Electronic version of an article published in Int. Journal of Information Acquisition, Vol. 5, No. 3 (2008) 209–221
 doi:10.1142/S0219878908001612 http://www.worldscinet.com/ijia

MULTIMODAL PEOPLE TRACKING AND IDENTIFICATION FOR SERVICE ROBOTS

NICOLA BELLOTTO and HUOSHENG HU
Dept. of Computing and Electronic Systems, University of Essex
Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom

In order for a service robot to approach humans and provide the services it has been designed for, an efficient system for people tracking and identification must be developed. This paper presents a novel solution to the problem that makes use of different sensors and data fusion techniques. The robot utilizes a laser device and a PTZ color camera to detect, respectively, human legs and faces. The relative information is integrated, in real-time, using a sequential implementation of Unscented Kalman Filter. Furthermore, thanks to an histogram comparison with a measure based on the Bhattacharyya coefficient, people are also identified and labelled according to their clothes. This measure is also used to improve the robustness of the data association process. The effectiveness of the proposed method is shown by experiments with a real mobile robot in challenging situations.

Keywords: People tracking, sensor fusion, histogram-based identification, Unscented Kalman Filter, data association.

1. INTRODUCTION

It is a common belief, within the research community, that the 21st century will be the robot century, as so as the previous one was for computers. After the recent advances of computing and robotics technology, human-centred and service robots will be soon ready to serve us in our home, hospital, office and everywhere. These robots are autonomous, interactive and intelligent. They should be aware of the human presence and then act properly. This means, for example, finding people in the surrounding area who are willing to interact with these robots, but also keeping track of other humans in order to avoid possible collisions. An efficient and robust people tracking system is therefore necessary for practical applications of service robots.

Many solutions developed for tracking people with a mobile robot are single-sensor based. For instance, a stereo vision system was used to track and follow a single person with a mobile robot at short distances [Beymer and Konolige, 2001]. A laser-based solution, which implements heuristic algorithms to detect and keep track of moving entities, was also developed [Lindström and Eklundh, 2001]. Some other work reports the results achieved using a robot equipped with two laser range sensors, one pointing forward and another backward, that can track several people using a combination of particle filters and probabilistic data association [Schulz et al., 2003]. Another computational demanding solution is the implementation of a particle filter for the data fusion of a laser and an omnidirectional camera [Chakravarty and Jarvis, 2006]. In a few cases, the tracking system integrates also an identification module to recognize and la-

bel people. Examples include classic histogram intersection for people identification and standard Kalman filter for laser-based tracking [Bennewitz et al., 2003]. Other systems using vision-based identification are often limited to single human targets [Cielniak and Duckett, 2003]. More recently, a robust vision-based tracking and identification solution, which makes use of a dynamic Bayesian network, has been presented [Zajdel et al., 2005], but unfortunately the system is limited by the narrow angle of view of the robot's camera.

The solution presented in this paper, which extends our previous work [Bellotto and Hu, 2007], uses multisensor data fusion techniques for tracking people with a mobile robot. The task is performed with a SICK laser, used for legs detection, and a PTZ camera, for face detection. The integration of these two devices improves the robustness of the tracking and augments the area covered by the detection system. The position of each human target is estimated with an efficient implementation of Unscented Kalman Filter, and nearest neighbor data association is used to handle multiple targets. A vision-based identification procedure, that makes use of color histogram comparison, is also adopted to label the persons being tracked. When available, this identity information contributes to the discrimination of different targets, improving the robustness of the data association. Finally, our approach takes into account the limitations of the robot's hardware and is a good solution for achieving real-time performances in case of computational constraints.

The remainder of the paper is organized as follows. Section 2 introduces the algorithms for legs and face detection. Section 3 explains the multisensor tracking algorithm. A metric for color histogram comparison and the relative human identification are illustrated in Section 4. Section 5 describes how to handle multiple tracks and improve the data association. Section 6 reports some experimental results and relative considerations. Finally, conclusions and future work are presented in Section 7.

2. HUMAN DETECTION

2.1. *Legs Detection*

The laser sensor of the robot, mounted a few decimeters from the floor, can be used to detect human legs in a range of several meters. Different implementations of people tracking, making use of this device, can be found in literature. However, in most of the cases the legs detection is simply based on the search of local minima [Bennewitz et al., 2003; Schulz et al., 2003], which are generally well distinguishable only in uncluttered environments, like corridors or empty rooms. Other solutions are based on motion detection [Lindström and Eklundh, 2001; Chakravarty and Jarvis, 2006], missing therefore static persons and often becoming unreliable because of the difficult robot's motion compensation. Our legs detection algorithm, instead, is based on the recognition of typical legs patterns extracted from a single scan. These patterns correspond to three possible postures: legs apart (LA), forward straddle (FS) and two legs together or single leg (SL).

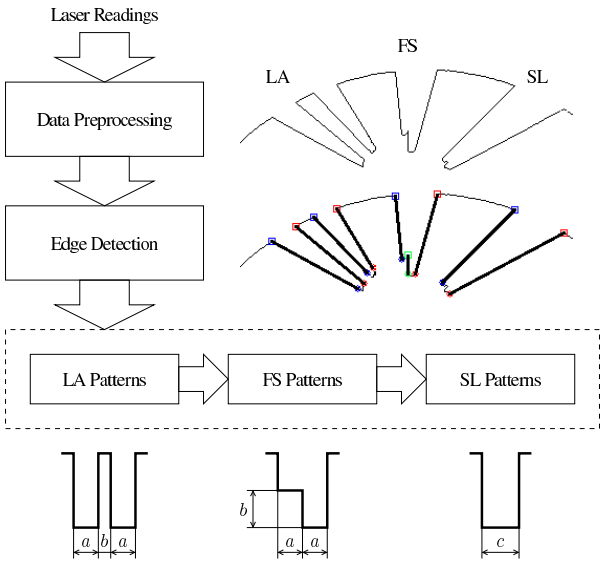


Fig. 1. Block-diagram of the legs detection algorithm. The three leg patterns are identified by sequences of vertical edges with the following constraints: leg width ($10 < a < 20$ cm), step length ($b < 40$ cm) and width of two legs together ($10 < c < 40$ cm). Patterns LA and FS are very seldom confused with other objects, while SL can be ambiguous for particular environments.

