} &= x^r + l_x \cdot \cos\phi^r \quad & l_{y0} &= y^r + l_x \cdot \sin\phi^r \end{aligned} \tag{17}$$

where c_x and l_x are the distances of camera and laser from the robot's centre, lying on its longitudinal axis (c_y and l_y are both null). Given also the height c_z of the camera from the ground, the whole observation model is the following:

$$\begin{cases} b_k^f = \arctan\dfrac{y_k^h - c_{y0}}{x_k^h - c_{x0}} - \phi_k^r - \psi_k^r \\ e_k^f = -\arctan\dfrac{z_k^h - c_z}{\sqrt{\left(x_k^h - c_{x0}\right)^2 + \left(y_k^h - c_{y0}\right)^2}} - \theta_k^r \\ b_k^l = \arctan\dfrac{y_k^h - l_{y0}}{x_k^h - l_{x0}} - \phi_k^r \\ r_k^l = \sqrt{\left(x_k^h - l_{x0}\right)^2 + \left(y_k^h - l_{y0}\right)^2} \end{cases} \tag{18}$$

The two models above (16) and (18) are respectively the practical implementations of the systems $f(\cdot)$ and $h(\cdot)$ in (1), from which we derive the necessary equations for the UKF. Please note also that the innovation is always normalized between $[-180^\circ, 180^\circ)$ for all the angular components of the observation vector in order to avoid the divergence of the filter.

IV. Data Association

In order to perform multiple tracking, each reading coming from the sensors must be correctly assigned to the proper human. A general scheme of data association is illustrated in [13] and can be summarized in the following steps: 1) update the states of the candidate tracks to the observation time and compute the predicted observations; 2) remove real observations which cannot match any track (gating); 3) form and association matrix of similarities between each pair of

real/predicted observations; 4) assign the current real observations to the proper tracks. The whole process is described below.

At time k, the set of measurements $\mathcal{Y}_k$ contains the following elements (hereafter we omit the subscript k, being clear the time step is always the same):

$$\mathbf{y}_m^f = \begin{bmatrix} b_m^f \\ e_m^f \end{bmatrix} \quad \mathbf{y}_n^l = \begin{bmatrix} b_n^l \\ r_n^l \end{bmatrix} \tag{19}$$

where $\mathbf{y}_m^f$ includes bearing and elevation of the m^{th} face, while $\mathbf{y}_n^l$ includes bearing and range of the n^{th} legs pair.

The first action is to update the candidate tracks to the current time step. This correspond to a prediction of the human states $\mathbf{x}_{1\,k}^{h\,-} \dots \mathbf{x}_{H}^{h\,-}$. Then the expected observations $\mathbf{z}_j^f$ and $\mathbf{z}_j^l$ have to be predicted. Here the current robot and camera's state $\mathbf{x}^r$ have also to be considered, as explained in Section III-B.

We adopt a common gating approach that consists in excluding all the measurements $\mathbf{y}_i$ outside a *validation region* [14]. This region, constructed around the predicted observation $\mathbf{z}_j$, is determined by the relation $d(\mathbf{y}_i, \mathbf{z}_j) \leq \lambda$, where d is the Mahalanobis (or statistical) distance defined as follows:

$$\begin{aligned} d(\mathbf{y}_i, \mathbf{z}_j) &= \sqrt{(\mathbf{y}_i - \mathbf{z}_j)^T \mathbf{C}_{ij}^{-1} (\mathbf{y}_i - \mathbf{z}_j)} \\ \mathbf{C}_{ij} &= cov(\mathbf{y}_i - \mathbf{z}_j) \end{aligned} \tag{20}$$

and λ is a threshold. Under the assumption that d^2 is χ^2 (chi-squared) distributed with n degrees of freedom for an n-dimensional measurement vector, the value of λ can be determined from tables of the χ^2 distribution. In this case $\mathbf{z}_j^f$ and $\mathbf{z}_j^l$ are both 2-dimensional, so we choose $\lambda = 3.03$ in order to have a probability $P_G = 0.99$ that a measurement generated by a human target falls inside the validation region.