Fig. 1 shows the three patterns and a schematic representation of the legs detection algorithm. First of all, the laser data is filtered in order to smooth the readings, then all the edges lying on the directions of the laser scans are detected. Finally, groups of adjacent edges are identified according to simple geometric relations and spatial constraints that can correspond to legs. The method is quite robust even for cluttered environments and, besides being computationally inexpensive, it is not influenced by the robot motion. An example of legs detection is illustrated in Fig. 2.

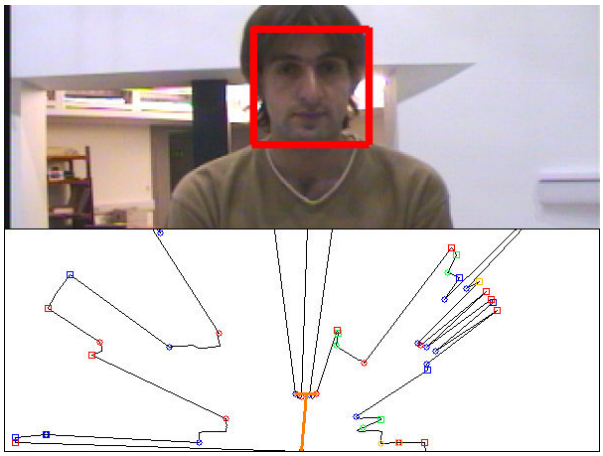


Fig. 2. Face and legs detection.

2.2. *Face Detection*

Several approaches used for face detection are based on the color segmentation of skin regions [Fritsch et al., 2003]. However, these solutions are not very robust to light variations and different skin tones. The face detection adopted in our system, instead, is based on the real-time algorithm of [Viola and Jones, 2004], which uses a cascade of classifier to extract simple but critical visual features. Besides being very fast, a useful quality of this solution is its color independence, which makes it robust to varying light conditions and not influenced by different skin tones. Compared to color segmentation techniques, the algorithm is more sensitive to head rotation and inclination, although this was not an issue in most of the tracking situations.

Given its position inside the current frame, the direction of the face, with respect to the camera, is expressed by the bearing α and the elevation β, calculated using a simple pin-hole camera model and the following transformations:

$$\begin{aligned} \alpha &= \tan^{-1}\left(\frac{W/2-u}{f}\right) \\ \beta &= \tan^{-1}\left(\frac{v-H/2}{f}\right) \end{aligned} \tag{1}$$

where (u, v) is the face's centre on an image $W \times H$ and f is the focal length in pixel units. Fig. 2 shows an example of face detection.

3. HUMAN TRACKING

Bayesian filtering is the most common solution to handle the uncertainty characterizing any tracking system. This consists in a recursive estimation of the target position, where the predicted state vector is corrected by the last sensor measurements. The evolution of the target state can be generally described by a deterministic model $\mathbf{x}_k = f(\mathbf{x}_{k-1}, \mathbf{w}_{k-1})$, where $\mathbf{x}_k$ is the state vector at the current time step k and $\mathbf{w}_{k-1}$ is white noise. The relative observations are described by another model $\mathbf{z}_k = h(\mathbf{x}_k, \mathbf{v}_k)$, where $\mathbf{z}_k$ is the observation vector and $\mathbf{v}_k$ is also white noise. These models, which are used respectively during the prediction and the correction steps of the Bayesian filter, are described in the next paragraphs for the human tracking of our robot.

3.1. *State Model*

When people walk, they often move in an unpredictable way, so tracking them becomes a very challenging task. Some researchers modelled human walking as a Brownian motion [Montemerlo et al., 2002], although a constant velocity (CV) model is a better choice in case of clutters [Beymer and Konolige, 2001; Schulz et al., 2003]. Another approach consists in learning the typical human motion patterns inside a particular environment using a network of distributed sensors [Bennewitz et al., 2003], but this is clearly an ad-hoc solution, not easy to implement in a

real scenario. We adopt instead a more general prediction model [Bellotto and Hu, 2006], which is basically an extension of the CV model. The relative equations are the following:

$$\begin{cases} x_k = x_{k-1} + x_{k-1}\ \Delta t_k\ \cos\phi_{k-1} \\ y_k = y_{k-1} + v_{k-1}\ \Delta t_k\ \sin\phi_{k-1} \\ z_k = z_{k-1} + n^z_{k-1} \\ \phi_k = \phi_{k-1} + n^\phi_{k-1} \\ v_k = |v_{k-1}| + n^v_{k-1} \end{cases} \tag{2}$$

where $\Delta t_k = t_k - t_{k-1}$. Besides position (x_k, y_k) and orientation ϕ_k, the component z_k is included for the estimation of the human height. Also, assuming a person is only walking forward, the velocity v_k is forced always positive. Noises n^z_{k-1}, n^ϕ_{k-1} and n^v_{k-1} are all zero-mean Gaussians.

3.2. *Observation Models*

The measurements provided by the laser are bearing b_k and range r_k of the legs. Since these depend also on the current state of the robot, the observation model includes its position and orientation $\left(x^R_k, y^R_k, \phi^R_k\right)$, as given by the odometry. Furthermore, we have to consider the displacement of the laser device with respect to the robot's center of rotation and given by the following equations:

$$\begin{array}{l} l^x_k = x^R_k + L_x \cos\phi^R_k \\ l^y_k = y^R_k + L_x \sin\phi^R_k \end{array} \tag{3}$$

where the constant L_x is the distances of the laser from the robot's centre, lying on its longitudinal axis (L_y is null). The laser observation model can be therefore written as follows:

$$\begin{cases} b_k = \tan^{-1}\left[\dfrac{y_k - l^y_k}{x_k - l^x_k}\right] - \phi^R_k + n^b_k \\ r_k = \sqrt{\left(x_k - l^x_k\right)^2 + \left(y_k - l^y_k\right)^2} + n^r_k \end{cases} \tag{4}$$

where the noises n^b_k and n^r_k are zero-mean Gaussians.

As for the laser, the displacement of the camera must also be calculated with the following equations:

$$\begin{array}{l} c^x_k = x^R_k + C_x \cos\phi^R_k \\ c^y_k = y^R_k + C_x \sin\phi^R_k \end{array} \tag{5}$$

In addition to the robot's position, the camera model must also take into account the pan ψ and tilt θ angles, plus its height C_z from the floor. The relative equations can be therefore written as follows:

$$\begin{cases} \alpha_k = \tan^{-1}\left[\dfrac{y_k - c^y_k}{x_k - c^x_k}\right] - \phi^R_k - \psi_k + n^\alpha_k \\ \beta_k = -\tan^{-1}\left[\dfrac{z_k - C_z}{\sqrt{\left(x_k - c^x_k\right)^2 + \left(y_k - c^y_k\right)^2}}\right] - \theta_k + n^\beta_k \end{cases} \tag{6}$$

Again, the noises n^α_k and n^β_k are zero-mean Gaussians.

3.3. *Sensor Fusion with Unscented Kalman Filter*

The Bayesian filter used for the state estimation of the tracking system plays a fundamental part. Kalman filters [Bar-Shalom and Li, 1995] provides an efficient way to integrate different sensor data and perform the estimation. In case of linear systems with Gaussian noises, a standard Kalman filter is known to be optimal, while an Extended Kalman Filter can be used to provide approximate solutions in case of non-linearities. Many of the solutions developed in the last years for tracking people with a robot, however, are mainly based on particle filters [Chakravarty and Jarvis, 2006; Montemerlo et al., 2002; Schulz et al., 2003] since their performances are not constrained by linear or Gaussian assumptions. Unfortunately, in terms of computational cost, such estimators are generally very demanding. Furthermore, the hardware requirements of these solutions increase with the number of targets to track.

A valid solution to perform human tracking is the Unscented Kalman Filter (UKF) [Julier and Uhlmann, 2004] that, instead of using a first-order linearization like in the EKF, captures mean and covariance of the probability distributions with carefully chosen weighted points, called "sigma points". Although this might sound similar to a particle filters, the two approaches differ on the fact that the sigma points are not random samples and their weights do not have to sum up to 1. Also, the number of points used by the UKF is small enough to make this estimator particularly indicated for

achieving real-time performances, even on mobile robots with limited hardware resources.

In case of asynchronous and uncorrelated measurements, a Kalman filter can be updated sequentially using only the observation available at the current step [Bar-Shalom and Li, 1995]. So, when new information is available from just one of the sensors (i.e. legs or face detection), only the relative observation model is used for the state correction. Moreover, when all the measurements are independent and available at the same time step, a sequential update of the filter, starting from the most to the least precise sensor data, gives a better estimate for non-linear systems and is also computationally more efficient. Under the same assumptions, the UKF can also be updated sequentially with the same benefits. In case both legs and faces are detected at the same time, the UKF filter is initially updated by the laser data, which is more precise, and then by the visual information.

4. HISTOGRAM-BASED IDENTIFICATION

Clothes can be used for a quick identification of different persons, and provide important information about human appearance. Cloth detection is an easy and effective way to distinguish people in case other more significant clues, like face or voice, are temporary unavailable. However, some care must be taken to handle varying light conditions and situations where people wear similar clothes.

4.1. *Color Histograms Comparison*

Different persons in an environment can be labelled according to the color histogram of their clothes, provided these are not completely identical. An efficient distance to compare color histograms is that one adopted by [Comaniciu et al., 2000] for a mean-shift visual tracking algorithm, which is based on the sample estimate of the Bhattacharyya coefficient. Given a discrete (normalized) density of reference $\hat{\mathbf{q}} = \{\hat{q}_u\}_{u=1\ldots m}$ from an m-bin histogram and another one at a given region of the image $\hat{\mathbf{p}} = \{\hat{p}_u\}_{u=1\ldots m}$, the sample estimate of the Bhattacharyya coefficient is so defined:

$$\rho\left(\hat{\mathbf{p}}, \hat{\mathbf{q}}\right) = \sum_{u=1}^{m} \sqrt{\hat{p}_u \hat{q}_u} \tag{7}$$

Using (7), the distance between the two distribution is calculated as follows:

$$d_h\left(\hat{\mathbf{p}}, \hat{\mathbf{q}}\right) = \sqrt{1 - \rho\left(\hat{\mathbf{p}}, \hat{\mathbf{q}}\right)} \tag{8}$$

Such a distance has some important properties, among which the scale invariance, since it uses discrete densities (differently from the classic histogram intersection), and the fact of being normalized between 0 and 1.

The region from which we extract the color histogram of a person is the torso, since this is the only part of the body that is almost always visible from the camera, either when the person is close or several meters far from the robot. As illustrated in Fig. 3, we consider the body proportions as in [Cielniak and Duckett, 2003], where the torso is 2/6 of the total height (and legs are 3/6).