For each pair of real/predicted observations a similarity measure is then calculated and used to form an association matrix. The similarity is given by the same Mahalanobis distance introduced in (20). We distinguish between observations relative to faces and legs, generating two different association matrices as follows:

$$\mathbf{S}^f = \begin{bmatrix} s_{11}^f & \dots & s_{1H}^f \\ \vdots & \dots & \vdots \\ s_{M1}^f & \dots & s_{MH}^f \end{bmatrix} \quad \mathbf{S}^l = \begin{bmatrix} s_{11}^l & \dots & s_{1H}^l \\ \vdots & \dots & \vdots \\ s_{N1}^l & \dots & s_{NH}^l \end{bmatrix} \tag{21}$$

where the elements are the similarities given by the following Mahalanobis distances:

$$s_{mj}^f = d\big(\mathbf{y}_m^f, \mathbf{z}_j^f\big) \quad s_{nj}^l = d\big(\mathbf{y}_n^l, \mathbf{z}_j^l\big) \tag{22}$$

Given the association matrices then, the simplest way to assign a real measurement to a predicted observation is the Nearest Neighbour (NN) technique [13], [14], which in practice consists of choosing the pairs $(\mathbf{y}_i, \mathbf{z}_j)$ with the highest similarities (i.e. lowest s_{ij}). Differently from other methods like JPDA [6] or MHT [7], the NN association is one-to-one, that is only one measurement is assigned to one prediction at each update step. This choice seems to be reasonable for most of the cases where the set of entities to track is not too dense [2], [5], as shown also in our experiments.

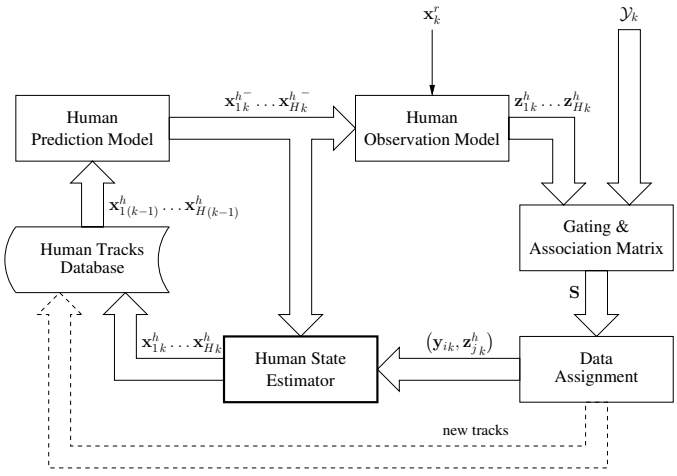

Fig. 3. Schematic representation of the system, including state estimators and data association. The process starts extracting candidates $\mathbf{x}_{1\,(k-1)}^h \dots \mathbf{x}_{H\,(k-1)}^h$ from the target database and assigns to each state the proper observations from the set $\mathcal{Y}_k$. The pairs $\left(\mathbf{y}_{i\,k}, \mathbf{z}_{j\,k}^h\right)$ are eventually used for the correction of the human state estimations.

The laser readings excluded by the gating procedure or the NN assignment are considered for the generation of new tracks[3]. Basically, parallel to the main human tracks database there is another list containing all the possible candidates. Each one of these is created by a sequence of readings falling inside a certain region, which is delimited by the maximum possible distance covered in the interval Δt_k (i.e. maximum human velocity). Every candidate is assigned a maximum lifetime during which, after a certain number of readings, in can be promoted to human track. Instead, if no further readings fall inside its region, the candidate is removed. Whereas the initial position, orientation and velocity of a new track can be simply obtained with the laser, the same is obviously not possible for the height. Therefore, our approach is to set initially an "average" height, for example 1.5m, and a flag indicating that such height is just temporary. As long as this flag is set, the variance of the relative height is kept very high. When the face is detected, the flag is unset and the height is normally updated by the estimator together with its variance.