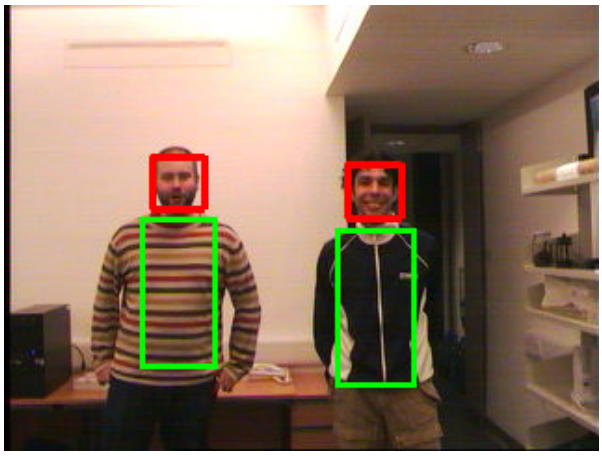


Fig. 3. Face detection and color histogram extraction. The area of the torso, from which the histogram is calculated, is 2/6 of the total human height.

4.2. *Human Identification*

Using the color histogram comparison, different people can be identified and distinguished from each other. First of all, if a target has been created but not yet labelled, the recognition system waits for his face to be detected, in order to estimate the region on the image where to get the histogram from. We select a window that covers

approximately the human torso, then the relative histogram is compared, using (8), to those ones already in a database. The target is labelled according to the histogram for which the distance is minimum, considering only the distances that fall below a certain threshold. If none of the histograms in the current database are "close" enough, the target is classified as unknown. In case the face is detected, however, a new label is assigned to the unknown person and the relative histogram is memorized.

From several tests in real situations, under slightly varying light conditions and with people wearing different clothes, the histogram comparison based on (8) has proved to be very selective, with almost zero cases of misclassification, but still robust enough to recognize the same person in different postures.

Also, one of the advantages of combining people tracking and identification is the possibility to adapt the recognition system to small changes on the current human targets or in the environment. In our system, once a human track has been created and the identity initialized with an existing or a new label, the relative color histogram is continuously updated as long as the person is tracked and, of course, he is inside the camera's field of view. To avoid wrong updates, we set a threshold on the distance between the new candidate histogram and the old one, so that the update is actually performed only when the difference between the two is very small.

5. HANDLING MULTIPLE TARGETS

A fundamental part of every multi-target system is the data association, that is, the assignment of the current measurements to the proper tracks. Many different data association techniques have been implemented for people tracking, using probabilistic approaches like JPDA or MHT [Bar-Shalom and Li, 1995]. In general, however, these methods are computationally quite expensive. Instead, if the set of entities to track is not too dense, a simple solution based on Nearest Neighbour (NN) data association proved to be a good and fast alternative [Montemerlo et al., 2002; Bellotto and Hu, 2006].

5.1. *Nearest Neighbour Data Association*

The NN is an intuitive one-to-one association algorithm [Bar-Shalom and Li, 1995]. For each candidate track, the observation $\mathbf{z}_k$ is initially predicted using the relative model. Then, a distance function d_{mn} is used to measure the similarity between every possible couple of predicted–real observation $(\mathbf{y}_m, \mathbf{z}_n)$ and this value is used to fill an association matrix $\mathbf{S}_{M\times N}$, where M is the number of sensor measurements and N is the number of predicted observations. Finally, the m-th measurement is chosen to update the n-th track if the relative element d_{mn} is the minimum distance (highest similarity) among all those in $\mathbf{S}_{M\times N}$; the relative m-th row and n-th column are excluded from further consideration. This last step is continuously repeated until there are no more available measurements or tracks to combine. The whole process is schematically illustrated in Fig. 4.

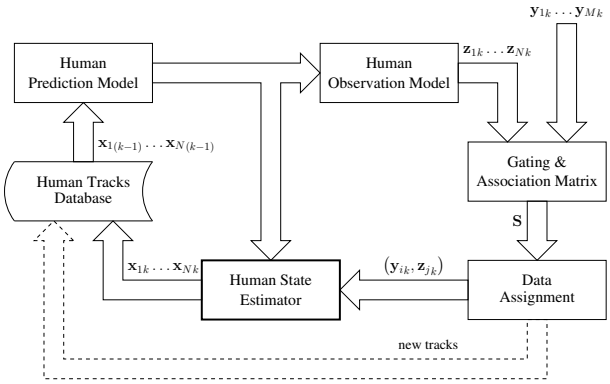


Fig. 4. Schematic representation of the system, including state estimation and data association. The process starts extracting candidates $\mathbf{x}_{1(k-1)} \dots \mathbf{x}_{N(k-1)}$ from the target database and assigns to each state the proper observations $\mathbf{y}_{ik}$. The pairs $(\mathbf{y}_{ik}, \mathbf{z}_{jk})$ are eventually used for the correction of the human state estimations.

Before the creation of the association matrix, a common gating approach is adopted to avoid unlikely assignments. In practice, the gating procedure consists in excluding all the measurements $\mathbf{y}_m$ outside a *validation region* [Bar-Shalom and Li, 1995]. This region is constructed around the predicted observation $\mathbf{z}_n$ according to the relation $d(\mathbf{y}_m, \mathbf{z}_n) \leqslant \lambda$, where λ is a threshold and d is the Mahalanobis distance de-

fined as follows:

$$d(\mathbf{y}_m, \mathbf{z}_n) = \sqrt{(\mathbf{y}_m - \mathbf{z}_n)^T \, \mathbf{\Sigma}_{mn}^{-1} \, (\mathbf{y}_m - \mathbf{z}_n)} \quad (9)$$

$\mathbf{\Sigma}_{mn}$ is the covariance matrix of the innovation $(\mathbf{y}_m - \mathbf{z}_n)$. The value of λ can be determined from tables of the *chi*-squared distribution ($\lambda = 3.03$ in our case).

5.2. *Integration of Histogram Information*

After the validation gate, the data association proceeds with the actual NN assignment. The most common measure of similarity between real and predicted observations is based on the same distance defined in (9). However, at this point we introduced an improvement that makes use of the identity information, i.e. the color histogram of the person. Using (8), we calculate the distance $d_h(\hat{\mathbf{p}}_m, \hat{\mathbf{q}}_n)$ between the histogram $\hat{\mathbf{p}}_m$ of the region relative to the current measurement and the histogram $\hat{\mathbf{q}}_n$ of the considered track. Such a distance is compared to a threshold γ (for example $\gamma = 0.8$) and, if greater, the measurement is discarded. This is basically an additional gating, although no *chi*-squared distributions are assumed in this case. If smaller than γ instead, d_h is multiplied to d and the result is used as a new similarity measure. The relative equations are the following:

$$d^*_{m,n} = \begin{cases} \infty, & \exists \hat{\mathbf{p}}_m, \hat{\mathbf{q}}_n \text{ and } d_h(\hat{\mathbf{p}}_m, \hat{\mathbf{q}}_n) > \gamma \\ d_h(\hat{\mathbf{p}}_m, \hat{\mathbf{q}}_n) d(\mathbf{y}_m, \mathbf{z}_n), & \exists \hat{\mathbf{p}}_m, \hat{\mathbf{q}}_n \text{ and } d_h(\hat{\mathbf{p}}_m, \hat{\mathbf{q}}_n) \leqslant \gamma \\ d(\mathbf{y}_m, \mathbf{z}_n), & \nexists \hat{\mathbf{p}}_m \text{ or } \nexists \hat{\mathbf{q}}_n \end{cases} \quad (10)$$

Clearly, the smaller $d^*_{m,n}$, the more similar real and predicted observations are. In practice, whenever the histograms $\hat{\mathbf{p}}_m$ and $\hat{\mathbf{q}}_n$ exist, respectively for the measurement and the track of reference, the measure is discarded if $d_h > \gamma$, otherwise d is weighted by d_h (with $0 \leqslant d_h \leqslant \gamma$). If $\hat{\mathbf{p}}_m$ is similar to $\hat{\mathbf{q}}_n$, the new similarity measure $d^*_{m,n}$ will be much smaller than the original one, otherwise it will tend to d. When one or both the histograms are missing, because the camera is pointing to another direction or the track is not labelled, the similarity is simply given by the Mahalanobis distance d.

Two different association matrices are created using (10), one for the legs and another for the face detections. In the first case, the histogram $\hat{\mathbf{p}}_m$ is computed from the region corresponding to the legs' direction, if covered by the camera's field of view; its size is a function of the legs' distance and *a priori* height of the considered track. In the second case instead, $\hat{\mathbf{p}}_m$ is computed from the region corresponding to the face's direction; its size is a function of the face's elevation and *a priori* position of the person.

5.3. *Creating and Removing Tracks*

To create new tracks we use the laser readings discarded by the gating procedure and the assignment. In particular, only the leg patterns LA and FS , explained in Section 2, are selective enough to be considered trustful for tracks creation. Faces are not taken into account at this stage as they cannot provide range information, indispensable for estimating the initial track's position. In parallel to the human tracks database, we keep then another list containing all the possible candidates. Each one of these is generated by a sequence of readings falling inside a certain region, delimited by the distance a person can cover in the interval Δt_k at a certain speed (we chose 1.5 m/s). Each candidate is assigned a maximum time interval, or "lifetime": if during this interval there are enough readings falling inside its region, the candidate is promoted to human track, otherwise it is considered a false positive and removed. Finally, normal tracks are deleted from the database if not updated for more than a certain time.

6. EXPERIMENTAL RESULTS

Several experiments have been conducted using a Pioneer robot equipped with a SICK laser and a PTZ camera, as shown in Fig. 5. The laser is located at approximately 30 cm from the floor and the camera is mounted on a special support, at about 1.5 m, in order to facilitate the face detection. The on-board PC is a Pentium III 800 MHz with 128 MB of RAM, running Linux OS.

Fig. 5. Robot equipped with SICK laser and PTZ camera. The camera used is the black one on the top.

The whole software has been written in C++ and runs in real-time on the robot's PC, although it is possible to use an external client, connect via wireless, for remote control and debug. The laser returns scans of 180° at 5 Hz, which is also the update frequency of our program, with range and angular resolution of 1 cm and 0.5° respectively. The camera has a field of view of about 49° and provides images with a resolution of 320 × 240 pixels at 10 Hz. Every track is created after at least 3 readings within the maximum time interval of 1 s and is removed if not updated for more than 2 s.

6.1. *Person Following with Clutters*

The experiment described next shows the robustness of the tracking system when following a person through cluttered environments. The path of the subject being tracked by the robot is shown in Fig. 6. This started from our laboratory (R1), passing through a corridor (R2) to finally reach an office (R3). In a couple of occasions, the person being tracked stopped a few moments to wait for the robot that was moving at about 40 cm/s. Despite several challenging situations, among which clutters, door passages and varying light conditions, the robot was able to track successfully the person along the whole path. Other similar tests, moving between different rooms and with other persons, were always accomplished correctly by the robot.

A few snapshots of the experiment are shown in Fig. 7. In the first one, the person was detected and labelled. The following snapshots show the same person moving and being tracked by the robot. In particular, some of the pictures demonstrate how the track could be successfully updated despite one of the two sensor information, legs or face detection, was missing. Indeed, the third snapshot in Fig. 7, for example, shows the tracking without visual detection, where the UKF was updated only by the legs observation. The opposite situation, instead, is illustrated in the last snapshot, where legs were not detected, so the tracking had to rely only on vision. Note also that clutters and false positives, often present along the path, were always correctly handled and did not prevent the tracking from working properly.