Another important issue is of course the tracks' deletion. First of all, from the human track database we continuously remove the elements which have not been updated by the Kalman estimator for more than a certain time. Also, we delete the tracks which are too close to each other, in order to avoid multiple tracking of the same human target. In practice, if two tracks are closer than a certain distance, the track with the highest covariance is removed.

The whole process is illustrated in Fig. 3. Although the observation vectors and the association matrix have been generalized in the figure, we remind that the process is actually split in two parts: one for the face and another for the

[3]Actually only laser patterns (a) and (b), explained in Section II-A, are considered trustful for candidate creation. Face readings are not taken into account as they cannot provide range information.

Fig. 4. Robot equipped with laser and PTZ camera for the experiments. The camera used is the black one on the top, about 1.5m high. The blue device is the laser, at approximately 0.4m from the floor.

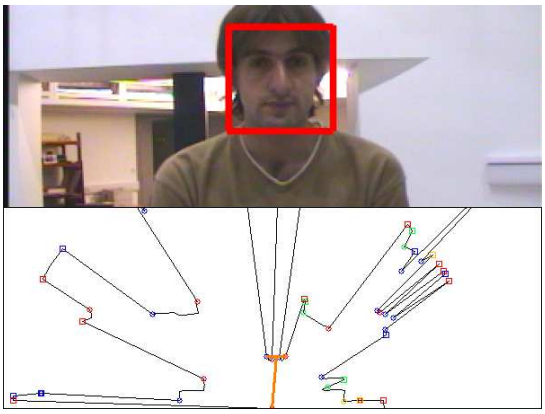

Fig. 5. A view of the experiments scenario, as seen by the robot, and a laser scan that shows the irregularity of the environment. The orange line on the laser scan indicates the detected legs.

legs detection. Once we have obtained the two sets of pairs $\left(\mathbf{y}_m^f, \mathbf{z}_j^f\right)$ and $\left(\mathbf{y}_n^l, \mathbf{z}_j^l\right)$, we have to consider that not only they can differ in size, but also may contain predictions which refer to different humans. For example, given three people $\{A, B, C\}$, we could have detected only the face of human A plus the legs of humans B and C. In this case we estimate the current human state using only "half" observation, $\mathbf{y}_i^f$ or $\mathbf{y}_i^l$, and setting to null the innovation of the missing component.

V. Experimental Results and Analysis

The robot used for our experiment is a Pioneer platform equipped with a laser and a PTZ camera, mounted on a proper support as shown in Fig. 4. The whole software has been implemented in C++ and runs in real-time on the robot's on-board computer, which is a PIII 800MHz. In particular, we derived the UKF from the Bayes++ library [15] and access to the hardware through the Player middleware [16]. Finally, the true position of the people was given by an external video-based tracking system that makes use of a ceiling mounted camera, as shown by the snapshot in Fig. 7.

We tested our systems in three different situations: a) with a single human and the robot moving; b) with several people and the robot stopped; c) with several people and the robot moving. The experiments were run in our robotic arena, in a scenario very challenging for both the face and the legs detection, as shown in Fig. 5. The observation frequency is approximately 10Hz and the time length of every trial is about 60s. Every track is created after 3 readings at maximum intervals of 0.5s and is removed if not updated for more than 1s. The results are reported in Table I. For each experimental case (A, B, C) and person (1, 2, 3) we indicate the Root Mean Square Error (RMSE), plus standard deviation (SD), minimum value and maximum value of the error.

A. Tracking a Single Human

In this experiment we evaluated the performance of the tracking system's accuracy when a person and the robot are both moving. The relative paths are illustrated in Fig 6 and the results in Table I (case A). We can see that the RMSE is relatively small, despite the fact that the robot was almost always moving and turning. In some cases, when the robot was not facing to the human for more than 1s, the track was lost; as soon as the robot could detect the person again, a new track was promptly recreated. Please note that the RMSE could be further reduced just decreasing the threshold of maximum time loss, for example from 1s to 0.5s, but this of course would reduce also our capability to predict the human position in case of clutters.