6.2. *People Tracking and Identification*

The scope of the following experiment was to test the performance of the tracking and identification system with several people. We set up a little environment inside the robotic arena (room R0 in Fig. 6) where three persons could walk in front of the robot or hide behind a cardboard wall.

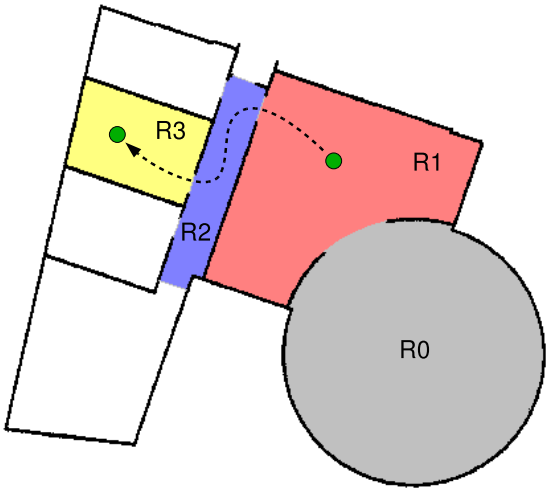


Fig. 6. Path of the person followed by the robot from room R1 (laboratory) to R3 (office).

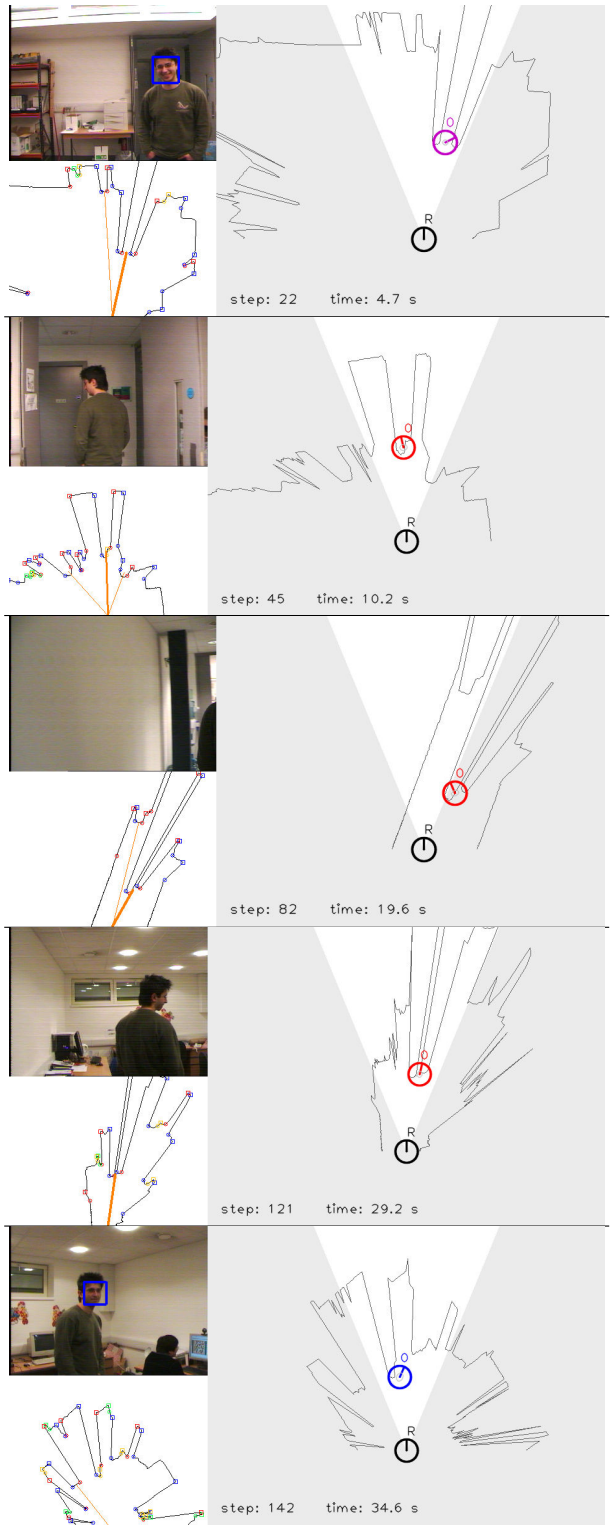


Fig. 7. A few snapshots of the robot following a person. The face and legs detection are shown on the left, while the estimated position of the person "0", as seen by the robot "R", are illustrated on the right.

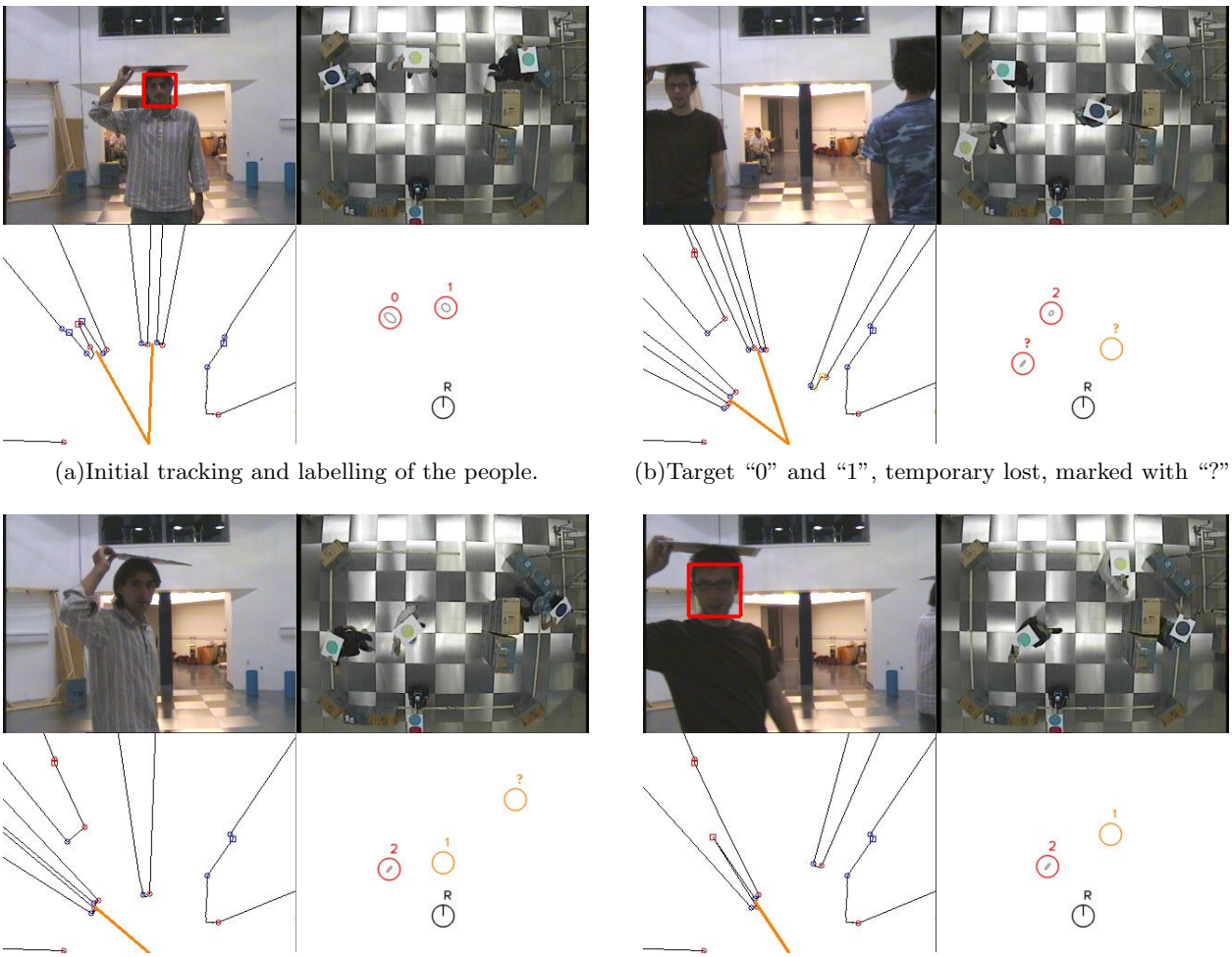
From a camera mounted on the ceiling we could observe the absolute positions of the three human targets, each one keeping a different color marker on the head. The pan/tilt control of the camera was kept fixed, since the robot in this case could almost always detect at least one face just pointing straight.

Initially, all the persons were hiding behind the wall. Then they came out, one by one, facing to the robot's camera in order to be labelled a first time. Finally, they started moving together at normal walking speed, sometimes going back behind the wall, and sometimes performing circular paths in front of the robot. Several occlusions made the tracking task particularly difficult, in particular when someone was passing right in front of the robot, preventing the laser and the camera to detect the other participants.

A few situations are illustrated in Fig. 8. For each of them, an image from the robot's camera and a laser scan are shown on the left, while a bird-eye-view shows the persons on the right together with the estimated tracks. During the experiment, which was approximately 60 s long, the estimated positions and the labels of the target were correct most of the time. In Fig. 8(b), target "0" and "1" were temporary lost after a sharp turn, so they were labelled with "?". Unfortunately, once visible again a few instants later, only target "1" was correctly re-labelled. Target "0" instead was not facing the robot and could not be recognized, so in Fig. 8(c) he was still keeping a "?" label.

6.3. *Data Association With and Without Identification*

This section presents a typical case of data association problem and compare the results with and without the integration of the histogram-based identification. Fig. 9 shows a sequence of images and laser scans, taken at intervals of 0.4 s, where a person was walking in front of the robot, and where another one appeared a few instants later. In the first frame, the legs of the closest person were detected; in the second one, no detections were available at all; finally, in the third frame, only the face of the other person was detected. Note that, in the last case, the



(a)Initial tracking and labelling of the people.

(b)Target "0" and "1", temporary lost, marked with "?".

(c)Only target "1" reacquires the correct label.

(d)The target behind the wall is removed after 2s.

Fig. 8. People tracking and identification. Three persons are labelled, in order, from "0" to "2" and tracked during paths in front of the robot. A sequence of four different moments, at intervals of about 5s, is shown.

detected face was inside the validation region of the track.

The top right part of each sub-figure shows the track of the person in front of the robot without using the histogram-based identification. The track was initially correct, but in Fig. 9(c) there was an estimation error because the system used the other face to update the track. In this case, the gating process, based only on positional information, was not good enough to discard the wrong measurement. In the bottom right part, instead, which represents exactly the same situation, the gating process performed correctly thanks to the identity information. As shown by Fig. 9(c) indeed, the track estimate with identification was still correct, since the color histogram of the second person was different from the initial one, and the relative face was therefore discarded.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a novel approach for the multimodal detection, tracking and identification of people with a service robot. The solution combines the information of a laser and a camera using a sequential implementation of Unscented Kalman Filter. An efficient histogram comparison is also illustrated for the identification and labelling of human tracks. This has also

been used to improve the robustness of the data association procedure, so that several people can be tracked simultaneously. The effectiveness of the proposed approach has been demonstrated experimentally with a real mobile robot in dynamic environments.