B. Tracking People from a Fixed Position

The complexity of the task increased considerably with the presence of more persons. Indeed, the three people involved in the experiment often crossed and sometimes touched each other, so there were many situations in which a person covered another one or they were so close to be confused as a single entity. It is clear then that for this experiment the performance of the data association is crucial. From the results in Table I (case B) we can see that the tracking was still accurate enough to be compared to the previous single-human case. We have to note however that a couple of times the data

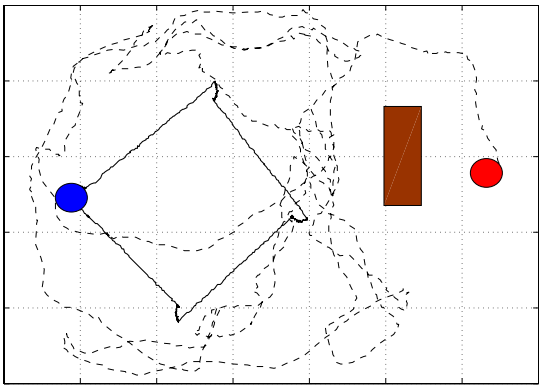

Fig. 6. Paths of the robot (thick line) during the tracking of a single human (dashed line). The circles indicate their starting points and the rectangle is a 2m high wooden wall.

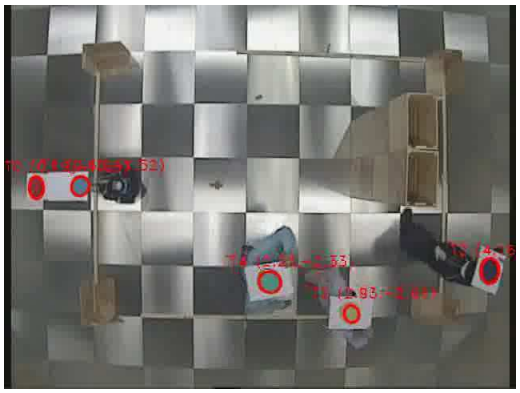

Fig. 7. Image from the ceiling camera. Each target wears a color marker for absolute position tracking. The robot (on the left) has two markers to retrieve its orientation. Here the two human targets on the bottom of the image are very close to each other and the laser cannot distinguish their legs.

TABLE I
TRACKING ERROR

Case	Person	RMSE [m]	SD [m]	Min [m]	Max [m]
A	(1)	0.26	0.13	0.01	0.64
B	(1)	0.22	0.11	0.03	0.42
	(2)	0.29	0.16	0.07	1.08
	(3)	0.31	0.15	0.05	1.07
C	(1)	0.40	0.28	0.03	2.83
	(2)	0.29	0.14	0.03	0.90
	(3)	0.27	0.09	0.07	0.42

association failed and the track originally generated for one person switched to a different one. These cases happened when the two people walked very close, out of the camera's field of view, and the laser data was not enough to distinguish the two different pairs of legs, like the situation illustrated in Fig. 7. Such errors could be probably reduced pointing the camera torwards targets which are moving very close to each other, trying to spot faces or other features: this is a possible solution we are currently investigating.

C. Tracking People while Moving

The last experiment was performed moving the robot in the environment with the same people wandering around. The results in Table I (case C) show the tracking error was still low and comparable to the previous cases. Like before, there have been a few data association errors. This is the main reason, for example, of the big maximum error in case C-(1), caused by a track which has been generated by person (1) and then updated with some incorrect readings not belonging to him. It seems instead that the robot's motion does not influence much the performance of the tracking, even in the worst case when the robot is turning around. We are pretty confident this will be also demonstrated with future experiments in a bigger environment.

VI. CONCLUSION AND FUTURE WORK

The work described in this paper illustrates a system for tracking multiple humans with a mobile robot. Differently from other solutions, we do not rely only on one device, that is camera or laser, but we perform data fusion in order to integrate the information provided by both of them. An implementation of the UKF for human state estimation has been proposed, together with data association and maintenance of multiple tracks. The system has been tested experimentally and considerations drawn from the results, showing the good performance of our solution.

We are currently improving our system to include the uncertainty of the robot motion in the human state estimation. We are also extending it in order to perform, in real-time, concurrent people tracking and recognition.

REFERENCES

[1] D. Beymer and K. Konolige, "Real-time tracking of multiple people using continous detection," in *Proc. of the Int. Conf. on Computer Vision*, Kerkyra, Greece, 1999. [Online]. Available: http://www.ai.sri.com/~beymer/vsam/index.html

[2] M. Bennewitz, G. Cielniak, and W. Burgard, "Utilizing learned motion patterns to robustly track persons," in *Proc. of Joint IEEE Int. Workshop on VS-PETS*, Nice, France, 2003, pp. 102–109.

[3] M. Lindström and J.-O. Eklundh, "Detecting and tracking moving objects from a mobile platform using a laser range scanner," in *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, vol. 3, Maui, HI, USA, 2001, pp. 1364–1369.

[4] D. Beymer and K. Konolige, "Tracking people from a mobile platform," in *Proc. of IJCAI-2001 Workshop on Reasoning with Uncertainty in Robotics*, Seattle, WA, USA, 2001. [Online]. Available: http://www.aass.oru.se/Agora/RUR01/proceedings.html

[5] M. Montemerlo, W. Whittaker, and S. Thrun, "Conditional particle filters for simultaneous mobile robot localization and people-tracking," in *Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA)*, Washington DC, USA, 2002, pp. 695–701.

[6] D. Schulz, W. Burgard, D. Fox, and A. B. Cremers, "People tracking with mobile robots using sample-based joint probabilistic data association filters." *Int. Journal of Robotic Research*, vol. 22, no. 2, pp. 99–116, 2003.

[7] J. Bobruk and D. Austin, "Laser motion detection and hypothesis tracking from a mobile platform," in *Proc. of the 2004 Australian Conference on Robotics & Automation*, Canberra, Australia, 2004. [Online]. Available: http://www.araa.asn.au/acra/acra2004/

[8] N. Bellotto and H. Hu, "Multisensor integration for human-robot interaction," *The IEEE Journal of Intelligent Cybernetic Systems*, vol. 1, July 2005. [Online]. Available: http://www.cybernetic.org.uk/ics

[9] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Proc. the IEEE Int. Conf. on Image Processing 2002*, vol. 1, New York, USA, 2002, pp. 900–903.

[10] G. Welch and G. Bishop, "An introduction to the kalman filter," University of North Carolina at Chapel Hill, Department of Computer Science, Tech. Rep. TR 95-041, 2004.

[11] S. J. Julier and J. K. Uhlmann, "A new extension of the kalman filter to nonlinear systems," in *Proc. of SPIE AeroSense Symposium*, FL, USA, 1997. [Online]. Available: http://www.cs.unc.edu/~welch/kalman/

[12] S. J. Julier, J. K. Uhlmann, and H. F. Durrant-Whyte, "A new method for the nonlinear transformation of means and covariances in filters and estimators," *IEEE Trans. on Automatic Control*, vol. 45, no. 3, pp. 477–482, March 2000.

[13] D. L. Hall, *Mathematical Techniques in Multisensor Data Fusion*. Artech House, 1992.

[14] Y. Bar-Shalom and X. R. Li, *Multitarget-Multisensor Tracking: Principles and Techniques*. Y. Bar-Shalom, 1995, vol. 3.

[15] M. Stevens, "Bayes++ the Bayesian Filtering Library." [Online]. Available: http://bayesclasses.sourceforge.net/Bayes++.html

[16] B. Gerkey, R. Vaughan, A. Howard, and N. Koenig, "The Player/Stage Project." [Online]. Available: http://playerstage.sourceforge.net/