Future directions of our research include an improved selection of the image region from where the histogram has to be extracted. Also, the integration of additional human identification modules, like face or voice recognition, will be included in the system. The final objective of our research is the implementation of a feasible real-time solution for joint people tracking and recognition with an interactive service robot.

References

Bar-Shalom, Y. and Li, X. R. (1995). *Multitarget-Multisensor Tracking: Principles and Techniques*. Y. Bar-Shalom.

Bellotto, N. and Hu, H. (2006). Vision and laser data fusion for tracking people with a mobile robot. In *Proc. of IEEE Int. Conf. on Robotics and Biomimetics*, pages 7–12, China.

Bellotto, N. and Hu, H. (2007). People tracking and identification with a mobile robot. In *Proc. of IEEE Int. Conf. on Mechatronics and Automation (ICMA)*, pages 3565–3570, Harbin, China.

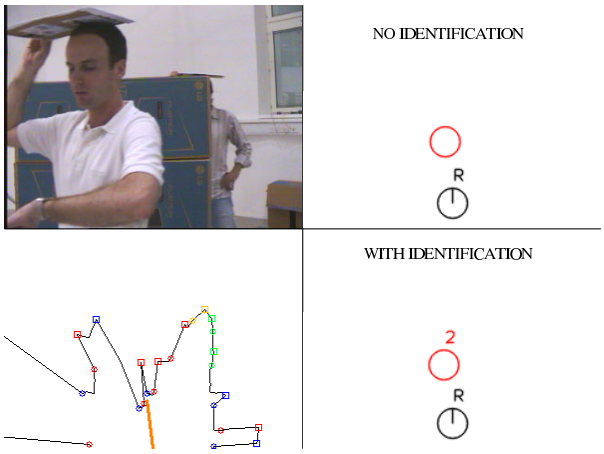
Bennewitz, M., Cielniak, G., and Burgard, W. (2003). Utilizing learned motion patterns to robustly track persons. In *Proc. of IEEE Int. W. on VS-PETS*, pages 102–109, France.

Beymer, D. and Konolige, K. (2001). Tracking people from a mobile platform. In *IJCAI-2001 Workshop on Reasoning with Uncertainty in Robotics*, Seattle, WA, USA.

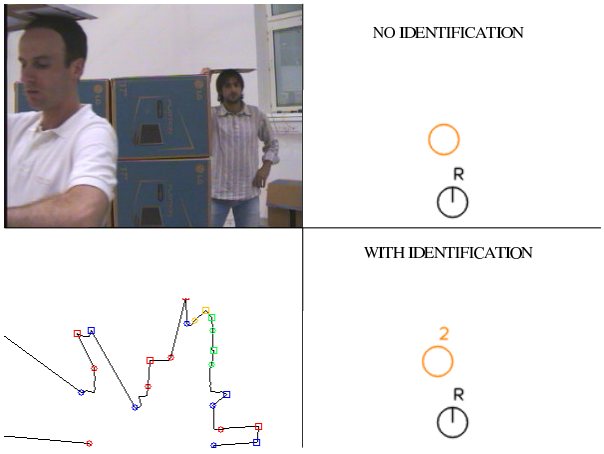
Chakravarty, P. and Jarvis, R. (2006). Panoramic vision and laser range finder fusion for multiple person tracking. In *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 2949–2954, Beijing, China.

Cielniak, G. and Duckett, T. (2003). Person identification by mobile robots in indoor environments. In *Proc. of the IEEE Int. Workshop on Robotic Sensing (ROSE)*, Örebro, Sweden.

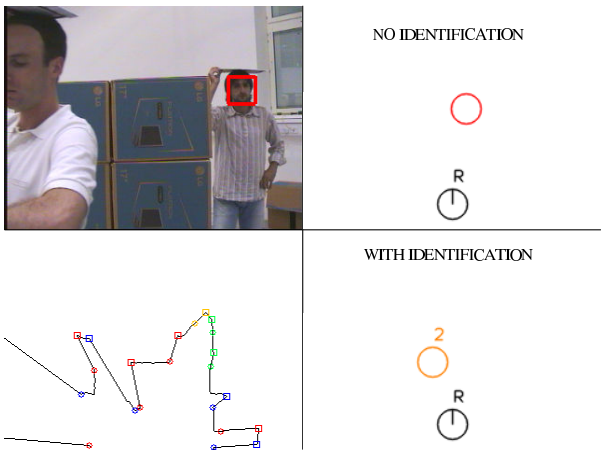
Comaniciu, D., Ramesh, V., and Meer, P. (2000). Real-time tracking of non-rigid objects using mean shift. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 142–149, South Carolina, USA.



(a)Initial track with relative legs detection.



(b)No legs or face detections.



(c)Face inside the validation region of the track.

Fig. 9. A person being tracked in front of the robot. The face of a different person is detected and it can be erroneously associated to the initial track.

Fritsch, J., Kleinehagenbrock, M., Lang, S., Plötz, T., Fink, G. A., and Sagerer, G. (2003). Multi-modal anchoring for human-robot-interaction. *Robotics and Autonomous Systems*, 43(2-3):133–147.

Julier, S. J. and Uhlmann, J. K. (2004). Unscented filtering and nonlinear estimation. *Proc. of the IEEE*, 92(3):401–422.

Lindström, M. and Eklundh, J.-O. (2001). 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)*, volume 3, pages 1364–1369, Maui, HI, USA.

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

Schulz, D., Burgard, W., Fox, D., and Cremers, A. B. (2003). People tracking with mobile robots using sample-based joint probabilistic data association filters. *Int. Journal of Robotics Research*, 22(2):99–116.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *Int. Journal of Computer Vision*, 57(2):137–154.

Zajdel, W., Zivkovic, Z., and Krŏse, B. J. A. (2005). Keeping track of humans: Have I seen this person before? In *Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 2093–2098, Barcelona, Spain.